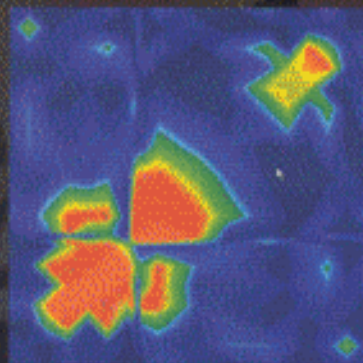
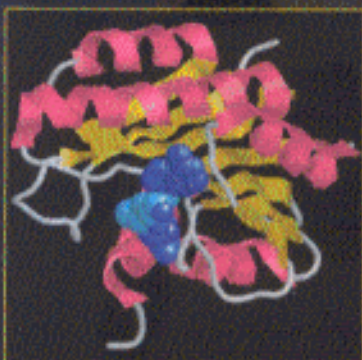
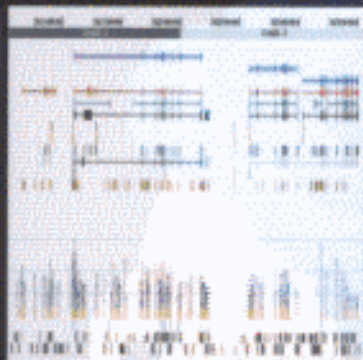


# Bioinformatics for Geneticists

Edited by  
Michael R. Barnes & Ian C. Gray



NTGV	AH	FE	YQ	IC	VT	CD	DY	YV	GF	GC
NTGI	AH	FE	YQ	IR	VT	CD	DY	YV	GF	GC
NTGI	AH	FE	YQ	IR	VT	CD	DY	YV	GF	GC
NGP	V	A	Q	F	E	Y	Q	I	R	V
SCH	V	A	H	L	E	L	Q	I	R	V
SCH	V	A	H	L	E	L	Q	I	R	V
SCH	V	A	H	L	E	L	Q	I	R	V
PCIT	A	H	I	E	Y	R	I	R	V	C
NGP	V	A	Q	F	E	Y	Q	I	R	V
NGP	V	A	N	F	E	Y	Q	I	R	V
NAG	M	T	Y	F	E	Y	Q	I	R	V
NTG	A	H	FE	YQ	IR	VT	CD	DY	YV	GF
IGR	N	A	R	I	T	Y	R	V	Y	Q



## CONTENTS

---

<i>List of contributors</i>	xi
<i>Foreword</i>	xiii
<b>SECTION I. AN INTRODUCTION TO BIOINFORMATICS FOR THE GENETICIST</b>	1
<b>Chapter 1 Introduction: The Role of Genetic Bioinformatics</b>	3
<i>Michael R. Barnes and Ian C. Gray</i>	
1.1 Introduction	3
1.2 Genetics in the post-genome era—the role of bioinformatics	6
1.3 Knowledge management and expansion	6
1.4 Data management and mining	6
1.5 Genetic study designs	8
1.6 Physical locus analysis	12
1.7 Selecting candidate genes for analysis	14
1.8 Progressing from candidate gene to disease-susceptibility gene	14
1.9 Comparative genetics and genomics	15
1.10 Conclusions	17
References	18
<b>Chapter 2 Internet Resources for the Geneticist</b>	21
<i>Michael R. Barnes and Christopher Southan</i>	
2.1 Introduction	22
2.2 Sub-division of biological data on the internet	23
2.3 Searching the internet for genetic information	24
2.4 Which web search engine?	24
2.5 Search syntax: the mathematics of search engine use	26
2.6 Boolean searching	27
2.7 Searching scientific literature—getting to ‘state of the art’	28
2.8 Searching full-text journals	29
2.9 Searching the heart of the biological internet—sequences and genomic data	30
2.10 Nucleotide and protein sequence databases	30
2.11 Biological sequence databases—primary and secondary	31
2.12 Conclusions	36
References	37

