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)

SECTION 2

THE IMPACT OF COMPLETE GENOME SEQUENCES ON GENETICS

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)

CHAPTER 5

Assembling a View of the Human Genome

COLIN A. M. SEMPLE

Bioinformatics
MRC Human Genetics Unit
Edinburgh EH4 2XU
UK

5.1 INTRODUCTION

The miraculous birth of the draft human genome sequence took place against the odds. It was only made possible by parallel revolutions in the technologies used to produce, store and analyse the sequence data and by the development of new large-scale consortia to organize and obtain funding for the work (Watson, 1990). The initial flood of sequence has subsided as the sequencing centres begin the task of converting the fragmented draft sequences into a finished, complete sequence for each chromosome. The steady progress of the cloned fragments of the human genome towards a finished state can be observed in the Genome Monitoring Table (Beck and Sterk, 1998; http://www.ebi.ac.uk/genomes/mot/), but although we can examine the sequences in public databases we have yet to comprehensively interpret them. There is a need to relate the raw sequence data to what we already know about human genetics and biology in general, this is the process of genome annotation. Preliminary annotation of a genome is a semi-automated process, with human curators interpreting the results of various computer programs. In practical terms, preliminary annotation currently consists of determining the position of known markers, known genes and repetitive sequence in combination with efforts to delineate the structure of novel genes. Eventually we would like to know much more, including the multifarious interactions of the genome's contents with one another and the environment, their expression in the biology of the cell and role in human physiology. These additional layers of annotation will come from the patient laboratory work of the next several decades but a prerequisite for this work is a complete (or nearly complete) genome sequence and an accurate preliminary annotation which is available to the total scientific community. This chapter will aim to describe the sources of freely available annotation, their strengths, their shortcomings and some likely future developments. All websites referred to in the text are listed in Table 5.1.

TABLE 5.1 The Websites Referred to in the Text

Site Description	URL
Genomic sequence assemblies	
CG Human Genome Assembly	http://public.celera.com
NCBI Human Genome Assembly	http://www.ncbi.nlm.nih.gov/genome/guide/human/
UCSC Human Genome Assembly	http://genome.ucsc.edu/
Annotation browsers	
Ensembl at EBI/Sanger Institute	http://www.ensembl.org/
Genome Channel at ORNL	http://compbio.ornl.gov/channel/
Human Genome Browser at UCSC	http://genome.ucsc.edu/
Map Viewer at NCBI	http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map_search
Data sources	
ArrayExpress at EBI	http://www.ebi.ac.uk/arrayexpress/
COGs database at NCBI	http://www.ncbi.nlm.nih.gov/COG/
dbSNP at NCBI	http://www.ncbi.nlm.nih.gov/SNP/index.html
DOTS at University of Pennsylvania	http://www.allgenes.org/
FlyBase at EBI	http://fly.ebi.ac.uk:7081/
Genome Monitoring Table at EBI	http://www.ebi.ac.uk/genomes/mot/
GEO at NCBI	http://www.ncbi.nlm.nih.gov/geo/
IHGMC FPC map at Washington University in St Louis	http://genome.wustl.edu/cgi-bin/ace/GSCMAPS.cgi?

TABLE 5.1 (*continued*)

Site Description	URL
InterPro at EBI	http://www.ebi.ac.uk/interpro/
Mouse Genome Database at Jackson Laboratory	http://www.informatics.jax.org/
Mouse Atlas Database at MRC Human Genetics Unit	http://genex.hgu.mrc.ac.uk/
OMIM at NCBI	http://www.ncbi.nlm.nih.gov/Omim/
Pfam at Sanger Institute	http://www.sanger.ac.uk/Software/Pfam/
Proteome Analysis at EBI	http://www.ebi.ac.uk/proteome/
RefSeq at NCBI	http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html
Saccharomyces Genome Database at Stanford University	http://genome-www.stanford.edu/Saccharomyces/
UniGene at NCBI	http://www.ncbi.nlm.nih.gov/UniGene/
Software	
ACEDB (Sanger Institute)	http://www.acedb.org/
Acembly (NCBI)	http://www.ncbi.nih.gov/IEB/Research/Acembly/help/AceViewHelp.html
Apollo (EBI)	http://www.ensembl.org/apollo/
BLAST (NCBI)	http://www.ncbi.nlm.nih.gov/BLAST/
BLAT (UCSC)	http://genome.ucsc.edu/cgi-bin/hgBlat?command=start)
DAS (Cold Spring Harbor Laboratory)	http://biodas.org/
EMBOSS (EMBnet)	http://www.uk.embnet.org/Software/EMBOSS/
Exofish (Genoscope)	http://www.genoscope.cns.fr/externe/tetraodon/
ePCR (NCBI)	http://www.ncbi.nlm.nih.gov/genome/sts/epcr.cgi
Fgenesh (Sanger Institute)	http://genomic.sanger.ac.uk/gf/Help/fgenesh.html
Gene Ontology Consortium	http://www.geneontology.org/
GENEWISE (Sanger Institute)	http://www.sanger.ac.uk/Software/Wise2/
GENSCAN (MIT)	http://genes.mit.edu/GENSCAN.html
GrailEXP (ORNL)	http://compbio.ornl.gov/grailexp/
HMMER (WUSTL)	http://hmmer.wustl.edu/
NIX at (HGMPRC)	http://www.hgmp.mrc.ac.uk/Registered/Webapp/nix/
Phrap (University of Washington)	http://bozeman.genome.washington.edu/index.html
RepeatMasker (Uni. of Washington)	http://ftp.genome.washington.edu/RM/RepeatMasker.html
SIM4 (Penn State University)	http://bio.cse.psu.edu/
Spidey (NCBI)	http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/
SSAHA (Sanger Institute)	http://www.sanger.ac.uk/Software/analysis/SSAHA/
Twinscan (WUSTL)	http://genes.cs.wustl.edu/

5.2 GENOMIC SEQUENCE ASSEMBLY

Any discussion of computational sequence annotation should begin with a consideration of the sequence data itself. Genomic sequence data has traditionally come from many sources: studies of transcribed sequences, individual genes and genetic/physical markers from mapping studies. Over the past decade we have entered the era of large-scale efforts to sequence entire genomes and the most abundant sources of sequence have become the sequencing vectors from these efforts. In practical terms this has meant that we acquire many fragments, from a few hundred bases to a few hundred kilobases in length, of a genome which must then be assembled computationally to produce a continuous sequence. In the case of the human genome, two unfinished 'draft' sequences have been produced using different methods, one by the International Human Genome Sequencing Consortium (IHGSC) and one by Celera Genomics (CG).

