**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 through the use of 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**

![Diagram

Description automatically generated]()Let be an matrix of either counts that measures abundance of OTUs for microbiota samples or a binary matrix that indicates presence of absence of those OTUs. Let be an vector denoting the independent variable and let denote the outcome variable. We assume that the microbiome mediates the effect of on through some unknown set of microbiome feature vectors for 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.

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

and the alternative hypothesis is that is violated for some

**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 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

,

where is the identity matrix and is a vector of ones. We further define and extract the eigen values and eigen vectors of The test statistic is defined as

.

We apply the permutation test to calculate the p-value based on For the th permutation we permute (to get ) and the residual of when regressed on (to get ). We calculate the following statistics:

,

,

.

The final test statistic under the ’th permutation is calculated as . The final p-value is obtained as the proportion of permuted statistics equal to or larger than the observed statistic

**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. 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. Simulation strategy**

We consider a model where the effect is mediated through OTU abundance. The independent variable affects the abundance of some OTU set, which in turn affects the outcome The microbiome counts are generated from a Dirichlet-Multinomial model