Statistics for Biology and Health

Daniel O. Stram

Design, Analysis, and Interpretation of Genome-Wide Association Scans

Springer

Statistics for Biology and Health

Daniel O. Stram

Design, Analysis, and Interpretation of Genome-Wide Association Scans

Springer

Statistics for Biology and Health

Series Editors:
M. Gail
K. Krickeberg
J. Samet
A. Tsiatis
W. Wong

For further volumes:
http://www.springer.com/series/2848

Daniel O. Stram

Design, Analysis, and Interpretation of Genome-Wide Association Scans

Springer

Daniel O. Stram
Department of Preventive Medicine
University of Southern California
Keck School of Medicine
Los Angeles, CA, USA

ISSN 1431-8776
ISBN 978-1-4614-9442-3 ISBN 978-1-4614-9443-0 (eBook)
DOI 10.1007/978-1-4614-9443-0
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013954411

© Springer Science+Business Media New York 2014
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

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

To Pavlova, Alex, and Douglas

Acknowledgments

Acknowledgments are due to Malcolm Pike, Larry Kolonel, and Brian Henderson for their vision in initiating the Multiethnic Cohort Study in Hawaii and Los Angeles. My work on genetic association studies began with a 6-month sabbatical spent at the Whitehead Institute/MIT Center for Genome Research in 2002 which was itself an offshoot of the early collaborations developed between the MEC and the Brave New World of twenty-first-century human genetics. Thanks to David Altshuler, Mark Daly, David Reich, Nick Patterson, Noel Burtt, Itsik Pe'er, and the many others from whom I learned enormously at that time. Additional sabbatical support in 2010 initiated the development of the material presented here into its current form. Special thanks to valued colleagues including Chris Haiman, Loic Le Marchand, Iona Cheng, Leigh Pearce, Peter Kraft and Duncan Thomas, as well as many students, with whom it has been a pleasure to work with during hugely exciting times.

Jianqi Zhang helped assess the text and homework problems from a student's point of view. Diane Lemasters assisted me with graphics and charts.

Contents

1 Introduction 1
1.1 Historical Perspective 1
1.2 DNA Basics 3
1.2.1 Organization of Chromosomes 3
1.2.2 Organization of DNA 4
1.2.3 DNA and Protein 5
1.3 Types of Genetic Variation 6
1.3.1 Single-Nucleotide Variants and Polymorphisms 6
1.3.2 Insertions/Deletions 11
1.3.3 Larger Structural Variants 11
1.3.4 Exonic Variation and Disease 12
1.3.5 Non-exonic SNPs and Disease 12
1.3.6 SNP Haplotypes 13
1.3.7 Microsatellites 13
1.3.8 Mitochondrial Variation 14
1.4 Overview of Genotyping Methods 14
1.4.1 SNP Calling 16
1.5 Overview of GWAS Genotype Arrays 18
1.6 Software and Data Resources 18
1.7 Web Resources 19
1.7.1 Basic Genomics 19
1.7.2 GWAS Associations 20
1.7.3 Annotation 20
1.8 Hardware and Operating Systems 22
1.9 Data Example 23
1.9.1 Save Your Work 24
References 27

2 Topics in Quantitative Genetics 31
2.1 Distribution of a Single Diallelic Variant in a Randomly Mixing Population 32

2.1.1 Hardy–Weinberg Equilibrium 32
2.1.2 Random Samples of Unrelated Individuals 33
2.1.3 Joint Distribution Between Relatives of Allele Counts for a Single SNP 33
2.1.4 Coefficients of Kinship and of Inbreeding 36
2.2 Relationship Between Identity by State and Identity by Descent for a Single Diallelic Marker 37
2.3 Estimating IBD Probabilities from Genotype Data 38
2.4 The Covariance Matrix for a Single Allele in Nonrandomly Mixing Populations 39
2.4.1 Hidden Structure and Correlation 39
2.4.2 Effects of Incomplete Admixture on the Covariance Matrix of a Single Variant 42
2.5 Direct Estimation of Differentiation Parameter *F* from Genotype Data 42
2.5.1 Relatedness Revisited 43
2.5.2 Estimation of Allele Frequencies 45
2.6 Allele Frequency Distributions 45
2.6.1 Initial Mutations and Common Ancestors 45
2.6.2 Mutations and the Coalescent 48
2.6.3 Allelic Distribution of Genetic Variants 48
2.6.4 Allele Distributions Under Population Increase and Selection 51
2.7 Recombination and Linkage Disequilibrium 52
2.7.1 Quantification of Recombination 54
2.7.2 Phased Versus Unphased Data and LD Estimation 56
2.7.3 Hidden Population Structure 57
2.7.4 Pseudo-LD Induced by Hidden Structure and Relatedness 60
2.8 Covering the Genome for Common Alleles 62
2.8.1 High-Throughput Sequencing 65
2.9 Principal Components Analysis 65
2.9.1 Display of Principal Components for the HapMap Phase 3 Samples 67
2.10 Chapter Summary 72
Data and Software Exercises 74
References 75

3 An Introduction to Association Analysis 79
3.1 Single Marker Associations 79
3.1.1 Dominant, Recessive, and Co-dominant Effects 80
3.2 Regression Analysis and Generalized Linear Models in Genetic Analysis 82
3.3 Tests of Hypotheses for Genotype Data Using Generalized Linear Models 84

3.3.1 Test of Hypothesis regarding Genotype Effects Testing Using Logistic Regression in Case–Control Analysis 85
3.3.2 Interpreting Regression Equation Coefficients 88
3.4 Summary of Maximum Likelihood Estimation, Wald Tests, Likelihood Ratio Tests, Score Tests, and Sufficient Statistics 90
3.4.1 Properties of Log Likelihood Functions 91
3.4.2 Score Tests 93
3.4.3 Likelihood Ratio Tests 94
3.4.4 Wald Tests 95
3.4.5 Fisher's Scoring Procedure for Finding the MLE 95
3.4.6 Scores and Information for Normal and Binary Regression 96
3.4.7 Score Tests of $\beta = 0$ for Linear and Logistic Models 98
3.4.8 Matrix Formulae for Estimators in OLS Regression 100
3.5 Covariates, Interactions, and Confounding 102
3.6 Conditional Logistic Regression 104
3.6.1 Breaking the Matching in Logistic Regression of Matched Data 105
3.6.2 Parent Affected-Offspring Design 109
3.7 Case-Only Analyses 111
3.7.1 Case-Only Analyses of Disease Subtype 111
3.7.2 Case-Only Analysis of Gene × Environment and Gene × Gene Interactions 111
3.8 Non-independent Phenotypes 114
3.8.1 OLS Estimation When Phenotypes Are Correlated 114
3.9 Needs of a GWAS Analysis 115
3.9.1 Hardware Requirements for GWAS 116
3.9.2 Software Solutions 116
3.10 The Multiple Comparisons Problem 118
3.11 Behavior of the Bonferroni Correction with Non-Independent Tests 119
3.12 Reliability of Small p-Values 121
3.12.1 Test of a Single Binomial Proportion 122
3.12.2 Test of a Difference in Binomial Proportions 124
3.13 Chapter Summary 125
Appendix 130
References 131

4 Correcting for Hidden Population Structure in Single Marker Association Testing and Estimation . . . 135
4.1 Effects of Hidden Population Structure on the Behavior of Statistical Tests for Association . . . 136
4.1.1 Effects on Inference Induced by Correlated Phenotypes . . . 136
4.1.2 Influences of Latent Variables . . . 140
4.1.3 Hidden Structure as a Latent Variable . . . 140
4.1.4 Polygenes, Latent Structure, Hidden Relatedness, and Confounding . . . 141
4.1.5 Hidden Non-mixing Strata . . . 142
4.1.6 Admixture . . . 147
4.1.7 Polygenes and Cryptic Relatedness . . . 150
4.2 Correcting for the Effects of Hidden Structure and Relatedness . . . 154
4.2.1 Genomic Control . . . 154
4.2.2 Regression-based Adjustment for Leading Principal Components . . . 156
4.2.3 Implementation of Principal Components Adjustment Methods . . . 157
4.2.4 Random Effects Models . . . 162
4.2.5 Retrospective Methods . . . 166
4.3 Comparison of Correction Methods by Simulation . . . 170
4.3.1 Comparison of the Mixed Model and Retrospective Approach for Binary (case–control) Outcomes . . . 171
4.3.2 Conclusions . . . 172
4.4 Behavior of the Genomic Control Parameter as Sample Size increases . . . 173
4.5 Removing Related Individuals as Part of Quality Control, Is It Needed? . . . 176
4.6 Chapter Summary . . . 177
Data and Software Exercises . . . 179
References . . . 179

5 Haplotype Imputation for Association Analysis . . . 183
5.1 The Role of Haplotypes in Association Testing . . . 184
5.2 Haplotypes, LD Blocks, and Haplotype Uncertainty . . . 185
5.3 Haplotype Frequency Estimation and Imputation . . . 185
5.3.1 Small Numbers of SNPs . . . 185
5.3.2 Haplotype Uncertainty . . . 186
5.4 Haplotype Frequency Estimation for Larger Numbers of SNPs . . . 190
5.4.1 Partition-Ligation EM Algorithm . . . 190
5.4.2 Phasing Large Numbers of SNPs . . . 190

5.5 Regression Analysis Using Haplotypes as Explanatory Variables 191
5.5.1 Expectation Substitution 191
5.5.2 Fitting Dominant, Recessive, or Two Degrees of Freedom Models for the Effect of Haplotypes 193
5.6 Dealing with Uncertainty in Haplotype Estimation in Association Testing 195
5.6.1 Full Likelihood Estimation of Risk Parameters and Haplotype Frequencies 195
5.6.2 Ascertainment in Case–Control Studies 197
5.6.3 Example: Expectation-Substitution Method 198
5.7 Haplotype Analysis Genome-Wide 201
5.7.1 Studies of Homogeneous Non-admixed Populations 201
5.7.2 The Four-Gamete Rule for Fast Block Definition 202
5.7.3 Multiple Comparisons in Haplotype Analysis 204
5.8 Multiple Populations 205
5.9 Chapter Summary 207
References 209

6 SNP Imputation for Association Studies 213
6.1 The Role of Imputed SNPs in Association Testing 214
6.2 EM Algorithm and SNP Imputation 214
6.3 Phasing Large Numbers of SNPs for the Reference Panel 217
6.4 Brief Introduction to Hidden Markov Models 217
6.4.1 The Baum–Welch Algorithm 220
6.5 Large-Scale Imputation Using HMMs 222
6.6 Using an HMM to Impute Missing Genotype Data when Both the Reference Panel and Study Genotypes Are Phased 223
6.7 Using an HMM to Phase Reference or Main Study Genotypes 229
6.7.1 Initializing and Updating the Current List of Haplotypes 232
6.8 Practical Issues in Large-Scale SNP Imputation 235
6.8.1 Assessing Imputation Accuracy 236
6.8.2 Imputing Rare SNPs 237
6.8.3 Use of Cosmopolitan Reference Panels 238
6.9 Estimating Relative Risks for Imputed SNPs 239
6.9.1 Expectation Substitution 239
6.10 Chapter Summary 239
6.10.1 Links 241
References 241