A PERMUTATION TEST FOR MEDIATING EFFECT OF HUMAN MICROBIOME USING DISTANCE-BASED APPROACHES.

Reproducing the findings presented in (Zhang et al., 2018).

1. INTRODUCTION

Recent studies have revealed a complex interplay between environment, the human microbiome and health and disease (Grice et al., 2009; MetaHIT Consortium (additional members) et al., 2011). Mediation analysis of the human microbiome in these complex relationships could potentially provide insights into the role of the microbiome in the aetiology of disease and, more importantly, lead to novel clinical interventions by modulating the microbiome (Faith et al., 2013; Lozupone et al., 2012).

Quite a few statistical methods have been proposed to test the association between the microbiome compositions and covariates of interest (e.g. environmental factors or clinical outcomes) based on 16S data, where 16S rRNA gene sequence tags are clustered into operational taxonomic units (OTUs) based on sequence divergence (Chen et al., 2012; Zhao et al., 2015). Researchers usually utilize distance metrics, measuring the pairwise dissimilarity in the microbiome profiles, to compute test statistics, and employ permutation tests to calculate the p-value. The performance of these distance-based methods depends on the choice of the distance metric (Chen et al., 2012). However, none of the existing methods which utilise just one distance are suitable for testing such clinically important mediation effect. High dimensionality, sparsity, non-normality and phylogenetic structure of microbiome data add to the complexity of the problem. Recently, methods that accommodate multiple distances have been proposed (Tang et al., 2016). They exhibit controlled Type I error rate and yield good performance comparable to the best choice of distance metric. However, previous methods could only analyse and test bivariate relations.

Traditional mediation analysis, which tests if a single variable mediates the relationship between a known exposure and an outcome, has been widely applied in biomedical, behavioural, and psychosocial studies (Baron & Kenny, 1986). The authors (Zhang et al., 2018) propose a distance-based approach for testing the mediation effect of the human microbiome. In the framework, the nonlinear relationship between the human microbiome and independent/dependent variables is captured implicitly using sample-wise ecological distances, and the phylogenetic tree information is incorporated by using phylogeny-based distance metrics. Multiple distance metrics are utilized to maximize the power to detect various types of mediation effect. The authors notes that simulation studies demonstrate that this method has correct Type I error control and is robust and powerful under various mediation models. In this report, we will

attempt to partially reproduce the results presented, by means of one numerical study. Additionally, we examine how this method, when applied to a real gut microbiome dataset reveals that the association between the dietary fibre intake and body mass index was mediated by the gut microbiome.

2. METHODS

2.1. Mediation model

Let $\boldsymbol{M}$ be an $n \times m$ matrix of either counts that measures abundance of m OTUs for n microbiota samples or a binary matrix that indicates presence of absence of those m OTUs. Let $\boldsymbol{X}$ be an $n \times 1$ vector denoting the independent variable and let $\boldsymbol{Y}$ denote the $n \times 1$ outcome variable. We assume that the microbiome mediates the effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ through some unknown set of microbiome feature vectors $\boldsymbol{f_M^{(l)}}$ for $l = 1, 2, \dots, L$. Due to the multivariate nature of the microbiome data, it is possible to have multiple microbiome features that mediate the effect.

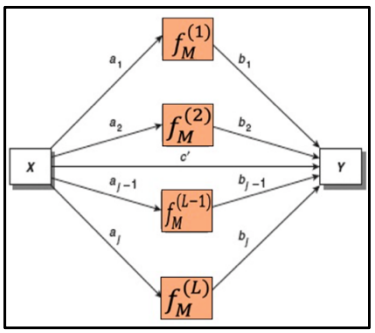

Figure 1: Graphical representation of the mediation model presented by Zhang et al. (2018).

We assume the following mediation model.

$$\boldsymbol{Y} = \boldsymbol{X}\gamma^* + \boldsymbol{\epsilon},$$
$$\boldsymbol{f_M^{(l)}} = \boldsymbol{X}a_l + \boldsymbol{\epsilon_l'}, \quad [l = 1,2, \dots, L]$$
$$\boldsymbol{Y} = \sum_{l=1}^{L} \boldsymbol{f_M^{(l)}} b_l + \boldsymbol{X}\gamma + \boldsymbol{\epsilon''},$$

where γ^* and γ represent the total and direct effects of $\boldsymbol{X}$ on $\boldsymbol{Y}$ respectively; $\boldsymbol{\epsilon}, \boldsymbol{\epsilon'}$ and $\boldsymbol{\epsilon''}$ are independent white noise vectors. Figure 1 provides a graphical representation of the model presented above. To establish a mediation pathway from $\boldsymbol{X}$ to $\boldsymbol{Y}$ along the microbiome feature vectors $\boldsymbol{f_M^{(l)}}$, we need to investigate whether there is a significant relation of $\boldsymbol{X}$ to some mediating feature $\boldsymbol{f_M^{(l)}}$ and whether the mediating variable $\boldsymbol{f_M^{(l)}}$ is significantly related to the dependent variable $\boldsymbol{Y}$, when adjusted for variation in $\boldsymbol{X}$. The null hypothesis may be expressed as

$$H_0: a_l b_l = 0 \quad \forall\, l \in 1,2, \dots, L,$$

and the alternative hypothesis H_1 is that H_0 is violated for some $l \in 1,2, \dots, L$.

