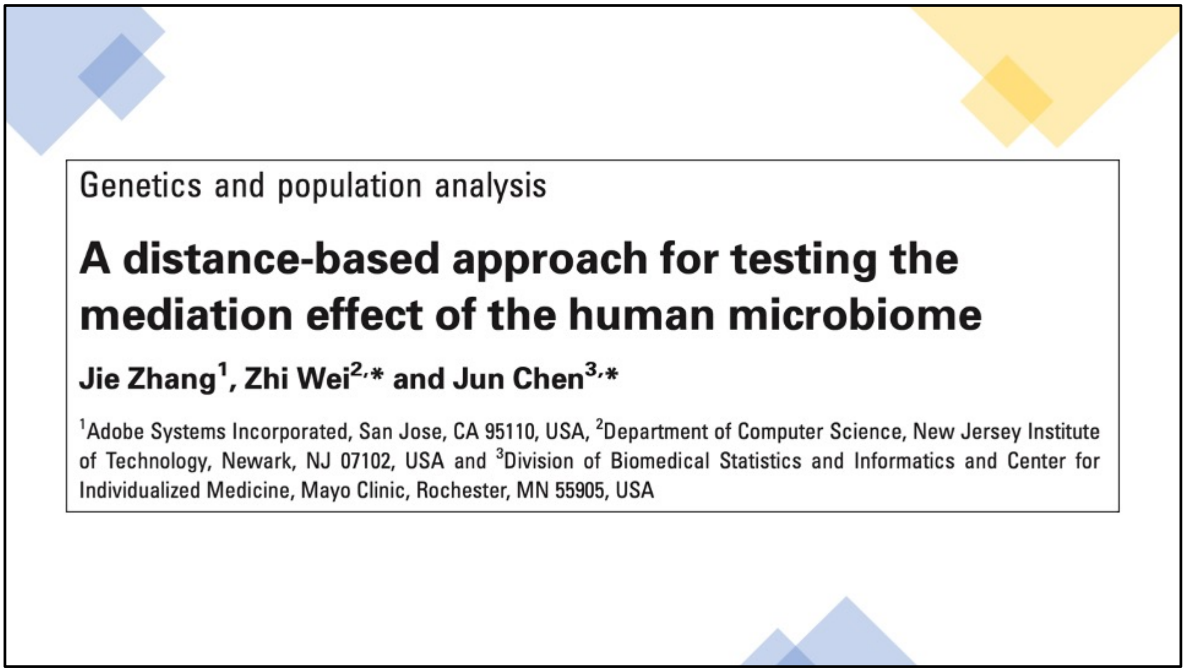


Genetics and population analysis

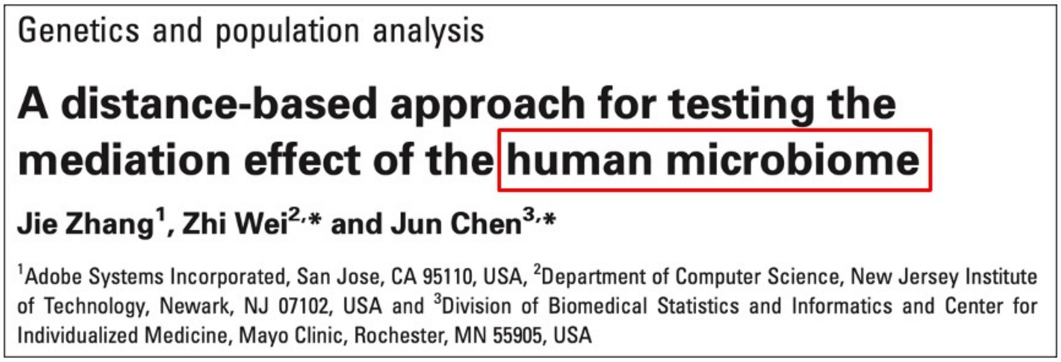
A distance-based approach for testing the mediation effect of the human microbiome

Jie Zhang[1], **Zhi Wei**[2,*] **and Jun Chen**[3,*]

[1]Adobe Systems Incorporated, San Jose, CA 95110, USA, [2]Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA and [3]Division of Biomedical Statistics and Informatics and Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA

Recent studies have revealed a complex interplay of environmental, genetic factors with the human microbiome in influencing clinical or biological outcomes. Mediation of the human microbiome could help better understand the role of the micriobiome in disease etiology and lead to clinical interventions by modulating the microbiome. High dimensionality, scarcity, non-normality and phylogenetic structure are key issues.

