

Genomic Hudson-Kreitman-Aguadé test

Innan (2006)

"Modified Hudson-Kreitman-Aguadé test and two-dimensional evaluation of neutrality tests"

Genetics 151: 1157-1164

Hideki Innan

The Graduate Univ. Advanced Studies, Japan

innan_hideki@soken.ac.jp

1, Prepare data files

The test requires variation (polymorphism and divergence) data for a test locus and multiple reference loci, which should be in the text files named "testlocus" and "refloci", respectively. The format is identical for the two files: a single line is for one locus data (tab-delimited), which consist of "locus name", "nucleotide length (L)", "sample size (n)", "No. of segregating sites (S)", "Average no. of pairwise differences (K)", and "divergence from an outgroup species (d)". The folder "genomicHKA" includes "testlocus" and "refloci", which were used Innan (2006). The data are from *Arabidopsis thaliana*, and the test was applied to the GD2A gene with six reference loci, AP3, F3H, PI, AP1, CAL and FAH.

2, Compile the C-codes

Compile the two sets of Hudson's ms-based C-codes as

```
>> gcc -O3 -o msTdivPDF msTdivPDF.c streec.c rand3.c -lm
>> gcc -O3 -o msGenomicHKA msGenomicHKA.c streec.c rand3.c -lm
```

Now, we have two compiled programs, "msTdivPDF", and "msGenomicHKA".

3, Run the programs

i) Run msTdivPDF

msTdivPDF is to obtain the probability distribution of the time to speciation, conditional on the data from the reference loci. msTdivPDF will read "refloci", and produce a result file, "div.infile". To run msTdivPDF, a regular command for ms would be acceptable. The process will be described along with the example of the GD2A gene in *Arabidopsis thaliana*. msGenomicHKA is run with the following command:

```
>> ./msTdivPDF 20 10000 -t 1 -r 0.33 1000
```

The first integer, 20, is the sample size, which should be the largest sample size in the list of the reference loci in "refloci". The next is the number of acceptances in the rejection-sampling algorithm (see Innan 2006), and 10000 is recommended. -t and -r specify the population mutation and recombination parameters. It is suggested to give -t 1. This setting is arbitrary because msTdivPDF will use the average population mutation parameter in the reference loci, which is 0.0062 for the example data. -r 0.33 is set because the recombination rate per adjacent sites may be about 1/3 of the mutation rate per site according to Hagenblad and Nordborg (2002 Genetics 161: 289-298). Note that only the ratio of r to t matters here. An identical result would be obtained with -t 10 -r 3.3. The last number would be any positive integer, which will not be used in the simulation. In the original ms software, this number is the length of simulated region, but msTdivPDF will replace this number by those in "refloci".

ii) Run msGenomicHKA

After msTdivPDF produced "div.infile", msGenomicHKA will read it to compute the P-value for the test locus, which is defined as the proportion of simulation replications with K/d less than the observed K/d at the test locus. With the example data, msGenomicHKA can be run with following command:

```
>> ./msGenomicHKA 16 10000 -t 3.3 -r 1.1 533
```

The first integer is again the sample size of the test locus, which has to be identical to that in "testlocus". The second is the number of simulation replications, and 10000 is suggested. The population mutation rate is set to be 3.3, which is derived by 0.0062×533 bp where 0.0062 is the average over the reference loci. The recombination rate is assumed to be 1/3 of the mutation rate, so -r 1.1 is given. The last number is the nucleotide length of the simulated region, which is 533 for the GD2A gene. The obtained one-tailed P-value will be written in "gHKA.out".