Handbook on Analyzing Human Genetic Data

Shili Lin • Hongyu Zhao
Editors

Handbook on Analyzing Human Genetic Data

Computational Approaches and Software

Springer

Editors
Professor Dr. Shili Lin
Department of Statistics
The Ohio State University
Columbus, Ohio 43210
USA
shili@stat.ohio-state.edu

Professor Dr. Hongyu Zhao
Department of Epidemiology and Public Health
Yale University
School of Medicine
60 College St.
New Haven, CT 06520-8034
USA
hongyu.zhao@yale.edu

ISBN 978-3-540-69263-8 e-ISBN 978-3-540-69624-5
DOI 10.1007/978-3-540-69264-5
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2009931713

Cover design: WMX Design GmbH, Heidelberg

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface and Introduction

The discipline of statistical genetics is highly computational. Be it exact computational methods, simulation based, or a hybrid of the two, computational packages are indispensable tools and constant companions of researchers in the field. This handbook is intended to provide human geneticists and other biomedical researchers with guidance on selections of appropriate computational methods and software packages for their specific genetic problems. It may also be used by students and other learners as a reference in conjunction with a more theoretical and/or methodologically oriented text book. This book tries to strike a balance between methodological expositions and practical guidelines for software selections. Wherever possible, comparisons among the competing methods and software are made to highlight the relative advantages and disadvantages of the approaches, so that the reader can make informed choices to best match their specific needs.

Human genetics has been undergoing an evolution in the past several years as new knowledge and technologies are transforming the field, leading to numerous new discoveries of genes associated with complex traits such as cancer, obesity, and diabetes. Many recent genome-wide association studies employ the case–control design, where the study subjects consist of unrelated affected individuals and normal controls. For each individual, a large number of genetic markers are queried. A genetic marker refers to a location in the human genome where people may differ in the genetic material they carry. Genetic markers can come in different forms, with the single nucleotide polymorphisms (SNPs) most commonly used due to their high abundance in the genome and the availabilities of reliable and affordable technologies to genotype them. For a SNP, two different forms (called alleles) generally exist at a single nucleotide position. Because each person carries two chromosomes, for a given SNP with two alleles A and a, there are three possible genotypes a person can have: AA, Aa, and aa. In this setting, a genetic association study amounts to identifying markers that are associated with disease status. This can be accomplished by examining whether there is a statistical association between the marker genotype and the disease status. Although this analysis resembles a standard epidemiological study where each marker can be treated as a potential risk factor, there are many issues that are unique to genetics studies that need to be addressed. For example, one major concern in these studies is sample heterogeneity in their genetic background, and ignoring this issue may result in many false positive findings that have nothing

to do with disease etiology. On the other hand, much research has been done to empirically characterize and theoretically model the distributions and dependencies of genetic markers, and such knowledge is very beneficial for association analysis. In fact, a thorough genetic association analysis is not possible without a good understanding of the basic principles in population genetics, a field devoted to the study of the allele frequency distribution and change under various factors that can impact them, including mutations, random sampling, migrations, and natural selections. The chapter by Dr. Weir provides an overview of the basic concepts of population genetics and serves as the starting point of the analysis of human genetics data.

Although current genotyping platforms can genotype up to one million markers, there are many more markers in the genome that are not queried on these platforms. The reason that these typed markers can provide a good coverage of the genome is the dependence among physically close markers, and such dependence is called linkage disequilibrium. For example, if one SNP has alleles A and a each with allele frequency 50%, and another marker with alleles B and b each with frequency 50%. If the two markers are independent of each other, we would expect that 25% of chromosomes carry both A and B in the population, and similarly for all other three possible combinations: Ab, aB, and ab. However, it is often the case that if these two markers are very close to each other on the same chromosome, the two alleles carried on the same chromosome are not independent. In the most extreme case, there are only two types of chromosomes, those carrying AB and those carrying ab, a phenomenon called perfect linkage disequilibrium. Haplotypes refer to the combination of alleles on the same chromosome, and the presence of such marker dependency is the key underlying recent successes of genetic association studies collecting the genotypes from only a small fraction of all known markers. There are many statistical challenges presented in the analysis of haplotypes, both for population genetics studies and for more effective genetic association studies. These topics are discussed in the chapter by Drs. Zhang and Niu focusing on population genetics and in the chapter by Drs. Epstein and Kwee in the context of disease association analysis.

