Comprehensive microbiome causal mediation analysis
using MiMed on user-friendly web interfaces

Hyojung Jang$^1$, Solha Park$^1$, Hyunwook Koh$^{1*}$

$^1$Department of Applied Mathematics and Statistics, The State University of New York, Korea,
Incheon, South Korea

$^*$These authors contributed equally to this work as first authors.

$^8$Corresponding author
Email: hyunwook.koh@stonybrook.edu (HK)
Abstract

It is a central goal of human microbiome studies to see the roles of the microbiome as a mediator that transmits environmental, behavioral, or medical exposures to health or disease outcomes. Yet, mediation analysis is not used as much as it should be. One reason is because of the lack of carefully planned routines, compilers and automatic computing systems for microbiome mediation analysis to perform a series of data processing, diversity calculation, data normalization, downstream data analysis and visualizations. Many researchers in various disciplines (e.g., clinicians, public health practitioners, biologists) are not also familiar with related statistical methods and programming languages on command-line interfaces. Thus, here we introduce a web cloud computing platform, named MiMed, that enables comprehensive microbiome mediation analysis on user-friendly web interfaces. We applied MiMed to the study on the mediating roles of oral microbiome in subgingival niches between e-cigarette smoking and gingival inflammation, and found significant mediating effects from two phyla (Proteobacteria, Spirochaetes), two classes (Flavobacterii, Betaproteobacteria), four orders (Flavobacterales, Burkholderiales, Neisseriales, Cardiobacterales), five families (Flavobacteriaceae, Burkholderiaceae, Neisseriaceae, Cardiobacteriaceae, Enterococcaceae), and eight genera (Bergeyella, Capnocytophaga, Actinomyces, Haemophilus, Kingella, Burkholderia, Cardiobacterium, Enterococcus).

Importance
The main features of MiMed are as follows. First, MiMed can survey the microbiome in various spheres (1) as a whole microbial ecosystem using different ecological measures (e.g., alpha- and beta-diversity indices) or (2) as individual microbial taxa (e.g., phyla, classes, orders, families, genera, species) using different data normalization methods. Second, MiMed enables covariate-adjusted analysis to control for potential confounding factors (e.g., age, gender), which is essential to enhance the causality of the results especially for observational studies. Third, MiMed enables a breadth of statistical inferences in both mediation effect estimation and significance testing. Fourth, MiMed provides flexible and easy-to-use data processing and analytic modules and creates nice graphical representations. Finally, MiMed employs ChatGPT to search for what have been known about the microbial taxa that are found significantly as mediators using AI technologies. MiMed is freely available on our web server (http://mimed.micloud.kr).

**Keywords.** Causal mediation analysis, Microbiome data analysis, Web cloud computing, Causal inference, Human microbiome

**Introduction**

The human microbiome is the totality of all microbes that live on and inside various organs (e.g., gut, mouth, skin, nose) of the human body. The advances in massively parallel metagenomic sequencing have dramatically lowered the cost of microbiome profiling with substantial increase in accuracy. Then, the microbiome field has not only become an active area of research, but has also rapidly grown in industry with the aim of identifying new ways to diagnose, treat and prevent human diseases.
Researchers have revealed a sophisticated interplay between microbiome and its host in various aspects. For instance, microbiome diversity and its taxonomic composition have been related to a variety of environmental, behavioral or medical exposures (e.g., diet [1], residence [2], smoking [3], preterm birth [4], delivery mode [5, 6], antibiotic/probiotic use [7, 8]). Researchers have also found that microbiome dysbiosis can lead to numerous disorders (e.g., obesity [9, 10], intestinal disease [11, 12, 13], cancers [14, 15, 16], diabetes [8, 17], brain disorders [18, 19]). However, beyond such separate discoveries, it is essential to understand if the microbiome transmits the effects of environmental, behavioral or medical exposures (say, treatment) to health or disease outcomes (say, outcome) as a mediator (Fig. 1), which can be surveyed through causal mediation analysis [20].

