

*Bioinformatics for Geneticists*. Edited by Michael R. Barnes and Ian C. Gray  
Copyright © 2003 John Wiley & Sons, Ltd.  
ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

---

# **BIOINFORMATICS FOR GENETICISTS**

---

---

# BIOINFORMATICS FOR GENETICISTS

---

Edited by

Michael R. Barnes

Genetic Bioinformatics

GlaxoSmithKline Pharmaceuticals, UK

and

Ian C. Gray

Discovery Genetics

GlaxoSmithKline Pharmaceuticals, UK



Copyright © 2003

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,  
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): [cs-books@wiley.co.uk](mailto:cs-books@wiley.co.uk)

Visit our Home Page on [www.wileyurope.com](http://www.wileyurope.com) or [www.wiley.com](http://www.wiley.com)

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the Publisher. Requests to the Publisher should be addressed to the Permissions Department, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, or emailed to [permreq@wiley.co.uk](mailto:permreq@wiley.co.uk), or faxed to (+44) 1243 770620.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

#### ***Other Wiley Editorial Offices***

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

#### ***British Library Cataloguing in Publication Data***

A catalogue record for this book is available from the British Library

ISBN 0-470-84393-4

ISBN 0-470-84394-2 (PB)

Typeset in 9/11pt Times by Laserwords Private Limited, Chennai, India

Printed and bound in Great Britain by TJ International, Padstow, Cornwall

This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

## 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>       | 39  |
| <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  |
| <br>   |     |
| <b>Chapter 4 Finding, Delineating and Analysing Genes</b>              | 71  |
| <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  |
| <br>   |     |
| <b>SECTION II. THE IMPACT OF COMPLETE GENOME SEQUENCES ON GENETICS</b> | 93  |
| <br>   |     |
| <b>Chapter 5 Assembling a View of the Human Genome</b>                 | 95  |
| <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 |
| <br>   |     |
| <b>Chapter 6 Mouse and Rat Genome Informatics</b>                      | 119 |
| <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        |
| <b>Chapter 7 Genetic and Physical Map Resources — An Integrated View</b>  |   | <b>143</b> |
| <i>Michael R. Barnes</i>  |   |            |
| 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        |
| <b>SECTION III. BIOINFORMATICS FOR GENETIC STUDY DESIGN</b>   |   | <b>163</b> |
| <b>Chapter 8 From Linkage Peak to Culprit Gene: Following Up Linkage Analysis of Complex Phenotypes with Population-based Association Studies</b> |   | <b>165</b> |
| <i>Ian C. Gray</i>  |   |            |
| 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        |
| <b>Chapter 9 Genetic Studies from Genomic Sequence</b>  |   | <b>179</b> |
| <i>Michael R. Barnes</i>  |   |            |
| 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>                                    | 273 |
| <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 |
| <br>   |     |
| <b>Chapter 14 Amino Acid Properties and Consequences of Substitutions</b>                                    | 289 |
| <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 |
| <br>   |     |
| <b>SECTION V. GENETICS/GENOMICS INTERFACES</b>   | 317 |
| <br>   |     |
| <b>Chapter 15 Gene Expression Informatics and Analysis</b>   | 319 |
| <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>   | 345 |
| <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 |
| <br>  |     |
| <b>Chapter 17 Concluding Remarks: Final Thoughts and Future Trends</b>          | 373 |
| <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 |
| <br>  |     |
| <b>Appendix I</b>   | 379 |
| <br>  |     |
| <b>Appendix II</b>  | 381 |
| <br>  |     |
| <b>Glossary</b>   | 387 |
| <br>  |     |
| <b>Index</b>  | 391 |

## LIST OF CONTRIBUTORS

---

### **Aruna Bansal**

Population Genetics  
GlaxoSmithKline Pharmaceuticals  
New Frontiers Science Park (North)  
Third Avenue  
Harlow, Essex CM19 5AW, UK

### **Michael R. Barnes**

Genetic Bioinformatics  
GlaxoSmithKline Pharmaceuticals  
New Frontiers Science Park (North)  
Third Avenue  
Harlow, Essex CM19 5AW, UK

### **Matthew J. Betts**

Bioinformatics  
DeCODE Genetics  
Sturlugötu 8  
101 Reykjavík, Iceland

### **Judith A. Blake**

The Jackson Laboratory  
600 Main Street  
Bar Harbor, ME 04609, USA

### **Peter R. Boyd**

Population Genetics  
GlaxoSmithKline Pharmaceuticals  
Medicines Research Centre  
Gunnels Wood Road,  
Stevenage, Herts, SG1 2NY

### **Carol J. Bult**

The Jackson Laboratory  
600 Main Street  
Bar Harbor, ME 04609, USA

### **H. N. Caron**

Department of Pediatric Oncology,  
Academic Medical Center,

University of Amsterdam  
Meibergdreef 9, 1105 AZ Amsterdam  
The Netherlands

### **Janan Eppig**

The Jackson Laboratory  
600 Main Street  
Bar Harbor, ME 04609, USA

### **Ian C. Gray**

Discovery Genetics  
GlaxoSmithKline Pharmaceuticals  
New Frontiers Science Park (North)  
Third Avenue  
Harlow, Essex CM19 5AW, UK

