A program to compare genetic differentiation statistics across loci using resampling of individuals and loci

Bryan D. Neff and Bonnie A. Fraser

Department of Biology, Biological & Geological Sciences Building, University of Western Ontario, London N6A 5B7, Canada (bneff@uwo.ca)
Phone: (519) 850-2532; Fax: (519) 661-2014

July 30, 2009

Description

This C++ program statistically compares population genetic differentiation statistics (F_{ST} or G'_{ST}) calculated from different loci. The program employs a routine that resamples either or both of individuals within populations and loci, and thereby allows comparison of genetic differentiation at individual loci. We use the F_{ST} formulation developed by Ronfort *et al*. (1998) for autotetraploid species (their equation 26), which can be used to calculate F_{ST} for either diploid or tetraploid species by defining the ploidy level. The program also conducts a comparison using values corrected for locus variability. This correction is accomplished by dividing the F_{ST} values by $G'_{ST(max)} = (1 - H_S)$ (see equation 3 in Hedrick 2005). The program begins by calculating F_{ST} or G'_{ST} for each of two locus-types and for each possible population pairwise comparison. The program then resamples individuals within each population with replacement until the original sample sizes for each population are produced and/or loci. The level of resampling is defined by the user. Using the resampled data, the two statistics are again calculated for each locus-type. For each population pair, a comparison is then made between the F_{ST} or G'_{ST} estimates for the two locus-types, and the locus-type associated with the higher or lower value is recorded. The routine is repeated for a total of 1000 replicates, from which the mean and median are calculated and the 95% confidence interval is determined from the 25^{th} and 975^{th} value in a ranked list. The proportion of comparisons in which one locus-type was either higher or lower than the other type is also reported. This proportion can serve as a one-tailed *p*-value for the null hypothesis that one locus-type is either higher or lower than the other type.

Directions

Before running the program two data files must be set-up. A sample of these data files can be downloaded from the website and these can be modified with your data. Microsoft Excel or equivalent software is recommended for modifying the data file. Each file must be saved as a **tab delimited** file with the ".txt" extension. The file names cannot be changed. The following data must be entered into each spreadsheet:

(1) MHCdata.txt

number of populations, number of alleles per individual (either 2 or 4), G'_{ST} correction (=0 for none, =1 to apply correction), resampling MHC by individuals (=0 no resampling, =1 to resample individuals)
number of individuals sampled from each population

MHC genotypes of each individual

Note, each allele appears in its own column and each individual on its own row, starting with the first individual from the first population consecutive through to the last individual in the final population. Note, alleles must be identified using a number between 0 and 350.

The following is a partial sample spreadsheet for MHCdata.txt (the entire spreadsheet can be downloaded from the website). The example assumes 10 populations, four alleles per individual, the G'_{ST} correction will be applied to the F_{ST} estimates, and individuals will be resampled to provide a variance estimate for the MHC G'_{ST} estimates. The second row provides the number of individuals genotyped in each of the 10 populations and the values range from 11 (population 3) to 17 individuals (population 8). The next two rows of numbers represent the MHC genotypes of the first two individuals from population 1. There are four alleles because the MHC gene is duplicated in this example. The complete spreadsheet has genotype data from a total of 142 individuals.

```
 10   4    1    1
 16  15   11   14  15  13  13  17  14  14

 90  90   10  220
 10  10  230  230
  .   .    .    .
```

(2) MICdata.txt

number of loci, resampling microsatellite data by individuals, resampling microsatellite data by loci (in both cases of resampling =0 for no resampling, =1 to resample at the desired level)

number of individuals sampled from each population

Microsatellite genotypes of each individual

Note, each allele appears in its own column and each individual on its own row, starting with the first individual from the first population consecutive through to the last individual in the final population. Note, alleles must be identified using a number between 0 and 350.

The following is a partial sample spreadsheet for MICdata.txt (the entire spreadsheet can be downloaded from the website). The example assumes six microsatellite loci have been used and there will be resampling at both the individual and locus levels. The second row provides the number of individuals genotyped in each of the 10 populations and the values range from 14 (population 8) to 33 individuals (population 5). The next two rows of numbers represent the genotypes of the first two individuals from population 1 for the six microsatellites. The first two numbers in a row are the alleles for the first microsatellite. The second two numbers are the alleles for the second locus, … The complete spreadsheet has genotype data from a total of 212 individuals.

6	1	1									
21	18	17	19	33	32	19	14	20	19		
191	222	162	167	210	210	150	164	212	261	257	257
179	187	162	164	202	206	150	150	277	281	257	262
.	.	.	.	.	.	.	.	.	.	.	.

Running the program

The program can handle up to 50 populations, 200 individuals per population and 50 loci for the MICdata file. The MHCdata file can have either one or two loci (e.g. two or four alleles per individual). To run the program, simply double click the Neff&Fraser_Fst.exe icon. All files must be in the same directory or an error message will appear indicating that one of the data files could not be opened. While the program is running, a counter will appear on the screen that counts down from 10 to 1. Depending on the number of populations, individuals sampled, and loci used, the program could take a few minutes to over an hour to execute.

Once the program has finished, the output is displayed on the screen and written to the file output.txt. The output file will appear in the same directory as the input and program files. It can be viewed using Excel or other similar programs. The output file contains 11 columns of data. The first two columns refer to the populations being compared. The third through sixth columns present the mean, median, 2.5% and 97.5% F_{ST} or G'_{ST} values for the MHC, respectively, the seventh through tenth columns present the mean, median, 2.5% and 97.5% F_{ST} or G'_{ST} values for the microsatellite loci, and the final column present the *p*-value for the comparison of the MHC and microsatellite F_{ST} or G'_{ST} values.