Mediation analysis aims to comprehend the underlying mechanism in an observed relationship between a treatment and an outcome through a third hypothetical variable, known as a mediator, indirectly. That is, in human microbiome studies, mediation analysis surveys two links jointly, (1) the effect of a treatment on microbiome (denoted as ‘treatment - microbiome’) and (2) the effect of microbiome on an outcome conditional on treatment status (denoted as ‘microbiome - outcome’) (Fig. 1). If we lose any one of these two links, microbiome does not serve as a mediator. That is, if we have ‘treatment - microbiome’ but do not have ‘microbiome - outcome’, the treatment alters microbiome, but the altered microbiome has no effect on the outcome. This means that the effect of the treatment on the outcome was made ‘directly’ or by some other unknown pathways, not through the microbiome. Similarly, if we do not have ‘treatment - microbiome’ but have ‘microbiome - outcome’, the treatment does not alter the microbiome, but only the variability in microbiome due to some other unknown sources influences the outcome. Thus, the roles of the microbiome as a mediator are satisfied only when we have both links [20], which we refer as the
presence of ‘indirect’ or ‘mediation’ effect. It substantially matters in a clinical context because if the microbiome is not in a causal pathway, any medical interventions to the microbiome do not fundamentally treat or prevent human diseases.

However, in human microbiome studies, mediation analysis is not used as much as it should be. One reason is because of the lack of carefully planned routines, compilers and automatic computing systems [21] for microbiome mediation analysis to perform a series of data processing, diversity calculation, data normalization, downstream data analysis and visualization. The microbiome data are highly complex and demand many data processing and analytic procedures. Many researchers in various disciplines (e.g., clinicians, public health practitioners, biologists) are not also familiar with related statistical methods and programming languages on command-line interfaces. Moreover, there are many other important issues that need to be addressed for microbiome causal mediation analysis as follows. First, we can view the microbiome as a whole community in an ecological context (referred in this paper for ‘community-level analysis’) or can focus on individual microbial taxa at various taxonomic hierarchies (i.e., phyla, classes, orders, families, genera, species) (referred in this paper for ‘taxonomy-level analysis’). Researchers usually survey the former using different ecological measures (e.g., alpha- and beta-diversity indices) [22, 23] and the latter using different data normalization methods (e.g., centered-log ratio (CLR) [24], arcsine-root [25]). Second, covariate-adjusted analysis is needed to control for potential confounding factors (e.g., age, gender), which is especially necessary for observational studies to enhance the causality of the results. Third, both mediation effect estimation and significance testing are important portions of statistical inference for better interpretability. Fourth, we need flexible and easy-to-use data processing and analytic modules as well as high-quality visualizations to be included in an academic paper. Finally, we need to figure out what have been
known about the microbes that we discovered as significant mediators. However, it is not easy in practice to figure it out all manually since there are too many microbial taxa [26] and related prior studies. Hence, we may need a well-trained artificial intelligence (AI) machine that can do such a job for us.

To tackle all those critical issues described above, here we introduce a web cloud computing platform, named MiMed, that enables comprehensive microbiome mediation analysis on user-friendly web interfaces. MiMed is the first web cloud computing platform for microbiome causal mediation analysis, which is distinguished from our prior platforms: (1) MiCloud for association analysis in cross-section or longitudinal microbiome studies [27]; (2) MiPair for design-based comparative analysis with paired microbiome data [28]; and (3) MiSurv for microbiome data analysis with survival responses [29]. Interestingly, MiMed is also built-in the popular AI language model, ChatGPT, to easily search for what have been known about the microbial taxa that are found significantly as mediators.

In the following Results section, we describe all the data processing and analytic modules one by one using an example study to see the mediating roles of oral microbiome between e-cigarette smoking and gingival inflammation [30]. Then, in the Discussion section, we summarize and discuss all the features and implications of MiMed. Finally, in the Methods section, we discuss the methodological ideas of causal mediation analysis methods, and describe our web server and local GitHub repository. MiMed is freely available on our web server (http://mimed.micloud.kr) or can alternatively run on a user’s local computer (https://github.com/yj7599/MiMedGit).

Results
Application Note: On the Roles of Oral Microbiome Between E-cigarette Smoking and Gingival Inflammation

To ease our demonstration, we use example data to survey the mediating roles of oral microbiome between e-cigarette smoking and gingival inflammation [30]. The original data are huge and can motivate a broad range of study orientations. We refer the original study paper [30] for all the details on study subjects, sample collection/processing, and sequencing/quantification procedures.

To describe only the portion of the data we use, the data are 16S oral microbiome data in subgingival niches obtained at the baseline visit of the subjects between 18 and 34 years in age. We employed a bioinformatic pipeline, QIIME2 [31], based on the expanded human oral microbiome database (eHOMD) [32] for raw sequence data processing, denoising, feature extraction/quantification, taxonomic annotation, and phylogenetic tree construction. We added detailed description on the use of each module using these example data at the end of each following section (see Application Note).