The IHGSC began with a BAC (bacterial artificial chromosome) clone-based physical map of the genome (IHGSC, 2001). This map was constructed by digesting each clone with restriction enzymes and deriving a characteristic pattern or fingerprint. All of the fingerprints are then processed by a program called FPC (Soderlund *et al.*, 2000) which produces BAC clone contigs on the basis of the shared fragments in their fingerprints (International Human Genome Mapping Consortium (IHGMC), 2001; http://genome.wustl.edu/cgi-bin/ace/GSCMAPS.cgi?). A selection of clones from this map covering the vast majority of the genome, were then 'shotgun sequenced' (Sanger *et al*, 1982). The fragments of each clone were then assembled into initial sequence contigs based upon overlaps between shotgun sequencing reads. The collection of initial sequence contigs from a single clone, make up the sequence data for a BAC clone in GenBank. As more shotgun sequencing of the clone is carried out, the initial sequence contigs are re-assembled with the new sequences and the database sequence entry for the clone is updated accordingly. Gradually the initial sequence contigs increase in length and decrease in number, until the sequence of the clone is finished and is represented by a single contig 100–200 kb in length. The program used to assemble the initial sequence contigs is called Phrap (Green, unpublished data; http://bozeman.genome.washington.edu/index.html) and takes sequencing quality estimates for each base into account. CG used the whole-genome shotgun method where the entire genome is randomly fragmented and each of the cloned fragments is sequenced (Venter *et al.*, 2001). Sequences from these cloned fragments are produced as mate-pairs: 150–800 bp sequencing reads from either end of the clone with known relative orientation and approximate spacing. A mixture of clones of different sizes was used: 2, 10, 50 and 100 kb. CG assembled their sequence data with that produced by the IHGSC and published an analysis of this early CG draft genome assembly (Venter *et al.*, 2001). Sequences from this assembly are available, under a variety of restrictions, from the CG draft genome publication site (http://public.celera.com), however the CG raw sequencing data and subsequent versions of the CG draft genome assembly are not publicly available. In spite of the differences between the two efforts to sequence the human genome, both groups had to address the fundamental problem of assembling incomplete data. In both cases the strategy was broadly to merge overlapping sequences into contigs and then to order contigs relative to one another using various types of mapping data.

The published IHGSC assembly was produced using a program called 'GigAssembler' devised at the University of California at Santa Cruz (UCSC) (Kent and Haussler, 2001). GigAssembler began with initial sequence contigs from GenBank at a given point (a 'freeze' dataset). All sequences were repeat masked using the RepeatMasker program (Smit and Green, unpublished data; http://ftp.genome.washington.edu/RM/RepeatMasker.html) to highlight known repetitive sequence. Within each IHGMC physical map contig (IHGMC, 2001) the initial sequence contigs from BAC clones belonging to it were assembled into consensus 'raft' sequences using sequence overlaps between fragments. The first joins were made between the best matching fragments. These rafts were ordered and orientated relative to one another using bridging sequences from other sources (mRNA, EST, plasmid and BAC end pairs) and FPC contig data. For instance the 5′ end of a single mRNA may be found within one raft while the 3′ end matches another raft. Repeated tracts of the letter 'N' were inserted between rafts to give a sequence for each IHGMC map contig. The published version of the UCSC assembly and all subsequent versions are freely available online (http://genome.ucsc.edu/).

The CG draft genome assembly was carried out by a program described as a 'compartmentalized shotgun assembler' (CSA) (Huson *et al*, 2001) using both CG sequence data and IHGSC initial sequence contigs from GenBank (as of 1 September 2000 for the published CG assembly) fragmented into smaller sequences a few hundred base pairs long. The CSA began by comparing all CG mate-pair fragments with all the initial sequence contig fragments and avoiding matches based upon repetitive sequence. Repetitive sequence was identified using comparisons to a library of known repeats (analogously to RepeatMasker) but also by additional procedures to detect sequence likely to represent unknown repeat sequences. The mate-pair fragment pairs matching more than one initial sequence contigs were then used as bridging sequences to order and orientate the initial sequence contig fragments within and between BAC clones. Essentially the paired CG fragments are used as high resolution mapping data to re-assemble both IHGSC BAC sequences and the broader genomic regions they originate from. The result was a set of 'scaffolds' consisting of ordered, oriented sequence contigs separated by gaps of estimated sizes. CG fragments not matching IHGSC initial sequence contigs were also assembled using a different algorithm (Myers *et al.*, 2000) to give additional scaffolds containing sequence not represented in IHGSC data. Scaffolds were then positioned relative to one another based upon sequence overlaps and bridging mate-pair fragments. The derived order of scaffolds was then manually curated to identify mistakes by examining sequence alignments by eye and confirming or rejecting orders based on external physical mapping data such as those from the IHGMC.

A third assembly method, using repeat masked data from the IHGSC, was produced by the National Centre for Biotechnology Information (NCBI) using a computational protocol (NCBI, unpublished data; http://www.ncbi.nlm.nih.gov/genome/guide/build.html) based upon the BLAST algorithm (Altschul *et al.*, 1997). The NCBI approach also began by finding an order for adjacent BACs but in this case it was derived from BAC sequence overlaps (detected using a variant of BLAST), fluorescence *in situ* hybridization (FISH) chromosome assignment and STS content. The sequence fragments from these overlapping BACs were then merged into consensus 'meld' sequences. As with the UCSC method, these melds were then ordered and orientated based on ESTs, mRNAs and paired plasmid reads before being combined into a single NCBI genomic sequence contig with melds separated by runs of the letter 'N'. NCBI contigs were ordered and oriented relative to one another according to matches to mapped STS markers and paired BAC end sequences.

The assembly protocols used by UCSC, CG and NCBI differ in terms of the amount and variety of input data and the algorithms used; it would therefore be surprising if they gave identical assemblies as output. Of particular interest are the relative rates of misassembly (sequences assembled in the wrong order and/or orientation) and the relative coverage achieved by the three protocols. Unfortunately the UCSC group are alone in having published assessments of the rate of misassembly in the contigs they produce. Using artificial datasets they found that on average □10% of assembled fragments were assigned the wrong orientation and □15% of fragments were placed in the wrong order by their protocol (Kent and Haussler, 2001). Two independent assessments of UCSC assemblies have come to similar conclusions. Katsanis *et al.* (2001) examined various UCSC consecutive draft genome assembly releases and reported that 10–15% of EST sequences identified within them appeared to be on wrongly assembled genomic sequences. In agreement with this, Semple *et al.* (2002) observed 19 and 11% of erroneously ordered marker sequences in two consecutive UCSC assemblies for a □5.8 Mb region of chromosome 4. The latter study also found wide variation in coverage (23–59% of the available IHGSC sequence data included) and rates of misassembly (2.08–4.74 misassemblies per Mb) between consecutive UCSC and NCBI assemblies and the published CG assembly for the same region. These analyses indicate that the lowest rate of misassembly is produced by the CG protocol, followed by the UCSC and lastly the NCBI protocols. However, the CG protocol also produced the lowest coverage, including only around half the sequence data recruited into the UCSC and NCBI assemblies. Olivier *et al.* (2001) compared orders of TNG radiation hybrid map STSs produced by UCSC and CG protocols. They found widespread differences, such that 36% of TNG STS pairs were present in orders that differed between UCSC and CG assemblies. The TNG order was consistent with the CG assembly order slightly more often than with the UCSC assembly order. The UCSC website provides a variety of comparisons of its assemblies to genetic, physical and cytogenetic mapping data and these comparisons represent a useful resource for users to assess the likely degree of misassembly in a region of interest.