### **Alexandre Hamburger**

Hybrigenics  
3-5 impasse Reille  
F75014 Paris, France

### **Pui-Yan Kwok**

University of California, San Francisco  
505 Parnassus Ave,  
Long 1332A, Box 0130  
San Francisco  
CA 94143-0130, USA

### **Gabor Marth**

National Center for Biotechnology Infor-  
mation, NLM, NIH  
8600 Rockville Pike  
Building 45, Room 5AS29  
Bethesda, Maryland 20894, USA

### **Ralph McGinnis**

Population Genetics  
GlaxoSmithKline Pharmaceuticals  
New Frontiers Science Park (North)  
Third Avenue  
Harlow, Essex CM19 5AW, UK

**J. M. Ruijter**

Department of Anatomy and Embryology,  
Academic Medical Center,  
University of Amsterdam  
Meibergdreef 9, 1105 AZ Amsterdam  
The Netherlands

**Robert B. Russell**

Structural & Computational Biology  
Programme  
EMBL, Meyerhofstrasse 1  
69117 Heidelberg, Germany

**Colin A. Semple**

Bioinformatics  
MRC Human Genetics Unit  
Edinburgh EH4 2XU, UK

**Christopher Southan**

Oxford GlycoSciences UK Ltd.  
The Forum, 86 Milton Science Park  
Abingdon OX14 4RY, UK

**B. D. C. van Schaik**

Bioinformatics Laboratory  
Academic Medical Center,  
University of Amsterdam  
Meibergdreef 9, 1105 AZ Amsterdam  
The Netherlands

**Antoine H. C. van Kampen**

Bioinformatics Laboratory  
Academic Medical Center,  
University of Amsterdam  
Meibergdreef 9, 1105 AZ Amsterdam  
The Netherlands

**R. Versteeg**

Department of Human Genetics  
Academic Medical Center,  
University of Amsterdam  
Meibergdreef 9, 1105 AZ Amsterdam  
The Netherlands

**Ellen Vieux**

Washington University School of Medicine  
660 S. Euclid Ave  
Box 8123  
St Louis  
MO 63110, USA

**Thomas Werner**

Genomatix Software GmbH  
Landsberger Str. 6  
D-80339 Muenchen, Germany

**Jérôme Wojcik**

Hybrigenics  
3–5 impasse Reille  
F75014 Paris, France

## FOREWORD

---

Despite a relatively short existence, bioinformatics has always seemed an unusually multidisciplinary field. Fifteen years ago, when sequence data were still scarce and only a small fraction of the power of today's pizza-box supercomputers was available, bioinformatics was already entrenched in a diverse array of topics. Database development, sequence alignment, protein structure prediction, coding and promoter site identification, RNA folding, and evolutionary tree construction were all within the remit of the early bioinformaticist.<sup>1,2</sup> To address these problems, the field drew from the foundations of statistics, mathematics, physics, computer science, and of course, molecular biology. Today, predictably, bioinformatics still reflects the broad base on which it started, comprising an eclectic collection of scientific specialists.

As a result of its inherent diversity, it is difficult to define the scope of bioinformatics as a discipline. It may be even fruitless to try to draw hard boundaries around the field. It is ironic, therefore, that even now, if one were to compile an intentionally broad list of research areas within the bioinformatics purview, it would often exclude one biological discipline with which it shares a fundamental basis: Genetics. On one hand, this seems difficult to believe, since the fields share a strong common grounding in statistical methodology, dependence on efficient computational algorithms, rapidly growing biological data, and shared principles of molecular biology. On the other hand, this is completely understandable, since a large part of bioinformatics has spent the last few years helping to sequence a number of genomes, including that of man. In many cases, these sequencing projects have focused on constructing a single representative sequence—the consensus—a concept that is completely foreign to the core genetics principles of variability and individual differences. Despite a growing awareness of each other, and with a few clear exceptions, genetics and bioinformatics have managed to maintain separate identities.

Geneticists need bioinformatics. This is particularly true of those trying to identify and understand genes that influence complex phenotypes. In the realm of human genetics, this need has become particularly clear, so that most large laboratories now have one or two bioinformatics 'specialists' to whom other lab members turn for computing matters. These specialists are required to support a dauntingly wide assortment of applications: typical queries for such people might range from how to find instructions for accessing the internet, to how to disentangle a complex database schema, to how to optimize numerically intensive algorithms on parallel computing farms. These people, though somewhat scarce, are essential to the success of the laboratory.