Data Processing: Data Input

Microbiome data can be composed of three data components: (1) a feature table (i.e., count data for operational taxonomic units (OTUs) or amplicon sequence variants (ASVs)), (2) a taxonomic table (i.e., taxonomic annotations at various taxonomic hierarchies, kingdom, phylum, class, order, family, genus, species), and (3) a phylogenetic tree (i.e., a rooted phylogenetic tree for evolutionary relationships across features, that are OTUs or ASVs). Of course, in addition to microbiome data, metadata on a treatment variable (e.g., environmental, behavioral or medical exposures), an outcome variable (e.g., health or disease status), and possibly covariates (e.g., age, gender) for
study subjects are needed. If we have all these data components, we can conduct microbiome causal mediation analysis comprehensively using all available functions of MiMed. However, researchers do not always have all these data components, but even in such a case, they can still want to conduct at least some parts of the analysis. Thus, we made the Data Input module flexible as follows.

(1) To fully perform all community-level analyses for all non-phylogenetic and phylogenetic alpha- and beta-diversity indices as well as all taxonomy-level analyses, users should upload a feature table, a taxonomic table, a phylogenetic tree, and metadata.

(2) To perform community-level analyses for only non-phylogenetic alpha- and beta-diversity indices as well as all taxonomy-level analyses, users can upload only a feature table, a taxonomic table, and metadata.

(3) To perform only community-level analyses for all non-phylogenetic and phylogenetic alpha- and beta-diversity indices, users can upload only a feature table, a phylogenetic tree, and metadata.

(4) To perform only community-level analyses for only non-phylogenetic alpha- and beta-diversity indices, users can upload only a feature table and metadata.

Users can upload their data components in a widely used unified format, called phyloseq [33], or as separate files.

Application Note: The example data we use can be downloaded in the Example Data section on the Data Input module. To help users to easily understand data components and their corresponding data analytic modules as described above, we uploaded four different sets of data components: (1) a feature table, a taxonomic table, a phylogenetic tree, and metadata; (2) a feature table, a taxonomic table, and metadata; (3) a feature table, a phylogenetic tree, and metadata; (4) a feature
table and metadata. Though since we aim in this paper to describe all available functions of MiMed, we uploaded the one with all data components (i.e., a feature table, a taxonomic table, a phylogenetic tree, and metadata).

**Data Processing: Quality Control**

MiMed performs quality controls (QCs) just as in MiCloud [27] and MiPair [28]. That is, users need to select (1) a kingdom of interest (default: Bacteria), (2) a minimum library size (i.e., total read count) for the study subjects to be rescued (default: 3,000), (3) a minimum mean relative abundance (i.e., proportion) for the features (OTUs or ASVs) to be rescued (default: 0.002%), and (4) erroneous taxonomic names in the taxonomic table to be removed.

MiMed displays the sample size, the number of features (OTUs or ASVs), the number of phyla, the number of classes, the number of orders, the number of families, the number of genera and the number of species using summary boxes before and after QCs. MiMed also visualizes library sizes across study subjects as well as mean proportions across features using interactive histograms and box plots before and after QCs. **Application Note:** We simply clicked the Run button to apply the default QC settings. Then, 147 subjects with 2,328 features, 11 phyla, 23 classes, 34 orders, 52 families, 99 genera and 215 species were retained in the following analyses (Fig. 2).

**Community-level Analysis: Diversity Calculation**

As in MiCloud [27], MiPair [28] and MiSurv [29], MiMed calculates 9 alpha-diversity indices (i.e., 8 non-phylogenetic indices: Observed, Shannon [34], Simpson [35], Inverse Simpson [35], Fisher [36], Chao1 [37], abundance-based coverage estimator (ACE) [38], incidence-based
coverage estimator (ICE) [39]; 1 phylogenetic index: phylogenetic diversity (PD) [40]) and 5 beta-
diversity indices (i.e., 2 non-phylogenetic indices: Jaccard dissimilarity [41], Bray-Curtis
dissimilarity [42]; 3 phylogenetic indices: Unweighted UniFrac distance [43], Generalized
UniFrac distance [44], Weighted UniFrac distance [45]). For reference, users can download all the
calculated alpha- and beta-diversity indices.

Application Note: We simply clicked the Run button to calculate all the alpha- and beta-diversity
indices.

