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

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
7 Design of Large-Scale Genetic Association Studies, Sample Size, and Power ................................. 243
  7.1 Design Considerations ..................................... 243
  7.2 Sample Size and Power for Studies of Unrelated Subjects .... 244
    7.2.1 Power for Chi-Square Tests ............................ 244
    7.2.2 Calculation of Non-centrality Parameters for Chi-Square Tests in Generalized Linear Models .............................................. 248
  7.3 QUANTO .................................................. 255
    7.3.1 Use of QUANTO to Compute Power to Detect Main Effects of Genetic Variants in Case–Control Studies .................. 258
  7.4 Alternative Designs ........................................... 258
    7.4.1 Sibling Controls ........................................... 260
    7.4.2 Power for Interactions .................................... 261
    7.4.3 Parent-Affected-Offspring Trios .......................... 262
    7.4.4 Power for Case-Only Analysis of Interactions .............. 262
  7.5 Control for Multiple Comparisons ............................ 264
    7.5.1 Single Marker Associations ............................... 264
    7.5.2 More Complex Marker Associations ........................ 264
    7.5.3 Reliability of Very Small p-Values ....................... 266
  7.6 Two-Staged Genotyping Designs ............................... 266
    7.6.1 Measured SNP Association Tests .......................... 267
    7.6.2 Optimal Two-Stage Case–Control Designs .................. 270
  7.7 Control of Population Stratification: Effects on Study Power .... 272
    7.7.1 Genomic Inflation and Study Power ........................ 273
    7.7.2 Correction for Admixture and Hidden Structure by Principal Components .............................................. 274
    7.7.3 A Retrospective Analysis of Study Power .................. 275
  7.8 Power of Multi-SNP Conditional Tests ........................ 280
  7.9 Chapter Summary ............................................... 280
References .................................................................. 283

8 Post-GWAS Analyses .................................................. 285
  8.1 Meta-analysis ..................................................... 285
  8.2 Meta-analysis of Linear or Logistic Regression Estimates ....... 287
    8.2.1 Random Effects Models .................................... 290
  8.3 Meta-analysis for the GWAS Setting ............................. 291
    8.3.1 Meta-analysis and Imputation ............................... 292
  8.4 Efficiency of Meta-analysis Versus Pooled Analysis .............. 293
  8.5 Sources of Heterogeneity in Meta-analysis of GWAS Data ........ 295
    8.5.1 LD Differences .............................................. 295
    8.5.2 Exposure Differences ....................................... 296
    8.5.3 Gene × Gene Interactions .................................... 296
8.6 Meta-analysis Based on Effect sizes, Z-scores, and \( P \)-Values ........................................... 297
   8.6.1 Z-Score Analysis .................................. 297
   8.6.2 Fisher’s Method of Combining \( P \)-Values .......... 298
   8.6.3 Meta-analysis of Score Tests ........................ 299
8.7 Multiethnic Analyses .................................... 299
8.8 Fine Mapping of Single-SNP Associations:
   Conditional Analyses .......................................... 301
8.9 Fine Mapping in Admixed Populations .................... 303
   8.9.1 The Role of Local Ancestry Adjustment ............. 304
8.10 Polygenes and Heritability ................................ 305
   8.10.1 Fraction of Familial Risk Explained
          by a Polygene Under a Multiplicative Model ........... 305
   8.10.2 Synergies Between Polygenes
          and Environmental Variables .......................... 306
8.11 GWAS Heritability Analysis ............................... 306
   8.11.1 Heritability Estimation in the Presence
          of Population Stratification ........................... 308
8.12 Analysis of Rare Variants .................................. 309
8.13 Contribution of Rare SNPs to Phenotypic Variance
       and Heritability Under a Polygenic (Additive) Model . 310
8.14 Moderately Rare Single-SNP Analysis .................... 312
8.15 Burden and Pathway Analysis ............................. 314
   8.15.1 A Weighted Sum Statistic ............................ 315
   8.15.2 Omnibus Tests, Variance Components,
          and Kernel Machines .................................. 316
8.16 Final Remarks ........................................... 319
8.17 Chapter Summary ......................................... 321
References .................................................. 324

Index .......................................................... 329
Design, Analysis, and Interpretation of Genome-Wide Association Scans
Stram, D.O.
2014, XV, 334 p. 39 illus., Hardcover
ISBN: 978-1-4614-9442-3