With the ever-increasing volume of sequence data, expression information and well-characterized structures, as well as the imminent genotype and haplotype data on large and diverse human populations, genetics laboratories now must move beyond singular dependence on the bioinformatics handyman. Some level of understanding and ability to use bioinformatics applications is becoming necessary by everyone in the lab. Fortunately, bioinformaticians have been particularly successful in developing user-friendly software that renders complex statistical methods accessible to the bench scientists who generated

and should know most about the data being analysed. To further these analyses, ingenious software applications have been constructed to display the outcomes and integrate them with a host of useful annotation features such as chromosome characteristics, sequence signatures, disease correlates and species comparisons<sup>3</sup>. With these tools freely available and undergoing continued development, mapping projects that make effective use of genetic and genomic information will naturally enjoy greater success than those less equipped to do so. Simply put, genetics groups that cannot capitalize on bioinformatics applications will be increasingly scooped by those who can.

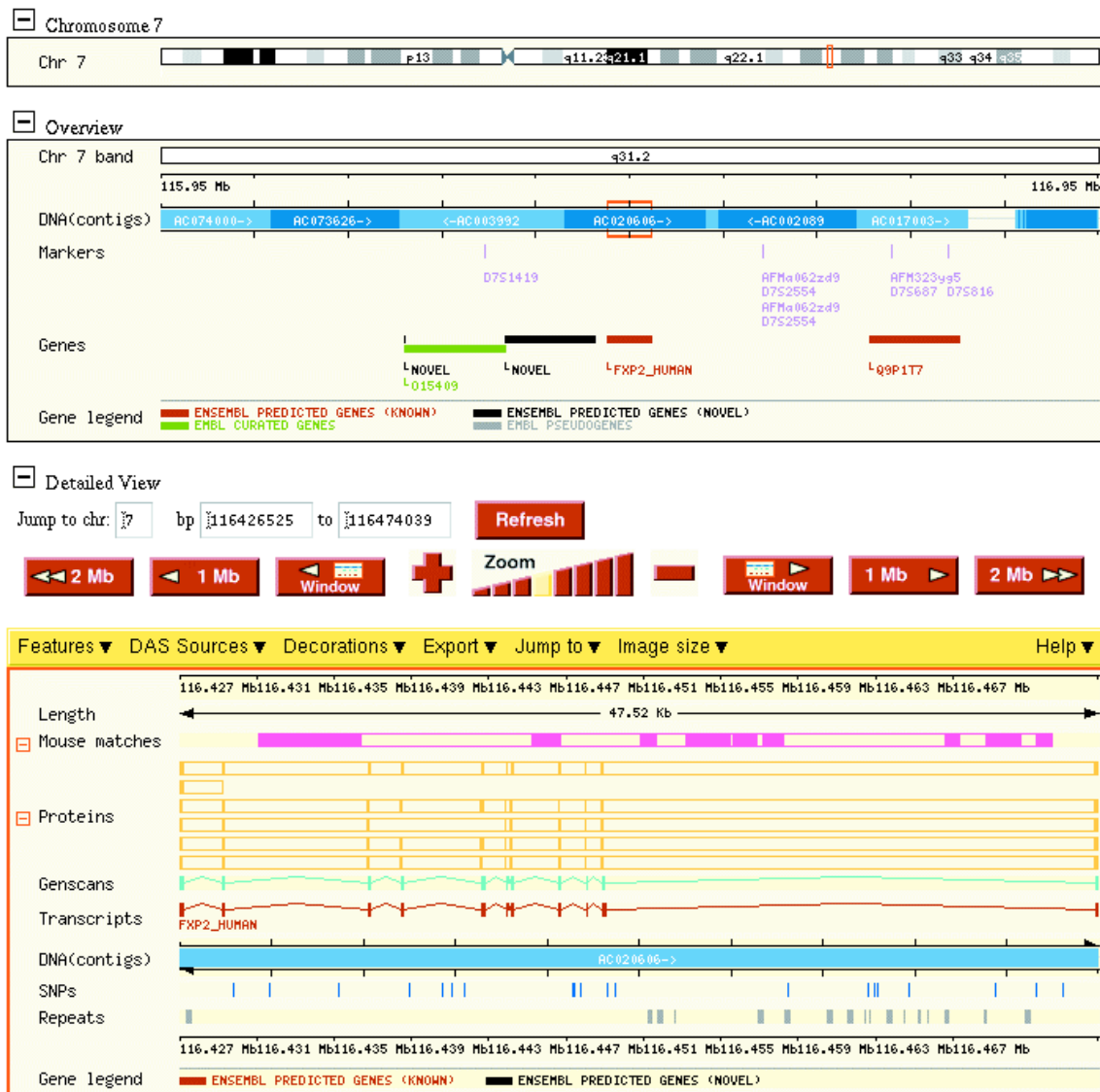
The emerging requirement of broader understanding of bioinformatics within genetics is the focus of this text, as easily appreciated by a quick glance at the title. Equally obvious is that geneticists are the editors' target audience. Still, one might ask 'toward what specific group of geneticists is this text aimed?' The software and computational backbone of bioinformatics is shared most noticeably with the areas of statistical and population genetics, so the statistical specialists would seem a plausible audience. By design, however, this text is not aimed at these specialists so much as at those with broader backgrounds in molecular and medical genetics, including both human and model organism research. The content should be accessible by skilled bench scientists, clinical researchers and even laboratory heads. Computationally, one needs only basic computing skills to work through most of the material. Biologically, appreciation of the problems described requires general familiarity with genetics research and recognition of the inherent value in careful use of *in silico* genetic and genomic information.

