Tutorial: testing microbiome mediation effect using miMediation

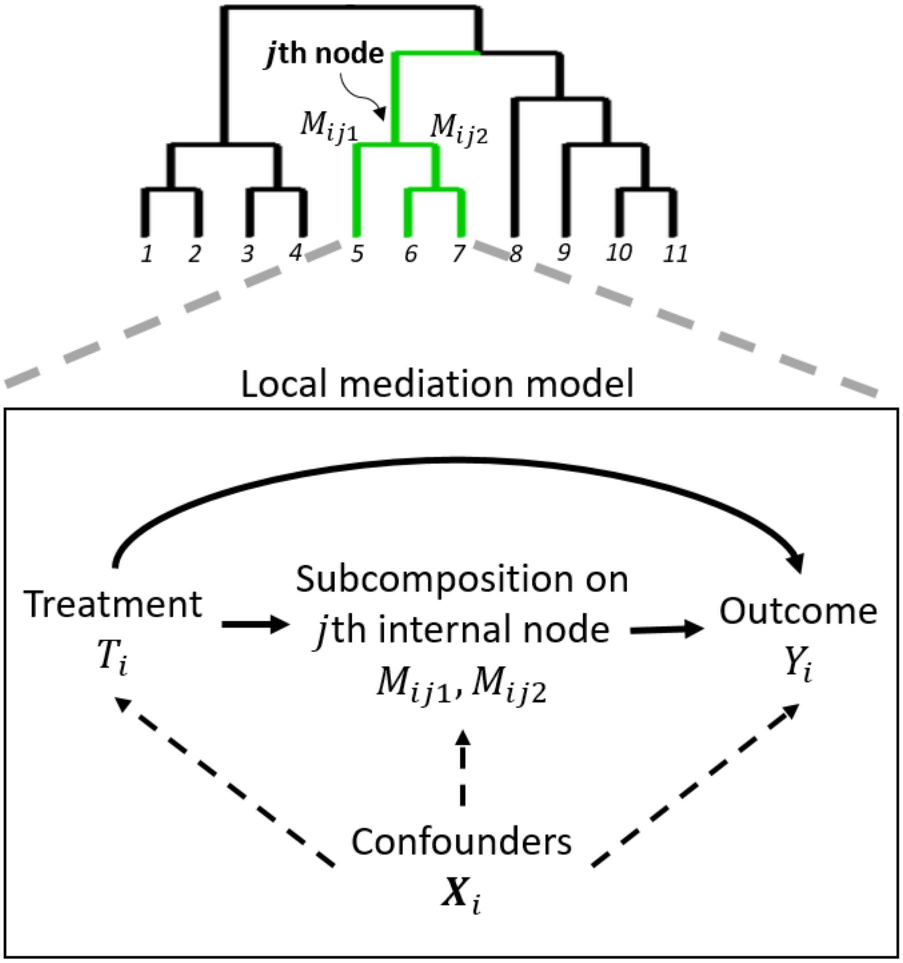
Qilin Hong

Last compiled on 20 December, 2022

This is a practical tutorial on the use of `miMediation` package, which introduces a phylogeny-based mediation test (PhyloMed) for high-dimensional microbial composition mediators. The methodology is described in detail in the Hong, Chen, and Tang (Manuscript).

A brief summary of the PhyloMed

PhyloMed models microbiome mediation effect through a cascade of independent local mediation models of subcompositions on the internal nodes of the phylogenetic tree. Each local model captures the mediation effect of a subcomposition at a given taxonomic resolution. The method improves the power of the mediation test by enriching weak and sparse signals across mediating taxa that tend to cluster on the tree. PhyloMed enables us to test the overall mediation effect of the entire microbial community and pinpoint internal nodes with significant subcomposition mediation effects.



