Preface

The twenty-first century is the time of excitement and optimism for biomedical research. Since the completion of the human genome project in 2001, we are entering into the postgenome era where the key research efforts are now interpreting and making sense of these massive genomic data, in order to translate into disease treatment and management. Over the past decade, DNA-based microarrays have been the assays of choice for high-throughput studies of gene expression. Microarray-based expression profiling was provided, for the first time, by means of monitoring genome-wide gene expression changes in a single experiment. Though microarray technology has been widely employed to reveal molecular portraits of gene expression in various cancers' subtypes and correlations with disease progression as well as response to drug treatments, it is not limited to measure gene expression. As the technology became established in early 2000, researchers began to use microarrays to measure other important biological phenomena. For example, (1) Microarrays are being used to genotype single-nucleotide polymorphisms (SNPs) by hybridizing the DNA of individuals to arrays of oligonucleotides representing different polymorphic alleles. The SNP microarray has accelerated genome-wide association studies over the last 5 years, and many loci that are associated with diseases have been discovered and validated. Similarly, another innovative application of the SNP microarray is to interrogate allele-specific expression for identifying disease-associated genes. (2) Array-comparative genomic hybridization (aCGH) is being used to detect genomic structural variations, such as segments of the genome that have varying numbers of copies in different individuals. (3) Epigenetic modifications such as methylation at CpG sites can also be assessed by microarray. (4) Using ChIP-chip assay, genome-wide protein–DNA interactions and chromatin modifications can be profiled by microarrays. (5) More recently, microarray has been used to measure genome-wide microRNA expression patterns to reveal the regulatory role of these noncoding RNAs in disease states. Obviously, the progress of microarray applications is tightly associated with the development of novel computational and statistical methods to analyze and interpret these data sets.

Recent improvements in the efficiency, quality, and cost of genome-wide sequencing have prompted biologists and biomedical researchers to move away from microarray-based technology to ultrahigh-throughput, massively parallel genomic sequencing (Next Generation Sequencing, NGS) technology. NGS technology opens up new research avenues for the investigation of a wide range of biological and medical questions across the entire genome at single base resolution; for example, sequencing of several human genomes, monitoring of genome-wide transcription levels (RNA-seq), understanding of epigenetic phenomena, DNA–protein interactions (ChIP-seq), and de novo sequencing of several genomes. Despite the differences in the underlying sequencing technologies of various NGS machines, the common output from them are the capability to generate tens of millions of short reads (tags) from each experimental run. Thus, NGS technology shifts the bottleneck in sequencing processes from experimental data production to computationally intensive informatics-based data analysis. As in the early days of microarray data analysis, novel computational and statistical methods tailored to NGS are urgently needed for drawing meaningful and accurate conclusions from the massive short reads. Furthermore, it is expected that NGS technology may eventually replace microarray technology in the

next decade, which will grow from a pioneering method applied by innovators at the cutting edge research to a ubiquitous technique that will allow researchers to investigate "big-picture" questions in biology at much higher resolution.

This book, Next Generation Microarray Bioinformatics, is our attempt to bring together current computational and statistical methods in analyzing and interpreting both microarray and NGS data. Here, we have compiled and edited 26 chapters that cover a wide range of methodological and application topics in microarray and NGS bioinformatics. These chapters are organized into five thematic sections: (1) Resources for Microarray Bioinformatics; (2) Microarray Data Analysis; (3) Microarray Bioinformatics in Systems Biology; (4) Next Generation Sequencing Data Analysis; and (5) Emerging Applications of Microarray and Next Generation Sequencing. Each chapter is a self-contained review of a specific methodological or application topic. Every chapter typically starts with a brief review of a particular subject, then describes in detail the computational and statistical techniques used to solve the biological questions, and finally discusses the computational results generated by these bioinformatics tools. Therefore, the reader need not read the chapters in a sequential manner. We expect this book would be a valuable methodological resource not only to molecular biologists and computational biologists who are interested in understanding the principle of these methods and designing future research project, but also to computer scientists and statisticians who work in a microarray core facility or other similar organizations that provide service for the high-throughput experiment community.

The first section of this book contains three important resource chapters of microarray and NGS bioinformatics community. The introductory chapter provides an overview on the current state of microarray technologies and is contributed by Kuo and colleagues. The second chapter is contributed by the KEGG group. The KEGG database represents one of the earliest databases to store, manage, integrate, and visualize genomics data. In this chapter, Kotera and colleagues provide the latest developments of the KEGG efforts in analyzing and interpreting omics data. The NCBI Gene Expression Omnibus (GEO) group writes the third chapter in this section, which is one of the major data repositories for high-throughput microarray and next-generation sequencing data. White and Barrett describe various strategies to explore functional genomics data sets in the GEO database.

