Tutorial: testing microbiome mediation effect using miMediation

Qilin Hong

Last compiled on 28 December, 2023

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. PhyloMed enables us to test the overall mediation effect of the entire microbial community and pinpoint internal nodes with significant subcomposition mediation effects.
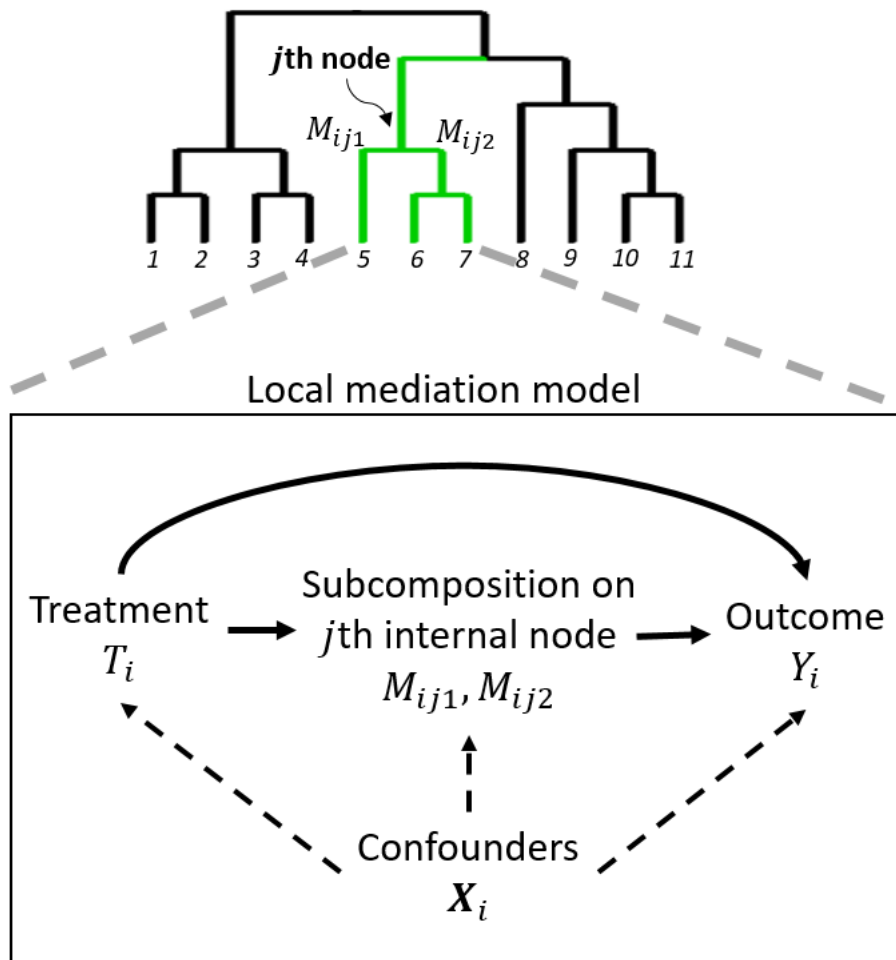
As depicted in the figure above, we propose to construct a local mediation model for the subcomposition at each internal node of the phylogenetic tree. The subcomposition on a given internal node consists of the relative abundance aggregated at its two child nodes. We apply the following robust linear 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 $\mathbf{X}_i$.

The local mediation null hypothesis is expressed as

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

, 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

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

The $P_{\alpha_j}, P_{\beta_j}$ represent the p-value for testing $\alpha_j = 0$ and $\beta_j = 0$, respectively. These two p-values could be obtained via asymptotic approach or permutation approach.

Thus, we obtain the p-value of mediation test in the jth local model using the following formula:

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

In this formula, we need to estimate three component probabilities $(\pi_{00}, \pi_{10}, \pi_{01})$ representing the proportion of three null hypotheses $(H_{00}^j, H_{10}^j, H_{01}^j)$. and two power functions evaluated at $p_{\max_j}$. We implement two methods (product, maxp) to estimate $\pi_{00}, \pi_{10}, \pi_{01}$.

- "product" method: $\hat{\pi}_{00} = \hat{\pi}_{0\bullet}\hat{\pi}_{\bullet 0}/\hat{\pi}_0$, $\hat{\pi}_{10} = (1 - \hat{\pi}_{0\bullet})\hat{\pi}_{\bullet 0}/\hat{\pi}_0$, 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: $\hat{\pi}_{00} = (\hat{\pi}_{0\bullet} + \hat{\pi}_{\bullet 0} - \hat{\pi}_0)/\hat{\pi}_0$, $\hat{\pi}_{10} = (\hat{\pi}_0 - \hat{\pi}_{0\bullet})/\hat{\pi}_0$, and $\hat{\pi}_{01} = (\hat{\pi}_0 - \hat{\pi}_{\bullet 0})/\hat{\pi}_0$.

