A tutorial on R package eBay

Tiantian Liu1,3, Hongyu Zhao 2,3, and Tao Wang 1,3,4,*

1 Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University, China,

2 Department of Biostatistics, Yale University, USA and

3 SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, China

4 MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University

November 26, 2019

1 Introduction

High-throughout sequencing technologies and advanced bioinformatics tools are now routinely applied to microbiome studies. Most of the studies focus on the detection of differential abundance features (DAF) in microbial communities. However, the process of analyzing DAF is complicated by several challenges – the varied sequencing depth and sparsity. We present a R package, eBay, that provides a novel normalization technique based on empirical bayes approach to addressed these challenges. We also extend our method by incorporating the phylogenetic tree into the normalization process. To model the real metagenomics data, we use Dirichlet-multinomial (DM) and Dirichlet-tree multinomial (DTM) distribution to simulate the count data. Functions of performing detection of DAF, several simulations and real dataset are provided.

2 Running the eBay

First, we need to install package dirmult, foreach, stats, doParallel, phyloseq, MGLM before we install eBay. To load eBay, type:

```
#install.packages("devtools")
#devtools::install_github("liudoubletian/eBay")
library(eBay)
```

2.1 A simulated example with data generated from DM model

The following function will simulate microbiome data according to the DM model with the sequence depth drawn uniformly from 5000 to 50000.

```
rm(list=ls())
set.seed(2016)
p <- 20 ### number of taxa
N <- 20 ### the number of samples in each group
rand_pi <- runif(p)
control_pi = case_pi = rand_pi/sum(rand_pi) ##the proportions of each taxa
control_pi[4]=control_pi[4]-0.01;control_pi[6]=control_pi[6]+0.01;
```

```
case_pi[4]=case_pi[4]+0.01;case_pi[6]=case_pi[6]-0.01;
##set OTU4 and OTU6 be different between case and control
control_theta = case_theta = 0.1  ## the dispersion parameter in DM model
group <- rep(c(0,1),each =20)
### an otu table with 40 rows and 20 colums
ntree_table <- simulation_dm(p,seed=1, N=20,control_pi, case_pi,control_theta,case_theta)
```

According to the above parameter settings, the simulation function returns an OTU table with two differential abundant OTUs. We can run the eBay function to implement the differential abundance testing.

```
ebay.res<-eBay(otu.data=ntree_table,group=group,test.method="t",cutf=0.05,adj.m="BH")
ebay.res
#> $final.p
#>            1            2            3            4            5            6
#> 0.8412635427 0.7975509485 0.8412635427 0.0092577865 0.7340110275 0.0002086701
#>            7            8            9           10           11           12
#> 0.8412635427 0.8412635427 0.8412635427 0.8412635427 0.8412635427 0.8412635427
#>           13           14           15           16           17           18
#> 0.9663613714 0.3245119193 0.5192511537 0.8412635427 0.7975509485 0.8412635427
#>           19           20
#> 0.8412635427 0.5192511537
#>
#> $dif.otus
#> 4 6
#> 4 6
```

Here, the return results of eBay contain the final.p and dif.otus.

2.2 A simulated example with data generated from DTM model

The following function will simulate microbiome data generated from DTM model. If a phylogenetic tree is provided, our eBay can incorporate the tree into the normalization process and implement differential abuance testing.

```
p <- 40
set.seed(1)
tree <- simulate_tree(p)  ###simulate a tree with 40 leaf nodes
set.seed(1)
control_pi = case_pi = c()
for(j in (p+1):(p+tree$Nnode)){
   set.seed(j)
   random_pi <- runif(1,0.2,0.4)
   control_pi[which(tree$edge[,1]==j)] <- c(random_pi, 1-random_pi)
   case_pi[which(tree$edge[,1]==j)] <- c(random_pi, 1-random_pi)
}### the proportions of each node
control_theta = case_theta = rep(0.1, tree$Nnode) ###the dispersion parameter
group <- rep(c(0,1),each =20)
tree_table<-simulation_dtm(p=40,tree,seed=1,N=20,control_pi, case_pi,control_theta,case_theta)
```

Run the eBay_tree function to normalize the data and return a set of differential abundant taxa.

```
ebay_tree.res<-eBay_tree(otu.data=tree_table,tree=tree,group=group,test.method="t", cutf=0.05,adj.m="BH
ebay_tree.res
#> $final.p
#>        42        45        43         4         1        44         2         3
#> 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085
#>        46        79        47        69        48        58        49        50
#> 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085
#>         5         6        51        53         7        52         8         9
#> 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085
#>        10        54        11        55        12        56        57        15
#> 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085
#>        13        14        59        64        60        61        16        17
#> 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085
#>        62        21        18        63        19        20        65        27
#> 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085
#>        22        66        23        67        24        68        25        26
#> 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085
#>        70        74        28        71        72        73        29        30
#> 0.9490085 0.9490085 0.9490085 0.9490085 0.9793070 0.9793070 0.9490085 0.9490085
#>        31        32        75        76        33        34        77        78
#> 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085
#>        35        36        37        38        39        40
#> 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085 0.9490085
#>
#> $dif.otus
#> NULL
```

