Use R!

Andrea S. Foulkes

Applied Statistical Genetics with R

For Population-based Association Studies

Springer

Use R!

Series Editors:
Robert Gentleman Kurt Hornik Giovanni Parmigiani

For other titles published in this series, go to
http://www.springer.com/series/6991

Andrea S. Foulkes

Applied Statistical Genetics with R

For Population-based Association Studies

Springer

Andrea S. Foulkes
University of Massachusetts
School of Public Health & Health Sciences
404 Arnold House
715 N. Pleasant Street
Amherst, MA 01003
USA
foulkes@schoolph.umass.edu

ISBN 978-0-387-89553-6 e-ISBN 978-0-387-89554-3
DOI 10.1007/978-0-387-89554-3
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: PCN applied for

To Rich, Sophie and Ella

Preface

This book is intended to provide fundamental statistical concepts and tools relevant to the analysis of genetic data arising from population-based association studies. Elementary knowledge of statistical methods at the level of a first course in biostatistics is assumed. Chapters 1–3 provide a general overview of the genetic and epidemiological considerations relevant to this setting. Topics covered include: (1) types of investigations, typical data components and features in genetic association studies, and basic genetic vocabulary (Chapter 1); (2) epidemiological principles relevant to population-based studies, including confounding and effect modification (Chapter 2); (3) elementary statistical methods for estimating and testing association (Chapter 2); (4) the overarching analytical challenges inherent in these investigations (Chapter 2); (5) basic genetic concepts, including linkage disequilibrium, Hardy-Weinberg equilibrium, and haplotypic phase (Chapter 3); and (6) quality control methods for assessing genotyping errors and population substructure (Chapter 3).

The remaining chapters are organized as follows. Chapters 4 and 5 deal primarily with methods that aim to identify single genetic polymorphisms or single genes that contribute individually to measures of disease progression or disease status. This includes testing concepts and methods for appropriately adjusting for multiple comparisons (Chapter 4) and approaches to the analysis of unobservable haplotypic phase (Chapter 5). Chapters 6 and 7 focus on methods for variable subset selection and particularly methods that simultaneously evaluate a large number of variables to arrive at the best predictive model for the complex disease trait under investigation. Notably, while all of these methods consider multiple polymorphisms concomitantly, some focus on conditional effects of these genetic variables, while other methods are specifically designed for identifying and testing potential interaction among genetic polymorphisms in their effects on disease phenotypes. This section covers classification and regression trees (Chapter 6), extensions of the tree framework—namely random forests, logic regression and multivariable adaptive regression splines—and a brief introduction to Bayesian variable selection (Chapter 7).

The field of statistical genomics includes a large array of methods for a wide variety of medical and public health applications. While the methods described herein are broadly relevant, this text does not directly address issues specific to family-based studies, evolutionary (population genetic) modeling, and gene expression analysis. This text also does not attempt to provide a comprehensive summary of existing methods in the rapidly expanding field of statistical genomics. Rather, fundamental concepts are presented at the level of an introductory graduate-level course in biostatistics, with the aim of offering students a foundation and framework for understanding more complex methods. Two application areas are considered throughout this text: (1) human genetic investigations in population-based association studies of unrelated individuals and (2) studies aiming to characterize associations between Human Immunodeficiency Virus (HIV) genotypes and phenotypes, as measured by *in vitro* drug responsiveness. Several publicly available datasets are used for illustration and can be downloaded at the book website (http://people.umass.edu/foulkes/asg.html). While data simulations are not described, emphasis is placed on understanding the implicit modeling assumption generally required for testing. An overarching theme of this text is that the application of any statistical method aims to characterize a *specific* relationship among variables. For example, just as an additive model of association can be used to evaluate additive structure, a classification or regression tree aims to characterize conditional associations. The array of methods that are applied to data arising from genetic association studies differ primarily in the types of associations that they are designed to uncover.

This text is also intended to complement the existing literature on statistical genetics and molecular epidemiology in two ways. First, this text offers extensive and integrated examples using R, an open-source, publicly available statistical computing software environment. This is intended both as a pedagogical tool for providing readers with a deeper understanding of the statistical algorithms presented and as a practical tool for applying the approaches described herein. Second, this text provides comprehensive coverage of both genetic concepts, such as linkage disequilibrium and Hardy-Weinberg equilibrium, from a statistical perspective, as well as fundamental statistical concepts, such as adjusting for multiplicity and methods for high-dimensional data analysis, relevant to the analysis of data arising from genetic association studies. Several excellent texts, including Thomas (2004) and Ziegler and Koenig (2007), provide in-depth coverage of genetic data concepts relevant to both population-based and family-based investigations. The present text presents these concepts within the context of familiar statistical nomenclature while providing coverage of several additional pertinent epidemiological concepts and statistical methods for characterizing association. This presentation is at a level that is accessible to the reader with a limited background in biostatistics and with an interest in public health or biomedical research. More advanced discussions of the underlying theory can be found in alterna-

tive texts such as Hastie *et al.* (2001) and Lange (2002), as well as the original manuscripts cited throughout this text.