Note that $\hat{\pi}_{0\bullet}, \hat{\pi}_{\bullet 0}, \hat{\pi}_0$ are estimated by applying Jin and Cai's method (Jin and Cai (2007)) to $P_{\alpha_j}, P_{\beta_j}, P_{\max_j}$. 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.

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 minimum requirement is to provide `treatment`, `mediators`, `outcome`, `tree` information. In the chunk below, we set FDR $= 0.1$ (`fdr.alpha=0.1`) in identifying mediating nodes and visualize the results in the tree plot (`graph="rectangular"`). Note that if `n.perm=1e4`, the function will output p-value calculated through permutation procedure as well and it will take ~ 3 minutes to output the result. In general, permutation procedure can provide more accurate result when sample size is small (e.g., sample size < 100). However, it is slower than asymptotic procedure. You can set `verbose=TRUE` to keep track of the process.

```
> # set random seed here so that you can get the same result every time you run the code
> set.seed(123)
> cecal.rsltlst <- phyloMed(Trt, M, Y, tree = tree, fdr.alpha = 0.1,
+                           n.perm = 1e4, graph = "rectangular")
Run phyloMed based on phylogenetic tree!
>
> # 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
    Node101     Node102     Node103     Node104     Node105     Node106
```

```
 0.030281982 0.087212850 0.785087274 0.201418675 0.621643166 0.398931998
      Node107      Node108      Node109      Node110      Node111      Node112
 0.978139609 0.854379789 0.189859663 0.016098017 0.744129812 0.802090213
      Node113      Node114      Node115      Node116      Node117      Node118
 0.982453199 0.420754447 0.919106969 0.249736872 0.256278130 0.382368257
      Node119      Node120      Node121      Node122      Node123      Node124
 0.258854190 0.172234807 0.718848156 0.601642392 0.362137412 0.068381195
      Node125      Node126      Node127      Node128      Node129      Node130
 0.656538756 0.641171215 0.524807997 0.922249833 0.912876455 0.888669021
      Node131      Node132      Node133      Node134      Node135      Node136
 0.792296692 0.353887654 0.378420707 0.147483826 0.678061992 0.291097804
      Node137      Node138      Node139      Node140      Node141      Node142
 0.735532041 0.088049675 0.157337728 0.198011861 0.775650836 0.647922241
      Node143      Node144      Node145      Node146      Node147      Node148
 0.843456006          NA 0.002121006 0.387749092 0.812096288 0.704785157
      Node149      Node150      Node151      Node152      Node153      Node154
 0.110456597 0.417617649 0.780344203 0.676221278 0.912876455 0.154203674
      Node155      Node156      Node157      Node158      Node159      Node160
 0.894615983 0.838084609 0.053767158 0.441361251 0.832772308 0.401819283
      Node161      Node162      Node163      Node164      Node165      Node166
 0.389802395 0.366394584 0.678061992 0.363950851 0.802090213 0.931790679
      Node167      Node168      Node169      Node170      Node171      Node172
 0.522471300 0.091543278 0.066833947 0.508831832 0.522471300 0.144818515
      Node173      Node174      Node175      Node176      Node177      Node178
 0.958083964 0.283995758 0.000834879 0.814631773 0.335736433 0.612268401
      Node179      Node180      Node181      Node182      Node183      Node184
 0.687398972 0.486436564 0.534343639 0.033309751 0.679911520 0.634543086
      Node185      Node186      Node187      Node188      Node189      Node190
 0.285157167 0.546704499 0.944779576 0.854379789 0.066696399 0.113809008
      Node191      Node192      Node193      Node194      Node195      Node196
 0.958083964 0.506620196 0.507723866 0.267913114 0.888669021 0.358559400
      Node197      Node198      Node199
 0.296433343 0.859934069 0.036027716

$sig.clade
$sig.clade$Node175
[1] "2470" "2941" "181"  "6640" "5035" "1842"


$null.prop
       H00         H10         H01
0.59817608  0.36254122  0.03928271

$global.pval
       HMP
0.06854347
```

4

The output consists of four components:

- `node.pval`: mediation *p*-values on each internal node of the phylogenetic tree.
- `sig.clade`: identified mediation node ids with their corresponding leaf-level descendant taxa name.
- `null.prop`: estimated proportion of three disjoint component null hypotheses.
- `global.pval`: global test *p*-value.