Community-level Analysis: Alpha Diversity

This module analyzes if a treatment alters alpha-diversity, and then the altered alpha-diversity, in
turn, influences an outcome, where the alpha-diversity can be surveyed using each of the 9 alpha-
diversity indices. Users first need to select (1) a treatment variable (e.g., diet, residence, smoking,
preterm birth, delivery mode, antibiotic/probiotic use), (2) an outcome variable (e.g., health or
disease status), (3) to include an interaction term between a treatment and a mediator (alpha-
diversity) in the model or not, and (4) covariates (e.g., age, gender) to be adjusted for. We set the
interaction term to be included (yes) as default since it is more natural to assume that the effect of
microbiome on an outcome can be modulated by a treatment. That is, in order words, the effect of
microbiome on an outcome can be different by treatment status. Ignoring the presence of such
interaction effects may cause potential bias in mediation analysis [46, 47]. The only available
analytic method that can address interaction effect is the Imai method [48]) [Table 1]. The Imai
method [48] in addition allows covariate adjustments, estimates mediation effects in both point
and interval estimation, and reports a $P$-value for significance testing. The other available analytic
methods are two traditional (but still in wide use) methods, the Sobel test [49] and Preacher-Hayes
approach [50, 51], and one recent method, named Divide-Aggregate Composite-null Test (DACT) [52]. We organized available functions for different mediation analysis methods in [Table 1]. We discuss them further later in the Methods section. MiMed visualizes the results from its alpha-diversity analysis using forest plots.

Application Note: We selected e-cigarette smoking as a treatment variable, gingival inflammation as an outcome variable, and age, sex and the frequency of brushing teeth as covariates to be adjusted for in the presence of interaction between e-cigarette smoking and alpha-diversity. Then, we found significant results using the Imai method [48] as e-cigarette smoking alters alpha-diversity of the oral microbiome in subgingival niches, and the altered alpha-diversity, in turn, influences gingival inflammation according to Observed, Shannon [34], InvSimpson [35], Fisher [36], Chao1 [37], ACE [38] and ICE [39] indices (Fig. 3).

Community-level Analysis: Beta Diversity

This module analyzes if a treatment alters beta-diversity, and then the altered beta-diversity, in turn, influences an outcome, where the beta-diversity can be surveyed using each of the 5 beta-diversity indices. As for alpha-diversity analysis, users need to select (1) a treatment variable (e.g., diet, residence, smoking, preterm birth, delivery mode, antibiotic/probiotic use), (2) an outcome variable (e.g., health or disease status), and (3) covariates (e.g., age, gender) to be adjusted for. MedTest [53] is currently the only available analytic method that can conduct causal mediation analysis for beta-diversity [Table 1]. While MedTest [53] allows covariate adjustments and reports a P-value for significance testing, it is purely a test for significance with no facilities for mediation effect estimation [Table 1]. Furthermore, it does not allow any interaction term to be included.
MiMed visualizes the results from its beta-diversity analysis using principal coordinate analysis (PCoA) plots [54].

**Application Note:** We selected e-cigarette smoking as a treatment variable, gingival inflammation as an outcome variable, and age, sex and the frequency of brushing teeth as covariates to be adjusted for. Then, we found significant results using MedTest [53] as e-cigarette smoking alters beta-diversity of the oral microbiome in subgingival niches, and the altered beta-diversity, in turn, influences gingival inflammation according to Jaccard dissimilarity [41], Bray-Curtis dissimilarity [42], Generalized UniFrac distance [44] and Weighted UniFrac distance [45] (Fig. 4).

**Taxonomy-level Analysis: Data Normalization**

MiMed normalizes taxonomic relative abundances using CLR [24] and arcsine-root [55] transformations. The CLR transformation is the most widely used normalization method in the microbiome field to relax the compositional constraint of the data [24]. The arcsine-root transformation is a traditional approach to stabilize the variance of relative abundances [55]. The arcsine-root transformation has also recently been often used in the microbiome field [25]. For reference, users can download all the original count and normalized taxonomic data for microbial taxa at various taxonomic hierarchies (i.e., phyla, classes, orders, families, genera, species).

