Tutorial: testing microbiome mediation effect using miMediation

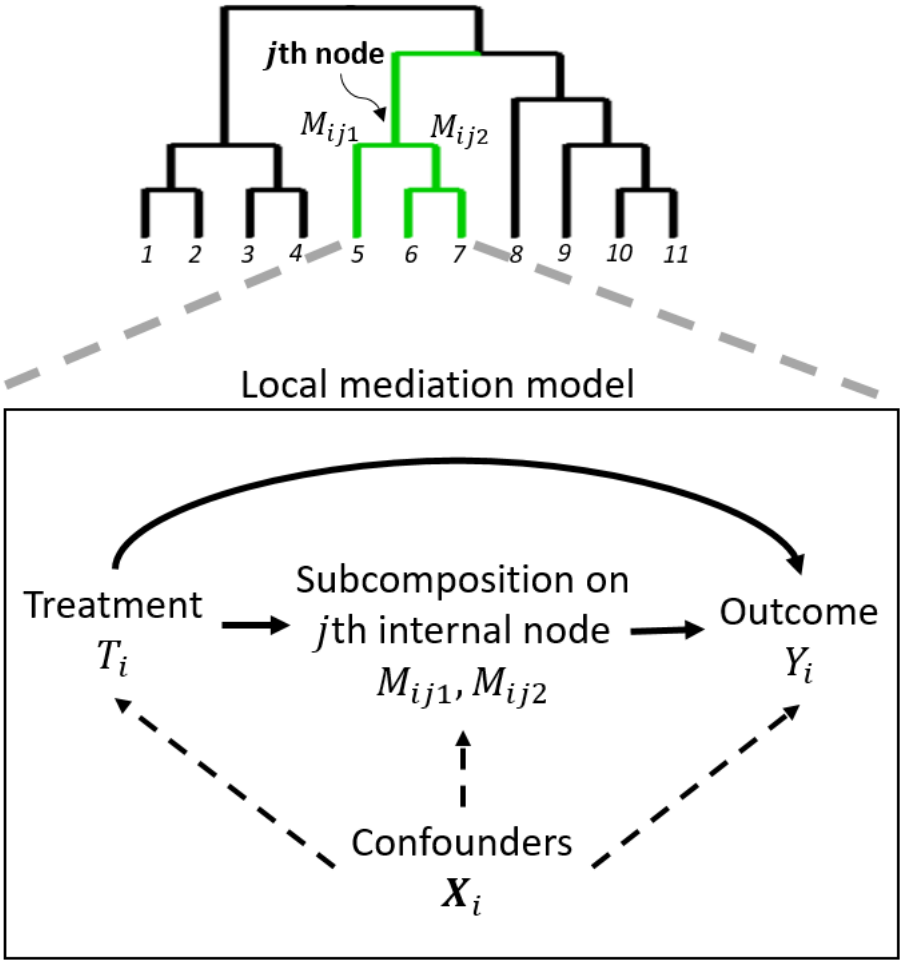
Qilin Hong

Last compiled on 23 August, 2021

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 (2021 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 linear and generalized linear regression models 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

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

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

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

In this formula, we need to estimate three component probabilities: π_{00}, π_{10}, and π_{01}, and two power functions evaluated at $p_{\max_j}$: $Pr(P_{\alpha_j} \leq p_{\max_j} \mid \alpha_j \neq 0)$ and $Pr(P_{\beta_j} \leq p_{\max_j} \mid \beta_j \neq 0)$. There are two various methods to estimate three component probabilities, the default method is JC's method:

- 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 $\beta_j = 0$). Then, 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 $\hat{\pi}_0 = \hat{\pi}_{0\bullet} + \hat{\pi}_{\bullet 0} - \hat{\pi}_{0\bullet}\hat{\pi}_{\bullet 0}$.
- Storey's method (Storey (2002), Dai, Stanford, and LeBlanc (2020)), which picks a tuning parameter λ from 0 to 1 to handle bias-variance trade-off and estimate $\pi_{0\bullet}, \pi_{\bullet 0}$ and π_{00}. Then, the estimates of π_{10}, π_{01} are $\hat{\pi}_{10} = \hat{\pi}_{\bullet 0} - \hat{\pi}_{00}$, $\hat{\pi}_{01} = \hat{\pi}_{0\bullet} - \hat{\pi}_{00}$.