As depicted in the figure above, we propose to construct a local mediation model for the subcomposition at each internal node of the phylogenetic tree. The subcomposition on a given internal node consists of the relative abundance aggregated at its two child nodes. We apply the following robust linear regression model and generalized linear regression model to represent the causal path diagram of the local mediation model at the jth internal node

$$E\left\{\log\left(\frac{M_{ij1}}{M_{ij2}}\right)\right\} = \alpha_{jX}^{\mathrm{T}}\mathbf{X}_i + \alpha_j T_i$$

$$g\{E(Y_i)\} = \beta_{jX}^{\mathrm{T}}\mathbf{X}_i + \beta_{jT}T_i + \beta_j \log\left(\frac{M_{ij1}}{M_{ij2}}\right)$$

where $g(\cdot)$ is the link function depending on the type of the outcome and we omit the intercept term in both models as it can be absorbed into $\mathbf{X}_i$.

Under potential outcome framework (VanderWeele (2015)) and assumptions of no unmeasured confounding variables, it leads to the null hypothesis

$$H_0^j = \alpha_j\beta_j = 0$$

, which is equivalent to the union of three disjoint component null hypotheses

$$H_{00}^j : \alpha_j = \beta_j = 0, \tag{1}$$

$$H_{10}^j : \alpha_j \neq 0, \beta_j = 0, \tag{2}$$

$$H_{01}^j : \alpha_j = 0, \beta_j \neq 0. \tag{3}$$

We define the mediation test statistic for H_0^j as

$$P_{\text{max}_j} = \max(P_{\alpha_j}, P_{\beta_j})$$

The $P_{\alpha_j}, P_{\beta_j}$ could be calculated via asymptotic approach or adaptive permutation approach when sample size is small.

In fact, P_{max_j} follows a mixture distribution with three components, each of which corresponds to one type of null hypothesis H_{00}^j, H_{10}^j, H_{01}^j. The p-value of mediation test in the jth local model is given by

$$Pr(P_{\text{max}_j} \leq p_{\text{max}_j}) = \pi_{00} p_{\text{max}_j}^2 + \pi_{10} p_{\text{max}_j} Pr(P_{\alpha_j} \leq p_{\text{max}_j} \mid \alpha_j \neq 0) + \pi_{01} p_{\text{max}_j} Pr(P_{\beta_j} \leq p_{\text{max}_j} \mid \beta_j \neq 0)$$

In this formula, we need to estimate three component probabilities: π_{00}, π_{10}, and π_{01}, and two power functions evaluated at p_{max_j}: $Pr(P_{\alpha_j} \leq p_{\text{max}_j} \mid \alpha_j \neq 0)$ and $Pr(P_{\beta_j} \leq p_{\text{max}_j} \mid \beta_j \neq 0)$. There are two various methods (product, maxp) to estimate three component probabilities. Both are derived from JC's method (Jin and Cai (2007)), which uses the empirical characteristic function and Fourier analysis to estimate $\pi_{0\bullet}$ (the proportion of null $\alpha_j = 0$) and $\pi_{\bullet 0}$ (the proportion of null $\beta_j = 0$).

- "product" method: The estimates of π_{00}, π_{10}, π_{01} are $\hat{\pi}_{00} = \hat{\pi}_{0\bullet}\hat{\pi}_{\bullet 0}/\hat{\pi}_0$, $\hat{\pi}_{10} = (1-\hat{\pi}_{0\bullet})\hat{\pi}_{\bullet 0}/\hat{\pi}_0$, and $\hat{\pi}_{01} = \hat{\pi}_{0\bullet}(1-\hat{\pi}_{\bullet 0})/\hat{\pi}_0$, where $\hat{\pi}_0 = \hat{\pi}_{0\bullet} + \hat{\pi}_{\bullet 0} - \hat{\pi}_{0\bullet}\hat{\pi}_{\bullet 0}$.
- "maxp" method: The estimates of π_{00}, π_{10}, π_{01} are $\hat{\pi}_{00} = (\hat{\pi}_{0\bullet} + \hat{\pi}_{\bullet 0} - \hat{\pi}_0)/\hat{\pi}_0$, $\hat{\pi}_{10} = (\hat{\pi}_0 - \hat{\pi}_{0\bullet})/\hat{\pi}_0$, and $\hat{\pi}_{01} = (\hat{\pi}_0 - \hat{\pi}_{\bullet 0})/\hat{\pi}_0$, where $\hat{\pi}_0$ is obtained from JC's method by using p_{max_j}.

