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

WILEY

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
Visit our Home Page on www.wileyeurope.com or www.wiley.com

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

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 Information, 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.[1,2]. 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 needs 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[3]. 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.

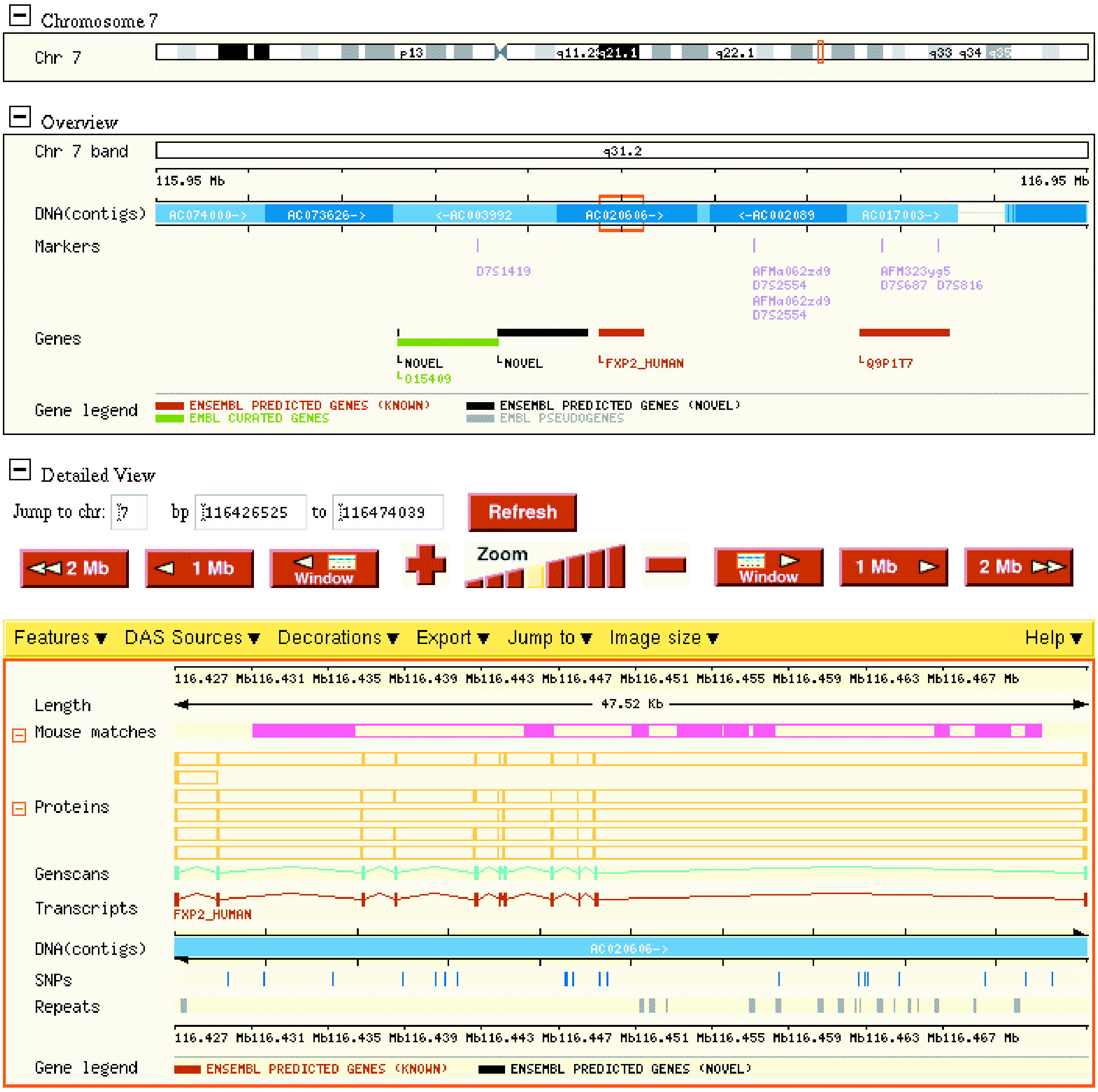
By necessity, the bioinformatics topics covered in this text reflect the diversity of the field. In order to obtain some order in this broad area, the editors have focused on computer applications and effective use of available databases. This concentration on applications means that descriptions of 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