Overview

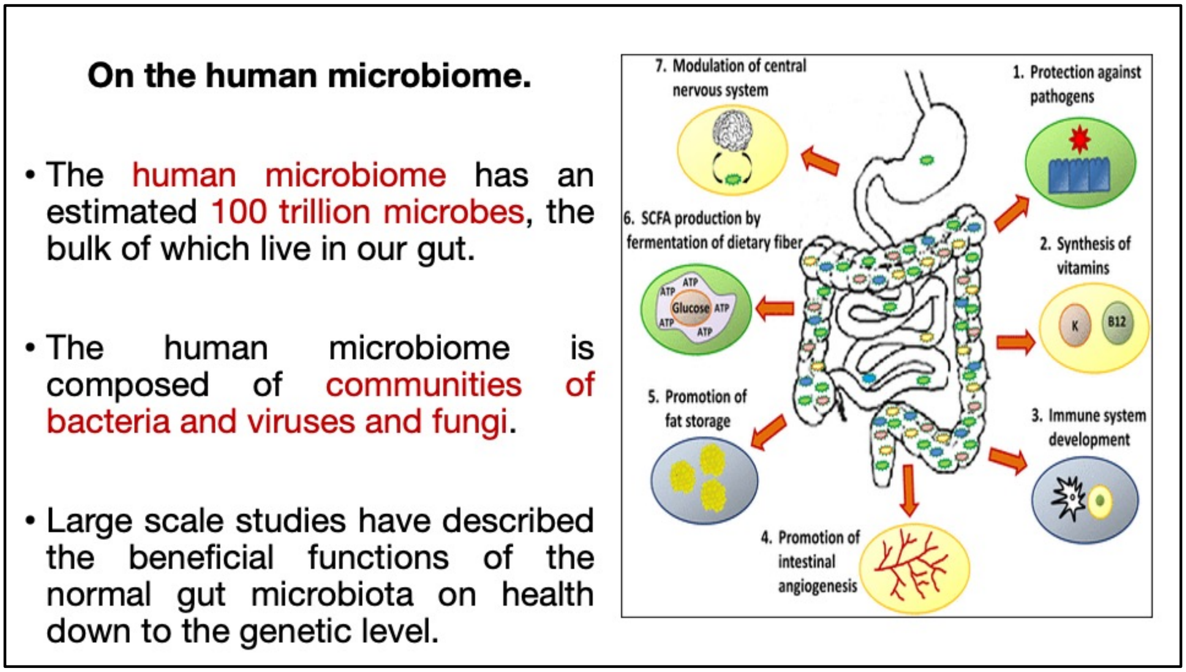


Genetics and population analysis

A distance-based approach for testing the mediation effect of the human microbiome

Jie Zhang[1], Zhi Wei[2,*] and Jun Chen[3,*]

[1]Adobe Systems Incorporated, San Jose, CA 95110, USA, [2]Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA and [3]Division of Biomedical Statistics and Informatics and Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA



On the human microbiome.

- The human microbiome has an estimated 100 trillion microbes, the bulk of which live in our gut.
- The human microbiome is composed of communities of bacteria and viruses and fungi.
- Large scale studies have described the beneficial functions of the normal gut microbiota on health down to the genetic level.

7. Modulation of central nervous system
1. Protection against pathogens
6. SCFA production by fermentation of dietary fiber
ATP
Glucose
2. Synthesis of vitamins
K
B12
5. Promotion of fat storage
3. Immune system development
4. Promotion of intestinal angiogenesis

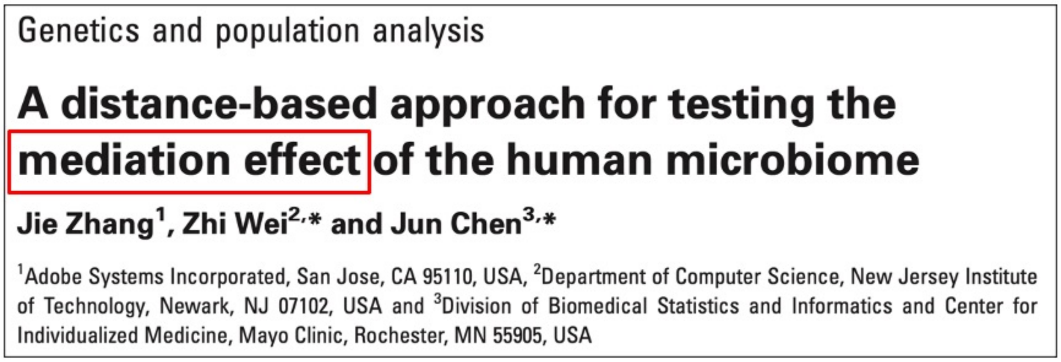
Whenever someone says microbiome – think of an obscenely large number of microbes – and they live inside you/ . that have a greater complexity than the human genome itself.

numerous microbiome studies have been conducted at different body sites, such as skin, gut and respiratory tract with the aim to understand the genetic and environmental forces shaping the human microbiome, and the relationship between the microbiome compositional variation and biological or clinical

An understanding of this complex ecological community is important as it affects our patients, and manipulation of the gut microbiome has the potential to be used in the treatment of childhood diseases in the future.

In terms of quantification – think of counts and proportions of various types of microbes!!