Unsurprisingly, it has been shown that differences between assemblies do indeed result in differences in annotation. Semple *et al.* (2002) found variable amounts of tandemly duplicated and interspersed repeat sequence between UCSC, NCBI and CG derived assemblies and more striking differences in annotation were also identified by Hogenesch *et al.* (2001) between CG and UCSC assemblies. Hogenesch *et al.* (2001) found large differences between the genes found in CG and UCSC assemblies, such that more than one-third of the genes identified in one assembly were not found in the other. Thus, genomic sequence annotation can only be as good as the underlying genomic sequence assembly and, as we have seen, accurate assembly of draft sequence fragments is far from error free.

The human genome is widely reported to be due for completion in 2003 but at the moment around one-quarter of publicly available human genome sequence is still categorized as ‘draft’ or unfinished. Relatively small, problematic regions of gapped draft sequence may well persist beyond 2003, since certain regions of the genome are simply not present within existing clone libraries and are also recalcitrant to subcloning (Hattori *et al.*, 2000). Specialized technologies are required to close such gaps in the clone map. It therefore seems likely that draft assemblies of some small regions of the human genome will be with us for some time to come. Also a fraction of the genome (perhaps 5%) consists of large (>10 kb) duplicated segments which share 90–98% sequence identity. Regions containing such duplicated segments are notoriously difficult to assemble accurately and are not only found in pericentromeric and subtelomeric regions but also across the rest of the genome, including the gene-rich regions that sequence annotators are primarily interested in (Eichler, 2001). A comparison of the completed sequence of chromosome 20 with the preceding public CG and UCSC draft assemblies of the same chromosome identified 'major discrepancies' (Hattori and Taylor, 2001). These authors concluded that the draft assemblies were probably confounded by large duplicated regions.

5.3 ANNOTATION FROM A DISTANCE: THE GENERALITIES

If some troublesome regions of the genome are set to continue as problems for cloning, sequencing and assembling, this is a minor concern in comparison to the comprehensive annotation of genomic sequence. At almost every level, computational annotation of genomic sequence is error prone and incomplete. Of course, the aim of computational annotation in common with much of bioinformatics, is to provide a preliminary set of predictions that must then be tested by 'wet' laboratory work. The aim is a rapid first pass or 'base line' annotation as the most comprehensive genomic annotation resource Ensembl (Hubbard *et al.*, 2002) puts it. From the computational point of view this enterprise is hugely successful: merely by considering the statistical qualities of the raw sequence data we can detect the presence of most protein-coding human genes. We can then identify the presence of known, structural domains within the conceptually translated products of these predicted genes and make informed guesses about functional roles and subcellular localization. When one looks at a raw BAC sequence entry from GenBank it is easy to appreciate the scale of these achievements but the view from the wet laboratory bench can be different. The broad success of computational gene prediction is little consolation to the bench geneticist who has to sift through numerous artifactual exon predictions only to find later that his gene of interest was not detected by any of the algorithms used. What is broadly impressive to the bioinformaticist can be just plain wrong to those dealing with specifics. In a recent excellent introduction to genomic sequence annotation Lincoln Stein has defined three, hierarchical levels of annotation: the most fundamental nucleotide level, followed by protein level and then process level (Stein, 2001).

5.3.1 Nucleotide Level

Nucleotide level is the point at which the raw genomic sequence is analysed and forms the basis for subsequent levels of interpretation. The first step is to identify as many known genomic landmarks as possible; these are generally markers from previous mapping studies, repeats and known genes already in public databases. This can be done quickly and accurately by a variety of programs. Markers from previous genetic, physical and cytogenetic maps are placed upon the genomic sequence by algorithms designed to find short, almost exact sequence matches such as the ePCR program (Schuler, 1997; http://www.ncbi.nlm.nih.gov/genome/sts/epcr.cgi), BLASTN (Altschul *et al.*, 1990), SSAHA (Ning *et al.*, 2001; http://www.sanger.ac.uk/Software/analysis/SSAHA/) and BLAT (Kent, unpublished data; http://genome.ucsc.edu/cgi-bin/hgBlat?command=start). Identifying these markers is essential to allow the genomic sequence to be seen in relation to the previous, pre-genome sequence literature, for example on human disease genetics. The newest type of markers, single nucleotide polymorphisms or SNPs, are also identified in the sequence to facilitate the next generation of disease gene mapping studies. Similar algorithms, extended to incorporate information on gene structure, are used to identify the positions of known mRNAs within the genomic sequence, examples include Spidey (Wheelan *et al*., 2001; http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/), SIM4 (Florea *et al*., 1998; http://bio.cse.psu.edu/) and est2genome which is available from the EMBOSS package (Rice *et al*., 2000; http://www.uk.embnet.org/Software/EMBOSS/). Just as the efforts to assemble genomic sequence take measures to identify and exclude repetitive sequence, an important part of annotation is to identify interspersed and simple repeats. The most widely used program for this task is RepeatMasker.

The central problem of nucleotide-level annotation is the prediction of gene structure. Ideally we would like to correctly delineate every exon of every gene but in large, repeat-rich eukaryotic genomes, liberally scattered with long genes with many exons, this task has turned out to be more difficult than expected. *Ab initio* gene prediction algorithms (that rely only on the statistical qualities of genomic sequence data) identify most protein coding genes reliably in prokaryotic genomes but the task is more complex in eukaryotic genomes (Burge and Karlin, 1998). Fundamentally the problem is gene density, whereas in prokaryotic genomes and yeast more than two-thirds of the genome is protein coding sequence only a few percent of the human genome fits this description. Additional problems are added by overlapping genes, alternatively spliced exons and the paucity of differences between intergenic sequence and introns. The gene prediction literature is full of metaphors involving needles and haystacks, and with good cause. The 13-Mb *S. cerevisiae* yeast genome provides a sobering example, completed in 1996 and initially thought to contain 6274 genes, the sequence has provided a steady trickle of additional genes that had been overlooked. Since publication of the yeast genome a further 202 genes have been discovered, most appear to have been missed because they are relatively short or overlap a previously annotated gene on the opposite strand (Kumar *et al*., 2002). At the same time, new analyses of these yeast sequences using a variety of statistical analyses and comparative genomics approaches have suggested that several hundred of the originally annotated genes may be spurious (Malpertuy *et al*., 2000; Zhang and Wang, 2000).

