Getting Started with Rhh

Rhh version 1.0.1 | June 30, 2010

This document was authored by Jussi Alho and Kaisa Välimäki

DISCLAIMER

Although Rhh has been tested and, to the best of our knowledge, works without problems, the authors take no responsibility for any bugs, errors, or any harm or damage caused by the software or the documentation. The software and the documentation come without any warranty; without even the implied warranty of merchantability or fitness for a particular purpose. Use the software and the documentation at your own risk.

INTRODUCTION

Rhh is a free, open source extension package for the statistical computing software R (*http://www.r-project.org*) for calculating multilocus heterozygosity measures and the heterozygosity-heterozygosity correlation in molecular ecological and evolutionary studies.

Estimates of individual multilocus heterozygosity have been used as proxy measures of inbreeding, but these estimates are expected to correlate only weakly with inbreeding coefficients calculated from the pedigree. Therefore, an inbreeding signal of a set of microsatellite markers should be tested before using multilocus heterozygosity measures in studies of inbreeding. This can be done by calculating the heterozygosity-heterozygosity correlation – i.e. repeatedly and randomly dividing the loci into two groups, calculating an estimate of individual multilocus heterozygosity for the two set of loci, and obtaining the mean correlation between two groups, which should be significant and positive if an inbreeding signal is present.

Rhh includes functions to estimate heterozygosity-heterozygosity correlation and to calculate three estimates of individual multilocus heterozygosity: homozygosity by loci, internal relatedness, and standardized heterozygosity.

HOMEPAGE

Rhh has a homepage at *http://www.helsinki.fi/biosci/egru/research/software*

CITING RHH

A computer note about Rhh has been published in an international journal, Molecular Ecology Resources. If you use Rhh in research, the package should be cited as:

Alho JS, Välimäki K, Merilä J (2010) Rhh: an R extension for estimating multilocus heterozygosity and heterozygosity–heterozygosity correlation. *Molecular Ecology Resources* **10**: 720–722.

INSTALLATION

Platform requirements

Rhh is an extension package for open source statistical computing software R (*http://www.r-project.org*) and requires that R is installed on your computer. R is freely available for several platforms, including UNIX, Windows and MacOS.

Rhh will work on any platform running R, but instructions for installation are given here only for Windows. Please refer to R website for installing extension packages on other platforms.

Installing R for Windows

Go to the R website (*http://www.r-project.org*) and click the *CRAN* link under *Download* in the left menu of the frontpage. A list of mirror servers of Comprehensive R Archive Network (CRAN) should appear. Choose one that is close to you and click the appropriate link.

You should now be on a page titled *The Comprehensive R Archive Network*. You want to download and install the latest precompiled binary distribution of the R base system for Windows, so click *Windows* link under *Download and Install R* on the top of the page.

On the page *R for Windows*, click *base* subdirectory. A page titled *R-2.10.0 for Windows* (or similar with later version number) opens. Click the link *Download R 2.10.0 for Windows* (or similar) and download the file *R-2.10.0-win32.exe* to your computer. After the download is completed, run the file. This will install R on your computer. The default options for the installation are fine.

Installing Rhh

When R is installed on your computer, launch R graphical user interface (RGui) from the Windows Start menu (*Start > All Programs > R > R 2.9.1* or something similar). Click *Packages > Set CRAN mirror…* in RGui and choose a server close to you. Click *Packages > Install package(s)…* and choose *Rhh* from the list that opens. The extension package will be installed when you click *ok*.

USING RHH

Data format

The input data file is assumed to be a tab or space delimited text file without a header line. Each row represents an individual. The first column is reserved for individual IDs. Every column-pair thereafter is assumed to contain the genotype information of a locus, i.e. two character strings each denoting either an allele or missing data. The user can decide the character string for missing data ("NA" or "0" are widely used).

Please note that an example data file and a script that runs the examples below are available for download from the program's homepage at *http://www.helsinki.fi/biosci/egru/research/software*. The script can be used as a template for your own analysis.

Getting multilocus heterozygosity estimates

Launch R graphical user interface (RGui) from the Windows Start menu (*Start > All Programs > R > R 2.9.1* or similar).

Load Rhh in RGui (*Package > Load package…*)

Change the working directory to the directory containing your input data file (*File > Change dir…* when R Console is the active, or selected, sub-window).

Type the following in the R Console (replacing the file names with your chosen input and output file names):

```
h <- mlh("example_input.txt", "example_output.txt", "0", 4)
```

This command will open the input data file *example_input.txt*; calculate homozygosity by loci, internal relatedness and standardized heterozygosity for each individual; save the results in the output file *example_output.txt* in your working directory; and save the result array in the variable *h*. Any genotype values of 0 in the input data file will be treated as missing data. *If the output file already exists, it will be overwritten, so be careful!*

The output file will contain four columns: individual IDs (column header: ID), internal relatedness (IR), standardized heterozygosity (SH) and homozygosity by loci (HL). The result array will contain the same columns. If a multilocus heterozygosity measure could not be calculated for an individual, the resulting value for that individual is NaN (not-a-number).

Getting heterozygosity-heterozygosity correlation

We assume you have launched RGui, loaded Rhh and changed the working directory as described in the previous section.

Type in the R Console (replacing the file name with your chosen input file name):

```
r <- h_cor("example_input.txt", "0", 1000, "hl")
```

This command will open input data file *example_input.txt* and save 1000 random samples of heterozygosity-heterozygosity correlation in the variable *r*. Any genotype values of 0 in the input data file will be treated as missing data. You can choose whether the correlation is calculated using internal relatedness ("ir"), standardized heterozygosity ("sh") or homozygosity by loci ("hl"). You can also change the number of samples. No output is produced at this stage. *Note that the execution of this command might take a very long time depending on the number of individuals, loci and samples – you might want to try with a low number of samples first.*

After running the previous command, use R commands below to obtain the mean correlation, 95% confidence interval and histogram, respectively.

```
mean(r)
quantile(r, probs=c(0.025, 0.975))
hist(r)
```

Number of alleles

When you run either the *mlh* or the *h_cor* function, the program produces a file called *number_of_alleles.txt* containing the number of alleles for each locus, and saves it in your

working directory. Each number is given on a separate row, so if you run the program with ten loci, you get a file with ten rows each representing the number of alleles for that particular locus. *If* number_of_alleles.txt *already exists in your working directory, it will be overwritten, so be careful!*

Further topics

Rhh has also some other functions that provide more flexibility than the functions introduced above. In addition to the functions introduced here, the package includes functions for calculating the three individual multilocus heterozygosity estimates separately, validating the input data file, and generating a randomized sample of heterozygosity-heterozygosity correlations. These are used internally by *mlh* and *h_cor* functions but people with R programming skills are able to take advantage of these directly as well. You might want to read the help files of the Rhh by typing to the R Console (after Rhh is loaded as described above):

```
?Rhh
```