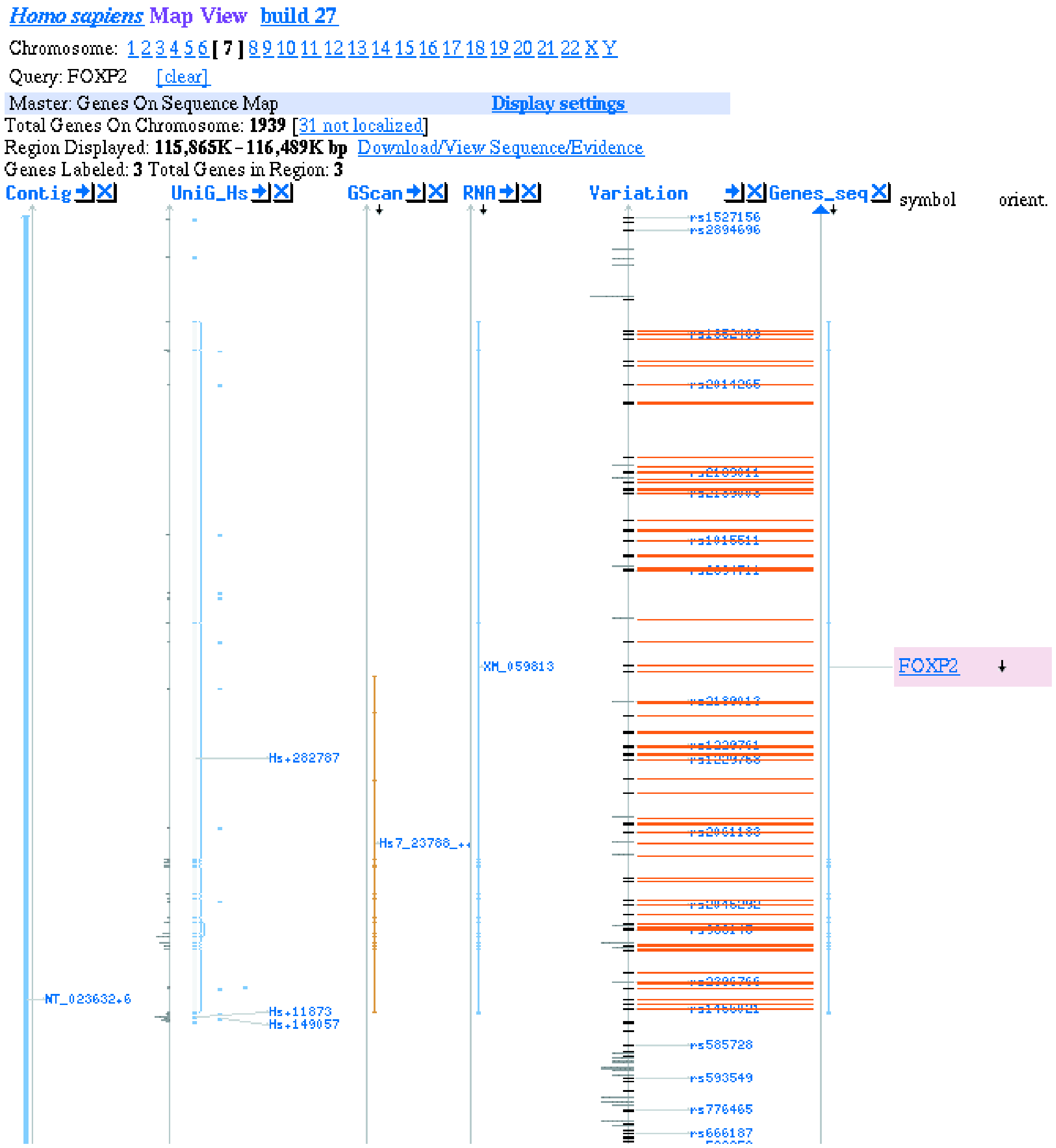
1. Doolittle, R. F. *Of URFs and ORFs: A primer on how to analyze derived amino acid sequences* (University Science Books, Mill Valley, California, 1987).
2. von Heijne, G. *Sequence analysis in molecular biology: Treasure trove or trivial pursuit* (Academic Press, London, 1987).
3. Wolfsberg, T. G., 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).

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)



Jump to chr: 7 bp 116426525 to 116474039

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.