**Application Note:** We simply clicked the Run button to normalize taxonomic relative abundances.

**Taxonomy-level Analysis: Taxonomic Analysis**

This module analyzes if a treatment alters microbial taxa, and then the altered microbial taxa, in turn, influence an outcome. Users first need to select a data format, CLR [24] or arcsine-root transformed data. There has been a long debate on which data normalization method is the best,
but it is beyond the scope of this paper to make any resolute judgement on it. We set CLR [24] as default and arcsine-root as a user option based on their popularities. As for alpha-diversity analysis, users then need to select (1) a treatment variable (e.g., diet, residence, smoking, preterm birth, delivery mode, antibiotic/probiotic use), (2) an outcome variable (e.g., health or disease status), (3) to include an interaction term between a treatment and a mediator (taxon) in the model or not, and (4) covariates (e.g., age, gender) to be adjusted for. Again, the only available analytic method that can address interaction effect is the Imai method [48] [Table 1]. Importantly, the Imai method [48] is a non-parametric method based on a bootstrap approach [56]. Thus, it is highly robust against the high skewness of microbiome data, especially the rare microbial taxa with excessive zeros [48]. The other available analytic methods are two parametric methods, the Sobel test [49] and DACT [52] [Table 1]. We set the Imai method [48] as default and the Sobel test [49] and DACT [52] as user options [Table 1], which is because of the robust performance of the Imai method [48] as well as its broad range of functionalities [Table 1]. We discuss methodological details further later again in the Methods section. To control for false discovery rates (FDR), MiMed applies the Benjamini-Hochberg (BH) procedure [57] to each taxonomic hierarchy. MiMed visualizes the results from its taxonomic analyses using forest plots and dendrograms.

Ask ChatGPT: In this sub-module, users can ask ChatGPT a question: What is known about (discovered taxon) on (treatment) and (outcome)? For this, users first need to insert a ChatGPT APT key that can be freely obtained on the website (https://platform.openai.com/account/api-keys). Then, users need to select a taxonomic rank (i.e., phylum, class, order, family, genus, species) and a taxon that is discovered as a significant mediator, and then rename the treatment and outcome variables using a human language (not code names). Then, ChatGPT will answer your question.
Application Note: We selected CLR as a normalization method, e-cigarette smoking as a treatment variable, gingival inflammation as an outcome variable, and age, sex and the frequency of brushing teeth as covariates to be adjusted for. Then, we found 21 significant taxa at the taxonomic hierarchies from phylum to genus (i.e., two phyla: *Proteobacteria, Spirochaetes*, two classes: *Flavobacteriia, Betaproteobacteria*, four orders: *Flavobacteriales, Burkholderiales, Neisseriales, Cardiobacterales*, five families: *Flavobacteriaceae, Burkholderiaceae, Neisseriaceae, Cardiobacteriaceae, Enterococcaceae*, and eight genera: *Bergeyella, Capnocytophaga, Actinomyces, Haemophilus, Kingella, Burkholderia, Cardiobacterium, Enterococcus*) using the Imai method [48] as e-cigarette smoking alters their relative abundances, and the altered relative abundances, in turn, influence gingival inflammation (Fig. 5 and Fig. 6).

We also asked ChatGPT a question, “What is known about *Bergeyella* on e-cigarette and gingival inflammation?”, selecting a taxonomic rank as genus and a discovered taxon as *Bergeyella*, and renaming the treatment and outcome variables as e-cigarette and gingival inflammation. Then, ChatGPT aided to reconfirm our results answering as “*Bergeyella* is a gram-negative bacteria that is commonly found in the oral microbiome. It has been associated with various oral diseases including gingivitis and periodontitis. There is some preliminary evidence suggesting that e-cigarettes may contribute to gingival inflammation and changes in the oral microbiome, including an increase in the abundance of *Bergeyella*. A study published in Scientific Reports in 2019 reported that e-cigarette users had higher levels of gingival inflammation and an altered oral microbiota compared to non-smokers or smokers of traditional cigarettes. Specifically, the study found that e-cigarette users had increased levels of *Bergeyella*, as well as other potentially pathogenic bacteria, in their oral microbiota. While the relationship between *Bergeyella* and e-
cigarettes is still being explored and more research is needed to fully understand the implications, these findings suggest that e-cigarette use may have negative effects on oral health.” (Fig. 7).

**Discussion**

Researchers are interested in discovering causal mechanisms through which environmental, behavioral or medical exposures influence health or disease outcomes. A promising approach has been to use mediation analysis, though it is highly demanding in the human microbiome field. The microbiome data are huge and highly complex, and many researchers are not familiar with dealing with such microbiome data. Thus, we need a well-designed ‘software’ that enables user-friendly operations for microbiome causal mediation analysis.