<b>Chapter 3 Human Genetic Variation: Databases and Concepts</b>	<b>39</b>
<i>Michael R. Barnes</i>	
3.1 Introduction	40
3.2 Forms and mechanisms of genetic variation	43
3.3 Databases of human genetic variation	50
3.4 SNP databases	51
3.5 Mutation databases	57
3.6 Genetic marker and microsatellite databases	60
3.7 Non-nuclear and somatic mutation databases	61
3.8 Tools for SNP and mutation visualization—the genomic context	63
3.9 Tools for SNP and mutation visualization—the gene context	63
3.10 Conclusions	67
References	67
<b>Chapter 4 Finding, Delineating and Analysing Genes</b>	<b>71</b>
<i>Christopher Southan</i>	
4.1 Introduction	71
4.2 The evidence cascade for gene products	72
4.3 Shortcomings of the standard gene model	75
4.4 Locating known genes on the Golden Path	76
4.5 Gene portal inspection	79
4.6 Locating genes which are not present in the Golden Path	80
4.7 Analysing a novel gene	81
4.8 Comprehensive database searching	88
4.9 Conclusions and prospects	90
References	90
<b>SECTION II. THE IMPACT OF COMPLETE GENOME SEQUENCES ON GENETICS</b>	<b>93</b>
<b>Chapter 5 Assembling a View of the Human Genome</b>	<b>95</b>
<i>Colin A. Semple</i>	
5.1 Introduction	95
5.2 Genomic sequence assembly	98
5.3 Annotation from a distance: the generalities	101
5.4 Annotation up close and personal: the specifics	105
5.5 Annotation: the next generation	113
Acknowledgements	114
References	114
<b>Chapter 6 Mouse and Rat Genome Informatics</b>	<b>119</b>
<i>Judith A. Blake, Janan Eppig and Carol J. Bult</i>	
6.1 Introduction	120
6.2 The model organism databases for mouse and rat	122
6.3 Mouse genetic and physical maps	124
6.4 Rat genetic and physical maps	127

6.5	Genome sequence resources	128
6.6	Comparative genomics	131
6.7	From genotype to phenotype	132
6.8	Functional genomics	135
6.9	Rodent disease models	137
6.10	Summary	137
	Acknowledgements	138
	References	138

## **Chapter 7 Genetic and Physical Map Resources — An Integrated View** 143

*Michael R. Barnes*

7.1	Introduction	144
7.2	Genetic maps	145
7.3	Physical maps	148
7.4	Physical contig maps	151
7.5	The role of physical and genetic maps in draft sequence curation	152
7.6	The human genome sequence — the ultimate physical map?	153
7.7	QC of genomic DNA — resolution of marker order and gap sizes	154
7.8	Tools and databases for map analysis and integration	155
7.9	Conclusions	159
	References	160

## **SECTION III. BIOINFORMATICS FOR GENETIC STUDY DESIGN** 163

### **Chapter 8 From Linkage Peak to Culprit Gene: Following Up Linkage Analysis of Complex Phenotypes with Population-based Association Studies** 165

*Ian C. Gray*

8.1	Introduction	165
8.2	Theoretical and practical considerations	166
8.3	A practical approach to locus refinement and candidate gene identification	173
8.4	Conclusion	176
	Acknowledgements	176
	References	177

### **Chapter 9 Genetic Studies from Genomic Sequence** 179

*Michael R. Barnes*

9.1	Introduction	180
9.2	Defining the locus	180
9.3	Case study 1: Identification and extraction of a genomic sequence between two markers	184
9.4	Case study 2: Checking the integrity of a genomic sequence between two markers	185
9.5	Case study 3: Definition of known and novel genes across a genomic region	188
9.6	Case study 4: Candidate gene selection — building biological rationale around genes	190

9.7 Case study 5: Known and novel marker identification	195
9.8 Case study 6: Genetic/physical locus characterization and marker panel design	199
9.9 Conclusions	201
References	201
<b>Chapter 10 SNP Discovery and PCR-based Assay Design: From <i>In Silico</i> Data to the Laboratory Experiment</b>	203
<i>Ellen Vieux, Gabor Marth and Pui Kwok</i>	
10.1 Introduction	204
10.2 SNP identification	205
10.3 PCR primer design	207
10.4 Broader PCR assay design issues	208
10.5 Primer selection	210
10.6 Problems related to SNP assay validation	212
10.7 Conclusion	213
References	213
<b>Chapter 11 Tools for Statistical Analysis of Genetic Data</b>	217
<i>Aruna Bansal, Peter R. Boyd and Ralph McGinnis</i>	
11.1 Introduction	218
11.2 Linkage analysis	218
11.3 Association analysis	223
11.4 Haplotype Reconstruction	226
11.5 Linkage disequilibrium	229
11.6 Quantitative Trait Locus (QTL) mapping in experimental crosses	235
Acknowledgements	240
References	240
<b>SECTION IV. BIOLOGICAL SEQUENCE ANALYSIS AND CHARACTERIZATION</b>	247
<b>Chapter 12 Predictive Functional Analysis of Polymorphisms: An Overview</b>	249
<i>Michael R. Barnes</i>	
12.1 Introduction	250
12.2 Principles of predictive functional analysis of polymorphisms	252
12.3 The anatomy of promoter regions and regulatory elements	257
12.4 The anatomy of genes	258
12.5 Pseudogenes and regulatory mRNA	264
12.6 Analysis of novel regulatory elements and motifs in nucleotide sequences	264
12.7 Functional analysis on non-synonymous coding polymorphisms	266
12.8 A note of caution on the prioritization of <i>in silico</i> predictions for further laboratory investigation	268
12.9 Conclusions	268
References	269