Note that p-value is NA at internal node 144. If we set `verbose=TRUE`, we could know the reason during the process. The underlying reason is that all values in the treatment variable equal to one after removing the subjects with subcomposition being zero. Thus, we skip this specific node.

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 taxonomy table. The `data.zeeviD` is a simulated dataset based on a real gut microbiome dataset from a healthy cohort (Zeevi et al. (2015)). The mediation signal was added at "Order.Clostridiales".

```
> # Load data
> data(data.zeeviD)
> # Take a look at the data
> Trt <- data.zeeviD$treatment
> table(Trt) # 0: control 1: treatment
Trt
  0   1
200 200
> M <- data.zeeviD$mediators
> dim(M)
[1] 400 100
> Y <- data.zeeviD$outcome
> summary(Y)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-238.401 -113.996  -58.528  -60.717   -6.791  183.129
> tree <- data.zeeviD$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"

> # run asymptotic result by default
> demo.rsltlst <- phyloMed(Trt, M, Y, tree = tree,
```

```
+                                    fdr.alpha = 0.1, graph = "circular")
No phylogenetic tree available, construct taxonomic tree!
> # take a look at phyloseq-class object
> demo.physeq <- demo.rsltlst$clean.data
> demo.physeq
phyloseq-class experiment-level object
otu_table()   OTU Table:        [ 100 taxa and 400 samples ]
sample_data() Sample Data:      [ 400 samples by 2 sample variables ]
tax_table()   Taxonomy Table:   [ 100 taxa by 7 taxonomic ranks ]
> demo.rsltlst$rslt$PhyloMed.A
$node.pval
            Genus.Alistipes              Genus.Bacteroides
                 0.170003884                    0.186275052
       Genus.Bifidobacterium                 Genus.Blautia
                 0.781201593                    0.576986566
          Genus.Clostridium             Genus.Coprococcus
                 0.163289948                    0.394528854
                Genus.Dorea              Genus.Eubacterium
                 0.079844820                    0.634174237
Genus.Lachnospiraceae_noname             Genus.Megamonas
                 0.385521850                    0.757961853
          Genus.Odoribacter          Genus.Parabacteroides
                 0.507963359                    0.232969190
        Genus.Paraprevotella              Genus.Prevotella
                 0.029002277                    0.435270744
            Genus.Roseburia             Genus.Ruminococcus
                 0.416151196                    0.244208259