Overview

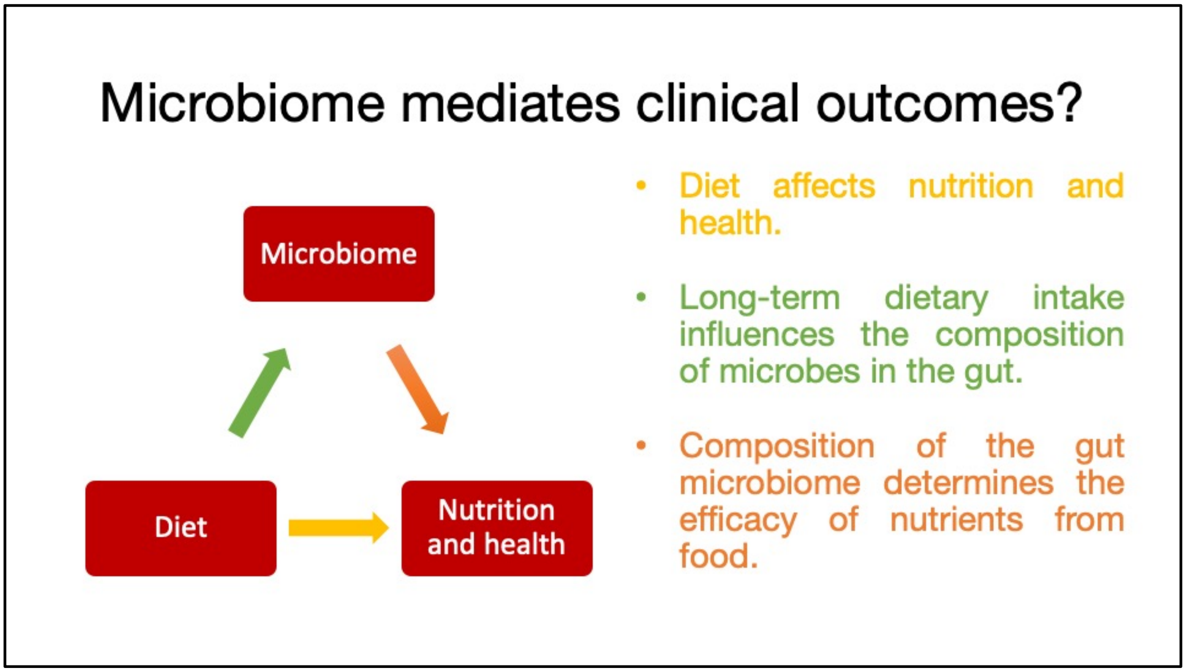


Genetics and population analysis

A distance-based approach for testing the mediation effect of the human microbiome

Jie Zhang[1], Zhi Wei[2,*] and Jun Chen[3,*]

[1]Adobe Systems Incorporated, San Jose, CA 95110, USA, [2]Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA and [3]Division of Biomedical Statistics and Informatics and Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA



Microbiome mediates clinical outcomes?

Microbiome

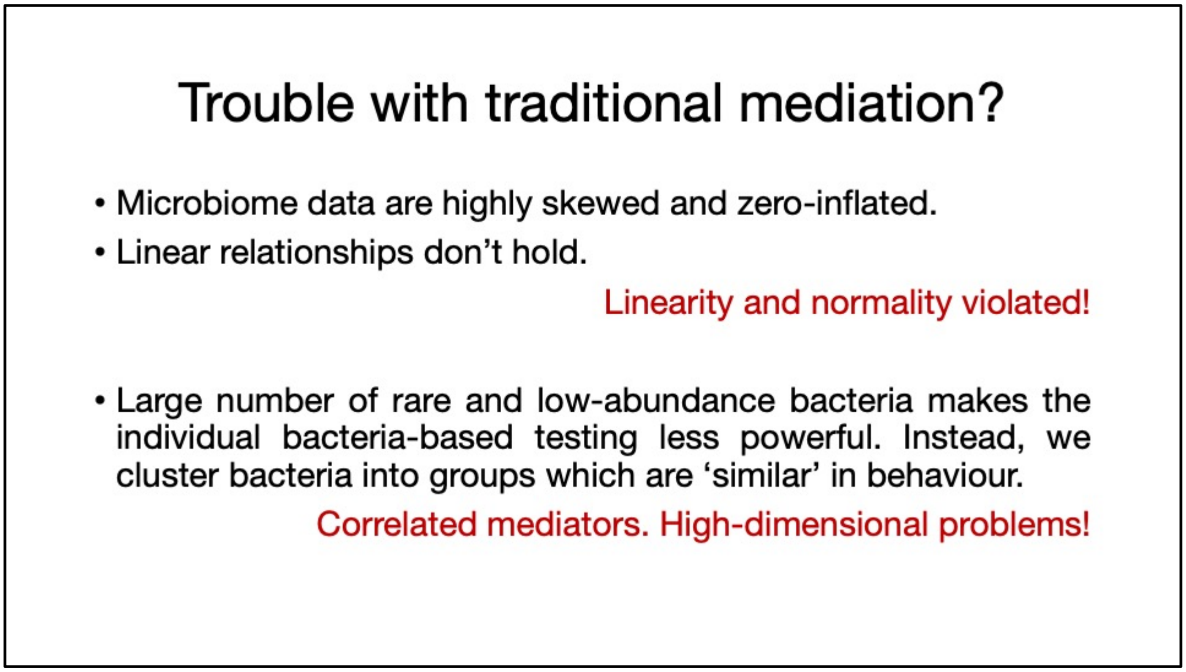
Diet

Nutrition and health

- Diet affects nutrition and health.
- Long-term dietary intake influences the composition of microbes in the gut.
- Composition of the gut microbiome determines the efficacy of nutrients from food.

there is a complex interplay among the environmental (or genetic) factors, the human microbiome and health.

microbiome may mediate the effect of long-term diet on human health? To establish mediation, we need to test whether there is a significant relation of the independent variable X
to some mediating variable M , and whether the mediating variable M is significantly related to the dependent variable Y, when adjusted by the independent variable.



Trouble with traditional mediation?

