May 27, 2015

GFE.cpp

This C++ program, written by Takahiro Maruki, estimates allele and genotype frequencies 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. For each site, ML estimates are obtained for the allele frequencies (under the assumption of no more than two alleles per site, these are by definition the two most abundant nucleotides in the population sample), error rate (due to all sources of error, not simply base quality), and disequilibrium/inbreeding coefficient. From the allele frequency and disequilibrium/inbreeding coefficient estimates, the program also estimates genotype frequencies. 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.

Input file. The input file is a tab-delimited text file, consisting of the reference nucleotide and individual nucleotide read quartets at every position (see the example input file In_GFE.txt). The first and second columns are the scaffold and position identifiers. The third column denotes the nucleotide of the reference genome. Thereafter, the nucleotide read quartet is presented in each column per individual.

Individual pro files in the 6-column format, which show the nucleotide quartet at each site, and the FASTA file (http://blast.ncbi.nlm.nih.gov/blastcgihelp.shtml) of the reference sequence, used for mapping sequence reads, are needed to make the input file. The individual pro files can be made from individual mpileup files (http://www.htslib.org/doc/samtools-1.2.html) using sam2pro written by Bernhard Haubold (http://guanine.evolbio.mpg.de/mlRho/sam2pro_0.6.tgz).

Output file. The output file is also a tab-delimited text file, in this case consisting of 22 columns. Column: 1) scaffold (chromosome) identifier; 2) site identifier (coordinate); 3) nucleotide of the reference sequence; 4, 5) nucleotides of major and minor alleles; 6) depth of coverage in the population sample (sum of the coverage over the individuals); 7) effective number of sampled chromosomes; 8) number of individuals with at least one read 9, 10) ML estimates of the major and minor allele counts; 11, 12) ML estimates of the major and minor allele frequencies ($\hat{p}$ and $\hat{q}$); 13 14) ML estimates of the error rate under the full model and null model assuming monomorphism; 15) ML estimate of the disequilibrium coefficient; 16) ML estimate of the inbreeding coefficient; 17, 18, 19) ML estimates of the frequencies of major homozygotes, heterozygotes, and minor homozygotes; 20) per-site heterozygosity estimate $2\hat{p}\hat{q}$; 21) likelihood-ratio test statistic for polymorphism; 22) likelihood-ratio test statistic for HWE deviation. These column definitions can be found in the Excel file: Meaning_Columns_GFE.xlsx.

The likelihood-ratio test statistics for polymorphism and HWE-deviation are expected to asymptotically correspond to chi-squared distributions with two and one degrees of freedom, respectively.

Reference

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

Maruki, T., and Lynch, M., Genotype-frequency estimation from high-throughput sequencing data. (in revision).

Instructions

Below are the specific procedures for using the software package:

1. Extract the nucleotide at every position of the reference sequence in the FASTA format using the Perl script

```
Ext_Ref_Nuc.pl
```

by typing the following command:

```
perl Ext_Ref_Nuc.pl reference_file_name output_file_name
```

- For example, to make an output file named RefNuc.txt from an input file named Reference.fa, type the following:

```
perl Ext_Ref_Nuc.pl Reference.fa RefNuc.txt
```

2. Make a file showing a nucleotide read quartet separated by slashes at each position for each individual from its pro file using the C++ program:

```
Make_InFile.cpp.
```

- Make an input file showing the name of the reference file (made in step 1), number of individuals in the analysis, names of the individual pro files, and individual IDs. See the example file (In_MIF.txt).

- The order of the individual IDs needs to correspond to that of the names of the individual profiles in the input file.

- Compile the program by the following command:

```
g++ Make_InFile.cpp -o Make_InFile -lm
```

- Run the program by the following command:

```
./Make_InFile $(< In_MIF.txt)
```

3. To make the input file for the genotype-frequency estimator, combine the reference file (made in step 1) and individual files (made in step 2) together using the UNIX command paste:

```
paste reference_file_name In_GFE_*.txt > In_GFE.txt
```

- This yields the input file for the genotype-frequency estimator In_GFE.txt, which can be viewed as an example provided.

4. Compile and run the genotype frequency estimator (GFE.cpp).

- Compile the program using the following command:

```
g++ GFE.cpp -o GFE -lm
```

- Run the program using the following command:

```
./GFE
```

- The names of the input and output files can be specified. Their default names are In_GFE.txt and Out_GFE.txt, respectively. For example, to make an output file named My_out_GFE.txt from an input file named My_in_GFE.txt , type the following:

```
./GFE -in My_in_GFE.txt -out My_out_GFE.txt
```

- The input file of the LD estimator (Maruki and Lynch 2014) can be prepared by specifying the "l" mode. For example, to prepare a file named My_out_l_GFE.txt from an input file named My_in_GFE.txt for the LD analysis, type the following:

```
./GFE -in My_in_GFE.txt -mode l -out My_out_l_GFE.txt
```

The output file in the l mode is conditioned on significant polymorphism, and is a tab-delimited text file, where 10 columns on the information on the site are followed by individual read quartets at each site. The meanings of the first 10 columns are as follows: 1) scaffold (chromosome) identifier; 2) site identifier (coordinate); 3) nucleotide of the reference sequence; 4, 5) nucleotides of major and minor alleles (1:A, 2:C, 3:G, and 4:T); 6) depth of coverage in the population sample (sum of the coverage over the individuals); 7) ML estimates of the major allele frequency ($\hat{p}$); 8) ML estimates of the error rate; 9) likelihood-ratio test statistic for polymorphism; 10) likelihood-ratio test statistic for HWE deviation. The chi-square critical value for the polymorphism test can be specified by adding the "-cv value" option to the command line, where value is the particular value specified. The default critical value is 5.991 (at the 5% level).

Copyright Notice

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.