In this paper, we introduced MiMed that is the first web cloud computing platform for microbiome causal mediation analysis. MiMed enables a long sequence of data processing and analytic operations on user-friendly web interfaces with widely extended flexibility and functionality. MiMed surveys the microbiome in various spheres as a whole ecosystem or as individual microbial taxa at various taxonomic hierarchies. MiMed also enables covariate-adjusted analysis and a breadth of statistical inferences in both mediation effect estimation and significance testing. MiMed also provides step-by-step data processing and analytic modules, and creates high-quality visualizations. Interestingly, MiMed is also built-in the recent popular chatbot, ChatGPT, to easily search for prior knowledge on discovered taxa using AI technologies. The ChatGPT module can be useful to re-confirm the analysis results from MiMed.

MiMed is comprehensive and built with many data processing and analytic approaches. It is usual in the human microbiome field that there is no consensus on which approach is always the
best. That is, there is not anything that is superior to the others in all contexts and situations. We are also curious about many different approaches. Thus, we left much room for our users to freely explore through many user options, while making a series of recommendations, as a developer, through default settings. For user’s convenience, MiMed also displays a list of references for the approaches that they use.

The human microbiome field is rapidly emerging, and the microbiome data are recently flooded. Yet, the microbiome data are demanding, and we are all so busy. Thus, MiMed can be attractive and useful in practice because it is user-friendly. MiMed will also provide new insights to the human microbiome field through causal mediation analysis that is “too important to abandon” [48].

Methods

Statistical Methods

This section is devoted to describing the methodological aspects of the causal mediation analysis methods that are available in MiMed. We describe only the conceptual ideas and terms to help our users to easily understand them, while referencing the original papers for all technical details.

To begin with the Sobel test [49], Preacher-Hayes approach [50, 51], and DACT [52], the Baron and Kenny’s two regression models [20] below can first be considered.

\[ M_i = \alpha_0 + \alpha_1 T_i + \epsilon_i \]
\[ Y_i = \beta_0 + \beta_1 M_i + \beta_1 T_i + \nu_i \]

where \( T_i \) is a treatment, \( M_i \) is a mediator (e.g., an alpha-diversity index or a microbial taxon), \( Y_i \) is a health or disease outcome, \( \alpha_0 \) and \( \beta_0 \) are intercepts, \( \alpha_1 \) and \( \beta_1 \) are slopes, and \( \epsilon_i \) and \( \nu_i \) are
independently distributed ransom errors with mean zero and variance $\sigma^2$ for the units $i = 1, \ldots, n$.

To ease our demonstration, we suppose in addition that $T_i$ is a binary treatment variable ($T_i = 0$ for control & $T_i = 1$ for treatment), and $Y_i$ is a continuous health or disease outcome variable. Yet, more extensions are available [Table 1]. Then, the null and alternative hypotheses below are considered.

$$H_0: \alpha_1 \beta_1 = 0 \text{ vs. } H_1: \alpha_1 \beta_1 \neq 0 \quad (3)$$

Here, $\alpha_1$ represents the effect of the treatment ($T_i$) on the mediator ($M_i$) as in Eq. (1), and $\beta_1$ represents the effects of the mediator ($M_i$) on the outcome ($Y_i$) conditional on treatment status ($T_i$) as in Eq. (2). Then, the null hypothesis, $H_0: \alpha_1 \beta_1 = 0$, states that at least one of $\alpha_1$ and $\beta_1$ equals to zero indicating no mediation effect, while the alternative hypothesis, $H_1: \alpha_1 \beta_1 \neq 0$, states that both $\alpha_1$ and $\beta_1$ are non-zero indicating the presence of mediation effect. The Sobel test [49] conducts significance testing for Eq. (3) using a parametric approach that assumes that $\epsilon_i$ and $\nu_i$ in Eq. (1) and Eq. (2) are normally distributed. In contrast, the Preacher-Hayes approach [50, 51] does it non-parametrically using a bootstrap method [56] without the normality assumption. As for the Sobel test [49], DACT [52] is a parametric approach, but considers the null hypothesis, $H_0: \alpha_1 \beta_1 = 0$, in Eq. (3) as a composite hypothesis that $H_0$: (1) $\alpha_1 = 0 \& \beta_1 \neq 0$; (2) $\alpha_1 \neq 0 \& \beta_1 = 0$; or (3) $\alpha_1 = 0 \& \beta_1 = 0$; to improve statistical power while rejecting $H_0$ for at least one of the three sub-statements.