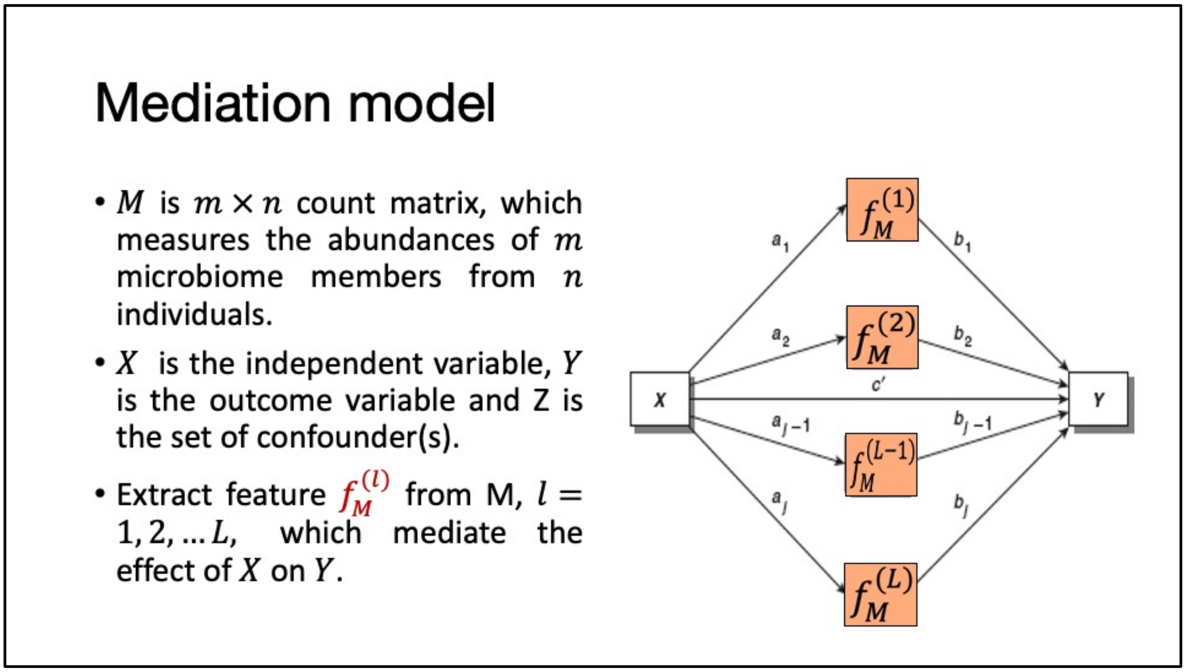
- Microbiome data are highly skewed and zero-inflated.
- Linear relationships don't hold.

Linearity and normality violated!

- Large number of rare and low-abundance bacteria makes the individual bacteria-based testing less powerful. Instead, we cluster bacteria into groups which are 'similar' in behaviour.

Correlated mediators. High-dimensional problems!

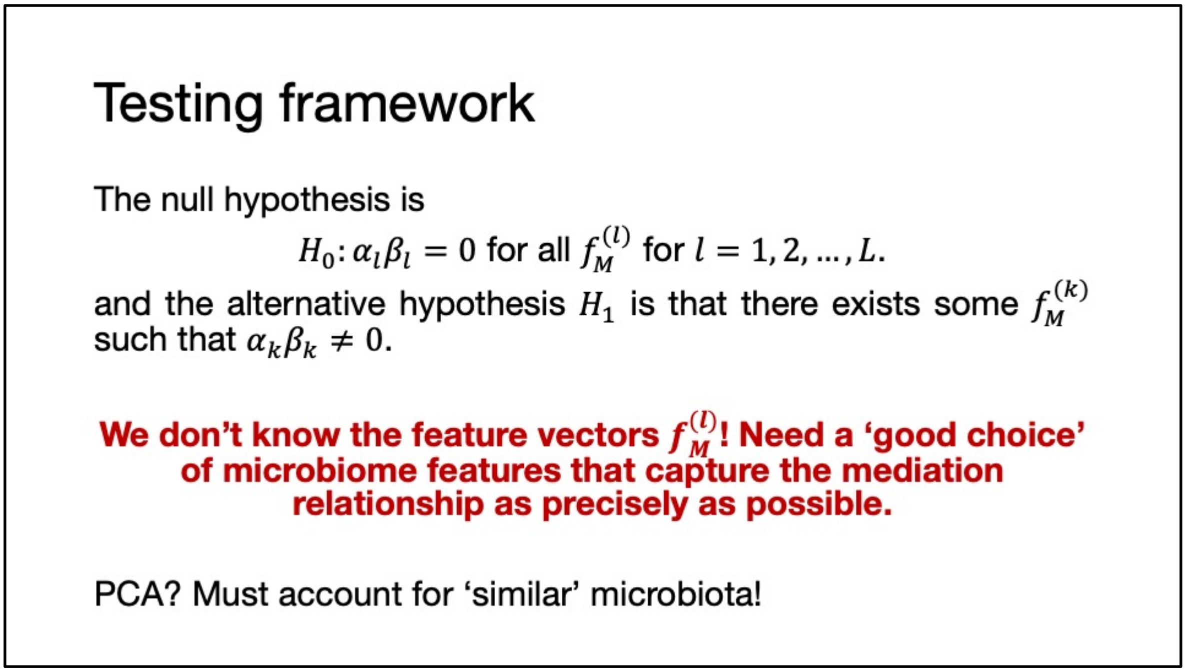
So what that means is the null distribution will be hard to obtain.