After obtaining the p-values on all internal nodes, we apply Benjamini-Hochberg (BH) false discovery rate procedure (Benjamini and Hochberg (1995)) to identify a collection of nodes on the phylogenetic tree with significant mediation effects. To test the global mediation null hypothesis $H_0 : \cap_{j=1}^J H_0^j$, we apply the harmonic mean p-value (HMP) method (Wilson (2019)) to combine local mediation p-values.

Specifically, the weighted harmonic mean of subcomposition mediation test p-values $p_1, \ldots, p_J$ is defined as

$$\mathring{p} = \frac{\sum_{j=1}^J w_j}{\sum_{j=1}^J w_j p_j},$$

where w_j's are weights that sum to 1 and we set $w_j = 1/J$ by default. The global test p-value can be obtained by calculating the tail probability from the $\mathring{p}$'s null distribution approximation.

Application with phylognetic information: Cecal data

It is well-known that low dose antibiotics have been used widely to stimulate weight gain in livestock. However, there is growing concern that antibiotic exposure may have long-term consequences. Several studies have shown that antibiotics can have great impact on the abundances of bacteria in the gut community. It is interesting to investigate whether the subtherapeutic antibiotic treatment effect on body weight is mediated through the perturbation of gut microbiome and study the underlying mechanisms.

The data here is from an experiment conducted by Cho et al. (2012), in which young mice were treated by different low-dose antibiotic and evaluated changes in body fat and compositions of the microbiome in cecal and fecal samples. The mice in antibiotic group were heavier than those in the control group. We will show how to perform `phyloMed` function by focusing on cecal samples.

```
> library(miMediation)
> # Load data
> data(data.cecal)
> # Take a look at the data
> Trt <- data.cecal$treatment
> table(Trt) # 0: control 1: antibotics
Trt
 0  1
```

```
10 38
> M <- data.cecal$mediators
> head(M[,1:6])
           3732 5004 4354 4432 3209 5058
cecal_C1      1    2   56   39   12   13
cecal_C10     1    7   60   42   34   31
cecal_C2      9    2   38   40   14    8
cecal_C3      4    4   41   53   16   18
cecal_C4      5    2  102   84   18   19
cecal_C5      5   13   83   62   29   29
> Y <- data.cecal$outcome
> summary(Y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17.20   20.55   21.80   22.32   23.38   32.10
> tree <- data.cecal$tree
```

To run phyloMed function, the parameter treatment, mediators, outcome, phylogeny tree and pi.method are required. Other inputs are optional. Note that if n.perm=1e5, the function will output p-value calculated through adaptive permutation procedure as well and it will take ∼ 6 minutes to output the result. You can set verbose=TRUE to keep track of the process. Here is an example described in the Hong, Chen, and Tang (Manuscript).

