The hfunmap Package

Version 0.0.0.9
Mu Shuaicheng
The Center for Computational Biolog at Beijing Forestry University, China

1. Introduction

The hfunmap package is developed to identify quantitative trait loci (QTL) for multiple longitudinal, or vectorized phenotypic trait as based on the multi-trait Funmap model which include multidimensional Structured Antedependence model(SAD) [1].This guide gives some brief instructions on how to perform the tasks of QTL detection by the hfunmap package. The outline of this guide is as follows.

Section 2: Installation
Section 3: Data Format
Section 4: Methodoligical Framework
Section 5: Examples
Section 6: Results
Section 7: Function List
Section 8: References

We refer to Ma et al. (2002), Zhao et al. (2005), Cao et al. (2017) and Mu et al. (2022) for the theoretical foundation of this package. If you benefited from this software, then please cite the following paper in your work:

1.Ma C, Casella G, Wu R. Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. Genetics (Austin), 2002. 161(4): 1751-1762.

2.Zhao W, Hou W, Littell R C, Wu R. Structured Antedependence Models for Functional Mapping of Multiple Longitudinal Traits. Statistical Applications in Genetics and Molecular Biology, 2005.17(4).

3.Cao J, Wang L, Huang Z, Gai J, Wu R. Functional Mapping of Multiple Dynamic Traits. Journal of Agricultural, Biological and Environmental Statistics, 2017. 22(1): 60-75.

4.Mu SC, Zhu XL, Ye MX, Wu RL. Application of multi-traits functional mapping to a growing QTL in Populus euphratica Oliv. Seedlings. Plant Science Journal

2. Installation

The hfunmap package depends on the ***DEoptim***,***Rcpp***,***RcppArmadillo*** package (available on CRAN), so these packages should be installed firstly. To install hfunmap, download the package file and type the appropriate command below or click the menu item “Install packages from local zip/tar.gz files”.

*Windows OS:

```
>install.packages("C:/yourpath/hfunmap0.0.0.9000.zip", repos=NULL)
```

*Linux/Mac OS:

```
>install.packages("C:/yourpath/hfunmap0.0.0.9000.tar.gz",repos=NULL)
```

*Or you can install the package from the Github from the LINUX command line:

```
$ git clone https://github.com/~/hfunmap.git
$ cd hfunmap
$ R CMD INSTALL hfunmap'
```

Before the package is used in R, the package importation is necessary by the following command:

```
>library(hfunmap)
```

3. Data format

The hfunmap package can identify pleiotropic QTL loci and perform hypothesis testing on the basis of data files containing the appropriately formated genotype, phenotype, phenotypic measurement time, number of traits, genotype deletion information and number of genotypes. The appropriate formatting for these files is normalization by function ***hfun_load_data***. ## 3.1 Genotype file

SNP site ID	Sample_1	Sample_2	…	Sample_N
1	0	0	…	1
2	0	1	…	2
3	0	-9	…	1
…	…	…	…	…

Three genotypes (aa=0,Aa=1,AA=2) and missing data (coded as -9) are valid marker values.

3.2 Phenotype file

Sample_ID	trait1-time_1	trait2-time_1	trait3-time_1	trait4-time_1	trait1-time_2	trait2-time_2	trait3-time_2	trait4-time_2	…	trait1-time_T	trait2-time_T
1	0.566416	2.116529	0.06627	0.483592	0.664886	2.630344	0.104492	0.696641	…	0.780314	3.266037
2	0.576517	2.955668	0.386421	0.44901	0.688055	3.377169	0.484008	0.651814	…	0.821059	3.858024
3	0.543166	1.374082	0.066051	0.369401	0.634388	1.589104	0.089472	0.476965	…	0.740816	1.837551
4	0.328811	3.450271	0.208096	0.207915	0.391629	3.883384	0.266004	0.301783	…	0.466392	4.370299
…	…	…	…	…	…	…	…	…	…	…	…

Sample_ID	trait1-time_1	trait2-time_1	trait3-time_1	trait4-time_1	trait1-time_2	trait2-time_2	trait3-time_2	trait4-time_2	...	trait1-time_T	trait2-time_T
N	0.732933542	4.212860756	0.284289267	0.599249354	0.8527134	4.801084984	0.365075716	0.839045952	...	0.991912428	5.46893147

The phenotype table should be filled with numerical values and without missing value.

