Review of Mediation Analysis on Microbiome

December 10, 2024

Abstract

Keywords: microbiome, mediation analysis

Background

The microbiome is a collection of various microbes (including bacteria, fungi and viruses). These microbes are not only found in the environment (e.g. soil, water and air), but the human body is also an important source of microbiome data. Although microbes are so small that we need a microscope to see them, the microbiome greatly affects human health. Disruption of the human microbiome caused by environmental factors has been linked to various diseases, including diabetes, obesity, and cardiovascular disease. With the advancement of tools for sequencing complex microbiome data, the study of microbiome data has attracted the interest of many biostatisticians and clinicians.

To study the microbiome, we first need to identify and analyze the composition of the microbiome and its related information by sequencing, this process is known as Microbiome Profiling[1]. Microbiome profiling reveals the structure and function of the microbial community in a given sample, and is the foundation for exploring how the microbiome affects human health and contributes to disease. Microbiome profiling begins with sample collection, which usually involves obtaining samples from a specific environment, such as the stool, mouth, or skin of a human being[2]. The goal is to derive and analyze information about the genetic material of microbes from these samples.

After we have finished collecting the samples and extracting the DNA from them, we need to use sequencing techniques[3] to get the original reads that can be used for analysis. For microbiome sequencing, several methods are available, introducing two of the most common and widely used methods. The first is Targeted Sequencing of Marker Genes, which includes 16S/18S/ITS sequencing[4]. Marker genes are typically genes that are widespread across different species and have specific functions. These genes perform the same basic functionality, but have a certain degree of variability among species, making them "fingerprints" for the identification of microbial species. Among these genes, 16S rRNA is the most commonly targeted gene and is widely used for identification and sequencing of bacteria and archaea[5, 6]. As a highly conserved gene, 16S rRNA plays a key role in the assembly of ribosomes and protein synthesis, which is essential for cell function and survival. At the same time, its relatively short gene size makes it ideally suited for targeted sequencing. Targeted marker gene sequencing only requires sequencing a small number of specific gene fragments, which greatly reduces computational costs, simplifies data processing and analysis, and enables excellent performance even with extremely large samples. In addition, 18S rRNA sequencing[7] is commonly used for microbiome analysis of eukaryotic microbes, whereas ITS sequencing[8] has a higher precision and is suitable for studying fungi at the genus or even species level. The second method is Metagenomic sequencing[9], which provides a comprehensive overview of the genetic information of the microbiome. This method is relatively easy to understand and unlike targeted sequencing, it does not require prior knowledge of specific microbes, but rather obtains DNA information for all microbes from a mixture of samples. This approach provides a comprehensive understanding of microbiome DNA information, microbial species composition, and more accurate identification of closely related microbial species when using advanced genome assembly tools such as metagenome-assembled genome (MAG)[10]. The amount of data required is large and the scope and precision of the analysis is significantly improved despite the computational complexity.

Through microbiome sequencing technology, we get original reads for analysis, these data lacks classification as well as processing and cannot be directly used to analyze the composition and functionality of microbial data. Therefore, after sequencing, necessary quality checking (QC)[11] is required. First, quality checking can effectively reduce errors and noise in the sequencing data to improve the accuracy of data analysis, and second, it can remove low-quality reads to prevent overestimation of community taxa diversity to avoid the generation of false-positive results. After cleaning the data, it is necessary to identify and categorize the sequences. There are also many methods for sequence identification, but most of the 16S rRNA gene sequences available today

(the primary source of microbiome data for our study) are essentially OTU-based methods[12, 4]. In general, operational taxonomic units with more than 97% similarity are considered to be the same taxonomic unit. OTU-based methods provide a fast way to assess microbial diversity by clustering these similar sequence fragments into groups, but they also have their drawbacks, as empirically set thresholds may underestimate the complexity of the phylogeny in reality. In recent years, Amplicon sequencing variants (ASVs)[13] have been proposed as an alternative to operational taxonomic units (OTUs) to overcome the limitations of OTU-based methods. Regardless of the sequencing or classification method being used, the Microbiome data obtained can be represented as an $n \times p$ matrix M, where M_{ij} denotes the relative abundance of the j-th feature in the i-th sample, since microbiome data are inherently compositional. The matrix M consists of a total of n samples and p features. Each sample i has a total sequencing depth or library size N_i defined as follows $\sum_{j=1}^{p} x_{ij} = N_i$.

