IFAA

IFAA offers a robust approach to make inference on the association of covariates with the absolute abundance (AA) of microbiome in an ecosystem. It can be also directly applied to relative abundance (RA) data to make inference on AA because the ratio of two RA is equal ratio of their AA. This algorithm can estimate and test the associations of interest while adjusting for potential confounders. High-dimensional covariates are handled with regularization. The estimates of this method have easy interpretation like a typical regression analysis. High-dimensional covariates are handled with regularization and it is implemented by parallel computing. False discovery rate is automatically controlled by this approach. Zeros do not need to be imputed by a positive value for the analysis. The IFAA package also offers the 'MZILN' function for estimating and testing associations of abundance ratios with covariates.

To model the association, the following equation is used:

$$\log(\mathcal{Y}_i^k)|\mathcal{Y}_i^k > 0 = \beta^{0k} + X_i^T\beta^k + W_i^T\gamma^k + Z_i^T b_i + \epsilon_i^k, \quad k = 1, ..., K+1,$$

where

- $\mathcal{Y}_i^k$ is the AA of taxa k in subject i in the entire ecosystem.
- X_i is the covariate matrix.
- W_i is the confounder matrix.
- Z_i is the design matrix for random effects.
- β^k is the regression coefficients that will be estimated and tested with the `IFAA()` function.

The challenge in microbiome analysis is that we can not oberve $\mathcal{Y}_i^k$. What is observed is its small proportion: $Y_i^k = C_i\mathcal{Y}_i^k$ where C_i is an unknown number between 0 and 1 that denote the observed proportion. The IFAA method successfuly addressed this challenge.

Package installation

To install, type the following command in R console:

```
install.packages("IFAA", repos = "http://cran.us.r-project.org")
```

The package could be also installed from GitHub using the following code:

```
require(devtools)
devtools::install_github("gitlzg/IFAA")
```

Input for IFAA() function

Most of the time, users just need to feed the first three inputs to the function: `experiment_dat`, `testCov` and `ctrlCov`. All other inputs can just take their default values. Below are all the inputs of the functions

- `experiment_dat`: A SummarizedExperiment object containing microbiome data and covariates (see example on how to create a SummarizedExperiment object). The microbiome data can be absolute abundance or relative abundance with each column per sample and each row per taxon/OTU/ASV (or any other unit). No imputation is needed for zero-valued data points. The covariates data contains covariates and confounders with each row per sample and each column per variable. The covariates data has to be numeric or binary. Categorical variables should be converted into binary dummy variables.