3.3 Measured time file

```
>c(1,3,5,7,9,11,13,16,19,23,28,32,39,47)
```

The measured time should be filled with numerical values and without missing values. All traits should have same measurement time, or you can standardize the measurement time points by mathematical function fitting before using ***hfun_load_data***

3.4 Number of trait

```
>int 4
```

You can set ranks by yourself or ***hfun_load_data*** can automatic setting by dim(phe)[2]/length(time).

3.5 Genotype deletion information

```
>logical TRUE/FALSE
```

In this framework only genotype file tolerate missing values.

3.6 Number of genotypes

```
>int 3
```

The parameter is automatic setting by ***hfun_load_data***.

4. Methodoligical Framework

Statistical model, hypothesis test and permutation will be explained in this section.

4.1 Statistical Model

The hfunmap is multi-trait QTL mapping methodology which has two fundamental points. The phenotypical trendency of the trait follows a logistic mathematic curve. In statistical view, the phenotypes of the trait at all time points follow a multivariate normal density. It can be described by the following equation. y is measure values at all time points, g can be considered a logistic mathematic curve.

$$f_j(y) = \frac{1}{(2\pi)^{T*ranks/2}|\Sigma|^{1/2}} exp[-(y-g_j)^T\Sigma^{-1}(y-g_j)/2]$$

The 2^{nd} point for the hfunmap is based on genetical law which QTL has different possibility for 2 or more genotypes depending on the position. The hfunmap tries to find a position where the minimum cumulation for the gap between measure value and curve value is achieved according to the genotype value. In calculation methods of the maximum-likelihood estimates are shown in the following equation, where δ_{ij} is the genotypic value for QQ, Qq, qq and so on, $f_j(y)$ is the gap between measure value and curve value.

$$logL(\Omega) = \sum_{i=1}^{N} log[\sum_{j=1}^{J} \delta_{ij} f_j(y_i)]$$

MLE algorithm is employed to estimate likelihood value and find QTL position for these curves. In addition, MLE algorithm also gives the estimation of the curve parameter for each genotype value. So the hfunmap can give the QTL position and estimated parameters at that QTL position. During the computation of MLE, multivariate normal distribution is involved to predict the probability of individual curve associated with the covariance structure which presents the correlation between the measured values at different traits or different time points. In the hfunmap package, we use multidimensional SAD1 model time dependent covariate matrix.

4.2 Hypothesis Test

The purpose of hypothesis test is to justify the existence of QTL and the difference between genotype expressions. The hfunmap implements the basic hypothesis test of QTL for all curves.

Hypothesis test
All parameters are identical for any genotype.
e.g. for logistic curve, with four traits,
$a_{11} = a_{12}, b_{11} = b_{12}, r_{11} = r_{12}$
$a_{21} = a_{22}, b_{21} = b_{22}, r_{21} = r_{22}$
$a_{31} = a_{32}, b_{31} = b_{32}, r_{31} = r_{32}$
$a_{41} = a_{42}, b_{41} = b_{42}, r_{41} = r_{42}$

In pratical computations, two source data sets, including phenotype data and genotype data, and loaded into a data object in R environment firstly by ***hfun_load_data***. Then hypothesis tests are performed on the data object to justify the existence of QTL position and do other evaluations. In other words, hypothesis test estimates likelihood ratio and all parameters for each genotype. According to likelihood ratio value, the hfunmap selects significant QTL positions which likelihood ratios are reached to maximum value. These significant QTL site and corresponding parameters will be made into a result object at last.

Data object and result object can visualization by ***plot*** or ***ggplot***, the form can be Manhattan and Fitting diagram. The hfunmap does not provide the contents of the visualization part of the result file, but it can be obtained by using functions as said before.

4.3 Permutation

The permutation is necessary to decide the significant threshold for likelihood ratio value (LR) calculated in the hypothesis test. From the consuming time to consider, you can set permutation loop by yourself, if you want to get the precise thresholds, 1000 permutations are recommended to obtain the p-value of 0.05.

Note that here hfunmap provide two calculation methods to determine the threshold. One is to take the maximum value of each permutation test and sort it from large to small, and then select the 50th(1000 permutation loop) as the threshold with p-value of 0.05; The other is to sort the LR value of each locus, select the 50th value, and then take the maximum value from the series of 50th value as the threshold with p-value of 0.05.