Genetic association studies can be performed on unrelated individuals using traditional epidemiological designs, for example, case–control design and cohort design, or designs unique to genetic studies, for example, family-based association design. Because sample heterogeneity in genetic background is one major concern in the validity of a genetic association study based on unrelated individuals, various statistical methods have been proposed to utilize genetic information in the collected marker genotypes to make appropriate adjustments in association analysis. For example, with enough marker information, it is possible to infer genetic background for each individual and such inferred background information can be incorporated in association analysis to make the results less susceptible to sample heterogeneity. This issue is thoroughly studied and addressed in the chapter by Drs. Zhu and Zhang.

With data from related individuals, genetic association tests may be conducted in a manner that is valid (i.e., not subject to bias due to sample heterogeneity) even without utilizing genetic markers to infer genetic background. The basic principle is to detect whether there is a departure from random marker segregation at a candidate

locus. For example, if a study population consists of affected children and their parents and a marker with two alleles A and a is studied for its potential involvement in the disease. If the marker has nothing to do with disease phenotype, we expect that a parent who is heterozygous Aa would have equal chance to transmit allele A or a to his/her affected offspring. On the other hand, if allele A increases disease risk, we would expect to observe a preferential transmission of allele A to the affected offspring. This testing procedure is robust to sample heterogeneity as the inference is conditional on each parent's genotype and the only genetic principle tested is random marker allele transmission from parents to offspring, the Mendel's first law. Many statistical developments along this research route are discussed in the chapter by Drs. Zhang and Zhao.

Both population-based and family-based association studies examine statistical associations between a phenotype and the genotypes at a marker. One implicit assumption is that the same marker genotype would exert the same or similar effects on a phenotype. While this is expected to be the case for most genetic markers that have direct functional impact, this assumption may well be violated for many markers. For example, consider a marker with two alleles A and a studied is not functional but rather is in linkage disequilibrium with a truly functional one with two alleles D and d. It is possible that A is positively associated with D in one population, that is, someone carrying A on one chromosome is also more likely to carry D on the same chromosome, but A is negatively associated with D in another population. In this case, an analysis using samples from these two populations together may not even be able to detect a genetic association. More importantly, when the markers are sparse and not expected to provide a good coverage of the genome, the association analysis paradigm discussed above will not be effective as a large proportion of the genome that likely harbors disease genes may be missed due to poor coverage. This was in fact the case only a few years ago when only fewer markers could be used for genetic analysis. In this scenario, although the markers were not dense enough to cover the genome for association analysis, they were more than adequate to allow geneticists to infer whether two relatives in an ascertained pedigree share a segment in the genome from the same ancestor. For example, if two siblings have the same marker genotypes across a set of closely linked markers on the same chromosome, then they likely have inherited the same genetic materials from both their parents. A genetic linkage analysis is to statistically assess whether there is a cosegregation of genetic materials within a candidate region and the phenotype within a family. For example, this can be done by studying whether there is a correlation between trait similarities and inheritance similarities at a candidate region among a set of individuals from the same family. Consider a study enrolling affected sib pairs. If majority of them share the same genetic materials from their parents in a region, then this region is likely involved in disease etiology. Note that in contrast to association analysis that is performed across all study subjects, linkage analysis is conducted within families and evidence is then summed over across individual families. Statistical methods for linkage analysis can be conducted for either qualitative traits (the chapter by Dr. Li and Abecasis) or quantitative traits (the chapter by Drs. Amos, Peng, Xu, and Ma).

