Statistical Analysis of Microbiome Data with R, Y. Xia, J. Sun and D.-G. Chen (2018). Singapore: Springer, 505 pages. ISBN: 978-981-13-1533-6.

Recent breakthroughs and advances in culture-independent techniques, such as whole genome shotgun metagenomics and 16S rRNA amplicon sequencing, have dramatically changed the way we can examine microbial communities. But there are many hurdles to tackle before we are able to identify and compare bacteria driving changes in their ecosystems. These challenges range from bioinformatics and statistical data processing to analysis. The data are complex with inherent characteristics: they are sparse, compositional and multivariate.

In *Statistical Analysis of Microbiome Data with R*, Yinglin Xia, Jun Sun and Ding-Gen Chen provide a comprehensive overview of microbiome data and how to analyse them appropriately using classical as well as recently proposed methodologies.

The first chapter reviews the technical processing steps necessary to understand where the data are coming from. This is one of the major hurdles when working with microbiome data: knowledge of the sequencing technologies that are used and of concepts in phylogeny and bioinformatics are required to ensure that data are ready to be statistically analysed. Chapter 2 describes in more detail how the count data look like, in the form of a table of Operational Taxonomy Units (OTU). The chapter gives a good overview of the challenges arising from microbiome data, namely their compositional nature, size (p >>n), over-dispersion and sparsity. These first chapters are essential to any data analyst interested in working in this field.

Chapter 3 presents the field of microbiome data analysis and how to frame statistical hypotheses in this context. The chapter gives an overview of the different types of statistical methods that can be applied, from univariate to multivariate, and touches on the challenges to anticipate with compositional data and longitudinal studies. Chapter 4 is a quick start with R and an introduction to some key packages for data wrangling and data visualisation (*dplyr*, *ggplot2*) that are illustrated on some case studies provided by the authors.

In Chapter 5, the authors address the important topic of power calculation in such studies, with a particular emphasis on hypothesis testing. They introduce useful functions for various types of analyses for estimating sample size and power. These functions are fully illustrated on real data provided by the authors.

Chapter 6 introduces the $Vdr^{-/-}$ mouse microbiome data set that will be further analysed in the remaining chapters. Alpha and beta diversity measures are presented and illustrated. Chapter 7 introduces the cigarette smoker data set. Exploratory analyses based on sample clustering and ordination are presented and illustrated using the key R package *phyloseq.*

Chapters 8 and 9 present the different methods of conducting univariate and multivariate community analyses. Standard univariate statistical methods are illustrated as well as strategies to correct for multiple testing. Multivariate analyses are based on permutation ANOVA and variants,

using the packages *RVAidememoire* and *vegan*. Some emphasis is given to the comparison of the different Unifrac distances (*GUnifrac* package), with an illustration using the breast milk case study.

Chapter 10 gives an excellent review of the field of compositional data and introduces the key principles of CoDa (compositional data analysis). Related challenges include the sparsity of the data and correlation analyses that lead to spurious results in this context. Graphical outputs (e.g. compositional biplots) are illustrated using the $Vdr^{-/-}$ case study with a step-by-step description of how to process the data in this context. R packages, including *zcomposition*, *composition, Aldex2* and *propr,* are illustrated.

The final two chapters present the current challenges of over-dispersed and zero-inflated data. Chapter 11 introduces the Poisson and Binomial models, which are typically used for the analysis of RNA-seq data. These models can be used in this context but with some limitations. The chapter describes step-by-step analyses of the cigarette smokers case study using the packages *edgeR* and *DESeq2* (binomial tests). Chapter 12 discusses the problem of zero-inflation, and where the zeroes may come from, in microbiome data (sampling or structural zeroes). Zero-inflation and zero hurdle Poisson and Negative Binomial models are introduced, then compared using a vaginal microbiota case study.

To summarise, this book not only gives theoretical but also practical bases for getting started with microbiome data analysis. Several microbiome case studies are analysed in depth, and the topical challenges in this field are described. The last chapters (9 – 12) provide future areas of investigation for statisticians. As such, *Statistical Analysis of Microbiome Data with R* represents a very good foundational resource for bioinformaticians and statisticians interested in this emerging area of research.

Dr Kim-Anh Lê Cao

School of Mathematics and Statistics

The University of Melbourne

Parkville, Australia

Email: kimanh.lecao@unimelb.edu.au