As for DACT [52], MedTest [53] considers the null hypothesis as a composite hypothesis, but it is a non-parametric significance test based on a permutation method. A more important distinction is that MedTest [53] formulates the mediator ($M_i$) in Eq. (1) and Eq. (2) as a function of beta-diversity (say, $f(M)_i$, where $f(.)$ is a function that transforms microbiome into a beta-diversity index); as such, it enables causal mediation analysis for beta-diversity [Table 1].
We can classify the Sobel test [49], Preacher-Hayes approach [50, 51], DACT [52] and MedTest [53] as ‘product-of-coefficients’ methods because of their shared hypothesis of Eq. (3) in the form of $\alpha_1\beta_1$ (i.e., the product of coefficients from Eq. (1) and Eq. (2)). However, the Imai method [48] in contrast is based on a potential outcomes framework of causal inference [58] $Y_i(T_i, M_i(T_i))$, where the level of health or disease outcome is a function of a treatment status (i.e., $T_i$) and the level of the mediator under a treatment status (i.e., $M_i(T_i)$). Then, the unit-level ‘total treatment effect’ can be defined as Eq. (4), the unit-level ‘direct effect (DE)’ on the mediator can be defined for each treatment status ($t = 0$ for control or $t = 1$ for treatment) as Eq. (5), and finally the unit-level ‘indirect effect or causal mediation effect (CME)’ can be defined for each treatment status ($t = 0$ for control or $t = 1$ for treatment) as Eq. (6),

$$\tau_i \equiv Y_i(1, M_i(1)) - Y_i(0, M_i(0)) \quad (4)$$

$$\zeta_i(t) \equiv Y_i(1, M_i(t)) - Y_i(0, M_i(t)) \quad (5)$$

$$\delta_i(t) \equiv Y_i(t, M_i(1)) - Y_i(t, M_i(0)) \quad (6)$$

Here, the unit-level total treatment effect in Eq. (4) was formulated by subtracting the level of health or disease outcome for the unit under control and the level of the mediator under control from the level of health or disease outcome for the same unit under treatment and the level of the mediator under treatment. The unit-level DE for each treatment status (i.e., for control or treatment) in Eq. (5) was formulated by subtracting the level of health or disease outcome for the unit with under control from the level of health or disease outcome for the same unit under treatment. Finally, the unit-level CME for each treatment status (i.e., for control or treatment) in Eq. (6) was formulated by subtracting the level of health or disease outcome for the unit with the level of the
mediator under control from the level of health or disease outcome for the same unit with the level of the mediator under treatment.

Then, the overall ‘average direct effect (ADE)’ can be found by the average between the average direct effect with the level of mediator under control, \( \frac{1}{n} \sum_{i=1}^{n} \zeta_i(0) \), the average direct effect with the level of mediator under treatment, \( \frac{1}{n} \sum_{i=1}^{n} \zeta_i(1) \). Finally, the overall ‘average causal mediation effect (ACME)’, that is the main result in causal mediation analysis, can be found by the average between the average causal mediation effect for control, \( \frac{1}{n} \sum_{i=1}^{n} \delta_i(0) \), and the average causal mediation effect for treatment, \( \frac{1}{n} \sum_{i=1}^{n} \delta_i(1) \).

The Imai method [48] conducts interval estimation for ACME (overall) (as well as ACME (control), ACME (treatment), ADE (overall), ADE (control), ADE (treatment)) using a bootstrap method [56] non-parametrically, and its significance testing follows accordingly.

There has been a long debate on parametric vs. non-parametric, but it is also beyond the scope of this paper to make any resolute judgement on it. However, it is usual that non-parametric approaches are more robust to highly skewed data (e.g., rare taxa with excessive zeros), while parametric approaches are well suited to less skewed data (e.g., alpha-diversity indices or common taxa). However, so long as the sample size is large, the skewness does not also substantially matter for parametric approaches. However, it does not also mean that non-parametric approaches are not suited to a large sample size. Parametric approaches are not well suited to high skewed data with a small sample size. Since the microbiome data are usually highly skewed (Fig. 2), we set non-parametric approaches as default, but we do not discourage the use of parametric approaches, which are also widely used and reasonable approaches for a large sample size [Table 1].
Web Server and Local GitHub Repository

As for MiCloud [27], MiPair [28] and MiSurv [29], we wrote all the user interfaces and server functions using R shiny (https://shiny.rstudio.com). We then developed our web server using ShinyProxy (https://www.shinyproxy.io) and Apache2 (https://httpd.apache.org) on the operating system, Ubuntu 20.04 (https://ubuntu.com). The web server currently runs on a computer with the specifications of Intel Core i7-12700T (12-core) processor and 36 GB DDR4 memory, and takes up to ten concurrent users. In case that the web server is busy, we also developed a local GitHub repository to enable to run MiMed using a user’s local computers. As usual, we, as a host, are responsible for and devoted to maintaining our web server and local GitHub repository reliable.