There are many scientific questions worth exploring in microbiome research, and one of the most common is to explore the role of the microbiome in forming causal relationships between exposure and outcome. Because mediation analysis of the microbiome may provide insights on how the microbiome affects health and disease, it may lead to the development of clinical interventions and medications that are targeted towards adapting the microbiome. Mediation analysis is a widely used statistical method in a diverse range of fields to separate the total exposure-outcome effect into a direct effect without a mediator variable and an indirect effect through a mediator variable. For example, the gut microbiome can be used as a mediator variable to study how high intake of fat and other nutrients can affect BMI[14]. Mediation analysis typically attempts to account for the part of the exposure-outcome effect that is not directly observed, which is the indirect effect, explained by the effect of the mediator variable.

The research on mediation analysis can be traced back to Baron & Kenny's[15] article in 1986, where a series of significance tests were conducted to test the presence of mediation effects. Throughout the years, most of the mediation analysis methods can be broadly categorized into two frameworks: Traditional Mediation Analysis[16] and Causal Mediation Analysis[17]. Traditional Mediation Analysis is based on linear regression models and mainly consists of two methods: the "product of coefficients method" and the "difference of coefficients method"[18]. The difference between these two methods is mainly in the different ways of calculating indirect effects, but both need to satisfy the same statistical assumptions and have the same limitations, such as the residuals of linear regression need to satisfy the assumption of normal distribution and the assumption of homogeneity of variance. Causal Mediation Analysis defines natural direct effect and natural indirect effect based on potential outcomes or Counterfactual Framework[19], which addresses the limitations of traditional mediation analysis for complex models. A detailed explanation of the two frameworks will be discussed in the subsequent method review section.

Although a tremendous number of mediation analysis methods have been developed, due to the properties of microbiome data, very few suitable mediation analysis methods are available. Microbiome data have several key properties: (1) Microbiome data are high-dimensional, and the number of features may be significantly larger than the sample size[20]. (2) Microbiome data are compositional, meaning that the sum of the values of all features within each sample is equal to a constant (usually 1), as a result of normalizing the depth of sequencing across different samples[20]. (3) Microbiome data are Sparsity and Zero Inflation[20]. The lack of sequencing depth and the fact that most sample species contain only a small number of microbiome features results in a large number of zero values in the data. (4) The microbiome data are based on the Phylogenetic Tree structure. The source of the tree structure is introduced by sequencing technology, providing additional information on the diversity, ecological functions of the microbiome[21]. The high-dimensionality, compositional, sparsity, and phylogenetic structure of microbiome data pose significant challenges to conventional mediation analysis methods[22, 23]. Our goal is to bridge the gap between these unique properties of microbiome data and the development or selection of appropriate mediation analysis methods.

The aim of this article is to review the mediation analysis methods for microbiome data published in the last decade and through numerical simulations to provide recommendations for the selection of mediation analysis methods for future research. In the remaining sections, the article is divided into two parts: method review and numerical simulations. In the method review section, we reviewed a total of 12 mediation methods for the microbiome, prioritizing methods that are innovative and have implemented codes and packages. The reader is expected to have a basic knowledge of statistics and does not need to know the microbiome or mediation analysis in depth. In the process of reviewing the method, we focus on the motivation, framework, ideas, and strengths as well as weaknesses of the method, aiming to provide a quick overview of the method. In the numerical simulations section, we use a variety of simulations to show the performance of different methods in different scenarios. We hope this article will help the reader to get a comprehensive understanding of the mediation analysis for microbiome data developed in the recent years, and be able to decide the appropriate method based on their requirements in future studies.