By necessity, the bioinformatics topics covered in this text reflect the diversity of the field. In order to obtain some order in this broad area, the editors have focused on computer applications and effective use of available databases. This concentration on applications means that descriptions of the statistical theory, numerical algorithms and database organization are left to other texts. The editors have intentionally bypassed much of this material to emphasize applications in widespread use—the focus is on efficient use, rather than development, of bioinformatics methods and tools.

The data behind many of the bioinformatics tools described here are rapidly changing and expanding. In response, the software tools and databases themselves tend to be (infuriatingly) dynamic. A consequence of this fluid state is that learning to use existing programs by no means guarantees a knack for using those in the future. Thus, one cannot expect long-term consistency in the tools and data-types described here (or in most any other contemporary bioinformatics text). By learning to use current tools more effectively, however, geneticists can not only capitalize on technology available, but perhaps engage more bioinformaticians in the excitement of genetics research. Bringing bioinformatics to geneticists is a crucial first step towards integrating the kindred fields and characterizing the frustratingly elusive genes that influence complex phenotypes.

*Lon R. Cardon*  
*Professor of Bioinformatics*  
*Wellcome Trust Centre for Human Genetics*  
*University of Oxford*

1. Doolittle, R. F. *Of URFs and ORFs: A primer on how to analyze derived amino acid sequences* (University Science Books, Mill Valley, California, 1987).
2. von Heijne, G. *Sequence analysis in molecular biology: Treasure trove or trivial pursuit* (Academic Press, London, 1987).
3. Wolfsberg, T. G., Wetterstrand, K. A., Guyer, M. S., Collins, F. S. & Baxevanis, A. D. A user's guide to the human genome. *Nature Genetics* **32** (suppl) (2002).



**Figure 5.1** The genomic region around the FOXP2 gene according to Ensembl.

[Homo sapiens Map View build 27](#)

Chromosome: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [**7**] [8](#) [9](#) [10](#) [11](#) [12](#) [13](#) [14](#) [15](#) [16](#) [17](#) [18](#) [19](#) [20](#) [21](#) [22](#) [X](#) [Y](#)

Query: FOXP2 [\[clear\]](#)