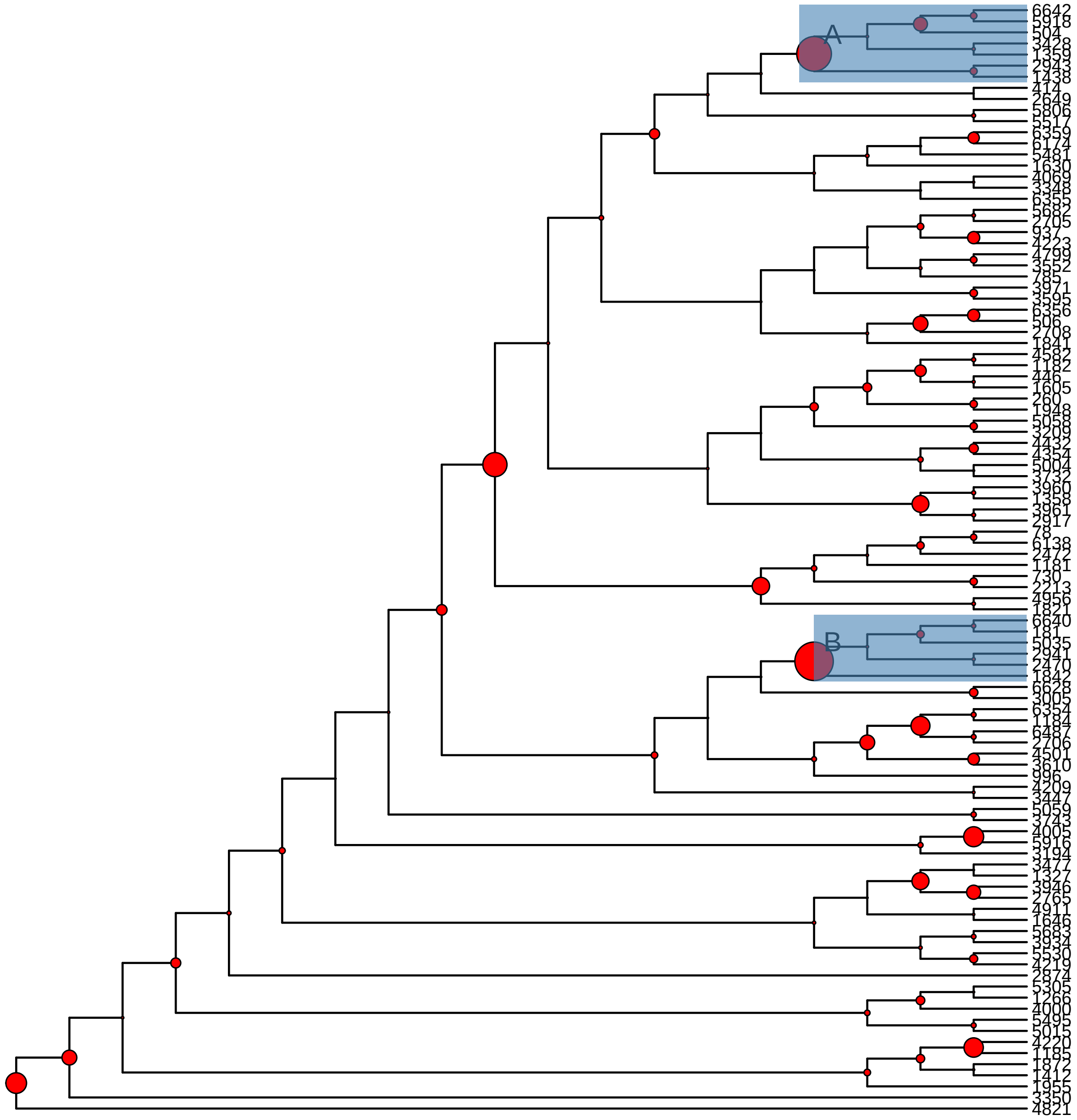
```
> # set random seed here so that you can get the same result every time you run the code
> set.seed(84)
> cecal.rsltlst <- phyloMed(Trt, M, Y, tree, fdr.alpha = 0.1, n.perm = 1e5, graph = TRUE)
>
> # take a look at phyloseq-class object
> cecal.physeq <- cecal.rsltlst$clean.data
> cecal.physeq
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 100 taxa and 48 samples ]
sample_data() Sample Data:       [ 48 samples by 2 sample variables ]
phy_tree()    Phylogenetic Tree: [ 100 tips and 99 internal nodes ]
> cecal.rslt <- cecal.rsltlst$rslt
> # take a look at rslt (PhyloMed.P)
> cecal.rslt$PhyloMed.P
$node.pval
 [1] 0.030249720 0.092335444 0.804658359 0.211165304 0.607828169 0.422528957
 [7] 0.982450877 0.814714539 0.169853475 0.015622998 0.771106344 0.862797494
[13] 0.969594333 0.484577282 0.871291406 0.259366125 0.269493336 0.376023190
[19] 0.269843492 0.176746122 0.748603126 0.634689441 0.328734323 0.075560518
[25] 0.658425404 0.604799852 0.566390083 0.877034873 0.891684145 0.909830655
[31] 0.735645104 0.374101141 0.385234590 0.136706734 0.685643960 0.292098093
[37] 0.775748956 0.091486313 0.154111793 0.197397182 0.862797494 0.603295673
[43] 0.862797494          NA 0.001977693 0.372196849 0.785181751 0.710874484
[49] 0.106154293 0.395567605 0.789973394 0.670886620 0.871291406 0.154603991
[55] 0.885773955 0.871291406 0.048660456 0.476565395 0.785181751 0.432264342
[61] 0.431437676 0.350083754 0.697090415 0.352930818 0.785181751 0.925446394
[67] 0.480541455 0.089538035 0.063109834 0.507901457 0.550680818 0.145076919
[73] 0.958103959 0.288488025 0.001168934 0.771106344 0.305193584 0.649771384
[79] 0.782804879 0.495994520 0.486618006 0.034923585 0.699030773 0.674523328
[85] 0.295792162 0.546873927 0.958103959 0.859998027 0.059727124 0.101071936
[91] 0.951414794 0.452994463 0.479541885 0.258716008 0.925446394 0.382561329
[97] 0.311923787 0.877034873 0.039373253
```

```
$sig.clade
$sig.clade$`145`
[1] "5918" "6642" "504"  "1359" "3428" "1438" "2943"

$sig.clade$`175`
[1] "2470" "2941" "181"  "6640" "5035" "1842"


$null.prop
       H00        H10        H01
0.59740074 0.36357720 0.03902206

$global.pval
       HMP
0.08532486
```