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 Simes' method (Simes (1986)) to combine local mediation p-values.

Real data application: Antibiotic∼Microbiome∼BodyFat(%)

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])
          1185 4220 1412 1872 1955 1266
cecal_C1    57   71    5    5   28    9
cecal_C10  167  227   19   11    0   10
```

```
cecal_C2    65  103    0    2    1    7
cecal_C3    63   89    2   15   11   12
cecal_C4    51   80    2    2    1    5
cecal_C5    55   66    4    6    7   13
> 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
> tree

Phylogenetic tree with 100 tips and 92 internal nodes.

Tip labels:
    3732, 5004, 4354, 4432, 3209, 5058, ...
Node labels:
    , 0.000, 0.825, 0.435, , , ...

Unrooted; includes branch lengths.
```

To run `phyloMed` function, the parameter treatment, mediators, outcome, phylogeny tree and 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 (2021 Manuscript).

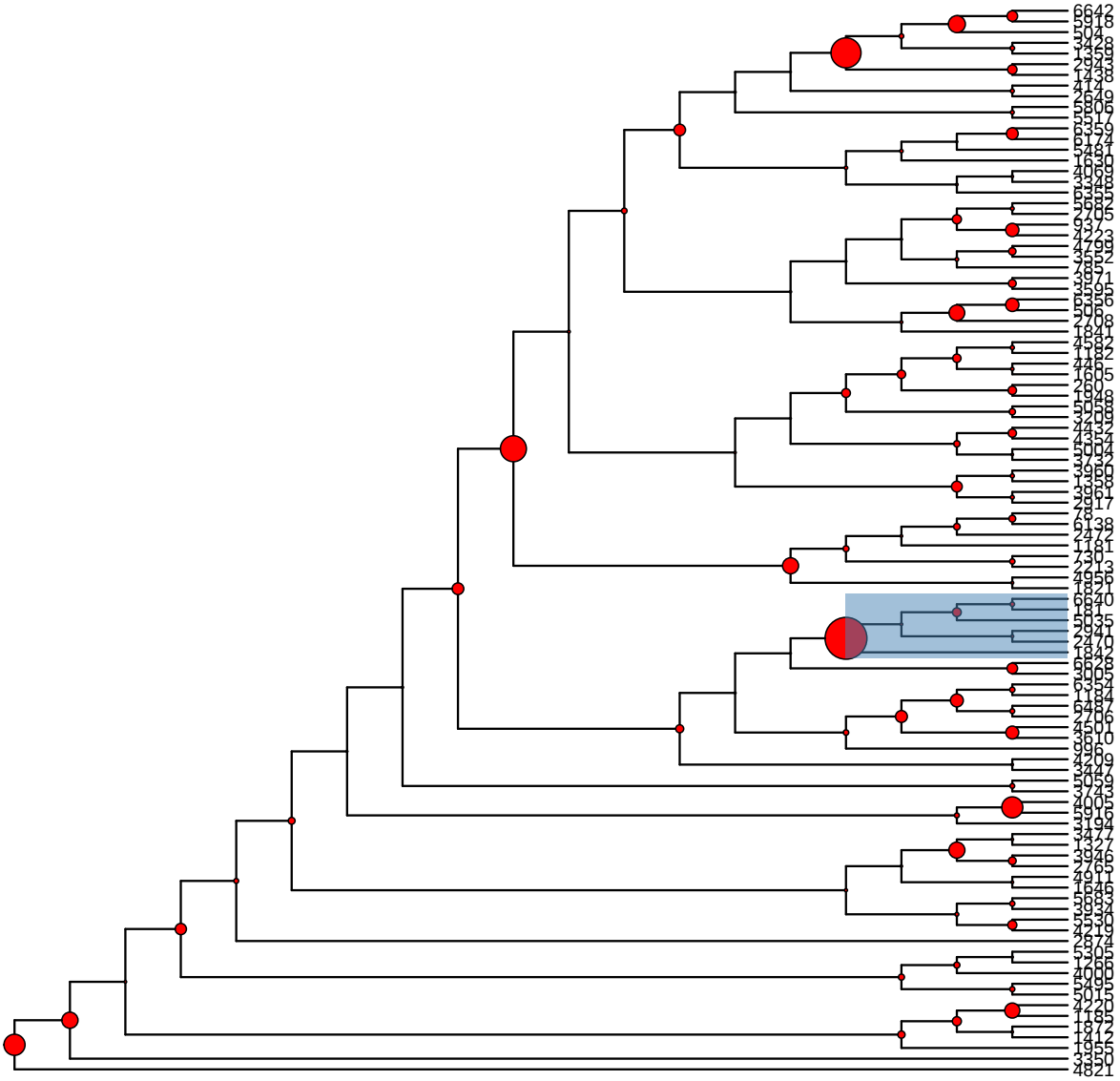
```
> # set random seed here so that you can get the same result every time you run the code
> set.seed(123)
> cecal.rsltlst <- phyloMed(Trt, M, Y, tree, fdr.alpha = 0.1, n.perm = 1e5, graph = TRUE)
>
> # take a look at phyloseq-class object
> cecal.physeq <- cecal.rsltlst$clean.data
> cecal.physeq
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 100 taxa and 48 samples ]
sample_data() Sample Data:       [ 48 samples by 2 sample variables ]
phy_tree()    Phylogenetic Tree: [ 100 tips and 99 internal nodes ]
> cecal.rslt <- cecal.rsltlst$rslt
> # take a look at rslt (PhyloMed.P)
> cecal.rslt$PhyloMed.P
$node.pval
 [1] 0.0327259826 0.0775136715 0.8199627641 0.1821254551 0.5531281386
 [6] 0.3911962796 0.9531539722 0.8884706255 0.1691043714 0.0140636945
[11] 0.7718676307 0.8675087784 0.9868810030 0.4303856080 0.8445642089
[16] 0.3046900019 0.2790741793 0.4303856080 0.2986896841 0.3028207684
[21] 0.7155519862 0.6253244811 0.2986896841 0.1986319779 0.6495228932
[26] 0.6320771996 0.4888107915 0.8854235733 0.9329601361 0.8870794598
[31] 0.6831043187 0.3585540577 0.2641918886 0.1285129265 0.6696961871
[36] 0.3310184175 0.7915145743 0.0833593363 0.1268843867 0.1715351244
[41] 0.8907697813 0.6603942832 0.9868810030 0.6442025581 0.0069566825
[46] 0.2523505364 0.5711969614 0.5755131012 0.0669011896 0.2046102031
[51] 0.7176678576 0.7030789415 0.9289998089 0.1653458425 0.8362252982
[56] 0.8334768217 0.0787774446 0.4599119750 0.8199627641 0.4143963486
[61] 0.3926707182 0.5102930657 0.7915145743 0.3199107623 0.8587814224
```

