PSA.cpp

November 24, 2021

Author: Takahiro Maruki

C++ program of the population structure analyzer (PSA)

This C++ program is for estimating Wright's fixation indices in the statistical framework by Weir and Cockerham (1984) using genotype frequencies estimated by GFE (Maruki and Lynch 2015).

Input file. The input file is a tab-delimited text file, consisting of the reference nucleotide and population-specific variables estimated by GFE at every position. The first and second columns are the scaffold and site identifiers. The third column denotes the nucleotide of the reference sequence. Thereafter, nine variables (major allele, minor allele, population coverage, effective number of sampled individuals (N_i), major-allele frequency estimate, minor-allele frequency estimate, error-rate estimate, heterozygote frequency estimate, and likelihood-ratio test statistic for polymorphism) are shown for each population.

Output file. The output file is also a tab-delimited file, in this case consisting of 12 columns. Column: 1) scaffold (chromosome) identifier; 2) site identifier (coordinate); 3) nucleotide of the reference sequence; 4) total coverage; 5) number of alleles; 6) number of populations with data; 7) F (F_{IT}) estimate; 8) θ (F_{ST}) estimate; 9) f (F_{IS}) estimate; 10) H_S (π) estimate; 11) H_T (IT) estimate; 12) minor-allele frequency estimate in the metapopulation.

Reference

If you use this program, please cite the following papers:

Maruki, T., and Lynch, M., (2015) Genotype-frequency estimation from high-throughput sequencing data. *Genetics* **201**:473-486.

Maruki, T., Ye, Z., and Lynch, M., (in revision) Evolutionary genomics of a subdivided species.

Weir, B. S., and C. C. Cockerham, 1984 Estimating F-Statistics for the analysis of population structure. *Evolution* **38:** 1358-1370.

Instructions

Below are specific procedures for using the program:

1. Run GFE_v3.0 in the `F' mode in each population. Include the reference information in the first population by setting the `-ref_info' option at one.

2. Make the input file by pasting the output files made in the previous step.

paste Out_F_GFE_pop1.txt Out_F_GFE_pop2.txt Out_F_GFE_pop3.txt Out_F_GFE_pop4.txt Out_F_GFE_pop5.txt Out_F_GFE_pop6.txt Out_F_GFE_pop7.txt Out_F_GFE_pop8.txt Out_F_GFE_pop9.txt Out_F_GFE_pop10.txt > In_PSA_10pops.txt

3. Compile the program by typing the following command:

g++ -o PSA PSA.cpp -lm

3. Run the program by typing the following command:

./PSA -in In_PSA_10pops.txt -out Out_PSA_10pops.txt

- In_PSA.txt and Out_PSA.txt are default names of the input and output files, respectively. The input and output file names can be specified by adding the `-in' and `-out' options, respectively.

- The minimum required effective number of sampled individuals in each deme can be specified by adding the ` -min_Ni' option. Its default value is 10.0.

- The chi-square critical value for the polymorphism test can be specified by adding the `-cv' option. The default critical value is 5.991 (at the 5% level).

- A usage help message explaining these options can be shown by typing the following command:

./PSA -h

Copyright notice

Contact

If you have difficulty using this software, please send the following information to Takahiro Maruki (tmaruki@asu.edu):

1. Brief explanation of the problem.

2. Command entered.

3. Part of the input file.

4. Part of the output file.