The second section of this book consists of eight chapters that describe methods to analyze microarray data from the top down approach. The first chapter, contributed by Van Loo and colleagues, that described a novel R-package ASCAT specifically designed to delineate genomic aberration in cancer genomes from SNP microarrays. Then Cheung, Meng, and Huang wrote the following two chapters of advanced machine learning methods in investigating disease classification and time-series microarray data analysis, respectively. Lin and colleagues provide a tutorial on a novel R-package, GeneAnswers, to perform gene-concept network analysis in the next chapter. Nair contributed the next chapter, which emphasizes the utility of R/Bioconductor, an open source software for bioinformatics, in the analysis and interpretation of splice isoforms in microarray. The next three chapters focusing on cross-platform comparisons of microarray data and integrative approaches for microarray data analysis were delivered by Li et al., Hovig et al., and Huttenhower et al., respectively.

The third section of this book concentrates on the bottom-up approaches for establishing different types of models based on microarray expression datasets in which the number of genes is much larger than that of samples. The first chapter written by Yu and colleagues discussed a general profiling method to estimate parameters in the ordinary differential

equation models from the time-course gene expression data. To deal with inhomogeneity and nonstationarity in temporal processes, Husmeier and colleagues described the inhomogeneous dynamic Bayesian networks which allow the network structure to change over time in the second chapter. Castelo and Roverato contributed the third chapter that introduced an R package of a graphic approach for inferring regulatory networks from microarray datasets. Wang and Tian contribute the final chapter of this section. They introduced a nonlinear model, which can be used to infer the transcriptional factor activities from the microarray expression data of the target genes as well as to predict the regulatory relationship between transcriptional factors and their target genes.

The fourth section of this book contains six chapters, specifically devoted to NGS data analysis. It starts from an overview of the NGS data analysis by Gogol-Döring and Chen, which includes the basic steps for analyzing NGS such as quality check and mapping to a reference genome. The second chapter is written by Sandber and colleagues, where the authors provide a detailed illustration of how to analyze gene expression using RNA-Sequencing data through several real examples. Lin and colleagues contributed to the third chapter that introduces the low level ChIP-seq data analysis such as preprocessing, normalization, differential identification, and binding pattern characterization. The fourth chapter is contributed by Xu and Sung, in which reader will find how to use Hidden Markov Model to identify differential histone modification sites from ChIP-seq data. The last two chapters describe two software packages (SISSRs developed by Narlikar and Jothi and ChIPMotifs developed by Jin and colleagues) that are designed to study protein–DNA interactions (e.g., peak finder and de novo motif discovery) by analyzing ChIP-based high-throughput experiments.

The final section of this book contains five methodological chapters that cover the emerging applications of microarray and next-generation sequencing in biomedical researchers. In Wei's chapter, it describes Hidden Markov Models for controlling false-discovery rate in genome-wide association analysis. Tan describes Gene Set Top Scoring Pairs (GSTSP), a novel machine learning method in identifying discriminative gene set classifier, based on the relative expression concept. In the next chapter, Wu and Ji focus on JAMIE, a software tool that can perform jointly analysis on multiple ChIP-chip experiments. In the chapter written by Pelligrini and Ferrari, they described an overview on bioinformatics methods in analyzing epigenetic data. The final chapter is a bioinformatics workflow for the analysis and interpretation of genome-wide shRNA synthetic lethal screen based on next-generation sequencing written by Kim and Tan.

We would like to acknowledge the contribution of all authors to the conception and completion of this book. We would like to thank Prof. John M. Walker, the Methods in Molecular Biology series editor, for entrusting and giving us this opportunity to edit this volume. We also like to thank the staff at the Humana Press and Springer publishing company for their professional assistance in preparing this volume. Finally, we would like to thank our families for their love and support.

Oslo, Norway — *Junbai Wang*
Aurora, CO, USA — *Aik Choon Tan*
Melbourne, VIC, Australia — *Tianhai Tian*

http://www.springer.com/978-1-61779-399-8

Next Generation Microarray Bioinformatics
Methods and Protocols
Wang, J.; Tan, A.C.; Tian, T. (Eds.)
2012, XVI, 401 p. 96 illus., 30 illus. in color., Hardcover
ISBN: 978-1-61779-399-8
A product of Humana Press