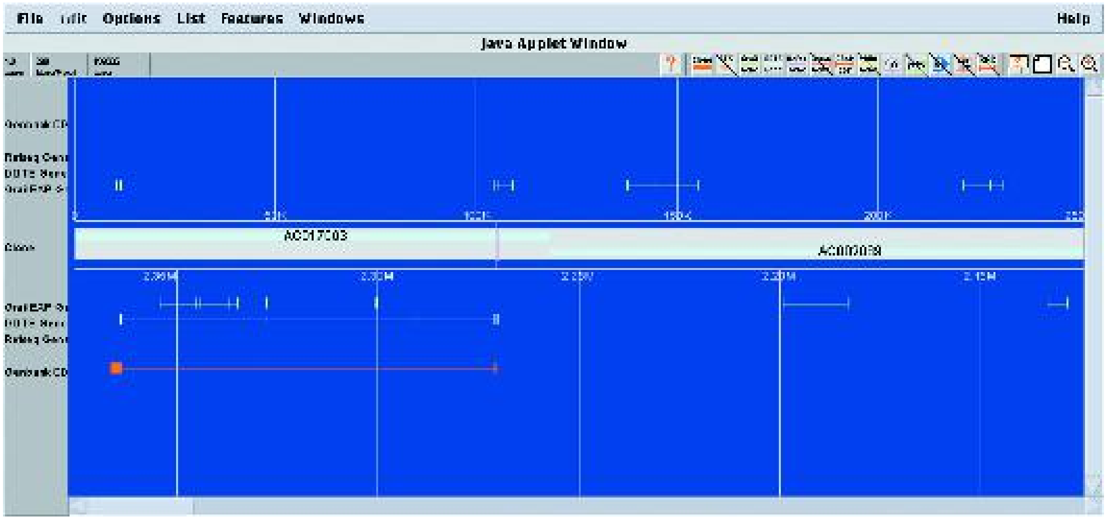


Figure 5.4 The genomic region around the FOXP2 gene according to ORNL Genome Channel.

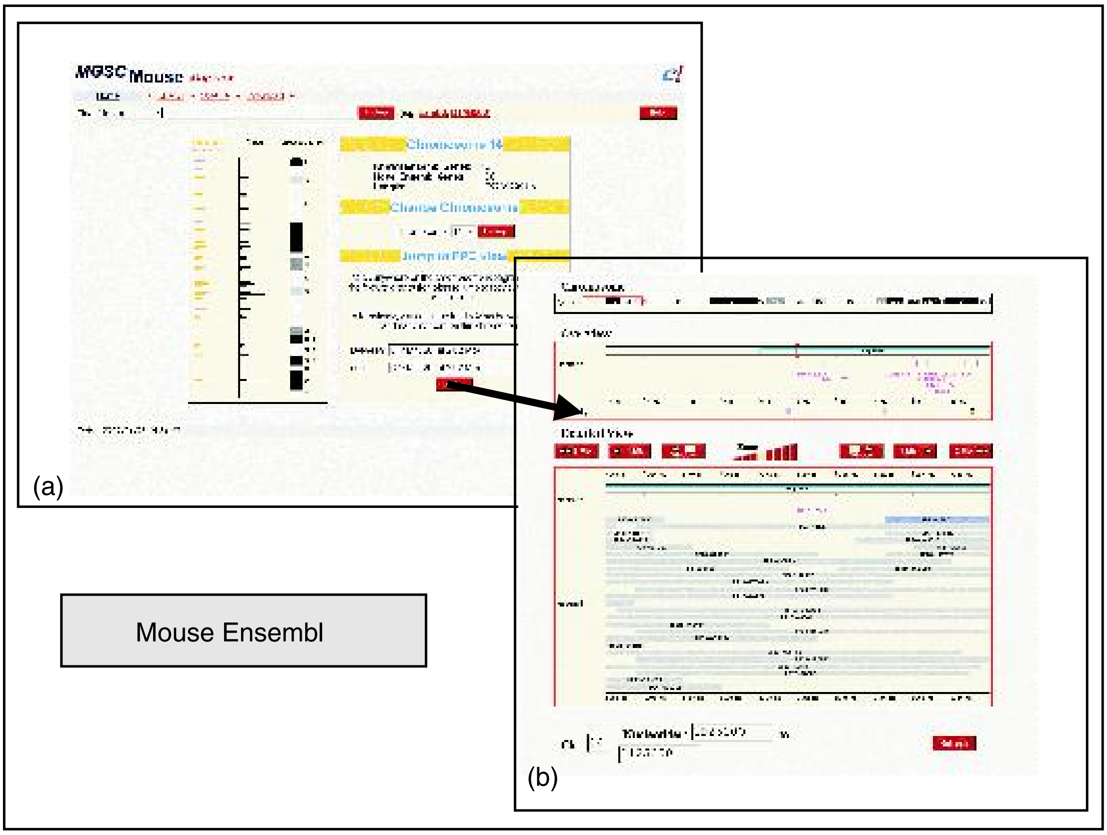


Figure 6.1 Mouse Ensembl. A graphical representation of the clone-based physical map for the proximal end of mouse chromosome 14 from Ensembl. This browser allows users to search for regions of a chromosome between two STS markers and to view the current clone coverage in the selected area. Because the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