It takes long time to do QTL scaning more than 100 times in permutation process. The significant threshold can also obtained by using boferroni correction, although this method is not as strict as permutation test.

5. Example

The main function for the hfunmap is *hfun_lr_cal* which can calculate the likelihood ratio value and parameters. *hfun_permutation* takes long time to do permutation test for whole genome, and final return significant cut likelihood ratio value.The typical syntax to run these functions is shown below.

```
# Load the pre-installed data for the example
data("test_data")

#calculate likelihood ratio value
r <- hfun_lr_cal(test_data,interval = c(1:2))

#calculate threshold
cutlevel <- hfun_permutation(test_data,permu_times = 2,interval = c(1:2),cutp = 0.05)

#find the significant QTL site
sig_sit1 <- which(r[,1] > cutlevel$cut1)
sig_sit2 <- which(r[,1] > cutlevel$cut2)
```

IF you want to get external data, you can use *hfun_load_data* as below,

```
test_data <- hfun_load_data(address.p = "../phe*.csv",
                address.g = "./geno*.csv",
                Times = c(1,3,5,7,9,11,13,16,19,23,28,32,39,47),
                missing = TRUE,
                log.p = FALSE,
                fit.check = FALSE)
```

6. Result

340 F1 hybrid populations of Populus euphratica were used as the research object. At 14 time points after seed germination, four related traits such as plant height, main root length, number of lateral roots and total length of lateral roots were measured as phenotypic data, sample sequencing result as genotype data. Then using hfunmap package to implement multi-trait functional mapping analysis, the results are as follows, ## 6.1 Phenotypic value fitting

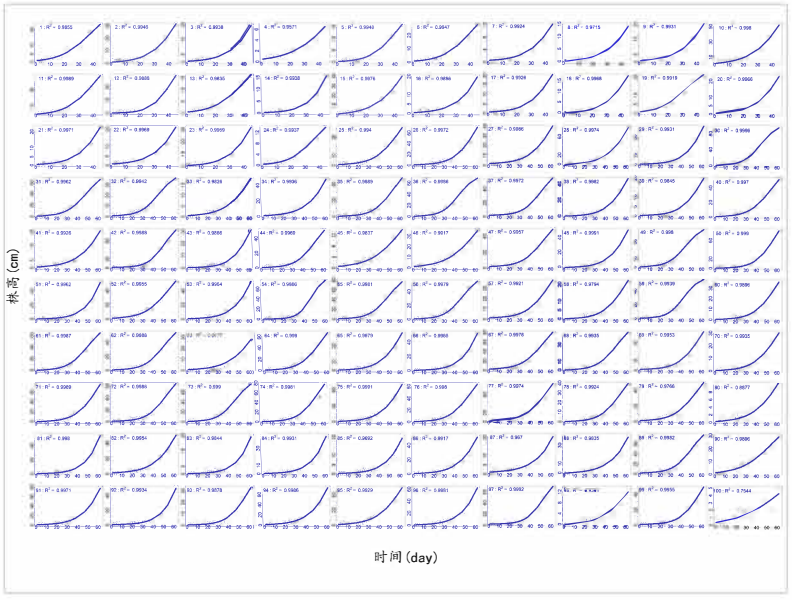

Figure 1: The figure with tiled curves.
The abscissa of figure 1 is time(day) and the ordinate is plant height(CM). Each picture correspinds to the fitting diagram of a sample, and the R^2 is marked