This brings us to the use of sequence similarity in gene prediction. In practice genome annotators use a combination of information to make predictions of gene structures: *ab initio* exon predictions (predictions of coding sequence made by a program on the basis of statistical measures of features such as codon usage, initiation signals, polyA signals and splice sites), repetitive sequence content and similarity to expressed sequences and proteins. These different strands of evidence are usually combined and evaluated by human annotators who use graphical interfaces, such as those provided by NIX (unpublished data; http://www.hgmp.mrc.ac.uk/Registered/Webapp/nix/) and ACEDB (Eeckman and Durbin, 1995; http://www.acedb.org/), to view all the evidence simultaneously. A recent trend in gene prediction is the design of programs that automatically incorporate evidence based on sequence similarity into their predictions. Among the best and most widely used *ab initio* algorithm is Genscan (Burge and Karlin, 1997; http://genes.mit.edu/GENSCAN.html). Guigo *et al.* (2000) tested its success in artificially produced sequence data designed to mimic human BAC sequences. At the same time they tested algorithms that use sequence similarity to make their predictions, such as GeneWise (Birney and Durbin, 2000; http://www.sanger.ac.uk/Software/Wise2/). The results showed a clear advantage to including evidence from sequence similarity where the similarity was strong. In such cases GENEWISE could correctly identify 98% of coding bases present while generating a comparatively low level of artifactual exons (2%) and missing 6% of real exons. Where levels of similarity were more modest however the performance of algorithms such as GENEWISE declined to below that of GENSCAN. GENSCAN was found to identify 89% of coding bases at the cost of a rather high level of artifactual exons (41%) and 14% of real exons missed. Guigo *et al.* (2000) suggest that the success of all the programs tested is expected to be lower in real genomic sequence. Another comparison of gene prediction programs using *D. melanogaster* genomic sequence identified similar levels of performance for the programs tested and also indicated an advantage to algorithms including similarity-based evidence in predictions (Reese *et al.*, 2000). Shortcuts to the structures of many genes may come from a large collection of full-length mouse cDNAs (Kawai *e al.*, 2001) and large human cDNA collections (Kikuno *et al.*, 2002), which are expected to grow rapidly over the next few years.

As we amass genomic sequence data from many organisms the reach of computational annotation based upon sequence similarity is increasing. New methods aimed at the prediction of non-coding features in the genome, such as regulatory regions and non-coding RNAs (ncRNAs) are evolving rapidly. Whereas protein coding exons have a distinctive statistical fingerprint ncRNAs do not, or at least they do not appear to from our present, limited knowledge of them (Eddy, 2001). For better understood classes of ncRNAs, such as tRNAs, prediction methods involving secondary structure prediction have been successful (Lowe and Eddy, 1997) but for novel ncRNAs the only effective methods are based on comparative genomics (Rivas *et al.*, 2001). The same is true for novel regulatory sequences, where only a fraction of transcription factor binding sites have been identified to date (Wingender *et al.*, 2001). Even incomplete, fragmentary sequence data from other organisms has been used with some success to predict putative regulatory regions (Chen *et al.*, 2001). This approach is examined in some detail in Chapter 7.

5.3.2 Protein Level

Once we have a gene prediction that we believe, the next step is to assign a possible function to the encoded protein; this is the central task of protein-level annotation. Most computationally assigned functions are derived from sequence similarity. A pair of proteins that align along 60% or more of their lengths with significant similarity (e.g. E <0.01 in a BLASTP search of a large public database) are very likely to be homologous — that is derived from a common ancestor. Such a pair of sister proteins may be paralogues, derived from a duplication event, or orthologues, that exist as a result of a speciation event. For every homologous pair identified in this way additional searches may verify that each member of the pair identifies the other member as the best match within the organism of interest. This makes it likely that the pairs identified are likely to be orthologues (Huynen and Bork, 1998), which is desirable since orthologues are likely to share the same function (Jordan *et al.*, 2001) whereas functional diversification between paralogues is thought to be common (Li, 1997). In most cases this strategy of reciprocal sequence similarity searches to identify orthologues is successful (Chervitz *et al.*, 1998) and is the rationale that underlies the construction of the Clusters of Orthologous Groups of proteins (COGs) database (Tatusov *et al.*, 2000; http://www.ncbi.nlm.nih.gov/COG/). However, caution is necessary when dealing with the results of such analyses. For example, a novel human gene may be directly descended from a common ancestor of a yeast gene (in which case the two genes are orthologues and are likely to share the same function), or it may be descended from a duplicated sister yeast gene (and the two genes are really paralogues) with a different function. Without a complete picture of the related family of proteins we are dealing with, it can be difficult to decide. Definitive evidence for orthology versus paralogy can come from comprehensive phylogenetic analysis but even then, when dealing with larger families and/or incomplete data, it can be difficult. As a result, it is not uncommon to find mistaken computational predictions of function that are not supported by further experiment (Iyer *et al.*, 2001).

In the absence of any detailed knowledge about the evolutionary pedigree of the protein under study, similarity may sometimes still imply functional similarity. For example two proteins only 30% identical may share much of their biochemistry but have different substrates (Todd *et al.*, 2001). In spite of their divergence they may share a common functional domain. There are a variety of protein domain databases and they are widely used in genome annotation. For example, version 7 of the Pfam database contains 3360 domains that match 69% of proteins in public sequence databases, with domains represented by alignments between regions of proteins containing them (Bateman *et al.*, 2002; http://www.sanger.ac.uk/Software/Pfam/). Statistical models of these alignments are constructed and searched using the elegant HMMER software package (Eddy, 1998; http://hmmer.wustl.edu/). The Interpro database (Apweiler *et al.*, 2000; http://www.ebi.ac.uk/interpro/), which amalgamates several databases (including Pfam) covering protein domains, families and functional sites, was used by the IHGSC to provide the publicly available annotation for the draft human genome. Interpro entries provide links to additional information including functional descriptions, references to the literature and structural data. Since the IHGSC draft genome publication, the EBI (European Bioinformatics Institute; http://www.ebi.ac.uk/proteome/) has maintained and updated annotation for the set of known and predicted human proteins using Interpro but their most recent analyses match only around 60% of the set. Thus even our most strenuous efforts to gain clues to protein function, often based upon rather distant homology, tell us nothing about 40% of human proteins.

5.3.3 Process Level

Ultimately the goal of genetics is to understand the relationship between genotype and phenotype. There is a large gap between annotation at the nucleotide or protein level and an understanding of how a given protein influences phenotype. Even in the best case, with a known gene coding for a protein containing well-studied domains, there are always questions that remain to be asked. How does the protein interact or complex with other proteins? Where does it localize within the cell? Which cellular processes and organelles is it involved with? In which tissues and at which developmental stages does it act? The answers to these questions provide process-level annotation. The most important applications of our knowledge about the human genome are in medicine, to discover the variations and aberrations that underlie disease. Process level annotation provides a rational way to select the best candidate genes for involvement in disease. For example, when it was first submitted to GenBank in 1997 a certain gene (accession number U80741) was annotated as '*Homo sapiens* CAGH44 mRNA' and 'polyglutamine rich'. Due to the painstaking work of Lai *et al.* (2001) on a region associated with speech disorders we now know this gene as FOXP2, the first gene found to be involved in human language acquisition disorders. Before their work FOXP2 appeared to be one of many transcription factors, expressed in many tissues and best studied in *D. melanogaster*. With better process level annotation FOXP2 may have been identified earlier as a good candidate for involvement in disease.