The primary focus of this text is on candidate gene studies that involve the investigation of polymorphisms at several genetic sites within and across one or more genes and their associations with a trait. In the past several years, technological advancements leading to development and widespread availability of "SNP chips" have led to an explosion of genome-wide association studies (GWAS) involving 500 thousand to 1 million single-nucleotide polymorphisms (SNPs). The methods presented in this text apply equally to candidate gene approaches and whole and partial GWAS. Notably, however, the latter setting requires additional consideration of the computational burden of associated analysis as well as data preprocessing and error checking, as discussed in Section 3.3 and throughout this text. While GWAS have gained a great deal of popularity in recent years, they do not obviate the need for candidate gene studies that further investigate the role of specific genes in disease progression as well as the potential confounding or modifying roles of traditional risk factors, including both clinical and demographic characteristics. Instead, GWAS provide investigators with a vastly improved body of scientific knowledge to inform the selection of candidate genes for hypothesis-driven research.

The term high-dimensional has taken on many meanings across different fields of research and over the past decade of rapid expansion in these fields. In this text, high-dimensional is defined simply as a large number of potentially correlated variables that may interact, in a statistical or a biological sense, in their association with the outcome under investigation. The term is used loosely to refer to any number of variables for which there is a complex, uncharacterized structure and the usual least squares regression setting may not be easily applicable. High-dimensional data methods including approaches to multiplicity and characterizing gene–gene and gene–environment interactions are addressed within the context of characterizing associations among genetic sequence data and disease traits. In these settings, the predictor variables are SNPs or corresponding amino acids and are categorical. Primary consideration is given to dependent variables that are either continuous measures of disease progression or binary indicators of disease status, though brief mention is also made of methods for multivariate and survival outcomes. Specific attention is given to the potential confounding and mediating roles of individual-level clinical and demographic data.

Implementation of all described methods is demonstrated using the R environment and associated packages, which are publicly available at the Comprehensive R Archive Network (CRAN) website (http://cran.r-project.org/). The decision to use R in this text over alternative programming languages is multifaceted. First, as a publicly available package, R is freely accessible to all readers and, importantly, students will continue to have access to R at all future personal and professional venues. As an open-source language, R also provides students with the opportunity to view code used to generate functions, serving as a valuable pedagogical tool for more programmatically

minded learners. Another key advantage of R is that investigators who develop new statistical methodology often provide an accompanying R package for implementation through the CRAN website, providing users with almost immediate access to implementation of the most recently developed approaches. Finally, with the availability of contributed packages, the choice of method to apply rests with the user rather than with what a core development team of the programming language chooses to release.

While strongly preferable for the reasons mentioned above, use of R in this text does have the drawback from a pedagogical perspective that both the versions and packages are updated frequently. That is, we see a clear trade-off between accessibility and stability. In the process of writing this text, several changes in the packages described herein occurred, resulting in inconsistent outputs. While these inconsistencies have been resolved as of the present date, several more are likely to arise over the next several years. The reader is encouraged to visit the textbook website for information on these changes. All of the programming scripts in this text were written and tested for R version 2.7.1. Ascii text files with complete R code used for the examples in this textbook can be found on the textbook website. The files can be downloaded, or read directly into R using the `source()` function. For example, to source the code from Example 1.1, we can write the following at the R prompt:

```
> source("http://people.umass.edu/foulkes/asg/examples/1.1.r")
```

Additionally specifying `print.eval=T` in this function call will print the corresponding output. While the programs presented within this text are comprehensive, the novice reader can begin with the appendix for a brief introduction to some fundamental concepts relevant to programming in R. Several, more comprehensive, introductions to R are available, and the reader is encouraged to reference these texts as well, including Gentleman (2008), Spector (2008) and Dalgaard (2002), for additional programming tools and background.

I am grateful for the advice and support I have received in writing this text from many colleagues, students, friends and family members. I would especially like to thank my students and postdoctoral fellows, M. Eliot, X. Li, Y. Liu, Dr. B.A. Nonyane and Dr. K. Au, who spent many hours checking for notational and programming consistency as well as sharing in helpful discussions. I am indebted to all of the students in the fall 2008 semester of public health 690T at the University of Massachusetts, Amherst for their helpful suggestions and for bearing with me in the first run of this text. I am grateful for having a long-term friend and colleague in Dr. R. Balasubramanian, whose support and encouragement were pivotal in my decision to write this text. I am also thankful for the many conversations with Dr. D. Cheng and her willingness to share her extensive knowledge in applied statistics. I am obliged to Dr. M.P. Reilly for an enduring collaboration that has fueled my interest and enhanced my knowledge in applied statistical genetics for medical research. I am grateful to Dr. A.V. Custer, whose dedication to the