Mediation model

- M is $m \times n$ count matrix, which measures the abundances of m microbiome members from n individuals.
- X is the independent variable, Y is the outcome variable and Z is the set of confounder(s).
- Extract feature $f_M^{(l)}$ from M, $l = 1, 2, \ldots L$, which mediate the effect of X on Y.

Due to the multivariate nature of the microbiome data, it is possible to have multiple microbiome features that mediate the effect



Testing framework

The null hypothesis is

$$H_0\colon \alpha_l \beta_l = 0 \text{ for all } f_M^{(l)} \text{ for } l = 1, 2, \ldots, L.$$

and the alternative hypothesis H_1 is that there exists some $f_M^{(k)}$ such that $\alpha_k \beta_k \neq 0$.

We don't know the feature vectors $f_M^{(l)}$! Need a 'good choice' of microbiome features that capture the mediation relationship as precisely as possible.

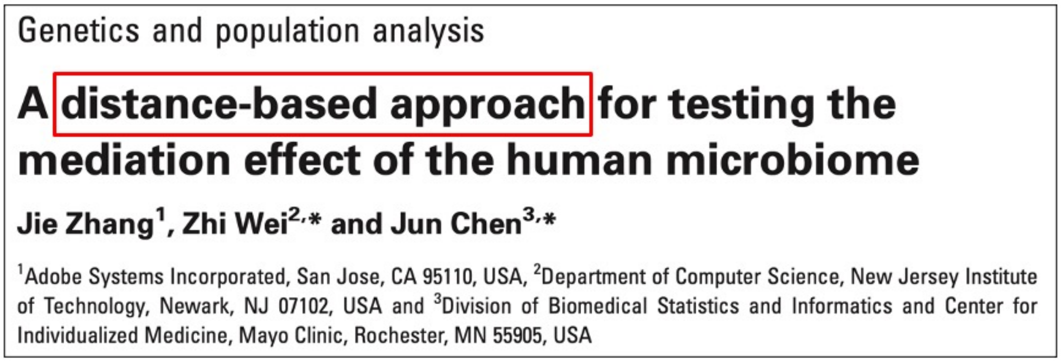
PCA? Must account for 'similar' microbiota!

(of no mediation for the microbiome quantified by M)

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, indi- vidual tests are usually underpowered. enrich signals and reduce multiple testing burden, community-level analysis is good!

PCA defines the micro- biome features based on the within-data correlations and is a good choice, but we need to factor in the 'similarity'

Overview

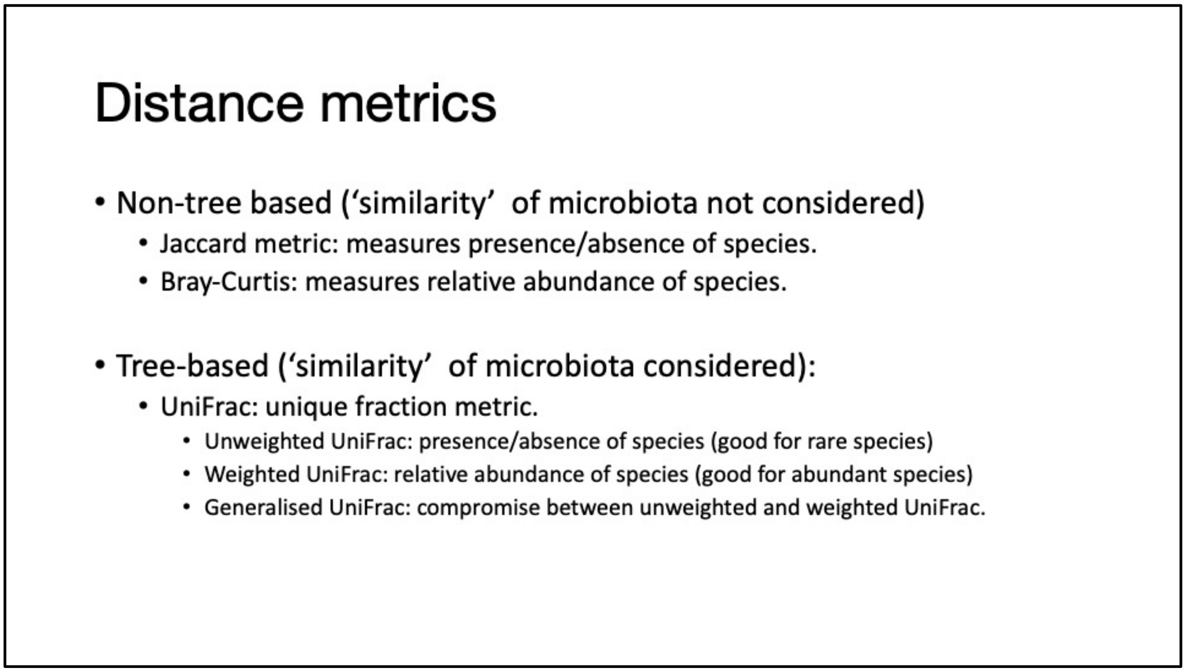


Genetics and population analysis

A distance-based approach for testing the mediation effect of the human microbiome

Jie Zhang[1], Zhi Wei[2,*] and Jun Chen[3,*]