The main source of process-level annotation is the scientific literature but, even with modern access through the web, the literature is a 20th century resource unsuited to 21st century needs. What we have is a dizzying array of terms for a single gene, function or process and no accepted way of organizing this information, added to this are all the vagaries and idiosyncrasies of human language. What is needed is a structured resource with a limited number of terms for genes and descriptions of their functions, organized so that it is easily processed automatically by computer programs. A recent initiative, called the Gene Ontology (GO) project has provided a framework to achieve this (Gene Ontology Consortium, 2001; http://www.geneontology.org/). GO consists of an hierarchical set of structured vocabularies to describe the molecular functions, biological processes, and cellular components associated with gene products. With the known and predicted genes in a genome annotated using GO it is possible to quickly retrieve, for example, all genes encoding transmembrane receptors, all genes involved in apoptosis, or all genes encoding products localized to the cytoskeleton. The hierarchical nature of GO means that subsets of these categories can also be retrieved, for example all G-protein coupled receptors within the transmembrane receptor category. GO annotation has already been adopted by databases for several model organism genomes, including the *Saccharomyces* Genome Database (Dwight *et al*., 2002; http://genome-www.stanford.edu/Saccharomyces/), FlyBase (FlyBase Consortium, 2002; http://fly.ebi.ac.uk:7081/) and the Mouse Genome Database (Blake *et al*., 2002; http://www.informatics.jax.org/). At the moment GO annotations are added to genes in these databases manually by trained biologist curators browsing the scientific literature. In the longer term, with the rapidly increasing number of completed genomes, this process will become increasingly automated. Efforts are already underway to develop software that will automatically extract information from the literature to be incorporated into the GO annotation of a gene (Raychaudhuri *et al*., 2002).

The scale of the problem of providing process-level annotation for every human gene is prompting the development of large-scale technologies to generate data on many genes at once. Large-scale parallel measurement of gene expression for entire genomes is now possible and should give good data on the developmental timing and tissue specificity of many human genes, from which it is possible to infer process-level annotation (Noordewier and Warren, 2001). An important step on the way to designating the processes a protein is involved in, is to define the proteins with which it interacts, and work is well underway to elaborate the web of interacting proteins and complexes that define the *S. cerevisiae* proteome (Gavin *et al*., 2002; Ho *et al*., 2002). However these high-throughput methods are known to generate false positives and negatives; that is they identify some artifactual interactions and miss some genuine interactions. Thus, high-throughput technologies may eventually provide useful process-level annotation for many, if not most, human genes but there will always be an indispensable role for conventional, detailed laboratory studies of smaller scale. New databases and analyses will be necessary to make sense of the network of genetic interactions that underlie the phenotype. A good example is the Mouse Atlas and Gene Expression Database Project (Baldock *et al*., 2001; http://genex.hgu.mrc.ac.uk/) which aims to describe the patterns of gene expression responsible for the emergence of anatomical structure during mouse development. It will enable gene expression data to be viewed in the context of three-dimensional embryo sections.

5.4 ANNOTATION UP CLOSE AND PERSONAL: THE SPECIFICS

Even given the difficulties and shortcomings in computational annotation discussed above, several well-resourced groups have undertaken the task of compiling, maintaining and updating freely accessible annotation for the entire human genome. There are now four well-designed websites offering users the chance to browse annotation of the draft human genome. All four sites offer a graphical interface to display the results of various analyses, such as gene predictions and similarity searches, for draft and finished genomic sequence. These interfaces are indispensable for allowing rapid, intuitive comparisons between the features predicted by different programs. For instance, one can see at once where an exon prediction overlaps with interspersed repeats or an SNP. But the four sites are not equivalent and there are important distinctions between them in terms of the data analysed, the analyses carried out and the way the results are displayed.

5.4.1 Ensembl

Ensembl is a joint project between the EBI and the Sanger Institute (http://www.sanger.ac.uk/). The Ensembl database (Hubbard *et al*., 2002; http://www.ensembl.org/), launched in 1999, was the first to provide a window on the draft genome, curating the results of a series of computational analyses. Until January 2002 (release 3.26.1) Ensembl used the UCSC draft sequence assemblies as its starting point but it is now based upon NCBI assemblies. The Ensembl analysis pipeline consists of a rule-based system designed to mimic decisions made by a human annotator. The idea is to identify 'confirmed' genes that are computationally predicted (by the GENSCAN gene prediction program) and also supported by a significant BLAST match to one or more expressed sequences or proteins. Ensembl also identifies the positions of known human genes from public sequence database entries, using GENEWISE to predict their exon structures. The total set of Ensembl genes should therefore be a much more accurate reflection of reality than *ab initio* predictions alone but it is clear that many novel genes are missed (Hogenesch *et al*., 2001). Of the novel genes that are detected many, if not most are expected to be incomplete for two main reasons. Firstly, as we have seen, while GENSCAN can detect the presence of most genes in a genomic sequence it is substantially less successful in predicting their correct exonic structures (as with other *ab initio* gene predictions). Secondly, any prediction is entirely dependent upon the quality of the genomic sequence and where the sequence is gapped or wrongly assembled the missing exons may not be present for the software to find.

Many other genomic features have been included as Ensembl has developed: different repeat classes, cytological bands, CpG island predictions, tRNA gene predictions, expressed sequence clusters from the UniGene database (Wheeler *et al*, 2002; http://www.ncbi.nlm.nih.gov/UniGene/), SNPs from the dbSNP database (Sherry *et al*., 2001; http://www.ncbi.nlm.nih.gov/SNP/index.html), disease genes found in the draft genome from the OMIM database (On-line Mendelian Inheritance in Man database; Wheeler *et al*., 2002; http://www.ncbi.nlm.nih.gov/Omim/) and regions of homology to mouse draft genomic sequences. GENSCAN-predicted exons that have not been incorporated into Ensembl-confirmed genes may also be viewed. This means that the display can be used as a workbench for the user to develop personalized annotation. For example, one may discover novel exons by finding GENSCAN exon predictions which coincide with good matches to a fragment of the draft mouse genome, or novel promoters by finding matches to the draft mouse genome that occur upstream of the 5′ end of a gene. Once you have identified a gene of interest you can link to a wealth of information at external sites such as the Interpro protein domains it encodes and its expression profile according to the SAGEmap repository (Lash *et al*., 2000). Eventually Ensembl aims to become a platform for studies in comparative genomics and already it is possible while browsing the human genome to jump to an homologous region of the mouse genome via a match to a mouse genomic sequence fragment. Substantial thought and effort has evidently gone into the Ensembl site design. The result is certainly a user-friendly experience, and not just by the standards of computational biology. The web interface to the database achieves the laudable aim of providing seamless access to the human genome. The user can sink down through cytogenetic ideograms of whole chromosomes, to large unfinished sequence contigs several Mb long and then to smaller fragments of individual BAC clones only kb long. Along the way a graphical display shows the relative positions of genes and the other features.