- `testCov`: Covariates that are of primary interest for testing and estimating the associations. It corresponds to X_i in the equation. Default is `NULL` which means all covariates are `testCov`.
- `ctrlCov`: Potential confounders that will be adjusted in the model. It corresponds to W_i in the equation. Default is `NULL` which means all covariates except those in `testCov` are adjusted as confounders.
- `sampleIDname`: Name of the sample ID variable in the data. In the case that the data does not have an ID variable, this can be ignored. Default is NULL.
- `testMany`: This takes logical value `TRUE` or `FALSE`. If `TRUE`, the `testCov` will contain all the variables in `CovData` provided `testCov` is set to be `NULL`. The default value is `TRUE` which does not do anything if `testCov` is not `NULL`.
- `ctrlMany`: This takes logical value `TRUE` or `FALSE`. If `TRUE`, all variables except `testCov` are considered as control covariates provided `ctrlCov` is set to be `NULL`. The default value is `FALSE`.
- `nRef`: The number of randomly picked reference taxa used in phase 1. Default number is `40`.
- `nRefMaxForEsti`: The maximum number of final reference taxa used in phase 2. The default is `2`.
- `refTaxa`: A vector of taxa names. These are reference taxa specified by the user to be used in phase 1 if the user believe these taxa are indepenent of the covariates. If the number of reference taxa is less than 'nRef', the algorithm will randomly pick extra reference taxa to make up 'nRef'. The default is `NULL` since the algorithm will pick reference taxa randomly.
- `adjust_method`: The adjusting method used for p value adjustment. Default is "BY" for dependent FDR adjustment. It can take any adjustment method for p.adjust function in R.
- `fdrRate`: The false discovery rate for identifying taxa/OTU/ASV associated with `testCov`. Default is `0.05`.
- `paraJobs`: If `sequentialRun` is `FALSE`, this specifies the number of parallel jobs that will be registered to run the algorithm. If specified as `NULL`, it will automatically detect the cores to decide the number of parallel jobs. Default is `NULL`.
- `bootB`: Number of bootstrap samples for obtaining confidence interval of estimates in phase 2 for the high dimensional regression. The default is `500`.
- `standardize`: This takes a logical value `TRUE` or `FALSE`. If `TRUE`, the design matrix for X will be standardized in the analyses and the results. Default is `FALSE`.
- `sequentialRun`: This takes a logical value `TRUE` or `FALSE`. Default is `FALSE`. This argument could be useful for debug.
- `refReadsThresh`: The threshold of proportion of non-zero sequencing reads for choosing the reference taxon in phase 2. The default is `0.2` which means at least 20% non-zero sequencing reads.
- `taxDropThresh`: The threshold of number of non-zero sequencing reads for each taxon to be dropped from the analysis. The default is `0` which means taxon without any sequencing reads will be dropped from the analysis.
- `SDThresh`: The threshold of standard deviations of sequencing reads for been chosen as the reference taxon in phase 2. The default is `0.05` which means the standard deviation of sequencing reads should be at least `0.05` in order to be chosen as reference taxon.
- `SDquantilThresh`: The threshold of the quantile of standard deviation of sequencing reads, above which could be selected as reference taxon. The default is `0`.
- `balanceCut`: The threshold of the proportion of non-zero sequencing reads in each group of a binary variable for choosing the final reference taxa in phase 2. The default number is `0.2` which means at least 20% non-zero sequencing reads in each group are needed to be eligible for being chosen as a final reference taxon.

Output for IFAA() function

A list containing 2 elements

- `full_results`: The main results for IFAA containing the estimation and testing results for all associations between all taxa and all test covariates in `testCov`. It is a dataframe with each row representing an association, and eight columns named as "taxon", "cov", "estimate", "SE.est", "CI.low", "CI.up", "adj.p.value", and "sig_ind". The columns correspond to taxon name, covariate name, association estimates, standard error estimates, lower bound and upper bound of the 95% confidence interval, adjusted p value, and the indicator showing whether the association is significant after multiple testing adjustment.
- `metadata`: The metadata is a list containing the following items: `covariatesData`: A dataset containing covariates and confounders used in the analyses. `final_ref_taxon`: The final 2 reference taxa used for analysis. `ref_taxon_count`: The counts of selection for the associations of all taxa with test covariates in Phase 1. `ref_taxon_est`: The average magnitude estimates for the associations of all taxa with test covariates in Phase 1. `totalTimeMins`: Total time used for the entire analysis. `fdrRate`: FDR rate used for the analysis. `adjust_method`: Multiple testing adjust method used for the analysis.