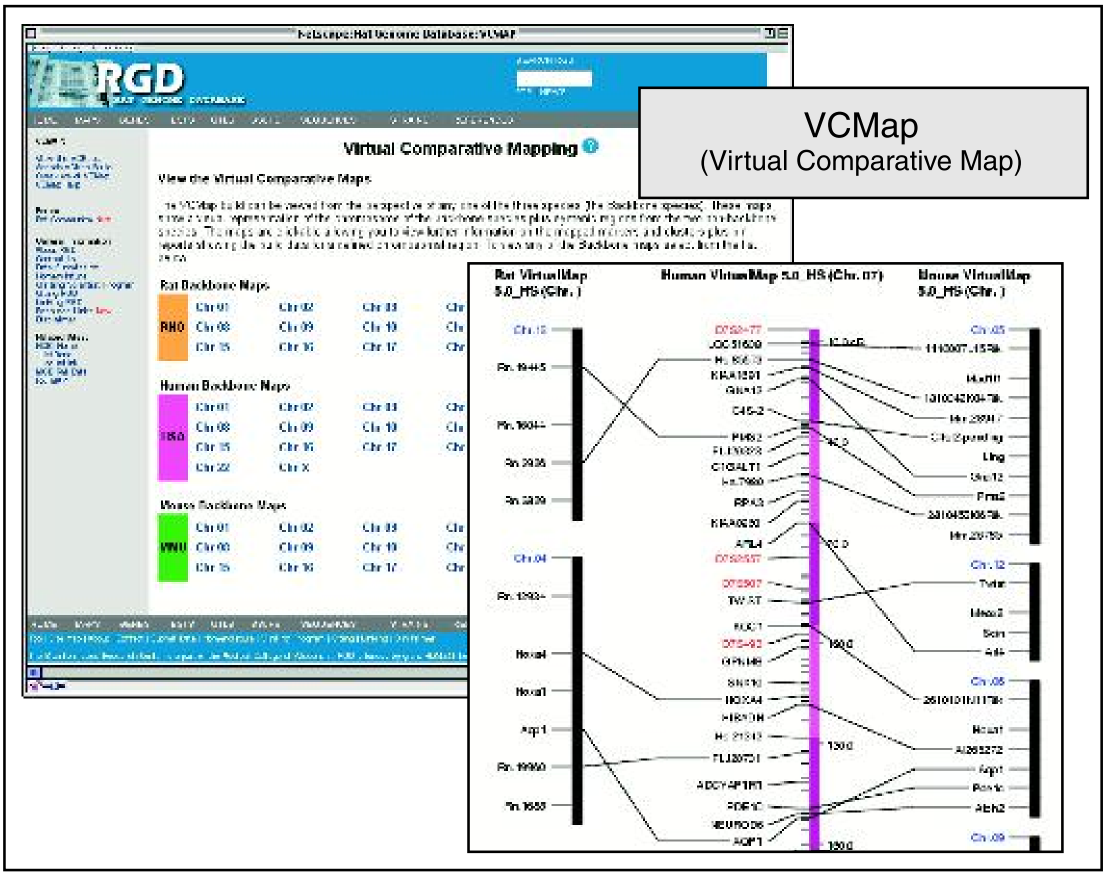


Figure 6.3 Virtual Comparative Map. The Virtual Comparative Map is generated using sequence-based algorithms that predict syntenic regions inferred from homology among mapped sequences. Sequence comparisons between ESTs and cDNAs from human, mouse and rat are combined with Radiation Hybrid map locations 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/.