Data Availability

We used public microbiome data, where the raw sequence data are deposited at the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under access number GSE201949. The processed data can also be found in the Example Data section on the Data Input module of MiMed (http://mimed.micloud.kr).

Code Availability

MiMed is freely available on our web server (http://mimed.micloud.kr) or can alternatively run on a user’s local computer (https://github.com/yj7599/MiMedGit).

Author Contributions
HJ wrote the programs, developed the shiny application, web server and local GitHub repository, performed data analysis, and contributed to the visualization and design. SP wrote the programs, developed the shiny application, performed data analysis, and contributed to the visualization and design. HK conceptualized and initiated the study, curated the data, contributed to the methodological aspects, overall architecture, and design, performed data analysis, and wrote the manuscript. All authors reviewed, revised, and approved the final manuscript.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgements

The work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (2021R1C1C1013861).
References


Table 1. Descriptive table for the functionalities of causal mediation analysis methods: Imai method, Sobel test, Preacher-Hayes approach, DACT and MedTest.

<table>
<thead>
<tr>
<th>Treatment Variable</th>
<th>Outcome Variable</th>
<th>Community-level Analysis</th>
<th>Taxonomy-level Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Alpha Diversity</td>
<td>Beta Diversity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Imai (Default)</td>
<td>Sobel</td>
</tr>
<tr>
<td>Binary</td>
<td>Interaction</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Covariates</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Point Estimation</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Interval Estimation</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>P-value</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td>Continuous</td>
<td>Interaction</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Covariates</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Point Estimation</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td></td>
<td>Interval Estimation</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>P-value</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>Binary</td>
<td>Interaction</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Covariates</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Point Estimation</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Interval Estimation</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>P-value</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>Continuous</td>
<td>Interaction</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Covariates</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Point Estimation</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td></td>
<td>Interval Estimation</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>P-value</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>
Figures

Fig. 1. A conceptual illustration for the roles of the microbiome as a mediator between a treatment/exposure and a health or disease outcome with potential covariate effects.
**Fig. 2. The status of the microbiome data after QCs.** The summary boxes below display the sample size, the number of features, the number of phyla, the number of classes, the number of orders, the number of families, the number of genera and the number of species after QCs. The histograms and box plots below visualize the library sizes across study subjects and the mean proportions across features.
Fig. 3. The results for alpha-diversity. We surveyed if e-cigarette smoking alters alpha-diversity of the oral microbiome in subgingival niches, and the altered alpha-diversity, in turn, influences gingival inflammation, adjusting for age, sex and the frequency of brushing teeth.

<table>
<thead>
<tr>
<th>Alpha Diversity</th>
<th>Estimate</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>0.091</td>
<td>0.003</td>
</tr>
<tr>
<td>Shannon</td>
<td>0.073</td>
<td>0.009</td>
</tr>
<tr>
<td>Simpson</td>
<td>0.034</td>
<td>0.278</td>
</tr>
<tr>
<td>InvSimpson</td>
<td>0.052</td>
<td>0.026</td>
</tr>
<tr>
<td>Fisher</td>
<td>0.091</td>
<td>0.005</td>
</tr>
<tr>
<td>Chao1</td>
<td>0.095</td>
<td>0.001</td>
</tr>
<tr>
<td>ACE</td>
<td>0.092</td>
<td>0.004</td>
</tr>
<tr>
<td>ICE</td>
<td>0.091</td>
<td>0.009</td>
</tr>
<tr>
<td>PD</td>
<td>0.035</td>
<td>0.217</td>
</tr>
</tbody>
</table>
Fig. 4. The results for beta-diversity. We surveyed if e-cigarette smoking alters beta-diversity of the oral microbiome in subgingival niches, and the altered beta-diversity, in turn, influences gingival inflammation, adjusting for age, sex and the frequency of brushing teeth.
Fig. 5. The results for microbial taxa. We surveyed if e-cigarette smoking alters the microbial taxa of the oral microbiome in subgingival niches, and the altered microbial taxa, in turn, influence gingival inflammation, adjusting for age, sex and the frequency of brushing teeth.
Fig. 6. A hierarchical visualization for the taxonomic discoveries. The numbers in circles are matched with the IDs in Fig. 5.
Fig. 7. The screenshot of the Ask ChatGPT module. We asked ChatGPT a question, “What is known about Bergeyella on e-cigarette and gingival inflammation?”. Then, ChatGPT answered the question.