GENHET: USER MANUAL v2.1

Aurélie Coulon, February 26th, 2009

GENHET is a function written for the program R (Ihaka and Gentleman 1996).

It **calculates 5 different individual heterozygosity estimates**:

- proportion of heterozygous loci (PHt) in an individual: PHt = number of heterozygous loci / number of genotyped loci

- standardized heterozygosity based on the mean expected heterozygosity (Hs_exp, Coltman 1999): Hs_exp = PHt / mean expected heterozygosity of typed loci

- standardized heterozygosity based on the mean observed heterozygosity (Hs_obs): Hs_obs = PHt / mean observed heterozygosity of typed loci

- internal relatedness (IR) (Amos 2001): $IR = (2H - \Sigma f_i) / (2N - \Sigma f_i)$, where H is the number of loci that are homozygous, N is the number of loci and f_i is the frequency of the *i*th allele contained in the genotype

- homozygosity by locus (HL) (Aparicio 2006): $HL = \Sigma E_h / (\Sigma E_h + \Sigma E_j)$, where E_h and E_j are the expected heterozygosities of the loci that an individual bears in homozygosis (*h*) and in heterozygosis (*j*), respectively.

This user manual is a complement to the comments already provided in the file GENHET.R, which contains the code of the function.
You can open GENHET.R with a text editor like Crimson Editor (freely downloadable at http://www.crimsoneditor.com) or, simply, Notepad.
This user manual is mainly aimed at people who are not familiar with R. Lines of command for running GENHET are marked in blue. Within those blue commands, the words ***in bold and italicized*** are examples of file names that have to be changed to your own file names.

1st step: preparation of the input files:

1st file: Individual genotypes:

You need to **create** a file containing the genotypes of each individual:

- the first column should contain the individuals' identifiers;
- the following columns contain the genotypes: 2 columns per locus: 1 for each allele; alleles have to be coded as **numeric values**; missing data must be coded as "NA";
- the first row may contain the names of the columns, but that's not mandatory (see below).

Save as a text file, separators = tabulators. See "exGENHETgenotinput.txt" for an example of genotype file with a header.

To **import your file in R:** open R, specify your working directory, i.e. the location where your input files are situated and/or where you want the output files to be saved (to do so: on a PC go to: file → change directory, and then browse; on a Mac go to: misc → change working directory, and then browse). Then, use the following command if you have a header, i.e. the first row of your file contains the names of the columns (don't forget to replace the file name in bold and italicized by the name of your input file):

datest=read.table("***exGENHETgenotinput.txt***",sep="\t",header=T)

If there is no header, use:

datest=read.table("***exGENHETgenotinput.txt***",sep="\t",header=F)

(Note: Once your data file is imported in R, it is called an "object". In the example above, the object created is called "datest".)

2nd file: locus names OR allele frequencies

To calculate individual heterozygosity estimates, GENHET needs allele frequencies; here, you have **two options**: the simplest one is to ask GENHET to calculate the allele frequencies in your file of genotypes; but in some cases (e.g. high number of related individuals in the dataset), one may want the estimates of heterozygosity to be based on allele frequencies of a different dataset. It is hence possible to provide GENHET with a file of allele frequencies.

A. If you choose the **first option** you have to provide a file giving the name of the loci (in the same order as in the file of genotypes). See "exGENHETallist.txt" for an example. To import this file in R, use the following commands:

locusname=scan("***exGENHETallist.txt***",what="character",sep="\t")

B. If you choose the **second option**, you don't need to provide the names of the loci; but instead you have to provide a file containing allele frequencies. The format of this file has to be the following:

same number of columns as the number of loci, and same number of rows as the total number of alleles (over the different loci); each cell contains either the frequency of the allele or 0 (when, in the dataset, the locus considered does not have the allele considered); the first column contains the names of the alleles; the first row contains the names of the columns, i.e. the word "alleles" for the first cell, and then the names of the loci.

See "exGENHETalfreqinput.txt" for an example.

