BGC_v1.1.cpp

August 30, 2017

Author: Takahiro Maruki

C++ program of a Bayesian genotype caller for diploid organisms useful for calling genotypes from low-coverage sequencing data

This C++ program is for calling genotypes from nucleotide-read quartets (read counts of A, C, G, and T) derived from individual high-throughput sequencing data for multiple diploid individuals from a population by a maximum-likelihood (ML) method. This genotype caller incorporates the genotype-frequency and sequencing-error rate estimates predetermined by an ML genotype-frequency estimator (GFE) (Maruki and Lynch 2015) to improve the accuracy of genotypes called from low-coverage sequencing data. At each of the significantly polymorphic sites pre-identified by GFE, the genotype for each individual is estimated by maximizing the likelihood of the observed data.

Input file. The input file is a tab-delimited text file, and can be prepared using GFE_v2.0.cpp, specifying the mode as ‘c`. 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) sequencing-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, nucleotide-read quartets are shown for each individual in each of the columns.

Output file. The output file is also a tab-delimited file. The meanings of the first five columns are: 1) scaffold (chromosome) identifier; 2) site identifier (coordinate); 3) major-allele frequency; 4) sequencing-error rate estimate; 5) depth of coverage in the population sample (sum of the coverage over the individuals. Thereafter, the called genotype is shown for each individual in each of the columns.

Reference

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

Maruki, T., and Lynch, M., (in press) Genotype-calling from population-genomic sequencing data. *G3: Genes | Genomes | Genetics*.

Instructions

Below are specific procedures for using the program:

1. Make the input file, using GFE_v2.1 in the `c' mode.

2. Compile the program.by typing the following command:

g++ -o BGC_v1.1 BGC_v1.1.cpp -lm

3. Run the program by typing the following command:

./BGC_v1.1 -in In_BGC.txt -out Out_BGC.txt

- In_BGC.txt and Out_BGC.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 coverage and maximum allowed coverage to call a genotype of an individual can be specified by adding the `-min_cov' and `-max_cov' options, respectively. Their default values are 1 and 2,000,000,000, respectively.

- The posterior probabilities of the genotypes for each individual can be shown in the output by adding the `-gp' option. When this value is set at one, the posterior probabilities of the major and minor homozygotes separated by a slash are shown for each individual.

- Genotypes can be called without conditioning on significant polymorphisms by setting the value of the `-as' option at one (1). For this purpose, the input file needs to be made using GFE_v2.1 in the c mode also setting the value of the `-as' option at one (1).

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

./BGC_v1.1 -h

Copyright notice

Update from BGC

• The output at all sites on the reference sequence can be shown by setting the value of the `-as' option at one (1), in which case the input file also needs to contain all sites on the reference. In this case, genotypes are called without conditioning on significant polymorphisms.

Contact

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

1. Brief explanation of the problem.

2. Command entered.

3. Part of the input file.

4. Part of the output file.