Statistics in Biosciences (2021) 13:197–199
https://doi.org/10.1007/s12561-021-09307-5

Introduction to Special Issue on Statistics in Microbiome and Metagenomics

Huilin Li[1] · Hongzhe Li[2]

Accepted: 1 March 2021 / Published online: 10 March 2021

Microbiome is the collection of all microbes that live in and on human body. It plays important roles in maintaining human health. Microbiome dysbiosis has been shown to be associated with many chronic diseases such as diabetes [1, 2], obesity [3, 4], Inflammatory bowel disease [5, 6], cardiovascular diseases [7, 8], and cancers [9, 10]. High-throughput sequencing technology makes it possible to perform either 16S rRNA sequencing [11] or shotgun metagenomic sequencing [12] on a large set of samples in different conditions. These raw sequencing read data can be further processed to provide the microbial community composition and microbial functional gene information. However, analyzing such data is challenging [11, 13] due to its unique data structures including high dimensionality, compositional nature, excessive zeros, and phylogenetic tree. Different from research in human genetics, microbiome is more susceptible to environmental confounding variables and is inherently dynamic and modifiable [14]. New statistical and computational methods are needed to better analyze such data, to reduce bias due to confounding, and to draw reproducible and stable conclusions.

This Special Issue of Microbiome and Metagenomics in Statistics in Biosciences consists of eight papers on various topics and covers the latest development of statistical methods in analyzing human microbiome data. Differential abundance analysis based on compositional data aims to identify the signature bacterial taxa that their relative abundances differentiate biological conditions such as disease and healthy conditions. Due to excessive zeros and compositional nature of the data, assumptions in many traditional statistical tests or regression analysis methods may not hold. Naïve application of such methods can either lead to false association or loss of power. To address these issues, Tang and Chen discuss several robust and powerful differential composition tests for clustered microbiome data that are often seen in family or longitudinal microbiome studies. The tests are based on estimating equations and do not require any distributional

✉ Huilin Li
Huilin.Li@nyulangone.or

[1] New York University, New York, NY, USA

[2] University of Pennsylvania, Philadelphia, USA

assumptions. Combettes and Müller discuss a general regression model for compositional data, where they present general log-contrast formulations to parameter estimation and introduce a proximal optimization algorithm with rigorous convergence guarantees to implement the method. To deal with excessive zeros in the data and satisfy the unit-sum constraint automatically, especially in longitudinal microbiome studies, Han et al. introduce a two-part linear mixed model with shared random effects, where they model the log-transformed standardized relative abundances using a random effect model. Wang et al. introduce a two-stage mixed effects models in order to link the microbiome longitudinal profiles to an outcome. They model the longitudinal microbial abundance count data as a function of time using the zero-inflated negative binomial mixed effects model in the first stage and use the estimated random intercepts and slopes from stage one in the second stage to link the temporal patterns with the outcome. Finally, distance-based methods have been very popular in microbiome data analysis due to its robustness and flexibility. Shaoyu Li presents a method that combines microbiome distances with quantile regression, an important alternative to existing distance-based regression methods that focus on modeling the mean.

Microbiome can serve as an important mediator in linking environmental exposure to clinical outcome or in linking treatment to its effectiveness. Mediation analysis aims to understand how treatment/environmental exposure shifts microbial composition and leads to clinical outcomes. Most existing mediation analysis methods only consider one or a few mediators. Zhang et al. present a mediation effect testing method developed particularly for microbiome compositional data, where they combine isometric log-ratio transformation with high dimensional regression using Lasso.

Understanding microbe–microbe interactions or microbiome–metabolite interactions is another important area in microbiome research. However, many models developed for Gaussian data do not apply to microbiome data directly. Jiang et al. discuss microbial interaction network estimation via bias-corrected graphical lasso based on a logistic normal multinomial distribution. The method corrects the bias of the naïve empirical covariance estimator arising from the different sequencing depths across samples. Jing Ma presents a joint microbial and metabolomics interaction network estimation using censored Gaussian graphical model where she treats the zero counts in the data as censored observation and extends the Gaussian graphical model to handle such a censoring in the data.

As the field of microbiome research progresses, we expect to see more research in linking gut microbiome with host genomics in order to gain functional insights into the role of microbiome in disease initiation and progression. We expect to see more diverse data sets that include gut microbiome, host gene expression, immune profiling, and metabolomics to be collected to investigate the interplay between microbe–host and microbe–microbe interactions. Analyzing such data raises many new statistical questions and challenges. We hope that more statisticians will work in this exciting area of research and make important scientific impact.

References

1. Vallianou NG, Stratigou T, Tsagarakis S (2018) Microbiome and diabetes: Where are we now? Diabetes Res Clin Pract 146:111–118
2. Tilg H, Moschen AR (2014) Microbiota and diabetes: an evolving relationship. Gut 63:1513–1521
3. Maruvada P, Leone V, Kaplan LM, Chang EB (2017) The human microbiome and obesity: moving beyond associations. Cell Host Microbe 22:589–599
4. John GK, Mullin GE (2016) The gut microbiome and obesity. Curr Oncol Rep 18:1–7
5. Manichanh C, Borruel N, Casellas F, Guarner F (2012) The gut microbiota in IBD. Nat Rev Gastroenterol Hepatol 9:599
6. Ni J, Wu GD, Albenberg L, Tomov VT (2017) Gut microbiota and IBD: causation or correlation? Nat Rev Gastroenterol Hepatol 14:573
7. Tang WW, Kitai T, Hazen SL (2017) Gut microbiota in cardiovascular health and disease. Circ Res 120:1183–1196
8. Witkowski M, Weeks TL, Hazen SL (2020) Gut Microbiota and Cardiovascular Disease. Circ Res 127:553–570
9. Vivarelli S, Salemi R, Candido S et al (2019) Gut microbiota and cancer: from pathogenesis to therapy. Cancers 11:38
10. Meurman JH (2010) Oral microbiota and cancer. J Oral Microbiol 2:5195
11. Johnson JS, Spakowicz DJ, Hong B-Y et al (2019) Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun 10:1–11
12. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N (2017) Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 35:833–844
13. Li H (2015) Microbiome, metagenomics, and high-dimensional compositional data analysis. Annu Rev Stat Appl 2:73–94
14. Rothschild D, Weissbrod O, Barkan E et al (2018) Environment dominates over host genetics in shaping human gut microbiota. Nature 555:210–215