Figure 5.1 shows the Ensembl display for the genomic region around the FOXP2 gene mentioned earlier. The region is shown at three levels of resolution. The upper panel shows the position of the region as a small red box on a cytogenetic ideogram of chromosome 7. The middle panel shows an exploded view of this box, including the structure of the draft genome assembly, the relative positions of various markers and a simple overview of the gene content. The bottom panel gives a detailed view of a subsection (indicated again by a red box) of the middle panel. This detailed view is the business end of the browser

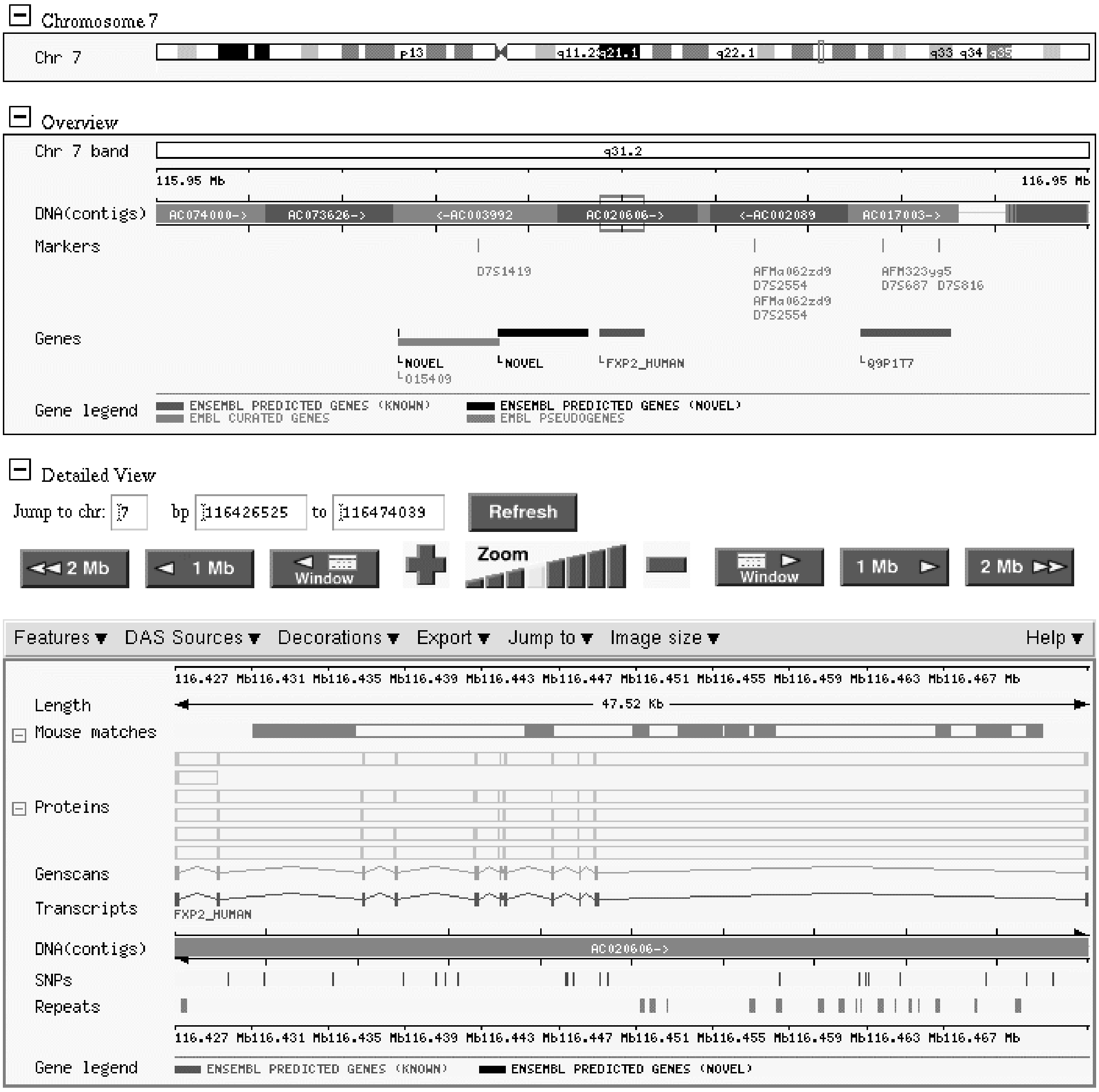


Figure 5.1 The genomic region around the FOXP2 gene according to Ensembl (See Colour Plates).

and is easily customized, via pull-down menus, to display any desired combination of the available features. In Figure 5.1 the combination chosen shows the positions of matches to the mouse genome in relation to GENSCAN-predicted exons and similarities to protein sequences, which allows a user to define non-coding conserved regions that may be of regulatory importance. Using this display one could also select SNPs that lie outside repetitive sequences; an important consideration for PCR-based SNP assays.

Data retrieval is extremely well catered for in Ensembl, with text searches of all database entries, BLAST searches of all sequences archived and the availability of bulk downloads of all Ensembl data and even software source code. Ensembl annotation can also be viewed and added to interactively on your local machine using the Apollo viewer (http://www.ensembl.org/apollo/).

5.4.2 UCSC Human Genome Browser (HGB)

The UCSC Human Genome Browser (HGB) bears many similarities to Ensembl, it too provides annotation of the NCBI assemblies (as well as UCSC assemblies) and it displays a similar array of features, including confirmed genes from Ensembl. The range of features displayed in HGB (and Ensembl) often change between releases but generally there are additional features of HGB that are not found in Ensembl. For example, at the time of writing HGB includes predictions from two *ab initio* gene prediction programs: GENSCAN and Fgenesh (Salamov and Solovyev, 2000; http://genomic.sanger.ac.uk/gf/Help/fgenesh.html). This should help the user to identify false positives (i.e. artifactual exons) from either program and concentrate on exons predicted by both programs that are most likely to be real. HGB also currently indicates regions with significant homology to the mouse genome as in Ensembl but also to the incomplete genome of the pufferfish *Tetraodon nigroviridis*. These HGB-specific features can provide useful information when one is dealing with gene predictions that are not well supported by similarity to expressed sequence. Another useful feature of HGB is the detailed description of the genomic sequence assemblies. Graphical representations of the fragments making up a region of draft genome can be displayed, showing the relative size and overlaps of each fragment and also whether any gaps between fragments are bridged by mRNAs or paired BAC end sequences. This means that one can get an idea of the likely degree of misassembly in a draft region. There is an increasing amount of data becoming available from large-scale gene expression studies and public repositories have emerged for their curation, such as the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) and ArrayExpress at the EBI (http://www.ebi.ac.uk/arrayexpress/). At the moment, the HGB is the only browser which incorporates such data, in the form of data from a microarray study exploring the variation in expression of several thousand genes in a screen for anti-cancer drugs (Ross *et al*., 2000). Undoubtedly the other browsers will develop to include similar data.