```
[66] 0.9211557885 0.5021892977 0.1627764250 0.1367538577 0.5585805914
[71] 0.5174232318 0.1366312943 0.9782803914 0.2055241219 0.0008864266
[76] 0.8042330015 0.2802818998 0.5903181897 0.7840476601 0.5655290602
[81] 0.5504365844 0.0329500690 0.7553515215 0.6549196233 0.2656514946
[86] 0.5284505972 0.8587814224 0.8120328116 0.0784125472 0.3354335112
[91] 0.8226355229 0.4465676740 0.5309570233 0.4493692936 0.8946185154
[96] 0.3655961298 0.2663867979 0.8587814224 0.0938189709

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


$null.prop
       H00        H10        H01
0.64701256 0.30100453 0.05198292

$global.pval
     Simes
0.08775624
```



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.

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.

Dai, James Y, Janet L Stanford, and Michael LeBlanc. 2020. "A Multiple-Testing Procedure for High-Dimensional Mediation Hypotheses." *Journal of the American Statistical Association*, 1–16.

Hong, Qilin, Guanhua Chen, and Zheng-Zheng Tang. 2021 Manuscript. "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.

Simes, R John. 1986. "An Improved Bonferroni Procedure for Multiple Tests of Significance." *Biometrika* 73 (3): 751–54.

Storey, John D. 2002. "A Direct Approach to False Discovery Rates." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (3): 479–98.

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