

Contents

Preface

xiii

1. Understanding Skin Color Variations as an Adaptation by Detecting Gene–Environment Interactions	1
<i>Sumiko Anno, Kazuhiko Ohshima, and Takashi Abe</i>	
1.1 Introduction	2
1.1.1 Human Skin Color as an Environmental Adaptation	3
1.1.2 Mechanism of Melanin Formation	3
1.1.3 Skin Color Diversity due to DNA Polymorphism	4
1.1.4 Objectives	4
1.2 A Linkage Disequilibrium–Based Statistical Approach to Detect Interactions between SNP Alleles at Multiple Loci That Contribute to Skin Pigmentation Variation between Human Populations	5
1.2.1 SNP Analysis of European and East Asian Cohorts	5
1.2.2 Genotype and Allele Frequencies for the 20 SNPs in the European and East Asian Population Groups	6
1.2.3 Cluster Analysis	8
1.2.4 Linkage Disequilibrium Generated by Gene–Gene Interactions Contributes to Differences between Racial Groups	11
1.3 SNP Analyses Reveal Natural Selection of Pigmentation Candidate Genes from Haplotypes	13
1.3.1 Background	13
1.3.2 Detecting Natural Selection in the Human Genome on the Basis of the Haplotype Structure	14

1.3.3	Discussion	15
1.3.4	Future Investigations of Natural Selection and Environmental Adaptability Regarding Skin Pigmentation	16
1.4	Elucidation of the UVR-Induced Selective Genetic Mechanisms Influencing Variations in Human Skin Pigmentation	17
1.4.1	Influence of UVR Levels on Skin Color Variation	17
1.4.2	Statistical Methods of Clarifying Gene-Environment Interactions, Viewing Skin Pigmentation as a Complex Trait	18
1.4.3	New Methods for Detecting Gene-Environment Interactions for Human Skin Pigmentation Variation	19
1.4.4	RS Data Processing and Spatial Analysis in GIS	20
1.4.5	Gene-Environment Interaction Analysis	21
1.4.6	Results	23
1.4.7	Discussion	27
1.5	Conclusions	29
2.	Information Theoretic Methods for Gene-Environment Interaction Analysis	39
	<i>Jonathan Knights and Murali Ramanathan</i>	
2.1	Introduction	39
2.2	Information Theoretic Metrics and Searching for GEIs	41
2.2.1	Entropy and Mutual Information	41
2.2.1.1	Entropy	41
2.2.1.2	Mutual information	42
2.2.1.3	The k -way interaction information	43
2.2.1.4	Total correlation information	45
2.2.1.5	Phenotype-associated information	45

2.3	How and Why Do the KWII and the PAI Measure Statistical Interaction?	46
2.4	Performance of KWII and PAI Search Algorithms on Simulated Data	47
2.4.1	KWII as an Interaction Metric	50
2.4.1.1	PAI removes the effects of LD on TCI	51
2.5	Algorithms	53
2.5.1	Noninformation Theoretic Methods	54
2.5.2	Information Theoretic Algorithms	56
2.6	Applications of Information Theory to Interaction Analysis	59
2.7	Critiques of Information Theory Approaches to Interaction Analysis	62
2.8	Conclusions	64
3.	Approaches for Gene-Environment Interaction Analysis: Practice of Regional Epidemiological Study	73
	<i>Mio Nakazato and Takahiro Maeda</i>	
3.1	Introduction	74
3.2	Community Investigation and Setting of the Research Objective	74
3.2.1	Study Background and Purpose: Homocysteine and Arteriosclerosis	77
3.3	Study Design and Methods	79
3.3.1	Survey Items	79
3.3.2	Sample Size	80
3.3.3	Sample Collection and Storage	80
3.3.3.1	Explanation of the community research survey to each organization and request for cooperation	80
3.3.3.2	Methods and flow of the research survey	81
3.3.3.3	Collection and storage of blood samples	82
3.3.4	Measurement and Analysis of Samples	83
3.3.5	Data Management	85
3.3.6	Ethics Committee	86