Example

The example generates an example data from scratch, with 10 taxon, 40 subjects, and 3 covariates.

```
library(IFAA)

set.seed(1)

## create an ID variable for the example data
ID=seq_len(20)

## generate three covariates x1, x2, and x3, with x2 binary
x1<-rnorm(20)
x2<-rbinom(20,1,0.5)
x3<-rnorm(20)
dataC<-data.frame(cbind(ID,x1,x2,x3))

## Coefficients for x1, x2, and x3 among 10 taxa.
beta_1<-c(0.1,rep(0,9))
beta_2<-c(0,0.2,rep(0,8))
beta_3<-rnorm(10)
beta_mat<-cbind(beta_1,beta_2,beta_3)

## Generate absolute abundance for 10 taxa in ecosystem.
dataM_eco<-floor(exp(10+as.matrix(dataC[,-1]) %*% t(beta_mat) + rnorm(200,sd=0.05)))

## Generate sequence depth and generate observed abundance
Ci<-runif(20,0.01,0.05)
dataM<-floor(apply(dataM_eco,2,function(x) x*Ci))
colnames(dataM)<-paste0("rawCount",1:10)

## Randomly introduce 0 to make 25% sparsity level.
dataM[sample(seq_len(length(dataM)),length(dataM)/4)]<-0

dataM<-data.frame(cbind(ID,dataM))

## The following steps are to create a SummarizedExperiment object.
```

```
## If you already have a SummarizedExperiment format data, you can
## ignore the following steps and directly feed it to the IFAA function.

## Merge two dataset by ID variable
data_merged<-merge(dataM,dataC,by="ID",all=FALSE)

## Seperate microbiome data and covariate data, drop ID variable from microbiome data
dataM_sub<-data_merged[,colnames(dataM)[!colnames(dataM)%in%c("ID")]]
dataC_sub<-data_merged[,colnames(dataC)]

## Create SummarizedExperiment object
test_dat<-SummarizedExperiment::SummarizedExperiment(
assays=list(MicrobData=t(dataM_sub)), colData=dataC_sub)
```

If you already have a SummarizedExperiment format data, you can ignore the above steps for creating a SummarizedExperiment object. In the generated data, rawCount1 is associated with x1, rawCount2 is associated with x2, and the sparsity level is 25%. Next we analyze the data to test the association between microbiome and the variable `"x1", "x2"` while adjusting for the variable (potential confounder) `"x3"`.

```
set.seed(123) # For full reproducibility
results <- IFAA(experiment_dat = test_dat,
                testCov = c("x1","x2"),
                ctrlCov = c("x3"),
                sampleIDname="ID",
                fdrRate = 0.05,
                nRef=2,
                paraJobs = 2)
#> Data dimensions (after removing missing data if any):
#> 20 samples
#> 10 taxa/OTU/ASV
#> 2 testCov variables in the analysis
#> These are the testCov variables:
#> x1, x2
#> 1 ctrlCov variables in the analysis
#> These are the ctrlCov variables:
#> x3
#> 1 binary covariates in the analysis
#> These are the binary covariates:
#> x2
#> 25 percent of microbiome sequencing reads are zero
#> Start Phase 1 analysis
#> 2 parallel jobs are registered for the analysis.
#> 33 percent of phase 1 analysis has been done and it took 0.18 minutes
#> 2 parallel jobs are registered for the analysis.
#> 100 percent of phase 1 analysis has been done and it took 0.35 minutes
#> Start Phase 2 parameter estimation
#> 2 parallel jobs are registered for analyzing reference taxa in Phase 2
#> 50 percent of Phase 2 is done and it took 0.113 minutes
#> 2 parallel jobs are registered for analyzing reference taxa in Phase 2
#> Entire Phase 2 parameter estimation done and took 0.241 minutes.
#> The entire analysis took 0.59 minutes
```

In this example, we are only interested in testing the associations with `"x1"` and `"x2"` which is why `testCov=c("x1","x2)`. The variables `"x3"` are adjusted as potential confounders in the analyses. The final

analysis results are saved in the list `full_result` and the significant results can be extracted as follows:

```
summary_res<-results$full_result
sig_results<-subset(summary_res,sig_ind==TRUE)
sig_results
#> DataFrame with 2 rows and 8 columns
#>         taxon         cov  estimate    SE.est    CI.low     CI.up adj.p.value
#>   <character> <character> <numeric> <numeric> <numeric> <numeric>   <numeric>
#> 1   rawCount1          x1 0.0810973 0.0189065 0.0440407  0.118154 3.89542e-04
#> 2   rawCount2          x2 0.2021147 0.0355637 0.1324098  0.271820 2.87484e-07
#>     sig_ind
#>   <logical>
#> 1      TRUE
#> 2      TRUE
```

After adjusting for multiple testing, the results found `"rawCount1"` associated with `"x1"`, and `"rawCount2"` associated with `"x2"`, while adjusting for `"x3"`. The regression coefficients and their 95% confidence intervals are provided. These coefficients correspond to β^k in the model equation.