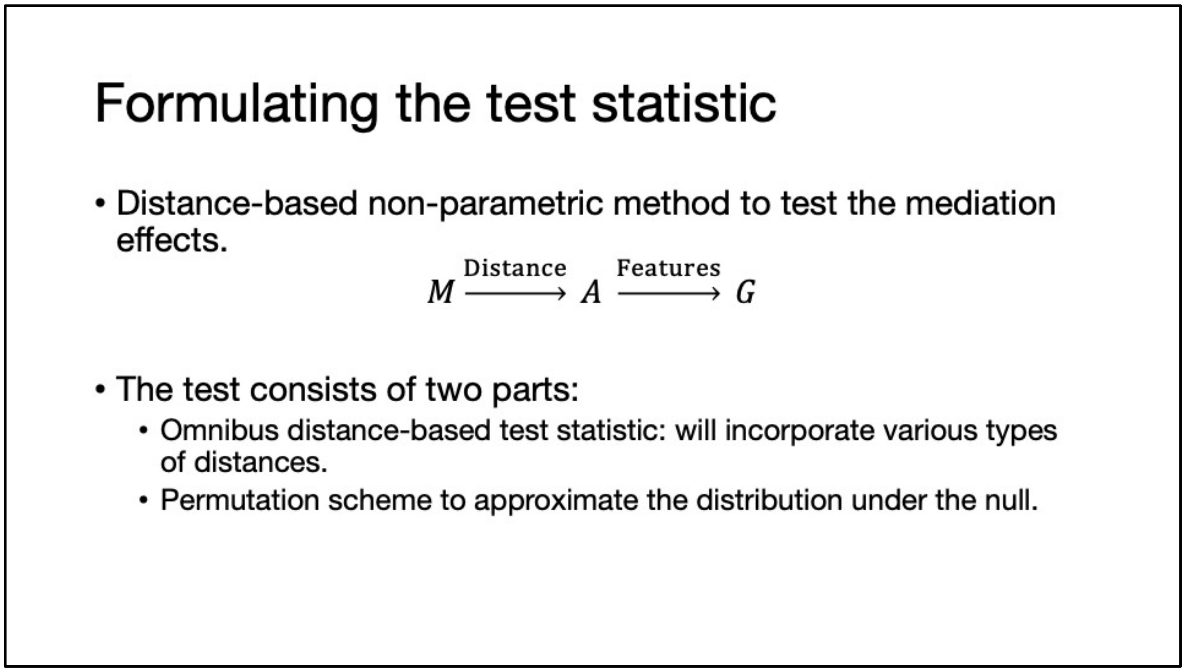
[1] Adobe Systems Incorporated, San Jose, CA 95110, USA, [2] Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA and [3] Division of Biomedical Statistics and Informatics and Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA



Distance metrics

- Non-tree based (‘similarity’ of microbiota not considered)
  - Jaccard metric: measures presence/absence of species.
  - Bray-Curtis: measures relative abundance of species.

- Tree-based (‘similarity’ of microbiota considered):
  - UniFrac: unique fraction metric.
    - Unweighted UniFrac: presence/absence of species (good for rare species)
    - Weighted UniFrac: relative abundance of species (good for abundant species)
    - Generalised UniFrac: compromise between unweighted and weighted UniFrac.

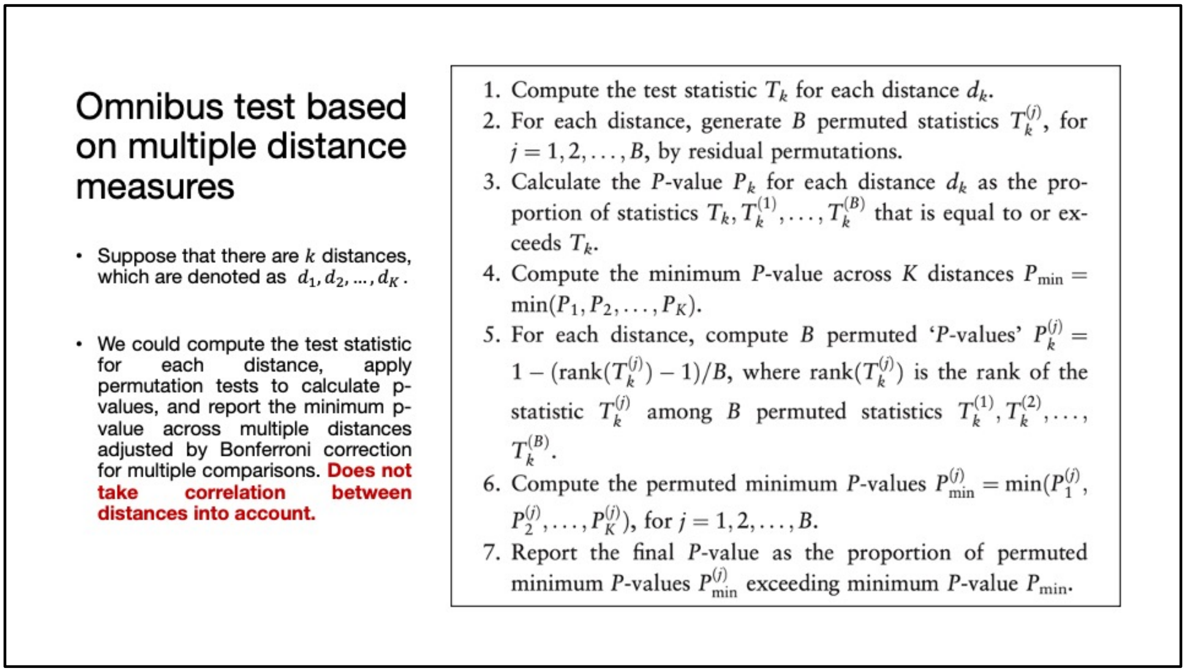


Formulating the test statistic

- Distance-based non-parametric method to test the mediation effects.

$$M \xrightarrow{\text{Distance}} A \xrightarrow{\text{Features}} G$$

- The test consists of two parts:
  - Omnibus distance-based test statistic: will incorporate various types of distances.
  - Permutation scheme to approximate the distribution under the null.