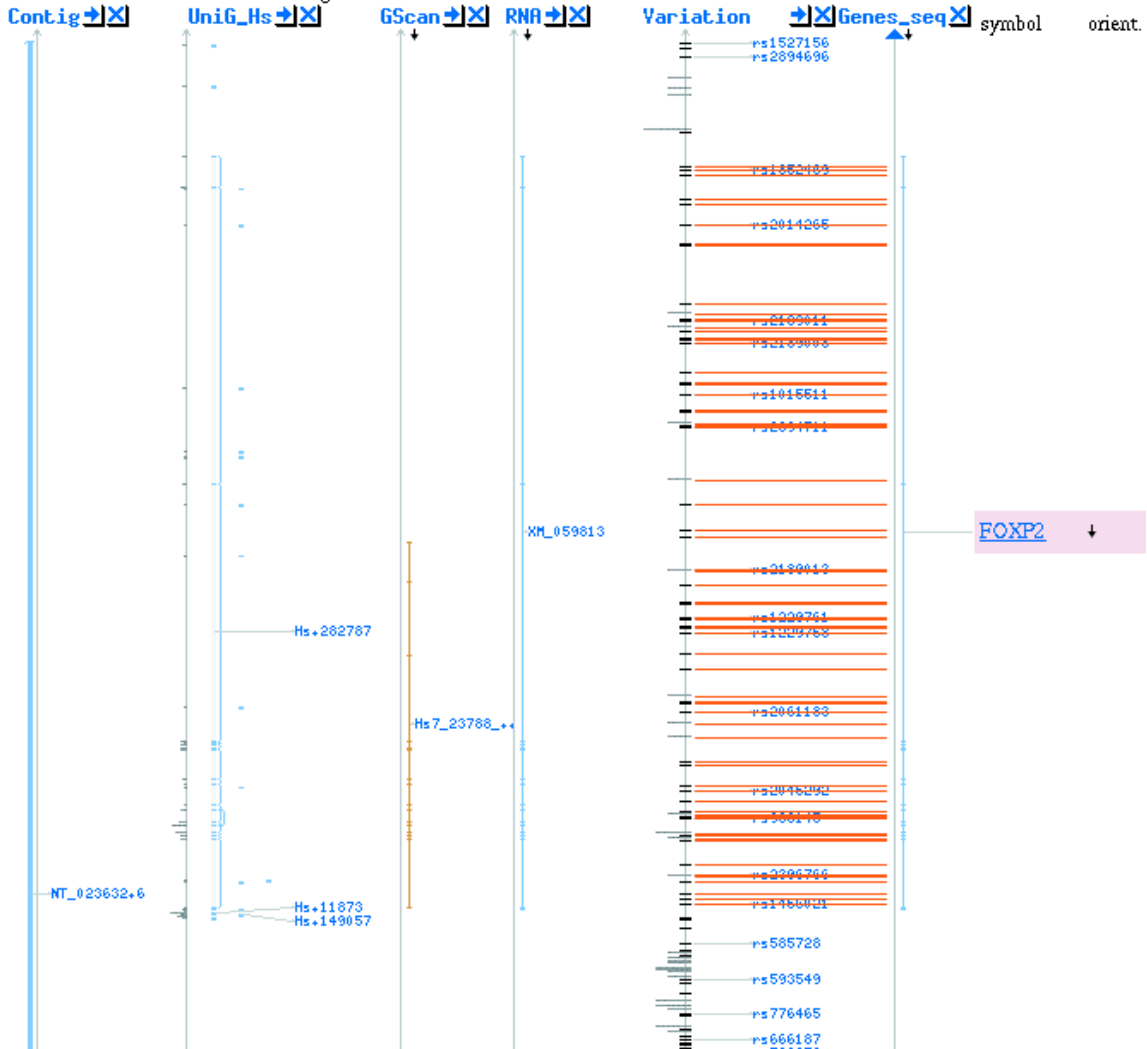
Master: Genes On Sequence Map

[Display settings](#)

Total Genes On Chromosome: 1939 [\[31 not localized\]](#)

Region Displayed: 115,865K – 116,489K bp [Download/View Sequence/Evidence](#)

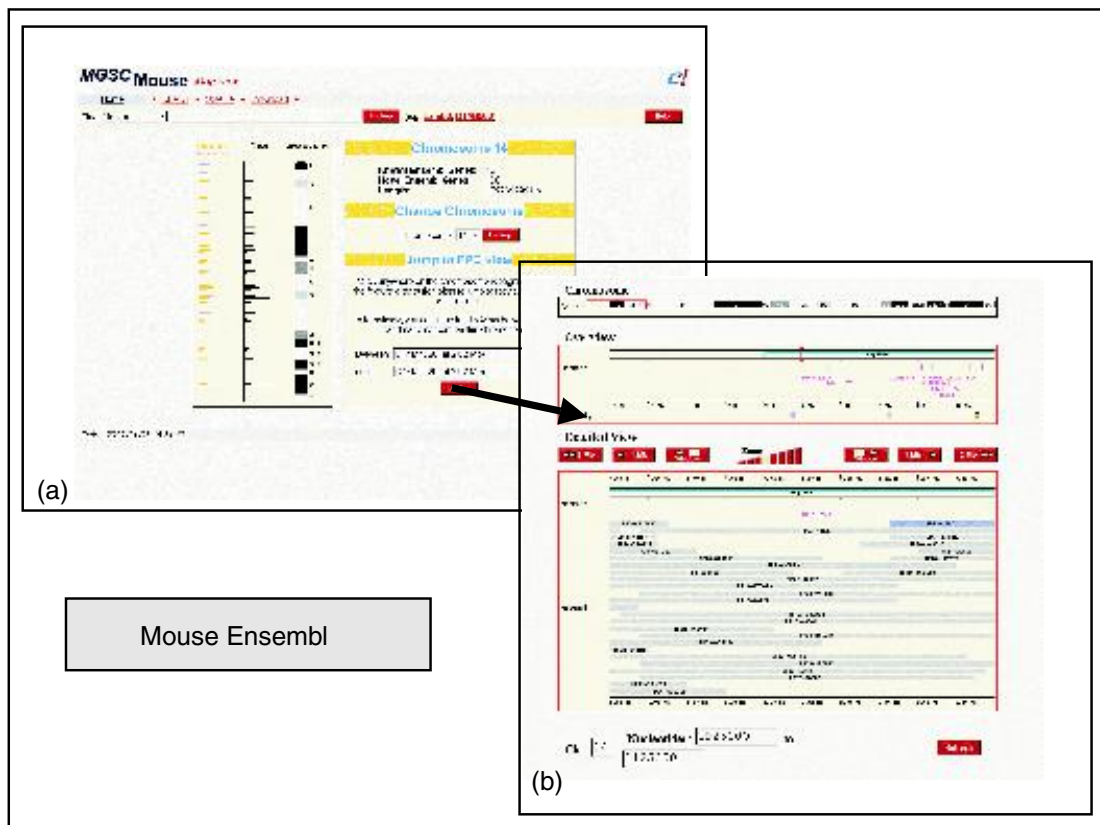
Genes Labeled: 3 Total Genes in Region: 3