        Genus.Streptococcus      Family.Enterobacteriaceae
                 0.147477523                    0.751206698
   Family.Erysipelotrichaceae      Family.Lachnospiraceae
                 0.449547728                    0.571919231
    Family.Porphyromonadaceae       Family.Prevotellaceae
                 0.496710495                    0.988395637
       Family.Ruminococcaceae       Family.Veillonellaceae
                 0.828254477                    0.727099201
           Order.Bacteroidales          Order.Clostridiales
                 0.147080796                    0.001585361
         Order.Lactobacillales        Order.Selenomonadales
                 0.576675904                    0.058233749
          Class.Actinobacteria  Class.Gammaproteobacteria
                 0.789683999                    0.673106970
            Phylum.Firmicutes         Phylum.Proteobacteria
                 0.142161075                    0.217352633
            Kingdom.Bacteria
                 0.700154726


$sig.clade
$sig.clade$Order.Clostridiales
 [1] "Ruminococcus_bromii"            "Faecalibacterium_prausnitzii"
 [3] "Eubacterium_siraeum"            "Ruminococcus_champanellensis"
 [5] "Ruminococcus_callidus"          "Coprococcus_catus"
 [7] "Butyrivibrio_crossotus"         "Eubacterium_eligens"
 [9] "Bacteroides_pectinophilus"      "Eubacterium_ventriosum"
```

```
[11] "Clostridium_sp_L2_50"                "Coprococcus_eutactus"
[13] "Coprococcus_sp_ART55_1"              "Eubacterium_hallii"
[15] "Lachnospiraceae_bacterium_5_1_63FAA" "Anaerostipes_hadrus"
[17] "Eubacterium_ramulus"                 "Eubacterium_rectale"
[19] "Roseburia_intestinalis"              "Roseburia_inulinivorans"
[21] "Roseburia_hominis"                   "Ruminococcus_obeum"
[23] "Ruminococcus_sp_5_1_39BFAA"          "Clostridium_nexile"
[25] "Coprococcus_comes"                   "Dorea_longicatena"
[27] "Dorea_formicigenerans"               "Ruminococcus_lactaris"
[29] "Lachnospiraceae_bacterium_1_1_57FAA" "Lachnospiraceae_bacterium_8_1_57FAA"
[31] "Ruminococcus_torques"                "Lachnospiraceae_bacterium_3_1_46FAA"
[33] "Ruminococcus_gnavus"                 "Clostridium_bartlettii"


$null.prop
      H00       H10       H01
0.3316206 0.1427697 0.5256097

$global.pval
       HMP
0.05735723
```
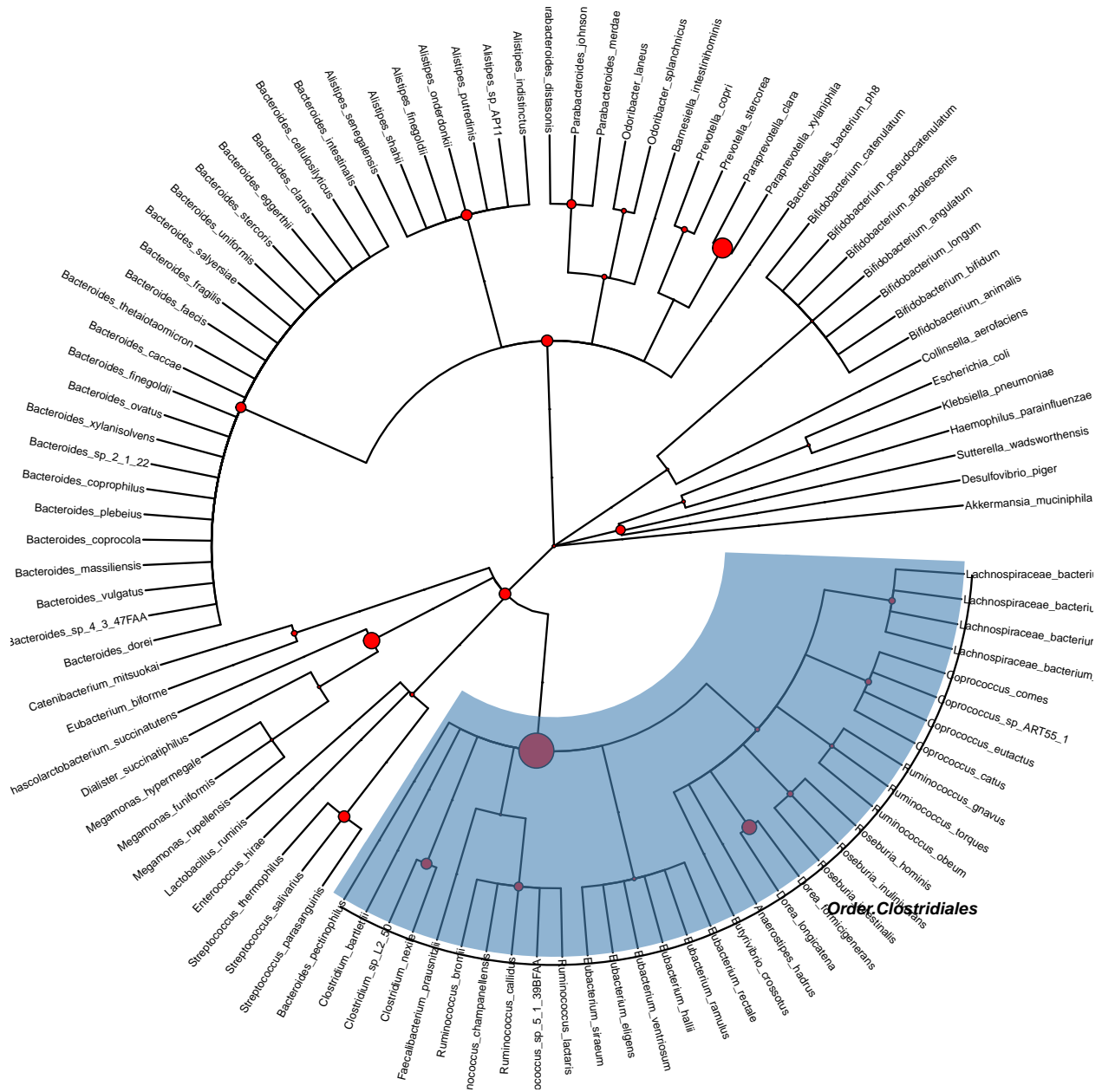
References

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.

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.