Exact inference of inheritance patterns within a pedigree is tractable either for a small pedigree or for a few markers, but such inference becomes computationally prohibitive for large pedigree with many genetic markers. In this case, the exact probabilities may be estimated by Monte Carlo simulations. In the chapter by Drs. Igo, Luo, and Lin, the principles and implementations underlying the simulation methods for linkage analysis in large complex pedigrees are discussed.

One central topic in statistical inference is the control of false positive results so as to minimize any consequences resulting from false leads. This issue has been well addressed when only one or a small number of statistical hypotheses are tested. However, hundreds of thousands of markers are tested for their associations with disease in a genome-wide association study, and false positive control at the individual marker levels will not be adequate. For example, if a study considers 500,000 markers and the statistical significance level is set at 0.01, we would expect to see 5,000 false positive results even when there is no association between disease status and any of the markers. Similar issue exists in the linkage analysis context, although not to the same great extent as association analysis. The chapter by Drs. Zhang and Ott presents some recent developments on appropriately controlling overall false positive results in genetic studies at the genome level.

The identifications of disease genes can lead to biological insights on pathways involved in disease etiology, and these findings can also be used to predict an individual's disease risk. In the chapter by Drs. Gail and Chatterjee, they discuss statistical methods that can be used to make use of findings from genetic studies to identify individuals at higher risks for disease.

The book concludes with the last chapter by Drs. Molony, Sieberts, and Schadt, where they discuss integrating genetics and genomics data to better delineate biological pathways underlying complex traits. In addition to disease status and possibly other clinical outcomes, they consider gene expression data that can now be routinely gathered to measure the expression levels of tens of thousands of genes simultaneously for each study subject. These gene expression data add another whole new dimension of statistical analysis and are very information rich. In principle, the expression level of each gene can be thought as a quantitative trait, and linkage/association analysis can be conducted to identity genes regulating a gene's expression level. Therefore, based on this perspective, we would be in a position to conduct genetic analysis for tens of thousands of traits. Some of these expression levels may be associated with disease outcome, and so it is natural to investigate how a genetic variation affects the expression levels as well as disease outcomes. Many biological questions on the underlying genetic networks relating genetic variations, expression variations, and phenotype variations can be posed and answered with these data. This chapter discusses topics falling into the domain of systems biology where the whole biological system is the focus of a study and genome-level data of different types are needed to dissect the networks.

We hope that this book will provide an overview of the most important areas in genetic data analysis methods. We focus on fundamental principles and, when possible, demonstrate these principles with real data examples. Despite our efforts, this is not an encyclopedia of statistical methods in human genetics, and some topics

are not included such as the experimental design of a genetic study, data preprocessing from high-throughput genotyping platforms, and copy number variations. Most importantly, this is a very rapidly developing field and new technologies are constantly introduced that demand novel statistical approaches to make the most use of the data collected. For example, the statistical methods discussed in this book may not be the most effective for inferring inheritance patterns in a pedigree using high density SNP data. On the other hand, the availabilities of re-sequencing data from a large number of study subjects lead to a new set of informatics and statistical challenges, such as the incorporation of SNP annotation information and the dealing of rare genetic variations. We hope the basic principles and statistical methods discussed in this book will motivate the readers to develop their own approaches if necessary to accelerate our progresses in mapping disease genes.