open-source software community was inspirational to me. Dr. V. De Gruttola's early mentorship continues to shape my research interests, and I am thankful for the passion and deep thinking he brings to our profession. I also value the strong encouragement and intellectual engagement of my early career mentors Dr. E. George and Dr. T. Ten Have. The efforts of Dr. E. Hoffman, Dr. H. Gorski and colleagues in providing the FAMuSS and HGDP data were extraordinary, and their commitment to public access to data resources is truly outstanding. I am also indebted to Dr. R. Shafer and colleagues for their remarkable effort in creating and maintaining the Stanford University HIV Drug Resistance Database, from which the Virco data were downloaded and several additional data sets can be accessed easily. I also greatly appreciate the insightful leadership of the R core development team and the individuals who wrote and maintain the R packages used throughout this text. All figures in this text were generated in R or created using the open-source graphics editor Inkscape (http://www.inkscape.org/). I value the many insightful comments and suggestions of the editors and anonymous reviewers. Support for this text was provided in part by a National Institute of Allergies and Infectious Disease (NIAID) individual research award (R01AI056983). Finally, thanks to my family for their tremendous love and support.

Andrea S. Foulkes
Amherst, MA
May 2009

Contents

List of Tables

List of Figures

Acronyms

AA: Amino acid

AIDS: Acquired immunodeficiency syndrome

ANOVA: Analysis of variance

BART: Bayesian additive regression tree

BSS: Between-group sum of squares

B-H: Benjamini and Hochberg (approach to multiple testing)

B-Y: Benjamini and Yekutieli (approach to multiple testing)

BMI: Body mass index

BVS: Bayesian variable selection

CART: Classification and regression trees

CV: Cross-validation

DNA: Deoxyribonucleic acid

EM: Expectation-maximization

FAMuSS: Functional SNPS Associated with Muscle Size and Strength

FDR: False discovery rate

FSDR: Free step-down resampling

FWEC: Family-wise error under the complete null

FWEP: Family-wise error under a partial null

FWER: Family-wise error rate

GLM: Generalized linear model

GWAS: Genome-wide association study

GWS: Genome-wide scan

HGDP: Human Genome Diversity Project

HTR: Haplotype trend regression

HWD: Hardy-Weinberg disequilibrium

HWE: Hardy-Weinberg equilibrium

IBD: Identical by descent

IBS: Identical by state

IDV: Indinavir

LD: Linkage disequilibrium

LOH: Loss of heterozygosity

LS: Learning sample

MARS: Multivariate adaptive regression splines

MCMC: Markov chain Monte Carlo

MDS: Multidimensional scaling

MI: Multiple imputation

MIRF: Multiple imputation and random forests

MSE: Mean square error

NFV: Nelfinavir

OOB: Out-of-bag

PCA: Principal components analysis

pFDR: Positive false-discovery rate

Pr: Protease

PRD: Positively regression dependent

QTL: Quantitative trait loci

RF: Random forest

RNA: Ribonucleic acid

RT: Reverse transcriptase

SAM: Significance analysis of microarrays

SNP: Single-nucleotide polymorphism

STP: Simultaneous test procedure

WSS: Within-group sum of squares

1

Genetic Association Studies

Recent technological advancements allowing for large-scale sequencing efforts present an exciting opportunity to uncover the genetic underpinnings of complex diseases. In an attempt to characterize these genetic contributors to disease, investigators have embarked in multitude on what are commonly referred to as *population-based genetic association studies*. These studies generally aim to relate genetic sequence information derived from unrelated individuals to a measure of disease progression or disease status. The field of genomics spans a wide array of research areas that involve the many stages of processing from genetic sequence information to protein products and ultimately the expression of a trait. The breadth of genomic investigations also includes studies of multiple organisms, ranging from bacteria to viruses to parasites to humans. In this chapter, two settings are described in which population-based genetic association studies have marked potential for uncovering disease etiology while elucidating new approaches for targeted, individualized therapeutic interventions: (1) complex disease association studies in humans; and (2) studies involving the Human Immunodeficiency Virus (HIV).

In both settings, interest lies in characterizing associations between multiple genetic polymorphisms and a measured trait. In addition, these settings share the essential need to account appropriately for patient-level covariates as potential confounders or modifiers of disease progression to make clinically meaningful conclusions. While these two settings are not comprehensive, together they provide a launching point for discussion of quantitative methods that address the challenges inherent in many genetic investigations. This chapter begins by describing types of population-based studies, which represent one class of investigations within the larger field of genomics research. Also discussed are the fundamental features of data arising from these investigations as well as the analytical challenges inherent in this endeavor.

A.S. Foulkes, *Applied Statistical Genetics with R: For Population-based Association Studies*, Use R, DOI: 10.1007/978-0-387-89554-3_1,