Preface

We are entering an era of *Big Data*. Big Data offer both unprecedented opportunities and overwhelming challenges. This book is intended to provide biologists, biomedical scientists, bioinformaticians, computer data analysts, and other interested readers with a pragmatic blueprint to the nuts and bolts of Big Data so they more quickly, easily, and effectively harness the power of Big Data in their ground-breaking biological discoveries, translational medical researches, and personalized genomic medicine.

Big Data refers to increasingly larger, more diverse, and more complex data sets that challenge the abilities of traditionally or most commonly used approaches to access, manage, and analyze data effectively. The monumental completion of human genome sequencing ignited the generation of big biomedical data. With the advent of ever-evolving, cutting-edge, high-throughput omic technologies, we are facing an explosive growth in the volume of biological and biomedical data. For example, Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) holds 3,848 data sets of transcriptome repositories derived from 1,423,663 samples, as of June 9, 2015. Big biomedical data come from government-sponsored projects such as the 1000 Genomes Project (http://www.1000genomes.org/), international consortia such as the ENCODE Project (http://www.genome.gov/encode/), millions of individual investigator-initiated research projects, and vast pharmaceutical R&D projects. Data management can become a very complex process, especially when large volumes of data come from multiple sources and diverse types, such as images, molecules, phenotypes, and electronic medical records. These data need to be linked, connected, and correlated, which will enable researchers to grasp the information that is supposed to be conveyed by these data. It is evident that these Big Data with high-volume, high-velocity, and high-variety information provide us both tremendous opportunities and compelling challenges. By leveraging

the diversity of available molecular and clinical Big Data, biomedical scientists can now gain new unifying global biological insights into human physiology and the molecular pathogenesis of various human diseases or conditions at an unprecedented scale and speed; they can also identify new potential candidate molecules that have a high probability of being successfully developed into drugs that act on biological targets safely and effectively. On the other hand, major challenges in using biomedical Big Data are very real, such as how to have a knack for some Big Data analysis software tools, how to analyze and interpret various next-generation DNA sequencing data, and how to standardize and integrate various big biomedical data to make global, novel, objective, and data-driven discoveries. Users of Big Data can be easily "lost in the sheer volume of numbers."

The objective of this book is in part to contribute to the NIH Big Data to Knowledge (BD2K) (http://bd2k.nih.gov/) initiative and enable biomedical scientists to capitalize on the Big Data being generated in the omic age; this goal may be accomplished by enhancing the computational and quantitative skills of biomedical researchers and by increasing the number of computationally and quantitatively skilled biomedical trainees.

This book covers many important topics of Big Data analyses in bioinformatics for biomedical discoveries. Section I introduces commonly used tools and software for Big Data analyses, with chapters on Linux for Big Data analysis, Python for Big Data analysis, and the R project for Big Data computing. Section II focuses on next-generation DNA sequencing data analyses, with chapters on whole-genome-seq data analysis, RNA-seq data analysis, microbiome-seq data analysis, miRNA-seq data analysis, methylome-seq data analysis, and ChIP-seq data analysis. Section III discusses comprehensive Big Data analyses of several major areas, with chapters on integrating omics data with Big Data analysis, pharmacogenetics and genomics, exploring de-identified electronic health record data with i2b2, Big Data and drug discovery, literature-based knowledge discovery, and mitigating high dimensionality in Big Data analysis. All chapters in this book are organized in a consistent and easily understandable format. Each chapter begins with a theoretical introduction to the subject matter of the chapter, which is followed by its exemplar applications and data analysis principles, followed in turn by a step-by-step tutorial to help readers to obtain a good theoretical understanding and to master related practical applications. Experts in their respective fields have contributed to this book, in common and plain English. Complex mathematical deductions and jargon have been avoided or reduced to a minimum. Even a novice,

with little knowledge of computers, can learn Big Data analysis from this book without difficulty. At the end of each chapter, several original and authoritative references have been provided, so that more experienced readers may explore the subject in depth. The intended readership of this book comprises biologists and biomedical scientists; computer specialists may find it helpful as well.

I hope this book will help readers demystify, humanize, and foster their biomedical and biological Big Data analyses. I welcome constructive criticism and suggestions for improvement so that they may be incorporated in a subsequent edition.

Shui Qing Ye
University of Missouri at Kansas City