Contents

Population Genetics 1
Bruce Weir

Haplotype Structure 25
Yu Zhang and Tianhua Niu

Linkage Analysis of Qualitative Traits 81
Mingyao Li and Gonçalo R. Abecasis

Linkage Analysis of Quantitative Traits 119
Christopher I. Amos, Bo Peng, Yaji Xu, and Jianzhong Ma

Markov Chain Monte Carlo Linkage Analysis Methods 147
Robert P. Igo and Yuqun Luo, Shili Lin

Population-Based Association Studies 171
Xiaofeng Zhu and ShuangLin Zhang

Family-Based Association Studies 191
Kui Zhang and Hongyu Zhao

Haplotype Association Analysis 241
Michael P. Epstein and Lydia C. Kwee

Multiple Comparisons/Testing Issues 277
Qingrun Zhang and Jurg Ott

Estimating the Absolute Risk of Disease Associated with Identified Mutations 289
Mitchell H. Gail and Nilanjan Chatterjee

Processing Large-Scale, High-Dimension Genetic and Gene Expression Data 307
Cliona Molony, Solveig K. Sieberts, and Eric E. Schadt

Index 331

Contributors

Gonçalo R. Abecasis Department of Biostatistics, Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI, USA, goncalo@umich.edu

Christopher I. Amos Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, 1155 Pressler Blvd, Unit 1340, Houston, TX, 77030, USA, camos@mdanderson.org

Nilanjan Chatterjee Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Rockville, MD 20852, USA

Michael P. Epstein Department of Human Genetics, Emory University School of Medicine, 615 Michael Street, Suite 301, Atlanta, GA 30322, USA, mepstein@genetics.emory.edu

Mitchell H. Gail Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Rockville, MD 20852, USA

Robert P. Igo Department of Epidemiology and Biostatistics, Division of Genetics and Molecular Epidemiology, Case Western Reserve University, Cleveland, OH, USA, rigo@darwin.EPBI.CWRU.edu

Lydia C. Kwee Department of Biostatistics, Emory University, Atlanta, GA, USA

Mingyao Li Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA, USA, mingyao@mail.med.upenn.edu

Shili Lin Department of Statistics, The Ohio State University, OH, USA, shili@stat.osu.edu

Yuqun Luo Department of Epidemiology and Biostatistics, Division of Genetics and Molecular Epidemiology, Case Western Reserve University, Cleveland, OH, USA, yuqun.luo@case.edu

Jianzhong Ma Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, 1155 Pressler Blvd, Unit 1340, Houston, TX, 77030, USA, jzma@mdanderson.org

Cliona Molony Rosetta Inpharmatics, LLC, (a wholly owned subsidiary of) Merck & Co., Inc., Seattle, WA 98109, USA

Tianhua Niu Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02215, USA

and Program in Molecular and Genetic Epidemiology, Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA, tniu@hsph.harvard.edu

Jurg Ott Beijing Institute of Genomics, Chinese Academy of Sciences, No. 7 Bei Tu Cheng West Road, Beijing 100029, China, ottjurg@yahoo.com

Bo Peng Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, 1155 Pressler Blvd, Unit 1340, Houston, TX, 77030, USA, bpeng@mdanderson.org

Eric E. Schadt Rosetta Inpharmatics, LLC, a wholly owned subsidiary of Merck & Co., Inc., Seattle, WA 98109, USA

Solveig K. Sieberts Rosetta Inpharmatics, LLC, (a wholly owned subsidiary of) Merck & Co., Inc., Seattle, WA 98109, USA

Bruce Weir Department of Biostatistics, University of Washington, Seattle, WA 98185.

Yaji Xu Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, 1155 Pressler Blvd, Unit 1340, Houston, TX, 77030, USA, yajixu@mdanderson.org

Yu Zhang Department of Statistics, the Pennsylvania State University, 422A Thomas Building, University Park, PA 16802, USA, yuzhang@stat.psu.edu

ShuangLin Zhang Department of Mathematical Science, Michigan Technological University, Houghton, MI, USA

Kui Zhang Section on Statistical Genetics, Department of Biostatistics University of Alabama at Birmingham, Birmingham, AL 35294, USA

Qingrun Zhang Chinese Academy of Sciences, Beijing Institute of Genomics, Beijing, China

Hongyu Zhao Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT, 06520, USA

Xiaofeng Zhu Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA