BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 8 2012, pages 1176–1177
doi:10.1093/bioinformatics/bts115

Genetics and population analysis
Advance Access publication March 7, 2012

rehh : an R package to detect footprints of selection in genome-wide SNP data from haplotype structure

Mathieu Gautier[1], and Renaud Vitalis[2]
[1]INRA and [2]INRA-CNRS, UMR CBGP (INRA – IRD – Cirad – Montpellier SupAgro), Campus International de Baillarguet, CS 30016, F-34988 Montferrier-sur-Lez Cedex, France

Associate Editor: Martin Bishop

ABSTRACT

Summary: With the development of next-generation sequencing and genotyping approaches, large single nucleotide polymorphism haplotype datasets are becoming available in a growing number of both model and non-model species. Identifying genomic regions with unexpectedly high local haplotype homozygosity relatively to neutral expectation represents a powerful strategy to ascertain candidate genes responding to natural or artificial selection. To facilitate genome-wide scans of selection based on the analysis of long-range haplotypes, we developed the R package *rehh*. It provides a versatile tool to detect the footprints of recent or ongoing selection with several graphical functions that help visual interpretation of the results.

Availability and implementation: Stable version is available from CRAN: http://cran.r-project.org/. Development version is available from the R-forge repository: http://r-forge.r-project.org/projects/rehh. Both versions can be installed directly from R. Function documentation and example data files are provided within the package and a tutorial is available as Supplementary Material. *rehh* is distributed under the GNU General Public Licence (GPL $\geq$ 2).

Contact: mathieu.gautier@supagro.inra.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 29, 2011; revised on February 9, 2012; accepted on March 5, 2012

1 INTRODUCTION

In a pioneering work, (Sabeti *et al.*, 2002) investigated the genetic footprints of recent positive selection in humans by analyzing long-range haplotypes at several candidate genes using a novel measure called the extended haplotype homozygosity (*EHH*). *EHH* is defined as the probability that two randomly chosen chromosomes carrying the same allele at a focal SNP (single nucleotide polymorphism) are identical by descent over a given distance surrounding it. If the core allele is under selection, then the *EHH* is expected to be close to one over a large distance upstream and downstream the focal SNP. However, testing the departure of *EHH* from neutral expectation remains difficult without making strong assumptions about the population demographic history. Voight *et al.* (2006) therefore proposed an empirical test based on the integral of the observed decay of *EHH*, which they defined as integrated *EHH* (*iHH*). They further defined a test statistic (*iHS*) as the log-ratio of *iHH* computed at the derived and the ancestral focal SNP alleles. The *iHS* is standardized using the average and standard deviation values over all SNPs with similar allele frequencies. Because such within-populations measures have low power when the frequency of the selected allele is high, Tang *et al.* (2007) developed a similar procedure to compare the *EHH* profiles between populations. Their approach consists in computing for each SNP in each population a weighted average of the *EHH* at both alleles, referred to as site-specific *EHH* (*EHHS*). The observed distribution of the standardized log-ratio of the integrated *EHHS* (*iES*) between pairs of populations (referred to as *Rsb*) are then used to detect signals of positive selection, *i.e.* genomic regions with unusually high *Rsb*.

EHH-based tests were proved remarkably efficient to identify relevant footprints of recent selection in humans (Tang *et al.*, 2007; Voight *et al.*, 2006) and other species (e.g. Gautier and Naves, 2011). Although the `sweep` software (Sabeti *et al.*, 2002) provides utilities to compute and visualize the decay of *EHH*, to our knowledge, however, no software package is available to compute all these statistics from large-scale datasets. To facilitate genome-wide scans for footprints of selection using *EHH*-based tests we, therefore, developed the package *rehh* for the statistical software package R (R Development Core Team, 2008). R is becoming a standard for the analysis of genetic data, and R packages are portable to most operating systems (Windows, Mac OS X and Linux). We briefly present in the following the main functionalities of the *rehh* package using data from a recently published study on cattle breeds (Gautier and Naves, 2011). Methods and functionalities implemented in the *rehh* package are described in more details in a tutorial available as online Supplementary Material.

2 DESCRIPTION

2.1 Input Data

rehh requires a SNP information file (SNP name, map positions and ancestral and derived alleles) and a genotype data file (with phased haplotypes) for each population(s) of interest. For this latter, *rehh* accepts two input formats (see the *rehh* tutorial), including *fastphase* (Scheet and Stephens, 2006) output files. The `data2haplohh()` function imports the data into an object of class `haplohh`.