In Figure 5.2 the genomic neighbourhood of the FOXP2 gene (represented by sequence U80741) according to HGB (as of 6 August 2001) is displayed. This provides the kinds of information available from the analogous Ensembl display and some interesting additional data. At the top of the display there are indications of the size, cytogenetic band and the genomic sequences corresponding to the region. Further down one can compare an Ensembl predicted transcript (ENST00000265436) and similar NCBI Acembly predictions with the original FOXP2 sequence entry (U80741). Notice that neither the Ensembl nor the Acembly predictions find all the FOXP2 exons that we know are present from U80741, at the same time both *ab initio* prediction algorithms (GENSCAN and Fgenesh) have split the gene into more than one prediction. These are all familiar problems in genomic sequence

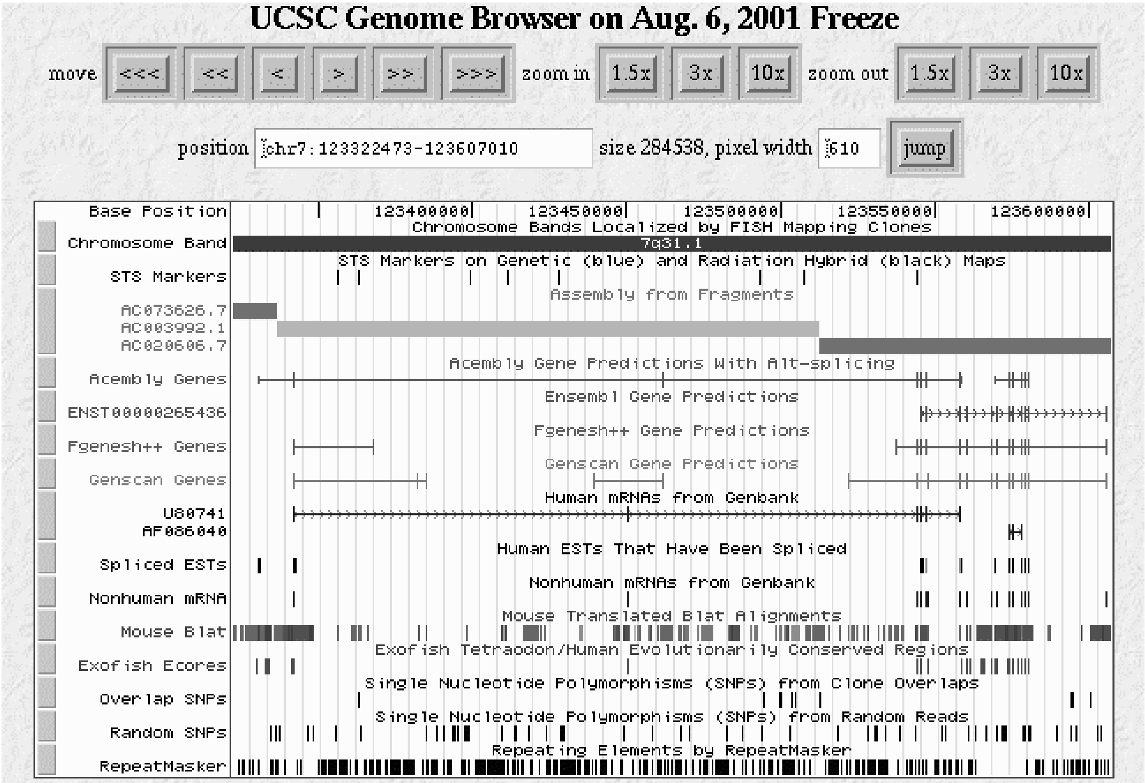

UCSC Genome Browser on Aug. 6, 2001 Freeze
move <<< << < > >> >>> zoom in 1.5x 3x 10x zoom out 1.5x 3x 10x
position chr7:123322473-123607010 size 284538, pixel width 610 jump
Base Position 123400000 123450000 123500000 123550000 123600000
Chromosome Band
Chromosome Bands Localized by FISH Mapping Clones
7q31.1
STS Markers
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
Assembly from Fragments
AC073626.7
AC003992.1
AC020606.7
Acembly Gene Predictions With Alt-splicing
Acembly Genes
Ensembl Gene Predictions
ENST00000265436
Fgenesh++ Gene Predictions
Fgenesh++ Genes
Genscan Gene Predictions
Genscan Genes
Human mRNAs from Genbank
U80741
AF086040
Human ESTs That Have Been Spliced
Spliced ESTs
Nonhuman mRNAs from Genbank
Nonhuman mRNA
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
Repeating Elements by RepeatMasker
RepeatMasker

Figure 5.2 The genomic region around the FOXP2 gene according to the UCSC Human Genome Browser.

annotation. Notice also that the Ensembl prediction has a number of additional exons 3′ of the last U80741 exon. This is because U80741 does not contain the full coding sequence of FOXP2 and the Ensembl prediction is based upon a later sequence entry (AF337817) which does. This illustrates another common problem: different annotation sources may be based upon different sequence data, depending on what is available at the time. As with Ensembl, the HGB display of the region shows regions of homology to the mouse genome but also to the pufferfish genome (identified by a program called Exofish, see http://www.genoscope.cns.fr/externe/tetraodon/). It is apparent that the evolutionary distance between humans and fish means that the Exofish results are more helpful in defining exons rather than regulatory regions. However there are still regions upstream of the first U80741 exon that appear to be well conserved across the human, mouse and pufferfish genomes. Such regions may define the promoter of the FOXP2 gene.

Data retrieval is facilitated by text, BLAT (a faster, less sensitive algorithm than BLAST) searches and bulk downloads of annotation or sequence data. As with Ensembl, the HGB website has been well designed and is sympathetic to the naive user, but the HGB graphical interface is more Spartan. If Ensembl is Disney then HGB is Southpark. The positive side of this is that HGB will usually display a region on your local web browser more quickly than Ensembl can. Both the Ensembl and HGB interfaces offer users the ability to jump between their respective views of a region and so, when they are both annotating the same version of the same NCBI assembly, they can easily be used as complementary resources.