The output consists of four components:

- `node.pval`: mediation p-values on each internal node of the phylogenetic tree.
- `sig.clade`: identified mediation nodes with their descendants.
- `null.prop`: estimated proportion of three disjoint component null hypotheses.
- `global.pval`: global test p-value.

In the figure above, the size of the circle on internal node is proportional to $-\log_{10}$(subcompostion mediation p-value p_j), where p_j lives in the `node.pval` output. The identified mediation node is highlighted by a blue rectangle.

Application with taxonomic information: ZeeviD data

When there is no phylogenetic information available, the `phyloMed` function could construct taxonomic tree based on the taxonomic information. Here we sampled 200 subjects out of 900 healthy subjects in a real microbiome dataset Zeevi et al. (2015) and divided them into two equal-sized treatment groups. We used the top 100 most abundant OTUs and the associated taxonomy table to run `phyloMed` function.

```
> # Load data
> data(data.zeeviD)
> # Take a look at the data
> Trt <- data.zeeviD$treatment
> table(Trt) # 0: control 1: treatment
Trt
  0   1
100 100
> M <- data.zeeviD$mediators
> dim(M)
[1] 200 100
> Y <- data.zeeviD$outcome
> summary(Y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-8.1661 -2.1830 -0.1615  0.4094  2.1486  8.8780
> tree <- data.zeeviD$tree
> head(tree)
Taxonomy Table:     [6 taxa by 7 taxonomic ranks]:
                                       Kingdom    Phylum       Class
s__Megamonas_hypermegale               "Bacteria" "Firmicutes" "Negativicutes"
s__Megamonas_funiformis                "Bacteria" "Firmicutes" "Negativicutes"
s__Megamonas_rupellensis               "Bacteria" "Firmicutes" "Negativicutes"
s__Phascolarctobacterium_succinatutens "Bacteria" "Firmicutes" "Negativicutes"
s__Dialister_succinatiphilus           "Bacteria" "Firmicutes" "Negativicutes"
s__Ruminococcus_bromii                 "Bacteria" "Firmicutes" "Clostridia"
                                       Order             Family
s__Megamonas_hypermegale               "Selenomonadales" "Veillonellaceae"
s__Megamonas_funiformis                "Selenomonadales" "Veillonellaceae"
s__Megamonas_rupellensis               "Selenomonadales" "Veillonellaceae"
s__Phascolarctobacterium_succinatutens "Selenomonadales" "Acidaminococcaceae"
s__Dialister_succinatiphilus           "Selenomonadales" "Veillonellaceae"
s__Ruminococcus_bromii                 "Clostridiales"   "Ruminococcaceae"
                                       Genus
s__Megamonas_hypermegale               "Megamonas"
s__Megamonas_funiformis                "Megamonas"
s__Megamonas_rupellensis               "Megamonas"
s__Phascolarctobacterium_succinatutens "Phascolarctobacterium"
s__Dialister_succinatiphilus           "Dialister"
s__Ruminococcus_bromii                 "Ruminococcus"
                                       Species
s__Megamonas_hypermegale               "Megamonas_hypermegale"
s__Megamonas_funiformis                "Megamonas_funiformis"
s__Megamonas_rupellensis               "Megamonas_rupellensis"
s__Phascolarctobacterium_succinatutens "Phascolarctobacterium_succinatutens"
s__Dialister_succinatiphilus           "Dialister_succinatiphilus"
s__Ruminococcus_bromii                 "Ruminococcus_bromii"
```