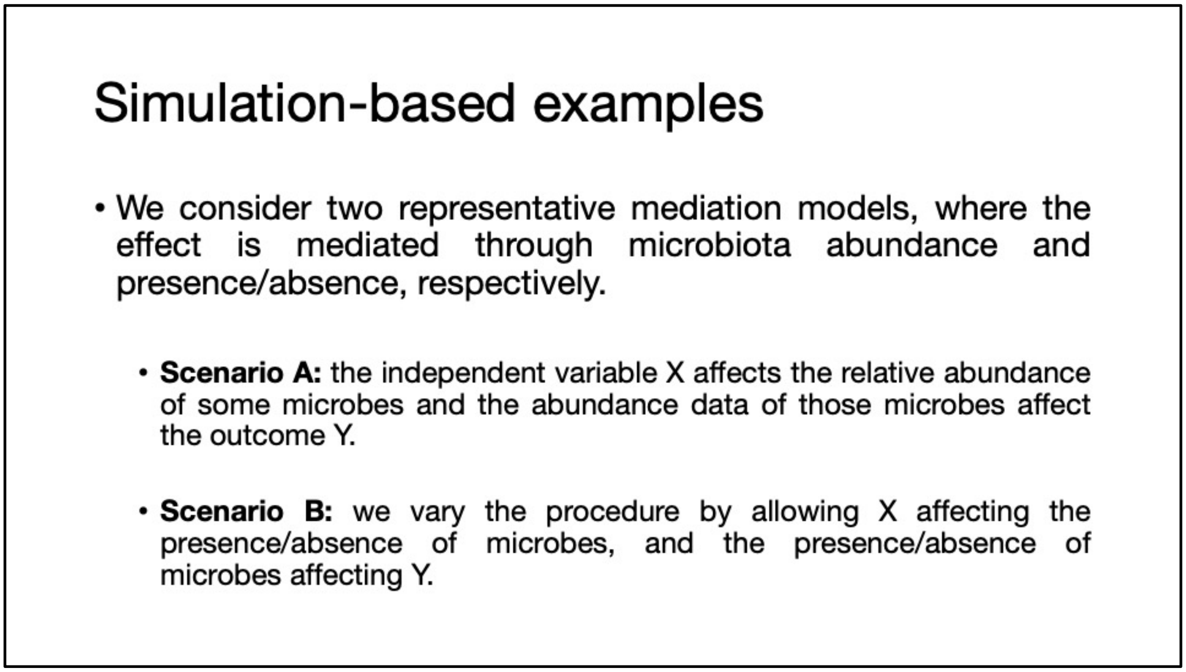
Omnibus test based on multiple distance measures

- Suppose that there are k distances, which are denoted as $d_1, d_2, \dots, d_K$.
- We could compute the test statistic for each distance, apply permutation tests to calculate p-values, and report the minimum p-value across multiple distances adjusted by Bonferroni correction for multiple comparisons. **Does not take correlation between distances into account.**

1. Compute the test statistic T_k for each distance d_k.
2. For each distance, generate B permuted statistics $T_k^{(j)}$, for $j = 1, 2, \dots, B$, by residual permutations.
3. Calculate the P-value P_k for each distance d_k as the proportion of statistics $T_k, T_k^{(1)}, \dots, T_k^{(B)}$ that is equal to or exceeds T_k.
4. Compute the minimum P-value across K distances $P_{\min} = \min(P_1, P_2, \dots, P_K)$.
5. For each distance, compute B permuted 'P-values' $P_k^{(j)} = 1 - (\text{rank}(T_k^{(j)}) - 1)/B$, where $\text{rank}(T_k^{(j)})$ is the rank of the statistic $T_k^{(j)}$ among B permuted statistics $T_k^{(1)}, T_k^{(2)}, \dots, T_k^{(B)}$.
6. Compute the permuted minimum P-values $P_{\min}^{(j)} = \min(P_1^{(j)}, P_2^{(j)}, \dots, P_K^{(j)})$, for $j = 1, 2, \dots, B$.
7. Report the final P-value as the proportion of permuted minimum P-values $P_{\min}^{(j)}$ exceeding minimum P-value $P_{\min}$.