2.2. A distance-based test for mediation

Had the microbiome feature vectors been known, we could apply traditional mediation tests. In practice, we have little knowledge about the specific microbiome features that mediate the effect. The power of the mediation test thus relies on a good choice of microbiome features that capture the mediation relationship as precisely as possible. One simple strategy is to treat the abundance of each OTU as the microbiome feature, perform tests on all the OTUs and apply Bonferroni correction to control the FWER. However, due to the extreme sparsity in the OTU data, individual tests are usually underpowered. To enrich signals and reduce multiple testing burden, community-level analysis, which considers all OTUs jointly, has been proposed to improve the power (Zhao et al., 2015).

Given the nature of microbiota, we are interested in defining the microbiome features based on the phylogenetic tree of OTUs. Environmental exposure or disease usually affects a cluster of phylogenetically related OTUs, which share a similar biological function. To accommodate the tree structure, we propose to form feature vectors while respecting the tree structure so that the PCs could capture the variation of evolutionarily related OTUs. One way to achieve this is through principal coordinate analysis (Jolliffe & Cadima, 2016) of a distance matrix, where the distance incorporates the tree structure information.

The authors thus propose a distance-based non-parametric method to test the mediation effects. The test consists of two parts: a distance-based test statistic and a permutation scheme to approximate the distribution under the null. Let $\boldsymbol{D} = \left(d_{ij}\right) \in R^{n\times n}$ be the distance be the distance matrix that measures the dissimilarity between the samples based on their microbiota profiles. The microbiome features could be formed by performing eigen-decomposition on the double centred matrix of squared distances, which is defined as

$$\boldsymbol{G} = \left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right)\boldsymbol{A}\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right),$$

where $\boldsymbol{I}$ is the identity matrix and $\mathbf{1}$ is a vector of ones. We further define $\boldsymbol{A} = \left(a_{ij}\right) = \left(-d_{ij}^2/2\right)$ and extract the eigen values $(\lambda_1, \lambda_2, \dots, \lambda_l)$ and eigen vectors $(\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_l)$ of $\boldsymbol{G}$. The test statistic is defined as

$$T = \sum_{l=1}^{L} \lambda_l \left|\hat{a}_l \hat{b}_l\right|.$$

We apply the permutation test to calculate the p-value based on T. For the j'th permutation we permute $\boldsymbol{X}$ (to get $\boldsymbol{X}^{(j)}$) and the residual of $\boldsymbol{Y}$ when regressed on $\boldsymbol{X}$ (to get $\boldsymbol{Y}_{\boldsymbol{X}}^{(j)}$). We calculate the following statistics:

$$T_X^{(j)} = \sum_{l=1}^{L} \lambda_l \left|< X^{(j)}, u_l >< Y_X, u_j >\right|,$$

$$T_Y^{(j)} = \sum_{l=1}^{L} \lambda_l \left| < X, u_l >< Y_X^{(j)}, u_j > \right|,$$

$$T_{X,Y}^{(j)} = \sum_{l=1}^{L} \lambda_l \left| < X^{(j)}, u_l >< Y_X^{(j)}, u_j > \right|.$$

The final test statistic under the j'th permutation is calculated as $T^{(j)} = \max\left\{T_X^{(j)}, T_Y^{(j)}, T_{X,Y}^{(j)}\right\}$. The final p-value is obtained as the proportion of permuted statistics equal to or larger than the observed statistic T.

2.3. Choice of distances

Each distance metric represents a distinct view of the microbiota and is expected to be the most powerful to detect a specific mediation pat- tern. However, in real applications, we may have little knowledge about the underlying mediation mechanism. Sticking to a single distance could miss important mediation effect(s). Therefore, considering different distance measures is key to the robustness and power of the test. We consider non-tree-based distances in Jaccard and Bray-Curtis distances, and unweighted, weighted and generalized unique fraction (UniFrac) distances, which account for the phylogenetic relationships between OTUs.

2.4. Numerical example: real data application

Diet strongly affects human health, partly by modulating gut microbiome composition (Wu et al., 2011). The authors study the association of long-term dietary and environmental variables with the gut microbiota. Ninety-eight healthy volunteers were enrolled in the cross-sectional study. The volunteers' long-term diet information was collected through food frequency questionnaire and converted to intake amounts of 214 nutrient categories. At the same time, their stool samples were collected, and the DNA samples were analyzed to yield rDNA gene segments. The data also has measurements of body mass index (BMI).