**Figure 5.3** The genomic region around the FOXP2 gene according to the NCBI Map Viewer.

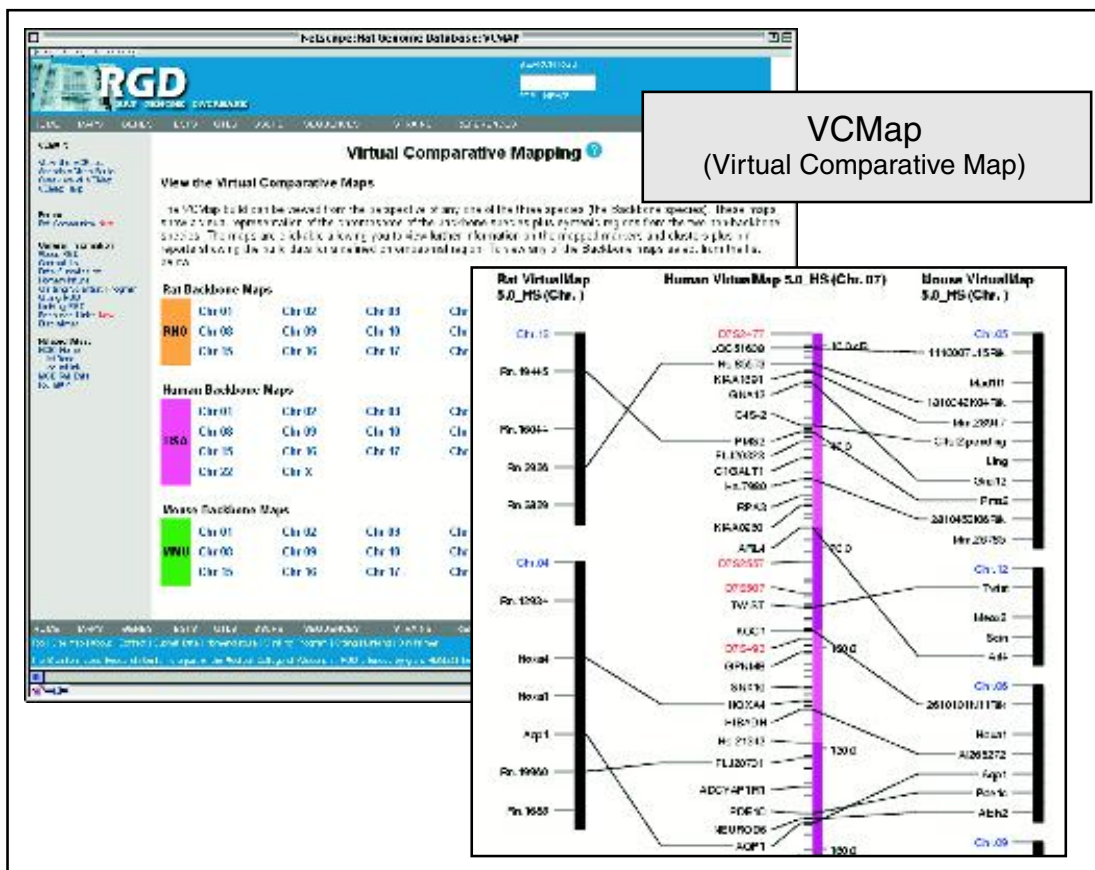


**Figure 5.4** The genomic region around the FOXP2 gene according to ORNL Genome Channel.



**Figure 6.1** Mouse Ensembl. A graphical representation of the clone-based physical map for the proximal end of mouse chromosome 14 from Ensembl. This browser allows users to search for regions of a chromosome between two STS markers and to view the current clone coverage in the selected area. Because the browser is Web-based, users do not have to download and install special software to view the BAC map.