To get this type of file you can, for example, use my function ALF (for R), provided on the same website as GENHET; or a program like Genalex (Peakall and Smouse 2006).

B1. If you use ALF, see the indications provided at the top of the code file (ALF.R). After copying and pasting the code of ALF in an R window, the commands you will have to use will look like:

```
locusname=scan("exGENHETallist.txt",what="character",sep="\t")
alfreqtest=ALF(geno=datest,locname=locusname)
```

The output object alfreqtest will be ready to be used by GENHET.

B2. If you use another program to build the file of allele frequencies, save it as a text file with tab separations, and import it in R:

```
alfreqtest=read.table("exGENHETalfreqinput.txt",sep="\t",header=T)
```

2nd step: loading of GENHET into R

You simply need to open the file containing the code of the function (GENHET.R) in a text editor like Crimson Editor (freely downloadable at http://www.crimsoneditor.com) or, simply, Notepad; copy and paste all the text in the R window.

3rd step: launching of GENHET

If it is the first time you are using GENHET, you will have to ask R to download the gtools package from the CRAN website (CRAN is the network of ftp and web servers around the world that store versions of code and documentation for R). To do so, on a PC go to Packages → Install Packages; and follow the instructions to choose and download gtools. On a Mac go to Packages and Data → Package Installer; and follow the instructions to choose and download gtools.

To launch the function, you have to tell R how to name your output object, and the value of the different arguments of the function. GENHET has four arguments:

- dat: this is the object containing the genotypes (called datest in our example)
- estimfreq: binary variable taking the value "T" (true) or "F" (false); as mentioned above, to calculate individual heterozygosity estimates, GENHET needs allele frequencies; if you want GENHET to estimate and use the allele frequencies of dat, use estimfreq="T"; if you want to use allele frequencies based on a different dataset, use estimfreq="F" (and provide a file of allele frequencies *via* the argument alfuser, see below)
- locname: this is a vector of the names of the different loci (in the same order as in dat) (called locusname in our example); only used if estimfreq="T" → if estimfreq="F" this argument does not have to be specified
- alfuser: if estimfreq="F", this is the object containing the allele frequencies (called alfreqtest in our example); if estimfreq="T" this argument does not have to be specified.

Below are two examples:

- example 1: allele frequencies estimated by GENHET:

Htest=GENHET(dat=datest,estimfreq="T",locname=locusname)

- example 2: allele frequencies provided by the user:

Htest=GENHET(dat=datest,estimfreq="F",alfuser=alfreqtest)

In the two examples above, the output object will be called Htest.

4th step: exporting the results

The output object is a matrix giving, for each individual, the 5 different estimates of individual heterozygosity. Each row starts with the name of the individual.

You can look at it in R by calling it, just by writing the name of the object:

in the example above:

Htest

will print the table on the screen.

But, more useful, you can export this output object as a text file:

write.table(Htest,"***Htest.txt***",sep="\t",row.names=F,quote=F)

Function restrictions

There is no restriction on the number of loci, alleles or individuals.

Enjoy!

Send me any comments or suggestions at acoulon[at]mnhn[dot]fr.

References

Amos W., Worthington Wilmer J., Fullard K., Burg T. M., Croxall J. P., Bloch D., Coulson T. 2001. The influence of parental relatedness on reproductive success. Proceedings of the Royal Society B: Biological Sciences. 268: 2021-2027.

Aparicio J. M., Ortego J., Cordero P. J. 2006. What should we weigh to estimate heterozygosity, alleles or loci? Molecular Ecology. 15: 4659-4665.

Coltman D. W., Pilkington J. G., Smith J. A., Pemberton J. M. 1999. Parasite-mediated selection against inbred Soay sheep in a free-living, island population. Evolution. 53: 1259-1267.

Ihaka R., Gentleman R. 1996. R. A language for data analysis and graphics. Journal of Computational and Graphical Statistics. 5: 299-314.

Peakall R., Smouse P. E. 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes. 6: 288-295.