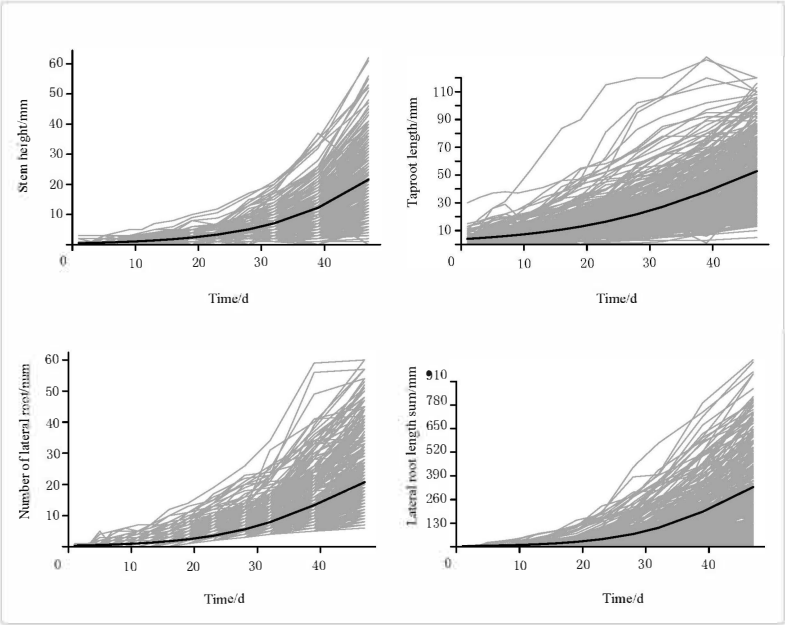

Figure 2:The figure with overlapped curves.
The four pictures in figure 2 correspond to the overall fitting of phenotypic values of four different traits.

6.2 Hypothesis test

The hypothesis test scans every SNP site. It takes a long time, after scanning, hfunmap identifies significant QTLs with thereshold building by function ***hfun_permutation***, final result plot to manhattan as follows:

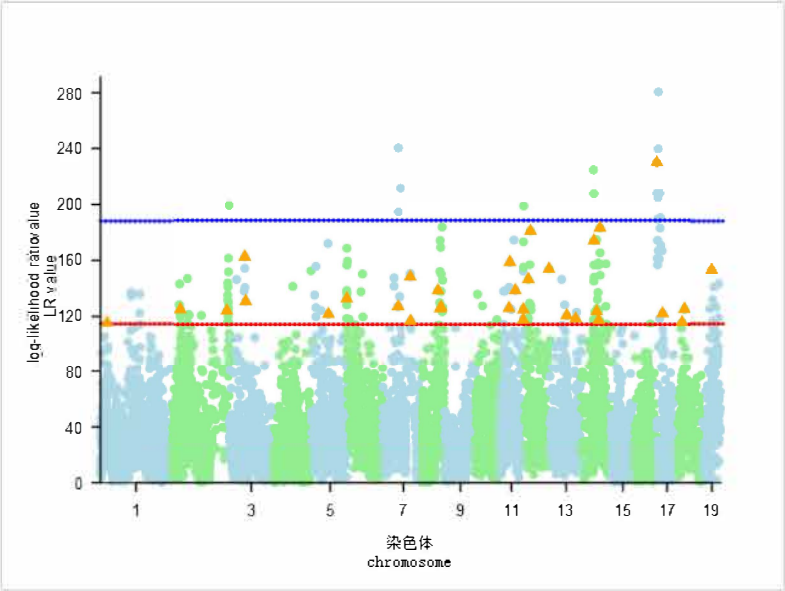

Figure 3:Manhattan output by multi-trait functional mapping.

Figure 3 abscissa is different chromosomes, ordinate is LR value, blue line and red line respectively correspond to the threshold line obtained by two threshold calculation methods mentioned above, and orange triangle markers are snp sites with gene annotation information.

7. Function list

NO	escription
1	**hfun_load_data(address.p,address.g,Times,ranks=NULL,missing=FALSE,log.p=TRUE,fit.check=FALSE)** Load phenotype file and genotype file.
2	**hfun_lr_cal(format_data,interval,useinv=TRUE,usecpp=TRUE,ini_control=NULL,LR_only=FALSE)** Perform the hypothesis test and return log likelihood ratio value of all SNP site.
3	**hfun_permutation(format_data,permu_times=5,interval,cutp=0.05)** Execute the permutation by the specified data object, return the thershold value corresponding to the two calculation methods.

8. References

1.Ma C, Casella G, Wu R. Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. Genetics (Austin), 2002. 161(4): 1751-1762.

2.Zhao W, Hou W, Littell R C, Wu R. Structured Antedependence Models for Functional Mapping of Multiple Longitudinal Traits. Statistical Applications in Genetics and Molecular Biology, 2005.17(4).

3.Cao J, Wang L, Huang Z, Gai J, Wu R. Functional Mapping of Multiple Dynamic Traits. Journal of Agricultural, Biological and Environmental Statistics, 2017. 22(1): 60-75.

4.Mu SC, Zhu XL, Ye MX, Wu RL. Application of multi-traits functional mapping to a growing QTL in Populus

euphratica Oliv. Seedlings. Plant Science Journal

5.R package: multi-dimensional functional mapping. Beijing Forestry University, 2022