<b>Chapter 13 Functional <i>In Silico</i> Analysis of Non-coding SNPs</b>	<b>273</b>
<i>Thomas Werner</i>	
13.1 Introduction	273
13.2 General structure of chromatin-associated DNA	275
13.3 General functions of regulatory regions	276
13.4 Transcription Factor binding sites (TF-sites)	276
13.5 Structural elements	276
13.6 Organizational principles of regulatory regions	277
13.7 RNA processing	279
13.8 SNPs in regulatory regions	279
13.9 Evaluation of non-coding SNPs	280
13.10 SNPs and regulatory networks	281
13.11 SNPs may affect the expression of a gene only in specific tissues	281
13.12 <i>In silico</i> detection and evaluation of regulatory SNPs	281
13.13 Getting promoter sequences	282
13.14 Identification of relevant regulatory elements	283
13.15 Estimation of functional consequences of regulatory SNPs	284
13.16 Conclusion	285
References	285
 <b>Chapter 14 Amino Acid Properties and Consequences of Substitutions</b>	 <b>289</b>
<i>Matthew J. Betts and Robert B. Russell</i>	
14.1 Introduction	291
14.2 Protein features relevant to amino acid behaviour	292
14.3 Amino acid classifications	296
14.4 Properties of the amino acids	298
14.5 Amino acid quick reference	299
14.6 Studies of how mutations affect function	311
14.7 A summary of the thought process	313
References	314
 <b>SECTION V. GENETICS/GENOMICS INTERFACES</b>	 <b>317</b>
 <b>Chapter 15 Gene Expression Informatics and Analysis</b>	 <b>319</b>
<i>Antoine H. C. van Kampen, Jan M. Ruijter, Barbera D. C. van Schaik, Huib N. Caron and Rogier Versteeg</i>	
15.1 Introduction	320
15.2 Technologies for the measurement of gene expression	322
15.3 The Cancer Genome Anatomy Project (CGAP)	324
15.4 Processing of SAGE data	325
15.5 Integration of biological databases for the construction of the HTM	334
15.6 The Human Transcriptome Map	336
15.7 Regions of Increased Gene Expression (RIDGES)	339
15.8 Discussion	340
References	341

<b>Chapter 16 Proteomic Informatics</b>	<b>345</b>
<i>Jérôme Wojcik and Alexandre Hamburger</i>	
16.1 Introduction	346
16.2 Proteomic informatics	347
16.3 Experimental workflow: classical proteomics	347
16.4 Protein interaction networks	351
16.5 Building protein interaction networks	354
16.6 False negatives and false positives	354
16.7 Analysing interaction networks	355
16.8 Cell pathways	356
16.9 Prediction of protein networks	359
16.10 Assessment and validation of predictions	363
16.11 Exploiting protein networks	366
16.12 Deducing prediction rules from networks	367
16.13 Conclusion	368
Acknowledgements	369
References	369
 <b>Chapter 17 Concluding Remarks: Final Thoughts and Future Trends</b>	 <b>373</b>
<i>Michael R. Barnes and Ian C. Gray</i>	
17.1 How many genes?	374
17.2 Mapping the genome and gaining a view of the full depth of human variation	375
17.3 Holistic analysis of complex traits	376
17.4 A final word on bioinformatics	376
Acknowledgements	376
References	376
 <b>Appendix I</b>	 <b>379</b>
 <b>Appendix II</b>	 <b>381</b>
 <b>Glossary</b>	 <b>387</b>
 <b>Index</b>	 <b>391</b>