The interpretation is:

- Every unit increase in `"x1"` is associated with approximately 8.1% increase in the absolute abundance of `"rawCount1"` in the entire ecosystem of interest (while controlling for `"x3"`); Every unit increase in `"x2"` is associated with approximately 20.2% increase in the absolute abundance of `"rawCount2"` in the entire ecosystem of interest (while controlling for `"x3"`), .

Reference

Li et al.(2021) IFAA: Robust association identification and Inference For Absolute Abundance in microbiome analyses. Journal of the American Statistical Association. 116(536):1595-1608

MZILN() function

The IFAA package can also implement the Multivariate Zero-Inflated Logistic Normal (MZILN) regression model for estimating and testing the association of abundance ratios with covariates. The `MZILN()` function estimates and tests the associations of user-specified abundance ratios with covariates. When the denominator taxon of the ratio is independent of the covariates, 'MZILN()' should generate similar results as 'IFAA()'. The regression model of 'MZILN()' can be expressed as follows:

$$\log\left(\frac{\mathcal{Y}_i^k}{\mathcal{Y}_i^{K+1}}\right)|\mathcal{Y}_i^k > 0, \mathcal{Y}_i^{K+1} > 0 = \alpha^{0k} + \mathcal{X}_i^T\alpha^k + \epsilon_i^k, \quad k = 1, ..., K,$$

where

- $\mathcal{Y}_i^k$ is the AA of taxa k in subject i in the entire ecosystem.
- $\mathcal{Y}_i^{K+1}$ is the reference taxon (specified by user).
- $\mathcal{X}_i$ is the covariate matrix for all covariates including confounders.
- α^k is the regression coefficients that will be estimated and tested.

Input for MZILN() function

Most of the time, users just feed the first three inputs to the function: `experiment_dat`, `refTaxa` and `allCov`. All other inputs can just take their default values. All the inputs for 'MZILN()' are:

- `experiment_dat`: A SummarizedExperiment object containing microbiome data and covariates (see example on how to create a SummarizedExperiment object). The microbiome data can be absolute

abundance or relative abundance with each column per sample and each row per taxon/OTU/ASV (or any other unit). No imputation is needed for zero-valued data points. The covariates data contains covariates and confounders with each row per sample and each column per variable. The covariates data has to be numeric or binary. Categorical variables should be converted into binary dummy variables.

- `refTaxa`: Denominator taxa names specified by the user for the targeted ratios. This could be a vector of names.
- `allCov`: All covariates of interest (including confounders) for estimating and testing their associations with the targeted ratios. Default is 'NULL' meaning that all covariates in covData are of interest.
- `sampleIDname`: Name of the sample ID variable in the data. In the case that the data does not have an ID variable, this can be ignored. Default is NULL.
- `adjust_method`: The adjusting method for p value adjustment. Default is "BY" for dependent FDR adjustment. It can take any adjustment method for p.adjust function in R.
- `fdrRate` The false discovery rate for identifying ratios associated with `allCov`. Default is `0.05`.
- `paraJobs`: If `sequentialRun` is `FALSE`, this specifies the number of parallel jobs that will be registered to run the algorithm. If specified as `NULL`, it will automatically detect the cores to decide the number of parallel jobs. Default is `NULL`.
- `bootB`: Number of bootstrap samples for obtaining confidence interval of estimates for the high dimensional regression. The default is `500`.
- `taxDropThresh`: The threshold of number of non-zero sequencing reads for each taxon to be dropped from the analysis. The default is `0` which means taxon without any sequencing reads will be dropped from the analysis.
- `standardize`: This takes a logical value `TRUE` or `FALSE`. If `TRUE`, the design matrix for X will be standardized in the analyses and the results. Default is `FALSE`.
- `sequentialRun`: This takes a logical value `TRUE` or `FALSE`. Default is `TRUE`. It can be set to be "FALSE" to increase speed if there are multiple taxa in the argument 'refTaxa'.