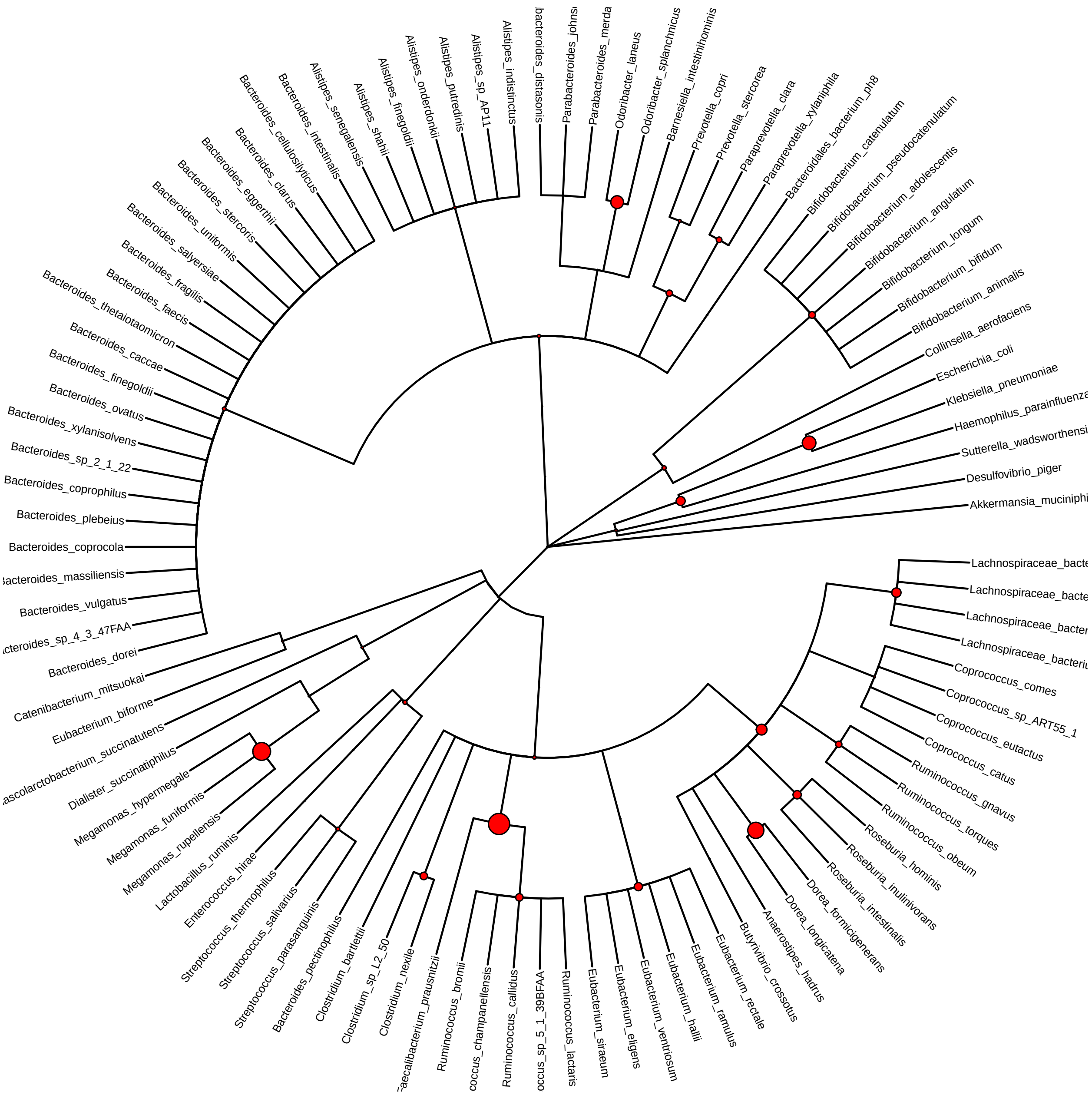
```
> # only show aysmptotic result
> demo.rsltlst <- phyloMed(Trt, M, Y, tree, graph = TRUE)
> # take a look at phyloseq-class object
> demo.physeq <- demo.rsltlst$clean.data
> demo.physeq
phyloseq-class experiment-level object
```

```
otu_table()   OTU Table:         [ 100 taxa and 200 samples ]
sample_data() Sample Data:       [ 200 samples by 2 sample variables ]
tax_table()   Taxonomy Table:    [ 100 taxa by 7 taxonomic ranks ]
> demo.rsltlst$rslt$PhyloMed.A
$node.pval
             Genus.Alistipes            Genus.Bacteroides
                  0.77601346                   0.63563292
       Genus.Bifidobacterium                Genus.Blautia
                  0.35594013                   0.38486173
           Genus.Clostridium            Genus.Coprococcus
                  0.30845559                   0.85310122
                 Genus.Dorea            Genus.Eubacterium
                  0.05920649                   0.26856782
Genus.Lachnospiraceae_noname              Genus.Megamonas
                  0.21775102                   0.04293845
           Genus.Odoribacter        Genus.Parabacteroides
                  0.11264747                   0.96369414
        Genus.Paraprevotella             Genus.Prevotella
                  0.44047726                   0.73153056
             Genus.Roseburia           Genus.Ruminococcus
                  0.26436600                   0.30956110
         Genus.Streptococcus    Family.Enterobacteriaceae
                  0.63083981                   0.09572732
  Family.Erysipelotrichaceae       Family.Lachnospiraceae
                  0.87936798                   0.16469013
   Family.Porphyromonadaceae        Family.Prevotellaceae
                  1.00000000                   0.38652436