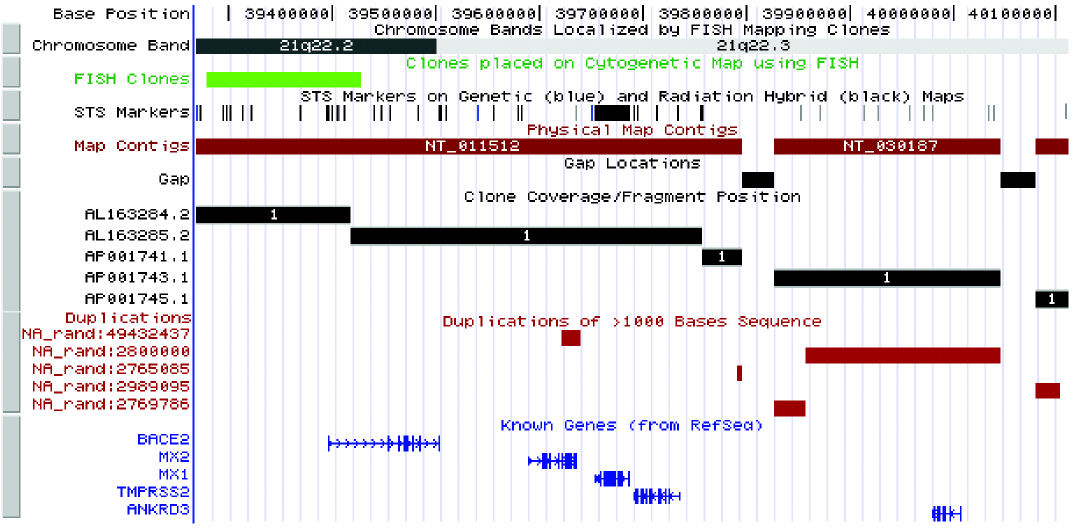


Figure 9.3 Definition of the D21S1245 and D21S1852 interval in genomic sequence using the UCSC human genome browser. The view immediately identifies five known genes, two gaps and several duplications across the region.

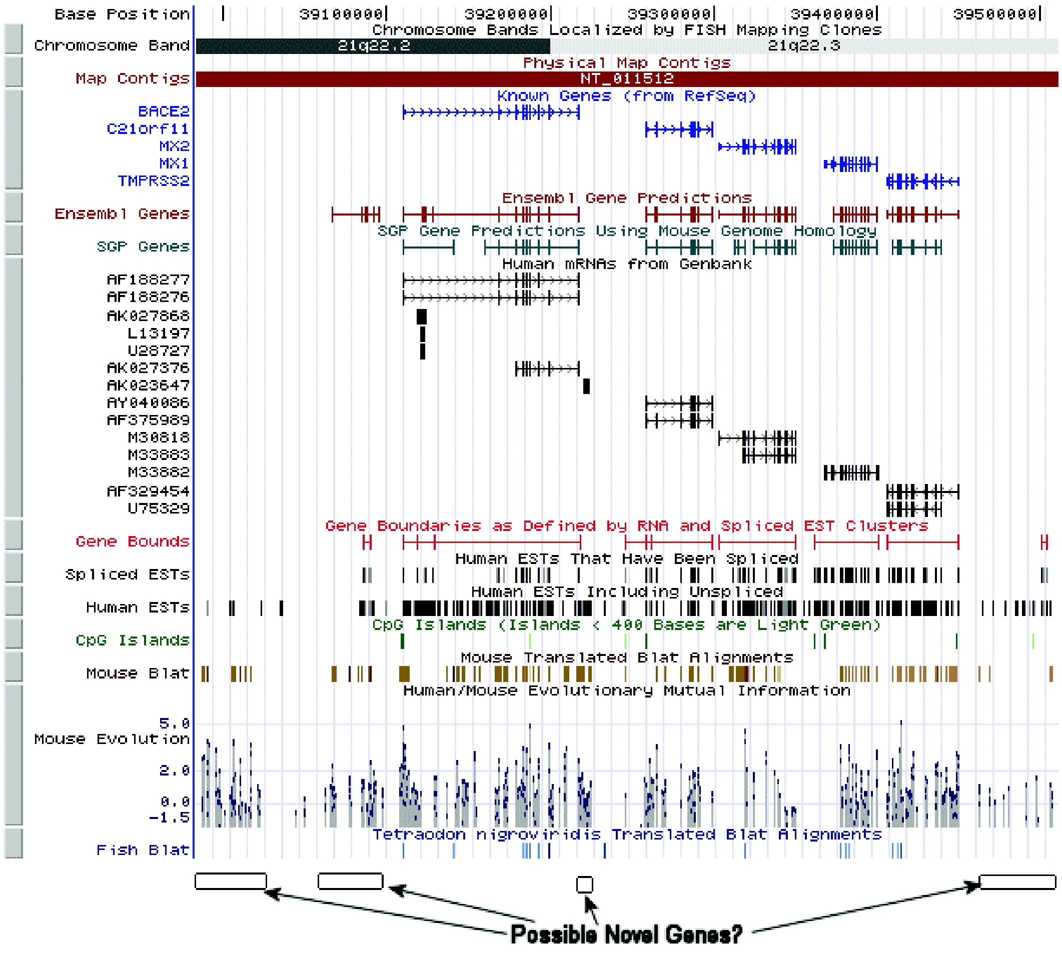


Figure 9.5 Using the UCSC human genome browser to identify known and novel genes. A range of evidence including mRNAs, ESTs and human–mouse homology supports the existence of five known genes and up to four novel genes.

