bmf.cpp

June 6, 2022

Author: Takahiro Maruki

C++ program of the Bayesian mutation finder (BMF).

This C++ program is for identifying mutations from BGC (Maruki and Lynch 2017) genotype calls inferred from high-throughput sequencing data from mutation accumulation experiments.

Input file. The input is a tab-delimited text file and can be made from a pro file of nucleotide read counts using GFE (Maruki and Lynch 2015; available at https://github.com/Takahiro-Maruki/Package-GFE) in the `c' mode. The meanings of the first twelve columns are: 1) scaffold (chromosome) identifier; 2) site identifier (coordinate); 3) nucleotide of the reference sequence; 4, 5) nucleotides of the major and minor alleles, respectively (1: A, 2: C, 3: G, 4: T); 6) depth of coverage in the population sample (sum of the coverage over the individuals; 7) error rate estimate; 8, 9, 10) ML estimates of the frequencies of major homozygotes, heterozygotes, and minor homozygotes, respectively; 11) likelihood-ratio test statistic for polymorphism; 12) likelihood-ratio test statistic for deviation from Hardy-Weinberg equilibrium. Thereafter, the nucleotide-read quartet is shown for each individual in each of the columns.

Output file. The output file is also a tab-delimited text file. The meanings of the first eight columns are: 1) scaffold (chromosome) identifier; 2) site identifier (coordinate); 3) nucleotide of the reference sequence; 4) error rate estimate; 5) depth of coverage in the population sample (sum of the coverage over the individuals); 6) inferred ancestral genotype; 7) number of genotype calls; 8) minor-allele count. Thereafter, the genotype call is shown for each individual in each of the columns.

Reference

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

Maruki, T, Ozere, A, and Cristescu, M. E., Systematic identification of single nucleotide mutations from genome-wide mutation accumulation data. In prep.

Instructions

Below are specific procedures for using the program:

1. Compile the program by typing the following command:

g++ -o bmf bmf.cpp -lm

2. Run the program by typing the following command:

./bmf -in Small_In_bmf.txt -out Small_Out_bmf.txt

- The `-in', and `-out' options specify the input file and output file name, respectively.

- The minimum and maximum coverage for calling individual genotypes can be specified by the `-min_cov' and `-max_cov', respectively. The default values for the minimum and maximum coverage are 1 and 2,000,000,000, respectively.

- The critical values for the heterozygous and homozygous cumulative binomial probabilities in the binomial test can be specified by `-cv_het' and `-cv_hom' options, respectively. The default values for the heterozygous and homozygous cumulative binomial probabilities are 0.025 and 0.05, respectively.

- Results at all sites in the reference sequence can be shown in the output file by setting the `-as' option at one (1). The input file also needs to contain all sites in the reference sequence in this case, which can be done by running GFE_v3.0 (available at https://github.com/Takahiro-Maruki/Package-GFE) in the `c' mode and setting the 'as' option at one (1).

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

./bmf -h

Copyright notice

Contact

If you have difficulty using this software, please send the following information to Takahiro Maruki (takahiro.maruki@mcgill.ca):

1. Brief explanation of the problem.

2. Command used.

3. Part of the input file.

4. Part of the output file.