README file for the program MLHKA.

Program summary
MLHKA is a program written by Stephen I. Wright.
It conducts maximum likelihood analysis of multilocus polymorphism and divergence data for testing for the action of natural selection on candidate genes. For any assistance running the program, and for feedback please email stephen.wright@utoronto.ca. To cite this program, please cite Wright, S.I. and Charlesworth B. 2004. The HKA test revisited: a maximum likelihood ratio test of the standard neutral model. Genetics 168:1071-1076.

Input File Format

The input file must be called 'infile.txt', and contains numloci+1 lines, where numloci is the number of loci used in the analysis. An example file is available for download,
which is the dataset analysed in the manuscript. The first line of the input file gives the following
information, separated by spaces:

numloci numselectedloci (selectedlocusID1 selectedlocusID2 .... selectedlocusIDn) startingT

numloci is the number of loci in the analysis, and numselectedloci is the number of loci hypothesized to be under selection in the model. This second number must be less than or equal to numloci-1, since at least one locus must be assumed to be neutral for there to be any power for parameter estimation.
If this value is zero, a strictly neutral model is run, assuming all loci follow the standard neutral model. If numselectedloci is greater than 0, the next entries are the locus numbers for the candidate selected loci, defined by the order of their appearance in the subsequent lines. In the example infile provided, there is one selected locus, and it is locus At1g36310. startingT is the starting value of the divergence time parameter T (in units of 2N generations), for the Markov chain. Provided enough chains are run, the starting values should not matter, but reasonable starting values could be obtained using a standard HKA test, or from independent information. The following lines in the input file give the information for each locus:

locusID L S n D startingtheta inheritancescalar

Where locusID is the locus name, L is the size of the region analysed (in bp), S is the observed number of segregating sites, n is the sample size, D is the pairwise number of between-species differences(from a random sequence
from each species), and startingtheta is the starting value of theta per base pair for the Markov chain. Inheritance scalar is a correction factor when loci differ in their effective population size (e.g. for an autosome, use 1, and an X chromosome locus use 0.75).

Running the program

When you run the program, you will be prompted for two values; the first is the random number seed, which is a random negative number. The second value is the 'chain length', or the number of cycles of the Markov chain. The appropriate number will depend on the number of loci examined, and it is important to run the program multiple times with different seeds (additionally, the starting parameter values can also be modified), to ensure that the chain length is long enough to obtain the same results. Chain lengths of at least 100,000 are

recommended for any numloci>2. While running, the program outputs its progress (the number of chains run in increments of 1000).

Output

The output will be found in a file called 'likelihoods.txt'. It will give the following output:

ML T theta(locusID1) k(locusID1) theta(locusID2) k(locusID2) ..... theta(locusIDn) k(locusIDn)

Where ML is the value of the maximum ln likelihood, T is the maximum likelihood estimate of the divergence time parameter, thetax is the maximum likelihood estimate of theta for locus x, and kx is the maximum likelihood estimate of the selection parameter k. If this locus is not a candidate selected locus in the model run, kx will always equal 1.

Testing for selection using a likelihood ratio

To test for selection, run the program under a neutral model, where numselectedloci=0, and then under a selection model, where numselectedloci>0. Significance can be assessed by the likelihood ratio test, where twice the difference in log likelihood between the models is approximately chi-squared distributed with df equal to the difference in the number of parameters (in this example, df=numselectedloci). Similarly, further nested likelihood ratio tests can be performed, for example to test for the presence of two vs. 1 locus under selection.