

Preface

Statistical genomics is a new interdisciplinary area of science, including statistics, genetics, computer science, genomics, and bioinformatics. The rapid advances in these areas have dramatically changed the amount and type of information available for characterization of genes. In many genomic applications, existing methods coupled with new computational technology have successfully directed the exploration of high-dimensional data. What remains to be accomplished is the successful statistical modeling of genomic data to support hypothesis-driven biological research. This will ultimately lead to the exploitation of the predictive wealth that much of the current and impending genomic data have the potential to offer. Statistical development will continue to significantly amplify and focus the molecular advances of the last decades toward general improvements in agriculture and human health.

Using advanced statistical technology to study the behavior of one or a few Mendelian loci defines the field of statistical genetics. For complex traits, such as grain yield in crops and cancers in human, one or two loci are rarely sufficient to explain majority of the trait variation. People then study the behavior of all genes influencing a trait without distinguishing the effects of individual genes, creating a field called quantitative genetics. Taking advantage of saturated DNA markers generated with advanced molecular technology, we are now able to localize individual genes on the genome that affect a complex trait, which leads to this new field of statistical genomics or quantitative genomics. In statistical genomics, we emphasize the notion of whole genome analysis and evaluate the joint effect of the entire genome on a quantitative trait.

Any genome study requires a sample of individuals from a target population and genomic data collected from this sample. Genomic data include (a) genotypes of molecular markers, (b) microarray gene expressions, and (c) phenotypes of clinic or economic traits measured from all individuals of the sample. Any particular genomic study may involve all the three types of data or any two of them. With the advanced biotechnology, molecular marker data will soon be replaced by whole genome sequences. In the narrow sense, phenotypic data are not genomic data but the ultimate purpose of genomic data analysis is to dissect these traits and

understand the genetic architectures of these traits. Therefore, phenotypic data are essential in genomic data analysis. This is why phenotypic data are included as genomic data. When a study involves phenotypes and marker genotypes, it is called QTL mapping where QTL stands for quantitative trait loci. A study involving phenotypes and microarray gene expressions is called differential expression (DE) analysis. If a study involves marker genotypes and microarray gene expressions, it is called expression quantitative trait locus (eQTL) analysis. The purpose of QTL mapping is to find the genome locations, the sizes, and other properties of QTL through associations of marker genotypes with the variation of a quantitative trait. In DE analysis, the phenotype of interest is usually binary such as case (represented by one) and control (represented by zero). The primary interest of DE is to find genes that express differently in case and control. The purpose of eQTL mapping is to find regulation pathways of the genes. Transcripts mapped to the same locus of the genome are considered in the same regulation pathway.

Many statistical models, methodologies, and computing algorithms are involved in the textbook. Major statistical models include the linear model (LM), the generalized linear model (GLM), the linear mixed model (LMM), and the generalized linear mixed model (GLMM). In a few places, the hidden Markov model (HMM) is required to infer the unobserved genotypes of QTL given observed marker genotypes. Another important model is the Gaussian mixture model for cluster analysis. Commonly used statistical methods include the least squares (LS) estimation, the maximum likelihood (ML) estimation, the Bayesian estimation implemented via the Markov chain Monte Carlo (MCMC) algorithm, and the Bayesian method via the maximum a posteriori (MAP) estimation. Optimization technologies include the Newton–Raphson algorithm, the Fisher scoring algorithm, and the expectation–maximization (EM) algorithm. For the Newton–Raphson algorithm, if the first- and second-order partial derivatives of the target function with respect to the parameters are easy to derive, an explicit form of the iteration equation will be given. Otherwise, numerical evaluations of the partial derivatives are calculated using some powerful numerical differentiation subroutines. In genomic data analysis, the number of parameters is often very large, updating all parameters simultaneously can be prohibitive. In this case, a coordinate descent approach may be taken, in which one parameter is updated at a time conditional on current values of all other parameters. This approach can improve the robustness of the optimization algorithm and save much computer memory but at the cost of computing time and risk of trapping to a local solution of parameters.

This book was compiled from a collection of lecture notes for the statistical genomics course (BPSC234) offered to UCR graduate students by the author. Approximately half of the material was collected from studies published by the author and his research team. A small proportion of the remaining half consists of some unpublished works conducted by the author. Much of the remaining half of the book represents a collection of the most updated statistical genomic methods published in various journals for the last couple of decades. The topics selected purely reflect the author’s choices for the course according to the level of understanding of the target students. The book is not an introduction to statistical

genomics because statistical genomic is a diversified area including many different topics, and this book only covers a proportion of the topics. However, the statistical technologies chosen represent the core of statistical genomics. Understanding the principles of these technologies, students will easily extend the methods to other analyses of genomic data generated from different experimental designs. Although the book narrowly focuses on a few topics, each topic introduced is provided with the derivation of the method or at least a direction leading to the derivation. Statistical genomics is a multidisciplinary area with a rapid development. Writing a comprehensive book in such an area is like shooting a moving target. For example, during the time between the completion of the first draft and the publication of this book, new technologies and methodologies may have already been developed. Therefore, the book can only focus on the principles of statistical genomics. Most recently developed methods may not be covered, for which the author owes an apology to those researchers whose works are relevant but not cited in the book.

The book consists of three parts. Part I contains Chaps. 1–4 and covers topics related to linkage map construction for DNA markers. Part II consists of Chaps. 5–16 and is the main part of the book. These chapters cover topics related to genetic mapping for quantitative trait loci using various designs of experiments. Part III (Chaps. 17–25) covers topics related to microarray gene expression data analysis. This book intends to be used as a textbook for graduate students in statistical genomics, but it can be used by researchers as a reference book. For advanced readers, they can choose to read any particular chapters as they desire. However, for junior researchers and graduate students, it is better to study from the beginning and not to escape any chapters because some of the methods introduced in early chapters will be used later in the book and they will only be referenced.

Former and current postdocs and graduate students in the lab all contributed to the material published by the UCR quantitative genetics team. Postdocs who contributed to the material relevant to this book include Damian Gessler, Chongqing Xie, Shaoqi Rao, Nengjun Yi, Claus Vogl, Chenwu Xu, Yuan-Ming Zhang, Lide Han, Zhiqiu Hu, and Fuping Zhao. Graduate students involved in the research include Lang Luo, Yun Lu, Hui Wang, Yi Qu, Zhenyu Jia, Xin Chen, Xiaohong Che, and Haimao Zhan. Without their hard work, the author would not have been able to publish this book. Their contributions are highly appreciated. In the main text, I choose to use the first person plural pronoun “we” instead of “I” for the very reason that the book material was mainly contributed by my research team. In the UCR quantitative genetics team, Nengjun Yi made the most contribution to the material included in the book and thus he deserves a special acknowledgement. A special appreciation goes to the three current members of the UCR quantitative genetics team, Zhiqiu Hu (postdoc), Haimao Zhan (student), and Xiaohong Che (student), for their help in drawing the figures, checking the accuracy of equations, and correcting errors occurred in an early draft of the book.



<http://www.springer.com/978-0-387-70806-5>

Principles of Statistical Genomics

Xu, S.

2013, XVI, 428 p., Hardcover

ISBN: 978-0-387-70806-5