MiHC Package

Hyunwook Koh, Ni Zhao

February 26, 2020

1. Overview

This R package, MiHC (v1.0), provides facilities for MiHC which tests the association between a microbial group (e.g., community or clade) composition and a host phenotype of interest. MiHC is a data-driven omnibus test taken in a search space spanned by tailoring the higher criticism test to incorporate phylogenetic information and/or modulate sparsity levels and including the Simes test for excessively high sparsity levels. Therefore, MiHC robustly adapts to diverse phylogenetic relevance and sparsity levels.

2. Installation

MiHC (v1.0) can be installed using the R functions below:

```
> library(devtools)
> install_github("hk1785/MiHC", force=T)
```

3. Requisite R packages

MiHC (v1.0) requires three existing R packages ("cluster", "permute", "phyloseq") to be pre-imported. "cluster" and "permute" are available in CRAN and "phyloseq" is available in Bioconductor, and they can be installed using the R functions below:

```
> install.packages("cluster")
> install.packages("permute")
> source("https://bioconductor.org/biocLite.R")
> biocLite("phyloseq")
```

4. MiHC

MiHC is the global omnibus test taken within the domain containing all the unweighted (i.e., uHC(h)'s) and weighted (i.e., wHC(h)'s) higher criticism tests and the Simes test (Simes 1986), while uHC(A) and wHC(A) are the two local omnibus test taken in each of the sub-domains: 1) uHC(h)'s and the Simes test and 2) wHC(h)'s and the Simes test. Here, uHC(h)'s and uHC(A) do not utilize phylogenetic tree information while wHC(h)'s and wHC(A) utilize phylogenetic tree information. Thus, uHC(h)'s and uHC(A) are more powerful when the OTUs associated with the host phenotype are randomly distributed while wHC(h)'s and wHC(A) are more powerful when the OTUs associated with the host phenotype are phylogenetically relevant. 'h' modulates the sparsity levels for both uHC(h)'s and wHC(h)'s. A small h value (e.g., h=1) suits high sparsity levels while a large h value (e.g., h=9) suits low sparsity levels. The Simes test (Simes 1986) suits high sparsity levels while not utilizing phylogenetic tree information.

5. Implementation

First of all, we need to import the three requisite R packages (“cluster”, “permute”, “phyloseq”) and MiHC (v1.0).

```
> library(cluster)
> library(permute)
> library(phyloseq)
> library(MiHC)
```

Then, we import an example data ‘phy’ attached in the MiHC package.

```
> data(phy)
> otu.tab <- otu_table(phy)
> tree <- phy_tree(phy)
> y <- sample_data(phy)$y
> covs <- data.frame(matrix(NA, length(y), 2))
> covs[,1] <- as.numeric(sample_data(phy)$x1)
> covs[,2] <- as.factor(sample_data(phy)$x2)
```

This example data ‘phy’ is in the phyloseq format (https://joey711.github.io/phyloseq, McMurdie and Homes, 2013) which is really fantastic to neatly organize all the necessary data (i.e., OTU table, taxonomic table, phylogenetic tree, sample information). Here, ‘y’ is disease status (0: non-diseased & 1: diseased) and ‘covs’ are the two confounding factors to be adjusted. We can format confounding factors as continuous ‘as.numeric()’ or categorical variables ‘as.factor()’ while preserving their characteristics in the ‘data.frame’.

Then, we can implement MiHC.

```
> set.seed(123)
> out <- MiHC(y, covs=covs, otu.tab=otu.tab, tree=tree, model="binomial")
```

set.seed() is to produce the same outcomes repeatedly. Please refer to the MiHC manual for all the arguments and options.

Then, we can draw the Q-Q plots between the expected and observed quantiles for the unweighted and weighted higher criticism tests.

```
> MiHC.plot(out)
```

These Q-Q plots also list the top 10 most influential OTUs. Blue dots represent individual OTUs and a red diagonal line represents no influential points; as such, the OTUs that fall along the diagonal line have no influence on the host phenotype while the OTUs that have larger deviations from the diagonal line are more influential on the host phenotype. Darker to lighter vertical lines represent more to less influential OTUs in rank order among the 10 most influential OTUs that correspond to the 10 largest deviations from the red diagonal line. Please refer to the MiHC manual for other graphical options.

6. Interpretation

We found significant association between the disease status and microbial composition based on MiHC (p-value < 0.05), and listed the top 10 most influential OTUs. Please refer to the real data applications in (Koh and Zhao, 2020) for other possible interpretations.

7. Reference

Koh and Zhao. A powerful microbial group association test based on the higher criticism analysis for sparse microbial association signals. (Under revision).

Donoho and Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. Annals of Statistics. 32(3):962-994

Simes (1986). An improved Bonferroni procedure for multiple tests of significance. Biometrika.73(3):751-754

McMurdie and Holmes (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE. 8(4):e61217