      Family.Ruminococcaceae       Family.Veillonellaceae
                  0.02260512                   0.89786580
         Order.Bacteroidales          Order.Clostridiales
                  0.69673427                   0.67296673
       Order.Lactobacillales        Order.Selenomonadales
                  0.58618988                   0.81632370
        Class.Actinobacteria    Class.Gammaproteobacteria
                  0.56921384                   0.23213703
           Phylum.Firmicutes        Phylum.Proteobacteria
                  1.00000000                   0.80143156
            Kingdom.Bacteria
                  0.92340636

$sig.clade
NULL

$null.prop
      H00       H10       H01
0.7177680 0.1811051 0.1011268

$global.pval
      HMP
0.5366665
```



References

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300.

Cho, Ilseung, Shingo Yamanishi, Laura Cox, Barbara A Methé, Jiri Zavadil, Kelvin Li, Zhan Gao, et al. 2012. "Antibiotics in Early Life Alter the Murine Colonic Microbiome and Adiposity." *Nature* 488 (7413): 621–26.

Hong, Qilin, Guanhua Chen, and Zheng-Zheng Tang. Manuscript. "PhyloMed: A Phylogeny-Based Test of Mediation Effect in Microbiome."

Jin, Jiashun, and T Tony Cai. 2007. "Estimating the Null and the Proportion of Nonnull Effects in Large-Scale Multiple Comparisons." *Journal of the American Statistical Association* 102 (478): 495–506.

VanderWeele, Tyler. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction.* Oxford University Press.

Wilson, Daniel J. 2019. “The Harmonic Mean p-Value for Combining Dependent Tests.” *Proceedings of the National Academy of Sciences* 116 (4): 1195–1200.

Zeevi, David, Tal Korem, Niv Zmora, David Israeli, Daphna Rothschild, Adina Weinberger, Orly Ben-Yacov, et al. 2015. “Personalized Nutrition by Prediction of Glycemic Responses.” *Cell* 163 (5): 1079–94.