3 Real data analysis

Let's try anohter example on the real data. To explore the association between severe acute malnutrition (SAM) and microbiota, [1] conducted a study of 996 stool samples collected monthly from 50 healthy Bangladeshi children during the first 2 years of life. We restricted our analysis to 12 to 18-month-old children which includs 20 healthy children and 27 children with SAM. We further filtered bacterial taxa with prevalance less than 20%, resulting in 50 taxa. Here, we run the eBay and eBay_tree function to the real data.

```
### load data
data(rep_tree)
data(sam_table)
tree <- rep_tree
ntree_table<- sam_table
p <- length(tree$tip.label)
colnames(ntree_table)=as.character(1:p)
group <- c(rep(0,27),rep(1,20))
eBay.res <- eBay(otu.data=ntree_table, group=group, cutf=0.05, test.methods="t",adj.m="BH")
eBay.res
#> $final.p
#>            1            2            3            4            5            6
#> 4.917519e-03 1.444788e-03 1.111233e-02 1.287588e-02 6.056851e-01 1.415185e-01
#>            7            8            9           10           11           12
#> 4.411180e-01 9.983770e-01 2.117890e-04 3.984101e-04 1.001214e-02 9.750349e-01
#>           13           14           15           16           17           18
```

```
#> 2.039866e-03 1.424158e-02 5.512623e-02 1.090689e-03 1.193852e-04 4.993131e-05
#>           19           20           21           22           23           24
#> 9.963223e-06 1.132821e-04 3.381460e-04 2.209375e-01 1.026392e-01 1.114485e-01
#>           25           26           27           28           29           30
#> 4.741916e-02 7.439108e-02 2.939951e-01 1.302934e-01 6.134489e-02 2.851672e-01
#>           31           32           33           34           35           36
#> 2.250762e-04 8.857246e-02 3.875911e-01 6.387854e-01 6.936066e-01 1.333721e-06
#>           37           38           39           40           41           42
#> 1.141872e-04 9.963223e-06 1.656484e-01 6.474250e-01 2.250762e-04 1.850195e-03
#>           43           44           45           46           47           48
#> 2.250762e-04 4.676481e-01 1.816910e-03 3.070945e-03 4.361597e-01 8.186547e-02
#>           49           50
#> 1.816910e-03 1.132821e-04
#>
#> $dif.otus
#>  1  2  3  4  9 10 11 13 14 16 17 18 19 20 21 25 31 36 37 38 41 42 43 45 46 49
#>  1  2  3  4  9 10 11 13 14 16 17 18 19 20 21 25 31 36 37 38 41 42 43 45 46 49
#> 50
#> 50
ebay_tree.t.res<-eBay_tree(otu.data=ntree_table,tree=tree,adj.m="BH",group=group, test.method="t",cutf=0
ebay_tree.t.res
#> $final.p
#>            1            2            3            4            5            6
#> 4.917519e-03 1.444788e-03 1.111233e-02 1.287588e-02 6.056851e-01 1.415185e-01
#>            7            8            9           10           11           12
#> 4.411180e-01 9.983770e-01 2.117890e-04 3.984101e-04 1.001214e-02 9.750349e-01
#>           13           14           15           16           17           18
#> 2.039866e-03 1.424158e-02 5.512623e-02 1.090689e-03 1.193852e-04 4.993131e-05
#>           19           20           21           22           23           24
#> 9.963223e-06 1.132821e-04 3.381460e-04 2.209375e-01 1.026392e-01 1.114485e-01
#>           25           26           27           28           29           30
#> 4.741916e-02 7.439108e-02 2.939951e-01 1.302934e-01 6.134489e-02 2.851672e-01
#>           31           32           33           34           35           36
#> 2.250762e-04 8.857246e-02 3.875911e-01 6.387854e-01 6.936066e-01 1.333721e-06
#>           37           38           39           40           41           42
#> 1.141872e-04 9.963223e-06 1.656484e-01 6.474250e-01 2.250762e-04 1.850195e-03
#>           43           44           45           46           47           48
#> 2.250762e-04 4.676481e-01 1.816910e-03 3.070945e-03 4.361597e-01 8.186547e-02
#>           49           50
#> 1.816910e-03 1.132821e-04
#>
#> $dif.otus
#>  [1] "1"  "2"  "3"  "4"  "9"  "10" "11" "13" "14" "16" "17" "18" "19" "20" "21"
#> [16] "25" "31" "36" "37" "38" "41" "42" "43" "45" "46" "49" "50"
```

Here we report the p values and differential abundant OTUs.