2.2 Analyses

At a given focal SNP, the `calc_ehh()` function computes the *EHH* statistics at both alleles for all neighbouring SNPs, as well as the corresponding *iHH* statistics. Likewise, the `calc_ehhs()` function computes the *EHHS* at a given focal SNP for all

*To whom correspondence should be addressed.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

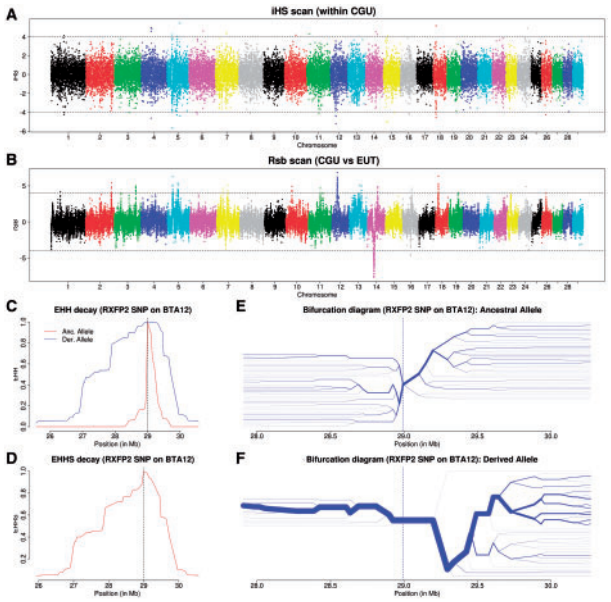

Fig. 1. Example of results obtained with the *rehh* package on a previously published dataset (see the main text).

neighbouring SNPs and the resulting *iES*. Both the `calc_ehh()` and the `calc_ehhs()` functions provide optional plots of the results, which depict, respectively, the decay of the *EHH* and that of the *EHHS* from the focal SNP. The `scan_hh()` function allows computing *iHH* statistics for both alleles and the *iES* for all the SNPs present in the chromosome under study. For efficiency, the computation of *EHH*-based statistics was encoded in C and wrapper functions have been developed to interface the resulting compiled code in R. As a result, scanning the example dataset consisting in 140 individuals genotyped at 1424 SNPs with the `scan_hh()` function only takes 3.3 s on a standard PC running with a 3.2 GHz processor. The `scan_hh()` function may therefore be used to analyze efficiently large SNP datasets. In order to perform genome-wide scans, we recommend to analyze each chromosome in turn using the `scan_hh()` function, and then to concatenate the resulting matrices of *iHH* and *iES* (see the *rehh* tutorial). The `ihh2ihs()` and `ies2rsb()` functions can then be used to compute the standardized statistics. Some options are available to fiddle the standardization in these two functions. To that end, departure from normality of the standardized score distributions can be visually inspected using the `distribplot()` function. Both the `ihh2ihs()` and `ies2rsb()` functions provide optional plots of the results, which depict, respectively, the ordered values of *iHS* and *Rsb* along the genome. Last, the `bifurcation.diagram()` function draws haplotype bifurcation diagrams (Sabeti *et al.*, 2002), which allow visualizing the breakdown of linkage disequilibrium at increasing distances from the focal core allele.

3 EXAMPLE

Figure 1 illustrates the results obtained using a previously published dataset (Gautier and Naves, 2011) consisting in 725 individuals from various cattle breeds genotyped at 44 057 SNPs spanning the 29 bovine autosomes. Estimates of *iHH* and *iES* at each SNP are provided in the *rehh* package for the CGU (Creole breed from Guadeloupe) and EUT (eleven different European breeds) populations. Figure 1A shows the outputs of the *iHS* scan within the CGU population using the `ihh2ihs()` function. In particular, Figure 1A suggests a footprint of selection on chromosome 12 (BTA12), around position 28.99 Mb. This signal is even more striking when the *Rsb* scores, computed between the CGU and EUT haplotypes using the `ies2rsb()` function, are examined (Fig. 1B). Figure 1C, which is an output of the `calc_ehh()` function, shows a smaller decay of *EHH* for the ancestral allele than the derived allele for the focal SNP located at the *iHS* peak position 28.99 Mb. This trend is also apparent from the decay of *EHHS* depicted in Figure 1D, which was obtained using the `calc_ehhs()` function. Finally, Figure 1E and F shows the bifurcation diagrams, which were obtained by running the `bifurcation.diagram()` function for both the ancestral and the derived core alleles at this same focal SNP. This graphical representation allows visualizing the breakdown of linkage disequilibrium on core haplotypes. The thickness of the lines is proportional to the frequency of each haplotype, which therefore informs on haplotype diversity. Comparing Figure 1E and F suggests that the favourable variant is associated with the SNP derived allele. Interestingly, this position is $<$10 kb upstream of the RXFP2 gene on BTA12 (Gautier and Naves, 2011). Note that BTA12 SNP map positions and the 280 CGU haplotypes are provided as example datasets in the *rehh* package.

4 CONCLUSION

Identifying genomic regions with unexpectedly high local haplotype homozygosity relatively to neutral expectation is a powerful strategy to ascertain candidate genes responding to natural or artificial selection. The availability of large SNP haplotype datasets in a growing number of both model and non-model species makes it possible to apply such a strategy. In this context, *rehh* provides a all-in-one user-friendly tool to detect the footprints of recent or ongoing selection using *EHH*-related statistics.

Funding: French ANR programme *EMILE* 09-BLAN-0145-01.

Conflict of Interest: none declared.

REFERENCES

Gautier,M. and Naves,M. (2011) Footprints of selection in the ancestral admixture of a new world creole cattle breed. *Mol. Ecol.*, **20**, 3128–3143.

R Development Core Team. (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Sabeti,P.C. *et al*. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.

Scheet,P. and Stephens,M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, **78**, 629–644.

Tang,K., *et al*. (2007) A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.*, **5**, e171.

Voight,B.F. *et al*. (2006) A map of recent positive selection in the human genome. *PLoS Biol.*, **4**, e72.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016