PooledAlleleFrequencyEstimator.cpp

This program, written in C++, uses a maximum-likelihood (ML) procedure to estimate the allele-frequencies from the numbers of the four nucleotides (quartets) observed at individual genomic sites. For each site, the major and minor nucleotides are identified, their frequencies are estimated by ML, and the polymorphism is tested for statistical significance.

The designated major nucleotide is simply the one with the highest rank, and the minor nucleotide is the one with the second highest rank. If the top three ranks are all equal, the site is treated as unresolvable, with both major and minor nucleotides designated in the output by a *. If the second and third ranks are equal but lower than the major-nucleotide count, the site is treated as monomorphic, with the minor nucleotide again designated by a *.

Input file. Consists of six tab delimited columns, one for each site: the first two entries are arbitrary identifiers (e.g., scaffold number, and site), and the final four are integer values for the number of times an A, C, G, and T was observed at the site.

If the user wishes to include additional upfront columns for data identification, the following line must be edited in the program:

```
while ( fscanf(instream,"%s\t%s\t%i\t%i\t%i", id1, id2, &n[1], &n[2], &n[3], &n[4]) != EOF) {
```

The default name of the input file is "datain.txt", and this file should be placed in the same location (directory) as the program. If an alternative file name is desired, the following line must be edited:

```
instream = fopen("datain.txt", "r");
```

Output file. Columns 1,2) = site identifiers; 3,4) = designated major and minor nucleotides; 5,6) = major- and minor-nucleotide frequencies; 7) = estimated error rate; 8) = total coverage at the site; 9) = likelihood-ratio test statistic for polymorphism. Output columns are tab delimited.

Under the assumption of a chi-square distribution for the test statistic with one degree of freedom, significance at the 0.05, 0.01, 0.001 levels requires that the likelihood-ratio test statistic exceed 3.841, 6.635, and 10.827, respectively.

In principle, the 95% support interval can be obtained by determining the changes in the estimate of the minor allele frequency in both directions required to reduce the log likelihood by the appropriate chi-square value (e.g., 3.841) although this is not currently implemented.

The default name of the output file is "dataout.txt", and this file will appear in the same location as the program. If an alternative file name is desired, the following line must be edited:

```
outstream = fopen("dataout.txt", "w");
```

Reference:

Lynch, M., D. Bost, S. Wilson, T. Maruki, and S. Harrison. 2014. Population-genetic inference from pooled-sequencing data. Genome Biol. Evol. 6: 1210-1218.