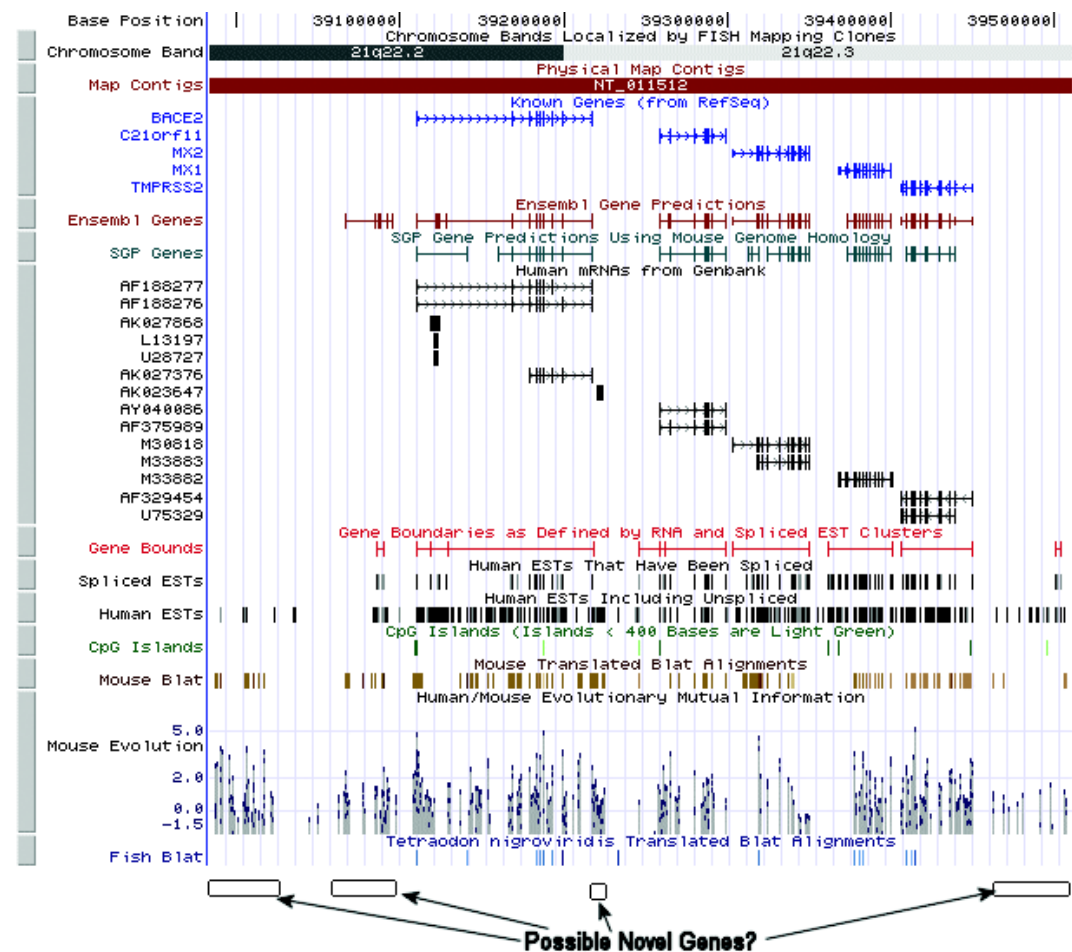




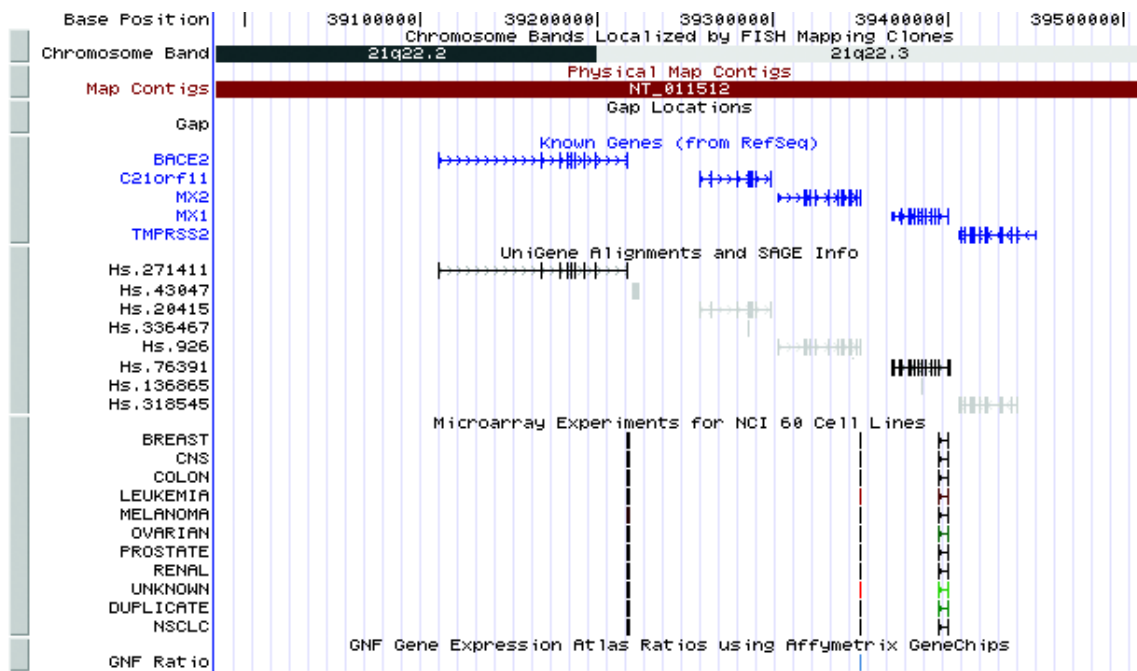
**Figure 6.3** Virtual Comparative Map. The Virtual Comparative Map is generated using sequence-based algorithms that predict syntenic regions inferred from homology among mapped sequences. Sequence comparisons between ESTs and cDNAs from human, mouse and rat are combined with Radiation Hybrid map locations to define regions of synteny. Locations for unmapped markers in a species are then predicted based on the map location of the orthologous marker in a syntenic region of another species. The forepanel shows a virtual comparative map using human as the backbone map (centre) and syntenic regions of rat (left) and mouse (right). Mapped genes, UniGenes and STSs are shown, with lines connecting predicted homologues among the species. Data sources for the virtual maps are RGD, NCBI and MGD. The virtual comparative maps are available at <http://rgd.mcw.edu/VCMAP/>.