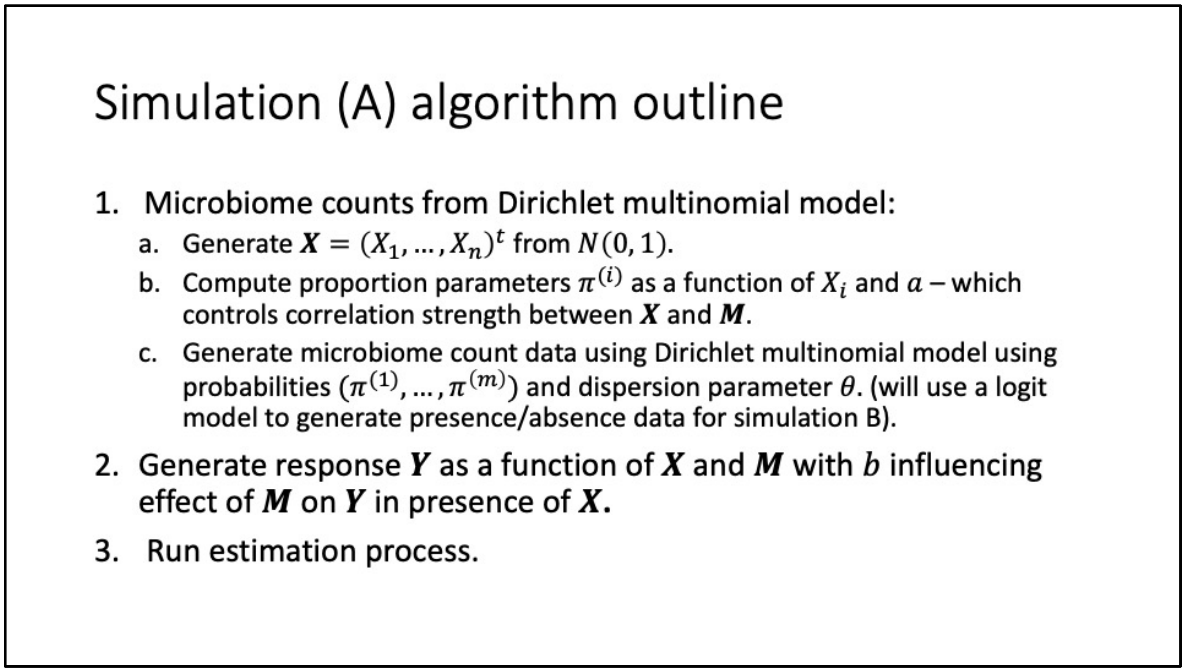
Each distance represents a distinct view of the microbiota and is ex- pected to be the most powerful to detect a specific mediation pat- tern. However, in real applications, we may have little knowledge about the underlying mediation mechanism. Sticking to a single dis- tance could miss important mediation effect. Therefore, considering different distance measures is key to the robustness and power of the test.

To integrate different distances simultaneously, we compute the minimum P-value across multiple distances as the test statistic and simulate its distribution under null hypothesis through permutation procedures.



Simulation-based examples

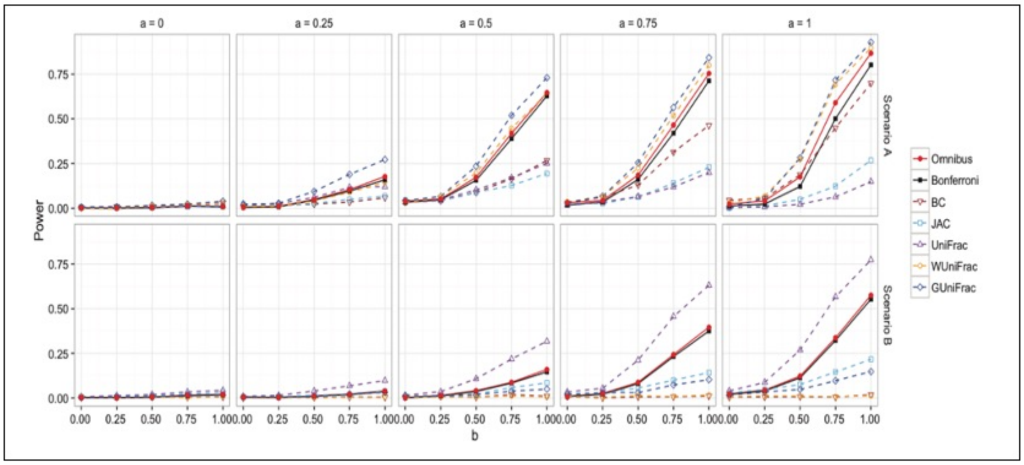
- We consider two representative mediation models, where the effect is mediated through microbiota abundance and presence/absence, respectively.
  - **Scenario A:** the independent variable X affects the relative abundance of some microbes and the abundance data of those microbes affect the outcome Y.
  - **Scenario B:** we vary the procedure by allowing X affecting the presence/absence of microbes, and the presence/absence of microbes affecting Y.



Simulation (A) algorithm outline

1. Microbiome counts from Dirichlet multinomial model:
   a. Generate $\boldsymbol{X} = (X_1, \dots, X_n)^t$ from $N(0, 1)$.
   b. Compute proportion parameters $\pi^{(i)}$ as a function of X_i and a – which controls correlation strength between $\boldsymbol{X}$ and $\boldsymbol{M}$.
   c. Generate microbiome count data using Dirichlet multinomial model using probabilities $(\pi^{(1)}, \dots, \pi^{(m)})$ and dispersion parameter θ. (will use a logit model to generate presence/absence data for simulation B).
2. Generate response $\boldsymbol{Y}$ as a function of $\boldsymbol{X}$ and $\boldsymbol{M}$ with b influencing effect of $\boldsymbol{M}$ on $\boldsymbol{Y}$ in presence of $\boldsymbol{X}$**.**
3. Run estimation process.

The paper compares the power of the omnibus test with tests based on some distance measures.


a = 0
a = 0.25
a = 0.5
a = 0.75
a = 1
Scenario A
Scenario B
Power
b
Omnibus
Bonferroni
BC
JAC
UniFrac
WUniFrac
GUniFrac