File S2 README of the computer program for simulated data analysis.

June 3, 2015

Author: Takahiro Maruki

GFESD.cpp

This C++ program, genotype-frequency estimator for simulated data (GFESD.cpp), estimates allele and genotype frequencies from simulated sequence-read data of multiple diploid individuals randomly sampled from a population by a maximum-likelihood (ML) method. It generates sequence read data for each diploid individual in the population sample given the frequencies of major homozygotes and minor homozygotes, number of sampled individuals, mean depth of coverage per site per individual, and error rate per site. The depths of coverage are assumed to be Poisson-distributed among the individuals. The errors are assumed to occur at the same rate from the true nucleotide to one of the other three. Then, it estimates the allele frequencies, disequilibrium/inbreeding coefficient, and error rate from the sequence-read data derived from the population sample. Genotype-frequency estimates are also obtained from the estimates of allele frequencies and disequilibrium/inbreeding coefficient. The statistical significance of the polymorphisms and their genotypic deviations from Hardy-Weinberg equilibrium (HWE) can also be tested, using the likelihood-ratio test statistics, with this program. To examine the performance of the proposed method given a specified set of parameter values, the generation of sequence-read data and estimation of the parameter values are repeated for a given number of replications.

Instructions

Below are the specific procedures for using the program.

1. Compile the program by typing the following command:

```
g++ GFESD.cpp -o GFESD –lm
```

2. Run the program by typing the following command:

```
./GFESD N e n reps P Q out_file_name
```

- In the above command, the meanings of the arguments are as follows:

N: number of sampled individuals (sample size)

e: error rate

n: mean depth of coverage

reps: number of simulation replications

P: major-homozygote frequency

Q: minor-homozygote frequency

out_file_name: name of the output file

- For example, to run 10,000 simulation replications with N = 100, e = 0.01, n = 3, P = 0.9, Q = 0.1, and output the results in a file named GFESD_N100e0.01n3P0.9Q0.1.txt, type the following:

```
./GFESD 100 0.01 3 10000 0.9 0.1 GFESD_N100e0.01n3P0.9Q0.1.txt
```

Output file. The output file is a tab-delimited text file.

At the top of the output file, the values of the major-homozygote frequency P, minor-homozygote frequency Q, disequilibrium coefficient D_A, inbreeding coefficient f, error rate, mean depth of coverage, sample size, and number of simulation replications are shown.

Then, the results in each simulation replication are shown with 18 columns in each line. Column: 1) replication number, 2) realized major-allele frequency in the population sample, 3) realized disequilibrium coefficient in the population sample, 4) realized inbreeding coefficient in the population sample, 5) realized major-homozygote frequency in the population sample, 6) realized minor-homozygote frequency in the population sample, 7) realized error rate among sequence reads, 8) estimated major-allele frequency, 9) estimated disequilibrium coefficient, 10) estimated inbreeding coefficient, 11) estimated major-homozygote frequency, 12) estimated minor-homozygote frequency, 13) estimated error rate, 14) maximum likelihood of the observed data under the HWE model, 15) maximum likelihood of the observed data under the null (monomorphism) model, 16) maximum likelihood of the observed data under the full model, 17) likelihood-ratio test statistic for polymorphism detection, 18) likelihood-ratio test statistic for HWE-deviation.

At the bottom of the output file, summaries of estimation results and realized values of the parameters in the simulation replications (simulation results) are shown. In the estimation results, the mean and standard deviation of the ML estimates of allele frequencies, disequilibrium coefficient, inbreeding coefficient, genotype frequencies, and error rate are shown. For the disequilibrium coefficient and inbreeding coefficient estimates, the root mean square deviation (RMSD) is also shown. The results for the disequilibrium coefficient and inbreeding coefficient are conditioned on significant polymorphism at the 5% level. Furthermore, the rates of significant polymorphism detection and significant HWE-deviation detection are also shown, with the statistical significance of both tests set at the 5% level. In the simulation results, the mean and standard deviation of the realized values of the corresponding parameters are shown.