Preface

The writing of this book is motivated by my research in the area of genetics and epigenetics and a course that I have been teaching recently with a focus on biostatistics methods used to analyze bioinformatics data. The rapidly advancing "omics" technologies have generated massive transcriptomics, metabolomics, proteomics, and epigenomics data and brought a strong potential to promote biomedical research. Raw "omics" data are filled with noise and biased by batches and technical variations, and furthermore, this data covers tens of thousands of genes which have complex relationships and interconnections. Students majoring in Biostatistics and researchers dealing with these types of data have a strong need to learn methods and analytical tools that have the ability to handle these high-dimensional data starting from the initial product, the raw data. These skills include but are not limited to data preprocessing, data mining, marker detection, and other related analyses. However, books having a "pipeline" of analytical methods with concrete examples are rather limited.

The book is an attempt to fill this gap and aims to provide a "pipeline" for gene expression and epigenetic data analysis starting from raw genome- and epigenome-scale data. For epigenetic data, we specifically focus on DNA methylation. The book is a combination of methodology introduction and R program implementation. Unlike many R programming books which focus on utilizing R packages, this book introduces each statistical method before moving to concrete examples and R packages. This structure is expected to substantially benefit biostatisticians as well as readers with interest in the underlying techniques in addition to implementation of R packages. A number of analytical methods and tools are discussed in the book, including methods to preprocess genome-scale gene expression and epigenetic data, methods for data mining to identify potentially informative factors, and methods for subsequent analyses after data mining, e.g., factor/variable selections, network construction, and testing for differential networks. Many methods introduced are recently developed and advanced, which are

expected to outperform existing ones. In the meantime, these advanced approaches will provide a platform for statisticians to develop more efficient approaches. For most methods introduced, they are followed by examples and R programs for users to practice and apply.

This book covers statistical methods and specific examples with R programs. It does not require strong statistical background to use the programs, and thus has a potential to apply to readers with different needs. The targeted readers of this book are graduate students with some statistics/biostatistics background and biostatisticians with interest in analyzing high-dimensional "omics" data, especially expressions of genes and DNA methylation, and researchers interested in analyzing "omics" data with limited knowledge in statistical methods. For graduate students, it is ideal if they are in the second year of their study. For PhD students, there is no limitation. Students will have a better understanding of the statistical methods and programs if they have some statistic background. For researchers, they are not required to have strong statistic backgrounds, but some understanding in statistic concepts will be very helpful.

In the process of writing this book, I was supported by a number of people. Especially, I would like to thank David Grubbs and his assistants for their patience with me and their flexibility to fit my academic schedule. The support from the School of Public Health at the University of Memphis made the whole writing process pleasant and productive. I am thankful to Luhang Han, Shengtong Han, Yu Jiang, and Meredith Ray for their help in proof-reading. My great appreciations also go to my husband and children for their understanding and generosity. I would like to dedicate this book to my parents, who are always supportive.