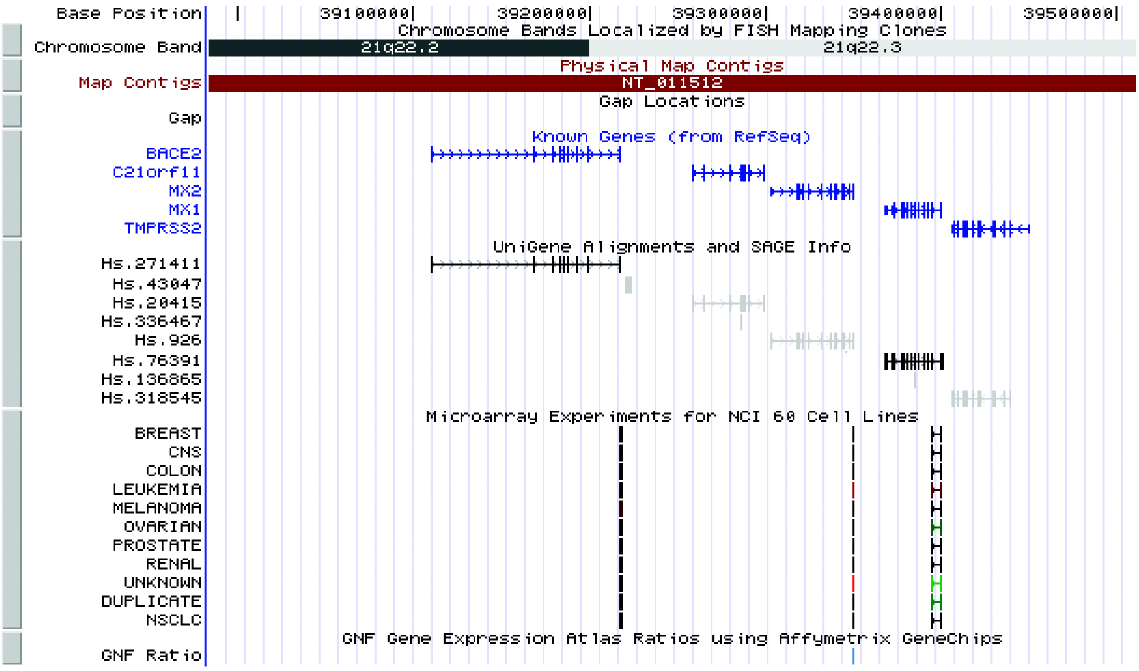


Figure 9.6 Using the UCSC human genome browser to evaluate gene expression across a locus. Four tracks provide information on gene expression. ESTs implicitly measure gene expression as each is derived from a specified tissue source. Unigene clusters link to SAGE expression profiles drawn from the SageMap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