3.3.7	Preparation of the Field of Regional Epidemiological Study	87
3.3.8	Study Limitations	88
3.4	Statistical Analysis	89
3.4.1	Preparation of Analytical Sheets	89
3.4.2	Close Investigation of Data	89
3.4.3	Selection of Statistical Analysis Methods	94
3.4.4	Practice of Analysis	95
3.5	Later Surveys	106
3.5.1	Homocysteine and Folic Acid	106
3.5.2	Adiponectin Polymorphism and Arteriosclerosis	108
3.5.3	Leukocyte Count and Arteriosclerosis	109
3.5.4	Current Study Purpose in Our Regional Epidemiological Research Field	110
3.5.5	Other Studies	110
3.6	Conclusion	111
4.	Use of Bioinformatics in Revealing the Identity of Nature's Products with Minimum Genetic Variation: The Sibling Species	121
	<i>K. Gajapathy, A. Ramanan, S. L. Goodacre, R. Ramasamy, and S. N. Surendran</i>	
4.1	Cryptic Species: An Introduction	122
4.2	Computer Power in Taxonomy of Sibling Species	124
4.2.1	DNA and Protein as Tools in Taxonomy: Alternatives or Auxiliary?	124
4.2.2	Protein-Based Assays	129
4.2.3	Karyotyping and Chromosome Banding Pattern	129
4.3	Automated Species Identification	130
4.4	Species Complex among Insect Vectors in Sri Lanka: Two Case Studies Where Bioinformatic Approaches Have Solved Taxonomic Problems	134
4.5	Conclusion	141

5. Integrated Bioinformatics, Biostatistics, and Molecular Epidemiologic Approaches to Study How the Environment and Genes Work Together to Influence the Development of Complex Chronic Diseases	151
<i>Alok Deoraj, Changwon Yoo, and Deodutta Roy</i>	
5.1 Introduction	152
5.2 Tools for Identifying Genetic, Environmental, and Stochastic Factors Relevant in Complex Human Diseases	156
5.2.1 Candidate Gene	157
5.2.2 Copy Number Variations	157
5.2.3 Single-Nucleotide Polymorphism	158
5.2.4 Haplotype Mapping	159
5.2.5 Genome-Wide Association Screen	160
5.2.6 Epigenetic Changes and Variation in Noncoding DNA Elements	161
5.2.7 Variations, Mutations, and Damages in the Mitochondrial Genome	163
5.2.8 Chronic Disease Bionetwork through Interactome, Phenome, and Systems Approaches	163
5.2.8.1 Transcriptomics	164
5.2.8.2 Proteomics	165
5.2.8.3 Interactomics	165
5.2.8.4 Metabolomics	166
5.2.8.5 Phenomics	166
5.3 Integration of Functional Genomic, Epigenetic, and Environmental Data of Molecular Epidemiological Studies Using Bioinformatics and Biostatistics Approaches	167
5.3.1 Molecular Epidemiologic Study Designs for $G \times G$ and GEI	168
5.3.1.1 Family-Based Studies	168
5.3.1.2 Studies of Unrelated Individuals	168
5.3.1.3 Retrospective Design	169
5.3.1.4 Prospective Design	170
5.3.1.5 Case-Only Design	171

5.3.2	Identification of Mutations, SNPs, Changes in CNVs, and Variations in a Significant Set of Over- and Underexpressed Genes	172
5.3.3	Enrichment Analysis of Significant Genes and Proteins	175
5.3.4	Analysis of Combined Effects of Modified Gene Expressions, CNVs, Mutations, and Molecular Interactions Influencing Biological Pathways and Network(s) That Contribute to the Susceptibility to, Resistance to, or Development of Complex Diseases	175
5.3.4.1	Multifactor Dimensionality Reduction	176
5.3.4.2	Bayesian Networks	176
5.3.5	Validation of Key Causal Genes/Proteins/Molecules/Environmental and Stochastic Factors Predicted to Be Involved in the Development of Disease by Statistical and Other Approaches	179
5.3.5.1	Validation by Statistical Methods	179
5.3.5.2	Literature-Based Validation of Key Causal Genes/Proteins/Molecules/Environmental and Stochastic Factors Involved in the Etiology of Chronic Diseases Using Models Generated by Empirical Data	180
5.4	Predictive Analysis of Lifetime Risk of Developing Disease	182
5.5	Technical Challenges	183
5.5.1	Sample Size	183
5.5.2	Complex Mixture of Covariates	184
5.5.3	Coordination in Data Collection and Their Meta-Analysis	184

	5.5.4 Lack of Computational Power	185
5.6	Conclusion	186
5.7	Online Resources	187
<i>Index</i>		193