Output for MZILN() function

A list with two elements:

- `full_results`: The main results for MZILN containing the estimation and testing results for all associations between all taxa ratios with refTaxan being the denominator and all covariates in `allCov`. It is a dataframe with each row representing an association, and ten columns named as "ref_tax", "taxon", "cov", "estimate", "SE.est", "CI.low", "CI.up", "adj.p.value", "unadj.p.value" and "sig_ind". The columns correspond to the denominator taxon, numerator taxon, covariate name, association estimates, standard error estimates, lower bound and upper bound of the 95% confidence interval, adjusted p value, and the indicator showing whether the association is significant after multiple testing adjustment.
- `metadata`: The metadata is a list containing a dataset of covariates and confounders used in the analyses, total time used in minutes, FDR rate, and multiple testing adjustment method used.

Examples

The example used data generated from IFAA example, with 10 taxon, 40 subjects, and 3 covariates.

Next we analyze the data to test the associations between the ratio "rawCount5/rawCount10" and all the three variables `"x1"`, `"x2"` and `"x3"` in a multivariate model where all `"x1"`, `"x2"` and `"x3"` are independent variables simultaneously.

```
set.seed(123) # For full reproducibility
results <- MZILN(experiment_dat=test_dat,
                 refTaxa=c("rawCount10"),
                 allCov=c("x1","x2","x3"),
                 sampleIDname="ID",
                 fdrRate=0.05,
                 paraJobs = 2)
#> Data dimensions (after removing missing data if any):
#> 20 samples
#> 10 taxa/OTU/ASV
#> 3 testCov variables in the analysis
#> These are the testCov variables:
#> x1, x2, x3
#> 0 ctrlCov variables in the analysis
#> 1 binary covariates in the analysis
#> These are the binary covariates:
#> x2
#> 25 percent of microbiome sequencing reads are zero
#> 2 parallel jobs are registered for analyzing reference taxa in Phase 2
#> Estimation done for the 1th denominator taxon: rawCount10 and it took 0.12 minutes
#> The entire analysis took 0.12 minutes
```

The full final analysis results can be extracted as follows:

```
summary_res<-results$full_results
```

The results for the log-ratio of "rawCount5" over "rawCount10" can extracted as follows:

```
summary_res[summary_res$taxon=="rawCount5",,drop=FALSE]
#> DataFrame with 3 rows and 10 columns
#>       ref_tax       taxon         cov   estimate    SE.est     CI.low
#>   <character> <character> <character>  <numeric> <numeric>  <numeric>
#> 1  rawCount10   rawCount5          x1  0.0099748 0.0228440 -0.0347994
#> 2  rawCount10   rawCount5          x2 -0.0932306 0.0386687 -0.1690213
#> 3  rawCount10   rawCount5          x3  1.5819462 0.0276923  1.5276693
#>        CI.up unadj.p.value adj.p.value   sig_ind
#>    <numeric>     <numeric>   <numeric> <logical>
#> 1  0.0547490     0.6623657     1.00000     FALSE
#> 2 -0.0174399     0.0159084     0.20252     FALSE
#> 3  1.6362230     0.0000000     0.00000      TRUE
```

The regression coefficients and their 95% confidence intervals are provided. These coefficients correspond to α^k in the model equation, and can be interpreted as the associations between the covariates and log-ratio of `"rawCount5"` over `"rawCount10"`.

