

Preface

Descending from a single ancestor, genomes became subject to Darwinian selection and adaptation. It is through mutations that functional innovation or loss of ancestral function could lead to an increase of genome complexity, and adaptation to the challenges of a continuously changing environment. Evolutionary genomics is a recently-emerged research field, with the ultimate goal of understanding the underlying evolutionary and genetic mechanisms for such processes. It stems from the links between high throughput data in functional genomics, statistical modeling and bioinformatics, and phylogeny-based analysis.

During the last decade, high throughput technologies in genomics have generated enormous genome sequences data in many organisms in all ranges of life forms. Indeed, these technological innovations have made evolutionary genomes a rapidly-growing field, witnessed by numerous research papers published in the literature. At this stage, inevitably these studies are not only highly heterogeneous in methodology, but also highly controversial in biological interpretation. The purpose of this book is to attempt to present a tentative framework of statistical theory and methods that are useful in the study of genomic evolution, and illustrate how to use them in real, large-scale data analyses. Since nowadays genomic data analyses are almost always software-based, our explanation of statistical methods is focused on biological foundations and model assumptions, rather than the details of computational algorithms and practical implementations.

The author wishes to unify the field of evolutionary genomics by building a theoretically inherent framework that can provide a basis for interpreting enormous genomic data by the means of evolution. This is a long-term, ambitious goal that cannot be accomplished within the volume of this book. This book discusses various models and methods, including those developed by the author and his collaborators. Some of them were investigated during the process of writing this book and have been published very recently or for the first time here. Preference was given to statistical models that are based on realistic biological assumptions and statistical methods that are practically useful. Since the author's laboratory has been engaged in this area of study since the emergence of genome sciences, this book includes a substantial portion of work conducted by his group. Yet, the author has acknowledged the impossibility of covering all important topics and references; that would be beyond his capability, as well as the mission of this book. Indeed, balance has to be made between personal preference and the different scientific views and approaches. For all these reasons, this book should be viewed as an effort from an individual scientist as a step toward the ultimate goal of synthesis.

This book is written for graduate students and researchers in the field of evolutionary genetics/genomics and related fields, as well as those scientists in mathematics,

statistics, or computer sciences who are interested in computational biology and bioinformatics. If this book is used in the classroom, the author encourages the instructor to select the contents that best fit their teaching purposes. As this book is intended to introduce statistical models and methods, the author expects readers with the necessary training in statistics and mathematics. In addition, readers should have some basic knowledge of genome sciences and evolutionary biology. Readers who wish to know more on some important topics that are discussed only briefly may refer to many excellent books, such as Li (1997) and Nei and Kumar (2000) for molecular evolution and phylogeny, Yang (2006) for statistical methods of DNA sequence analysis, Lynch (2009c) for genomic evolution, and Evens and Grant (2005) for statistical bioinformatics.

In this book, the author introduces the basics of molecular evolutionary theory and bioinformatics tools so that readers can comprehend related issues discussed in this book. Then follows a range of topics, chapter by chapter, including functional divergence of gene families, expression divergence between duplicate genes, tissue-driven evolution, gene pleiotropy, and gene content evolution. These topics were organized by the model-driven approach, with a few well-selected examples of data analyses. The author acknowledges that there are a huge number of relevant publications in the literature that cannot all be cited here. In particular, this book does not discuss in detail the widely-used data-driven approach that takes advantage of high computational capacity and high throughput genomic data. Indeed, after developing integrated, comprehensive databases, one can utilize various efficient IT technologies such as clustering or machine-learning to conduct extensive data exploration, in an attempt to find biologically meaningful patterns that can be further tested by experimentation or independent datasets. In the author's view, the model-driven approach and the data-driven approach are complementary; both are indispensable for having a deep understanding of some fundamental biological problems. The last chapter of this book tackles several evolutionary issues that have not been completely resolved in systems biology. The intent was to explore the field coined 'evolutionary systems biology' for what has been accomplished, and what may potentially lie ahead. Hopefully these explorations will benefit research in this field. Finally, it should be noted that most statistical methods discussed in this book had computer programs available when they were published. For instance, previously, the author's laboratory published a computer program (DIVERGE) to predict amino acid sites that are responsible for functional divergence between two duplicate genes. Nevertheless, the author's group plans to develop an updated, user-friendly computer software package that can cover these methods.

The author has many people to thank: Jiazhen Tan (C. C. Tan), the founder of Chinese modern genetics who has influenced the author's career in genetics and evolution profoundly since he was an undergraduate student in Fudan University, China; Zhudong Liu, his master's degree supervisor in Fudan University, who provided him with a unique opportunity to learn mathematical population genetics; Wen-Hsiung Li, his Ph.D supervisor in the University of Texas, who guided him to the frontline of molecular evolution and phylogeny; and Masatoshi Nei, his postdoc supervisor in Penn State University, who encouraged him to develop his own research niche when genome

science was emerging. The author would also like to thank all his friends, colleagues, and collaborators during his career development over the last two decades, especially Li Jin, Yunxin Fu, Ranajit Chakraborty, Sudhir Kumar, C-I Wu, Manyuan Long, Dan Graur, Mike Miyamoto, Ziheng Yang, Jianzhi Zhang, Zhenglong Gu, Gunter Wagner, Duane Enger, Dan Voytas, Tom Peterson, Patrick Schnable, Dan Nettleton, Karin Dorman, Hui-Hsien Chou, Xiaoqi Huang, Jonathan Wendel, Eric Gaucher, Yaping Zhang, Jun Yu, Ji Yang, and Yang Zhong. The author is grateful to former and current students, research associates, and collaborators in his laboratory, who have made great contributions to solving many challenging problems in the statistical study of genomic evolution, computer program development, and large-scale data analysis; especially to Yufeng Wang, Jianying Gu, Kent Vander Velden, Zhongqi Zhang, Shiquan Wu, Huaijun Zhou, Zhixi Su, Yong Huang, Yangyun Zou, Hongmei Zhang, and Xiujuan Wang. The author acknowledges support in all aspects from the Department of Genetics, Development and Cell Biology, Iowa State University, for giving him time to finish the book. The author also has special acknowledgement for his wife, Wei. Without her full support and encouragement over the past 17 years, this book would not have been possible.

Research activities conducted in the laboratory of the author have been partly supported by the National Institutes of Health, USA (NIH), the National Science Foundation, USA (NSF), the National Science Foundation in China (NSFC), DuPont Young Investigator Award, Iowa State University, and Fudan University. The author is grateful to these funding agencies for their generous support.

Xun Gu
Ames, Iowa USA
February, 2010