Knowing that researchers have established significant association between dietary fibre intake and BMI, as well as a significant association between gut microbiota and fibre intake, (Wu et al., 2011) we wish to examine if the association between fibre intake and BMI is mediated by the gut microbiota by means of the distance-based permutation test proposed by (Zhang et al., 2018). ***Table 1*** (taken from (Wu et al., 2011)) reports the p-values for association tests between fibre intake (X), gut microbiota (M) and BMI (Y). The p-value for $Y \sim X$ are calculated using the standard F-test, while those corresponding to $M \sim X$ and $Y \sim M$ are calculated by means of the Microbiome Regression-based Kernel Association Test (MiRKAT) (Zhao et al., 2015). ***Table 2*** summarises the results summarizes the results for the single distance-based mediation tests and the omnibus test.

3. RESULTS

From ***Table 1***, we note that X demonstrates significant associations with Y (p-value: 0.022), and M is significantly associated with both X and Y (p-value: 0.039 and 0.030 respectively). Thus, we consider the mediation test to check whether M mediates the effects of the X on Y. From ***Table 2*** we note only one of the distances could detect significant mediation effects – namely, the non-tree Jaccard distance (p-value: 0.007). The tree-based UniFrac detects a marginally significant effect (p-value: 0.088). These distances are known to be effective in capturing association patterns in presence/absence data. None of the other three distances yielded p-values less than the 0.05 threshold. Not knowing know the underlying mediation model in advance, using a single distance alone could miss important mediating effects. The proposed omnibus test, which simultaneously considers multiple distance metrics, achieves a single p-value of 0.033.

Table 1: p-values for association tests between fibre intake (X), BMI (Y) and microbiome (M). Reproduced from Wu et al. (2011).

	$Y \sim X$	$M \sim X$	$Y \sim M$
p-value	0.022	0.039	0.030

Table 2: p-values for single distance-based tests and the omnibus test. Single distance-based tests are based on the Bray-Curtis (BC), Jaccard (JAC), UniFrac, weighted UniFrac (WUniFrac) and generalised UniFrac (GUniFrac) distances.

	BC	**JAC**	**UniFrac**	**WUniFrac**	**GUniFrac**	**Omnibus**
p-value	0.539	0.007	0.088	0.780	0.573	0.033

4. DISCUSSION

The omnibus test presented by the authors is shown to be robust and powerful in detecting a mediation structure, both by means of simulation studies as well as real data analysis. The method is noted to be very general and may be extended to work for genetic data where linkage disequilibrium is of major significance. The proposed method focuses on detecting an overall mediation effect by using an ensemble of distance measures. Future directions of research may involve being able to identify those microbiota that are most important within the mediation structure. We note that this method suffers from the disadvantage of not being able to quantify direct and indirect effects, since direct mediation effects are not tested in this approach.

5. NOTE ON REPRODUCIBILITY

All the relevant code and data used in this report may be found in the Github repository: https://github.com/soumikp/bios815

REFERENCES

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., Collman, R. G., Bushman, F. D., & Li, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, *28*(16), 2106–2113. https://doi.org/10.1093/bioinformatics/bts342

Faith, J. J., Guruge, J. L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A. L., Clemente, J. C., Knight, R., Heath, A. C., Leibel, R. L., Rosenbaum, M., & Gordon, J. I. (2013). The Long-Term Stability of the Human Gut Microbiota. *Science*, *341*(6141), 1237439. https://doi.org/10.1126/science.1237439

Grice, E. A., Kong, H. H., Conlan, S., Deming, C. B., Davis, J., Young, A. C., NISC Comparative Sequencing Program, Bouffard, G. G., Blakesley, R. W., Murray, P. R., Green, E. D., Turner, M. L., & Segre, J. A. (2009). Topographical and Temporal Diversity of the Human Skin Microbiome. *Science*, *324*(5931), 1190–1192. https://doi.org/10.1126/science.1171700

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., & Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature*, *489*(7415), 220–230. https://doi.org/10.1038/nature11550

MetaHIT Consortium (additional members), Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., … Bork, P. (2011). Enterotypes of the human gut microbiome. *Nature*, *473*(7346), 174–180. https://doi.org/10.1038/nature09944

Tang, Z.-Z., Chen, G., & Alekseyenko, A. V. (2016). PERMANOVA-S: Association test for microbial community composition that accommodates confounders and multiple distances. *Bioinformatics*, *32*(17), 2618–2625. https://doi.org/10.1093/bioinformatics/btw311

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F. D., & Lewis, J. D. (2011). Linking long-term dietary patterns with gut

microbial enterotypes. *Science (New York, N.Y.)*, *334*(6052), 105–108. https://doi.org/10.1126/science.1208344

Zhang, J., Wei, Z., & Chen, J. (2018). A distance-based approach for testing the mediation effect of the human microbiome. *Bioinformatics*, *34*(11), 1875–1883. https://doi.org/10.1093/bioinformatics/bty014

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H., & Wu, M. C. (2015). Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. *The American Journal of Human Genetics*, *96*(5), 797–807. https://doi.org/10.1016/j.ajhg.2015.04.003