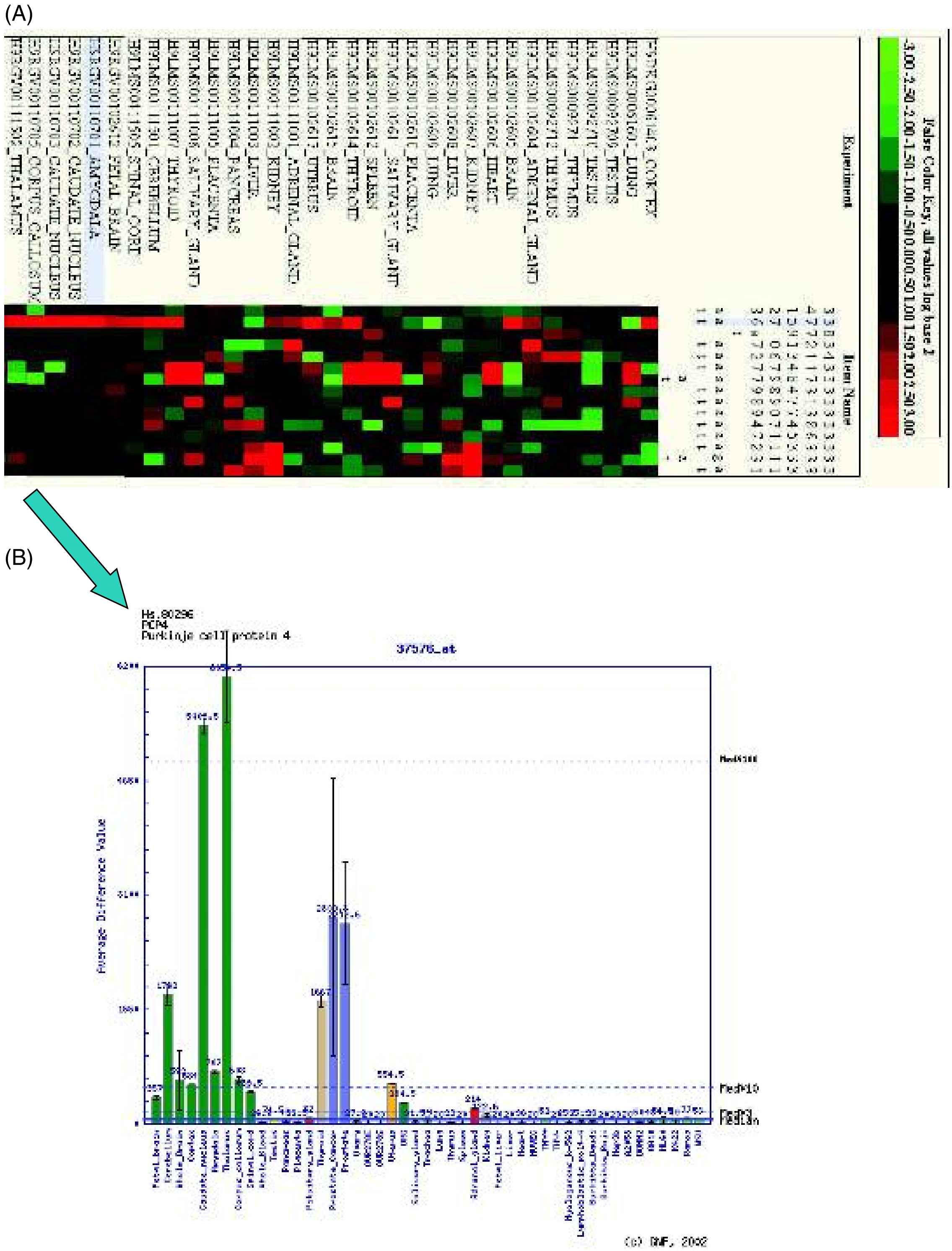


Figure 9.7 GNF gene expression atlas ratios displayed by the UCSC human genome browser. (A) The browser shows a view of the expression profiles of 15 genes across the wider locus. One gene, Purkinje cell protein 4 (PCP4, no. 37576), shows high expression in a wide range of neuronal tissues, including the thalamus. (B) Detailed gene expression profile for the PCP4 gene.