**Figure 9.3** Definition of the D21S1245 and D21S1852 interval in genomic sequence using the UCSC human genome browser. The view immediately identifies five known genes, two gaps and several duplications across the region.

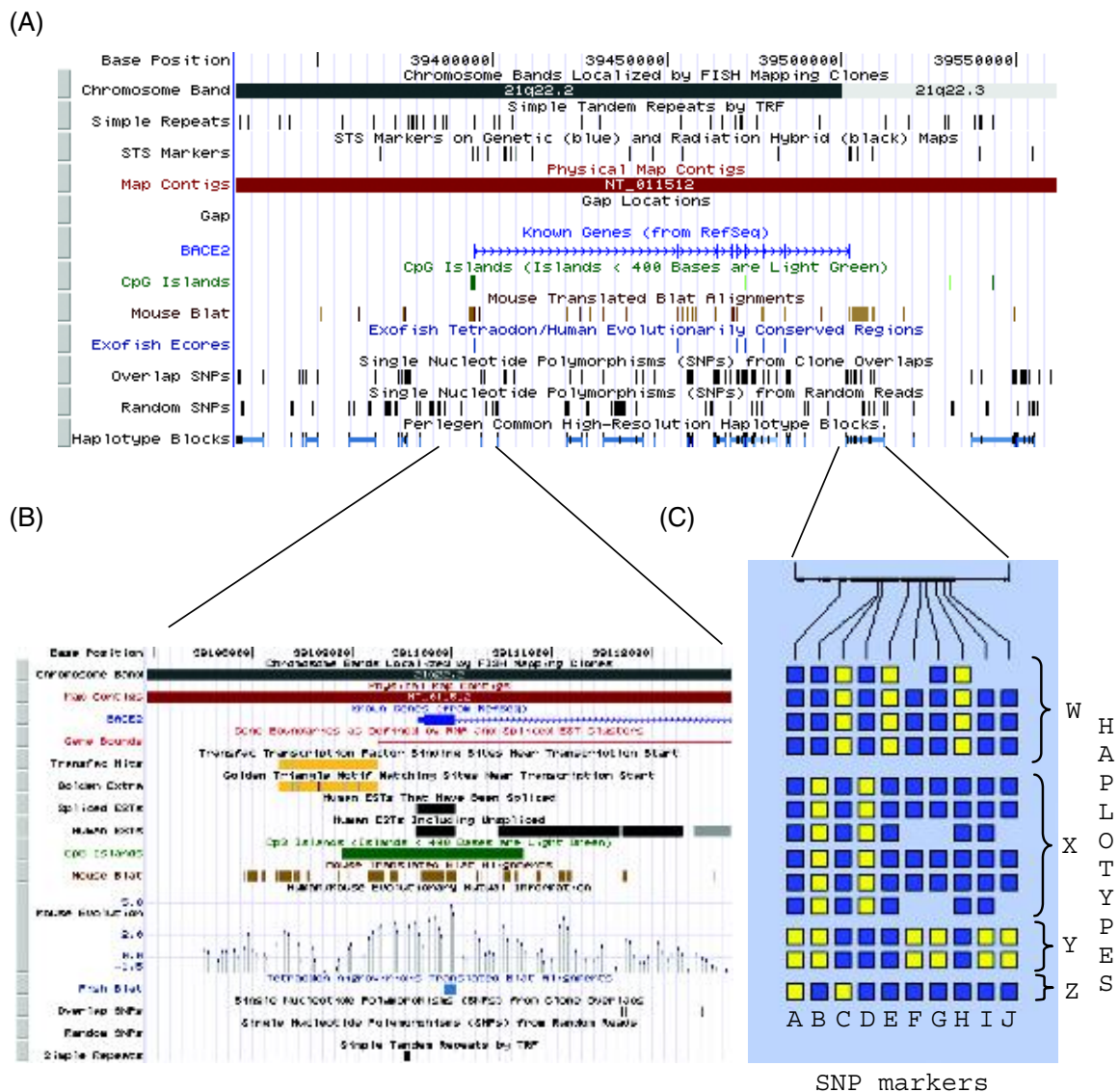


**Figure 9.5** Using the UCSC human genome browser to identify known and novel genes. A range of evidence including mRNAs, ESTs and human–mouse homology supports the existence of five known genes and up to four novel genes.



**Figure 9.6** Using the UCSC human genome browser to evaluate gene expression across a locus. Four tracks provide information on gene expression. ESTs implicitly measure gene expression as each is derived from a specified tissue source. Unigene clusters link to SAGE expression profiles drawn from the SageMap project. Finally data from two genome-wide microarray projects are presented, the NCI60 cell line project and GNF gene expression atlas ratios.





**Figure 9.8** SNP visualization in the UCSC human genome browser. (A) A detailed view of the BACE2 gene allows the user view a range of information, including SNP haplotype data. (B) Close viewing of the BACE2 locus allows the user to assess the functional context and genomic conservation of the region surrounding each SNP. (C) A detailed view of a Perlegen haplotype.