References

[1] Micah Hamady and Rob Knight. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. *Genome research*, 19(7):1141–1152, 2009.

[2] Kjersti Aagaard, Joseph Petrosino, Wendy Keitel, Mark Watson, James Katancik, Nathalia Garcia, Shital Patel, Mary Cutting, Tessa Madden, Holli Hamilton, et al. The human microbiome project strategy for comprehensive sampling of the human microbiome and why it matters. *The FASEB Journal*, 27(3):1012, 2013.

[3] Wilhelm J Ansorge. Next-generation dna sequencing techniques. *New biotechnology*, 25(4):195–203, 2009.

[4] Richa Bharti and Dominik G Grimm. Current challenges and best-practice protocols for microbiome analysis. *Briefings in bioinformatics*, 22(1):178–193, 2021.

[5] GC Baker, Jacques J Smith, and Donald A Cowan. Review and re-analysis of domain-specific 16s primers. *Journal of microbiological methods*, 55(3):541–555, 2003.

[6] Jill E Clarridge III. Impact of 16s rrna gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical microbiology reviews*, 17(4):840–862, 2004.

[7] Kenan Hadziavdic, Katrine Lekang, Anders Lanzen, Inge Jonassen, Eric M Thompson, and Christofer Troedsson. Characterization of the 18s rrna gene for designing universal eukaryote specific primers. *PloS one*, 9(2):e87624, 2014.

[8] IJFW Álvarez and Jonathan F Wendel. Ribosomal its sequences and plant phylogenetic inference. *Molecular phylogenetics and evolution*, 29(3):417–434, 2003.

[9] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *nature*, 464(7285):59–65, 2010.

[10] Robert M Bowers, Nikos C Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, TBK Reddy, Frederik Schulz, Jessica Jarett, Adam R Rivers, Emiley A Eloe-Fadrosh, et al. Minimum information about a single amplified genome (misag) and a metagenome-assembled genome (mimag) of bacteria and archaea. *Nature biotechnology*, 35(8):725–731, 2017.

[11] Ravi K Patel and Mukesh Jain. Ngs qc toolkit: a toolkit for quality control of next generation sequencing data. *PloS one*, 7(2):e30619, 2012.

[12] Robert C Edgar. Uparse: highly accurate otu sequences from microbial amplicon reads. *Nature methods*, 10(10):996–998, 2013.

[13] Benjamin J Callahan, Paul J McMurdie, and Susan P Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME journal*, 11(12):2639–2643, 2017.

[14] Yi Wan, Fenglei Wang, Jihong Yuan, Jie Li, Dandan Jiang, Jingjing Zhang, Hao Li, Ruoyi Wang, Jun Tang, Tao Huang, et al. Effects of dietary fat on gut microbiota and faecal metabolites, and their relationship with cardiometabolic risk factors: a 6-month randomised controlled-feeding trial. *Gut*, 68(8):1417–1429, 2019.

[15] Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.

[16] Douglas Gunzler, Tian Chen, Pan Wu, and Hui Zhang. Introduction to mediation analysis with structural equation modeling. *Shanghai archives of psychiatry*, 25(6):390, 2013.

[17] Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological methods*, 15(4):309, 2010.

[18] Theis Lange, Kim Wadt Hansen, Rikke Sørensen, and Søren Galatius. Applied mediation analyses: a review and tutorial. *Epidemiology and health*, 39, 2017.

[19] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[20] Hongzhe Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2(1):73–94, 2015.

[21] Jian Xiao, Hongyuan Cao, and Jun Chen. False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing. *Bioinformatics*, 33(18):2873–2881, 2017.

[22] Jun Sun. Impact of bacterial infection and intestinal microbiome on colorectal cancer development. *Chinese Medical Journal*, 135(04):400–408, 2022.

[23] Christine B Peterson, Satabdi Saha, and Kim-Anh Do. Analysis of microbiome data. *Annual Review of Statistics and Its Application*, 11, 2023.