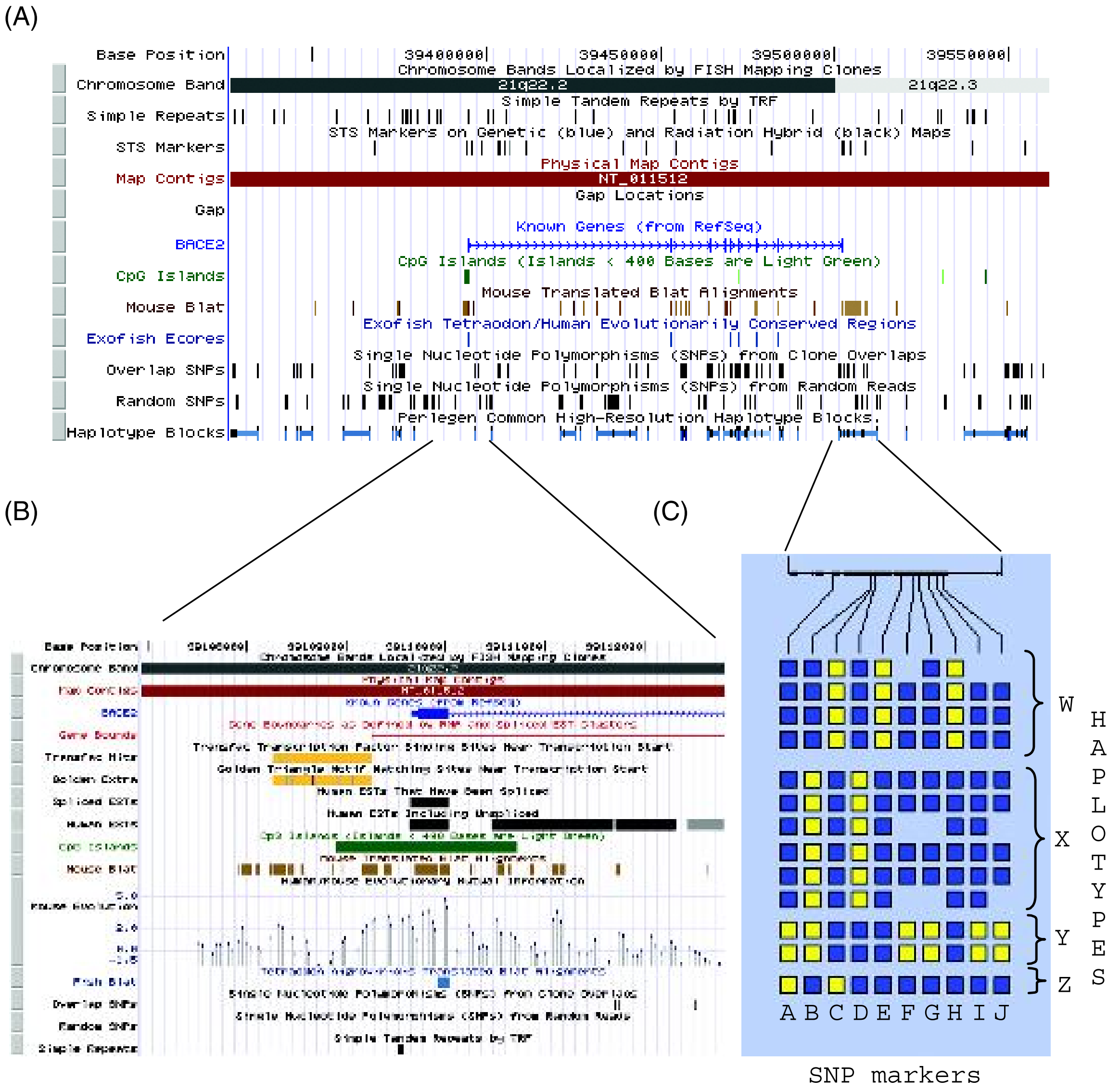

(A)
Base Position
39400000
39450000
39500000
39550000
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
21q22.2
21q22.3
Simple Tandem Repeats by TRF
Simple Repeats
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
Physical Map Contigs
Map Contigs
NT_011512
Gap Locations
Gap
Known Genes (from RefSeq)
BACE2
CpG Islands (Islands < 400 Bases are Light Green)
CpG Islands
Mouse Translated Blat Alignments
Mouse Blat
Exofish Tetraodon/Human Evolutionarily Conserved Regions
Exofish Ecores
Single Nucleotide Polymorphisms (SNPs) from Clone Overlaps
Overlap SNPs
Single Nucleotide Polymorphisms (SNPs) from Random Reads
Random SNPs
Perlegen Common High-Resolution Haplotype Blocks.
Haplotype Blocks
(B)
Base Position
Chromosome Band
Map Contigs
BACE2
Gene Bounds
Transfac Hits
Golden Extra
Spliced ESTs
Human ESTs
CpG Islands
Mouse Blat
Mouse Evolution
Fish Blat
Overlap SNPs
Random SNPs
Simple Repeats
Chromosome Bands Localized by FISH Mapping Clones
Physical Map Contigs
Known Genes (from RefSeq)
Transfac Transcription Factor Binding Sites Near Transcription Start
Golden Triangle Motif Matching Sites Near Transcription Start
Human ESTs That Have Been Spliced
Human ESTs Including Unspliced
CpG Islands (Islands < 400 Bases are Light Green)
Mouse Translated Blat Alignments
Human/Mouse Evolutionary Rate Information
Tetraodon Nigroviridis Translated Blat Alignments
Single Nucleotide Polymorphisms (SNPs) from Clone Overlaps
Single Nucleotide Polymorphisms (SNPs) from Random Reads
Simple Tandem Repeats by TRF
5.0
2.0
0.0
-1.5
(C)
W
X
Y
Z
HAPLOTYPES
A B C D E F G H I J
SNP markers

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.