The results show that the log-ratio is associated with `"x2"` and `"x3"` before adjusting for multiple testing. The interpretation is:

- Every unit increase in `"x2"` is associated with approximately 9.3% reduction in the abundance ratio of `"rawCount5"` over `"rawCount10"` (while controlling for `"x1"` and `"x3"`), but this is not significant after adjusting for multiple testing because the adjusted p value is 0.20252; Every unit increase in `"x3"` is associated with approximately 386.4% (=exp(1.582)-1) increase in the abundance ratio of `"rawCount5"` over `"rawCount10"` (while controlling for `"x1"` and `"x2"`), and this association remains significant after adjusting for multiple testing.

To extract the significant associations for all ratios with `"rawCount10"` being the denominator, one can do: `subset(summary_res,sig_ind==TRUE)`.

Reference

Li et al.(2018) Conditional Regression Based on a Multivariate Zero-Inflated Logistic-Normal Model for Microbiome Relative Abundance Data. Statistics in Biosciences 10(3): 587-608

Session Info

```
sessionInfo()
#> R version 4.2.1 (2022-06-23 ucrt)
#> Platform: x86_64-w64-mingw32/x64 (64-bit)
#> Running under: Windows 10 x64 (build 19044)
#>
#> Matrix products: default
#>
#> locale:
#> [1] LC_COLLATE=English_United States.utf8
#> [2] LC_CTYPE=English_United States.utf8
#> [3] LC_MONETARY=English_United States.utf8
#> [4] LC_NUMERIC=C
#> [5] LC_TIME=English_United States.utf8
#>
#> attached base packages:
#> [1] stats     graphics  grDevices utils     datasets  methods   base
#>
#> other attached packages:
#> [1] IFAA_1.0.9
#>
#> loaded via a namespace (and not attached):
#>  [1] Rcpp_1.0.7                  mvtnorm_1.1-3
#>  [3] lattice_0.20-45             class_7.3-20
#>  [5] glmnet_4.1-3                digest_0.6.29
#>  [7] RhpcBLASctl_0.21-247.1      foreach_1.5.1
#>  [9] parallelly_1.30.0           slam_0.1-49
#> [11] R6_2.5.1                    GenomeInfoDb_1.32.2
#> [13] cellranger_1.1.0            stats4_4.2.1
#> [15] evaluate_0.16               rootSolve_1.8.2.3
#> [17] e1071_1.7-11                httr_1.4.2
#> [19] zlibbioc_1.42.0             rlang_1.0.3
#> [21] Exact_3.1                   readxl_1.4.0
#> [23] rstudioapi_0.13             data.table_1.14.2
#> [25] S4Vectors_0.34.0            Matrix_1.4-1
#> [27] rmarkdown_2.14              mathjaxr_1.4-0
#> [29] splines_4.2.1               stringr_1.4.0
#> [31] RCurl_1.98-1.7              DelayedArray_0.22.0
#> [33] proxy_0.4-27                compiler_4.2.1
#> [35] xfun_0.29                   BiocGenerics_0.42.0
#> [37] shape_1.4.6                 DescTools_0.99.45
#> [39] htmltools_0.5.2             SummarizedExperiment_1.26.1
#> [41] lmom_2.9                    GenomeInfoDbData_1.2.8
#> [43] MatrixExtra_0.1.12          expm_0.999-6
#> [45] IRanges_2.30.0              codetools_0.2-18
#> [47] matrixStats_0.62.0          MASS_7.3-57
#> [49] bitops_1.0-7                grid_4.2.1
#> [51] magrittr_2.0.1              gld_2.6.4
#> [53] cli_3.3.0                   stringi_1.7.6
```

```
#> [55] XVector_0.36.0              doRNG_1.8.2
#> [57] doParallel_1.0.16           boot_1.3-28
#> [59] HDCI_1.0-2                  iterators_1.0.13
#> [61] tools_4.2.1                 float_0.3-0
#> [63] Biobase_2.56.0              rngtools_1.5.2
#> [65] MatrixGenerics_1.8.0        parallel_4.2.1
#> [67] fastmap_1.1.0               survival_3.3-1
#> [69] yaml_2.2.1                  GenomicRanges_1.48.0
#> [71] knitr_1.39
```