5.4.3 NCBI Map Viewer (NMV)

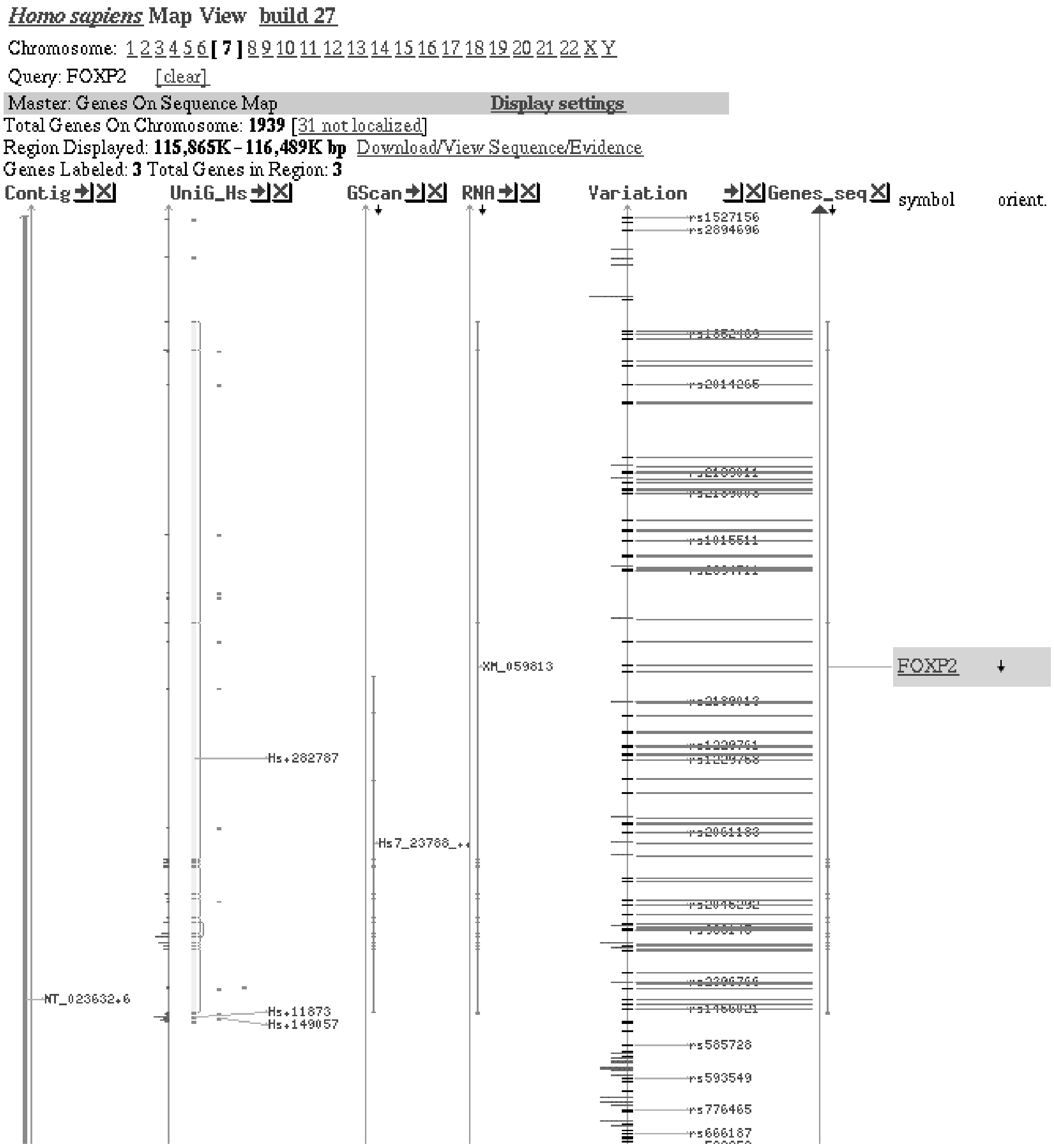
As the human genome nears completion the problems of dealing with draft sequence data will recede and the main task will be to curate the finished sequences representing each chromosome. This task will be undertaken at the NCBI. Whereas Ensembl and HGB both previously provided annotation of the UCSC draft genome assemblies the NCBI Map Viewer (NMV; http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map_search) has always displayed features present in the NCBI assemblies. As the name suggests, the NMV shows useful comparisons between a wide range of cytogenetic, genetic and radiation hybrid maps in parallel with NCBI draft and finished sequence contigs. The locations of genes, markers, and SNPs are indicated on the contig sequences. As with Ensembl, there is an analysis protocol which aims to predict gene structures based upon EST and mRNA alignments with the draft genome. This is carried out by a program called Acembly (unpublished data; http://www.ncbi.nih.gov/IEB/Research/Acembly/help/AceViewHelp.html) which aims to derive gene structure from these alignments alone. The program also attempts to give alternative splice variants of genes where its alignments suggest them. These gene structures and transcripts end up as records in the NCBI RefSeq database, which is slowly compiling a non-redundant curated dataset representing current knowledge of known genes (Wheeler *et al.*, 2002; http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html). Like the Ensembl protocol many Acembly-predicted structures (the NCBI estimate 42%) are incomplete. These structures can be displayed alongside *ab initio* gene models predicted by GenomeScan (a variant of GENSCAN) and matching UniGene clusters to allow users to make their own assessments about the likeliest gene structure.

Figure 5.3 shows the FOXP2 gene as it appears in the NMV which shows features on a vertical rather than horizontal display. The genomic sequence contig the gene occurs on (NT_023632) is shown in the leftmost column, followed by BLAST matches to three UniGene expressed sequence clusters. This gene is typical in having more than one UniGene cluster representing it, particularly at the $3'$ end as ESTs are more commonly sequenced from the $3'$ ends of mRNAs. In the next columns are a GenomeScan prediction which misses some exons and a depiction of XM_059813: the model of FOXP2 that Acembly has constructed by aligning expressed sequences with this region of the genome. SNPs from the NCBI dbSNP database are also displayed with those occurring within the gene highlighted, however there is no indication of repetitive sequence. In the rightmost column the FOXP2 gene structure is displayed according to the XM_059813 model.

The NMV offers tabulated downloads of data and it is possible to BLAST search the NCBI assembly (via the NCBI BLAST site: http://www.ncbi.nlm.nih.gov/BLAST/) and view the matching regions using the NMV. All annotated genes are connected to NCBI LocusLink which provides links to associated information such as related sequence accession numbers, expression data, known phenotypes and SNPs.

5.4.4 ORNL Genome Channel (GC)

The ORNL (Oak Ridge National Laboratory) Genome Channel (GC; http://compbio.ornl.gov/channel/) consists of a series of tools for visualizing and querying the NCBI human genome sequences and those of other organisms assembled and annotated by ORNL and collaborators. The GC browser provides the usual categories of nucleotide-level annotation: repetitive sequences, CpG islands, polyA sites and marker positions. The GC gene prediction protocol is pitched somewhere between Ensembl and HGB: GrailEXP (Uberbacher *et al.*, 1996; http://compbio.ornl.gov/grailexp/) and GENSCAN predictions are given where they are supported by BLAST matches to expressed sequence along with known genes from RefSeq or GenBank entries. Sequence similarity results are not viewable as independent features (as in the other browsers), only as evidence associated with predicted exons. This is rash considering the number of coding sequences missed by *ab initio* algorithms and unhelpful where one is interested in non-coding regions such as UTRs.



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 (See Colour Plates).

The only kind of sequence similarity results displayed independently are gene predictions derived from transcripts from the Database of Transcribed Sequences (DoTS; unpublished data; http://www.allgenes.org/) which clusters and assembles expressed sequences. On the platforms I tested (Netscape running in UNIX and Microsoft Internet Explorer in Windows NT), the graphical display itself also has a problem: several features (different classes of repeats, CpG islands and polya sites) appear on top of one another, which makes it difficult to see what is going on. On the positive side GC does allow users to submit their own sequences to the suite of BLAST searches and gene prediction programs underlying the GC analysis pipeline. None of the other sites allow this. Downloads of genomic DNA and the mRNA and peptide sequences for the predicted genes in GC are available. The GC browser's view of the FOXP2 gene and flanking regions is provided in Figure 5.4. The central horizontal band displays the clones making

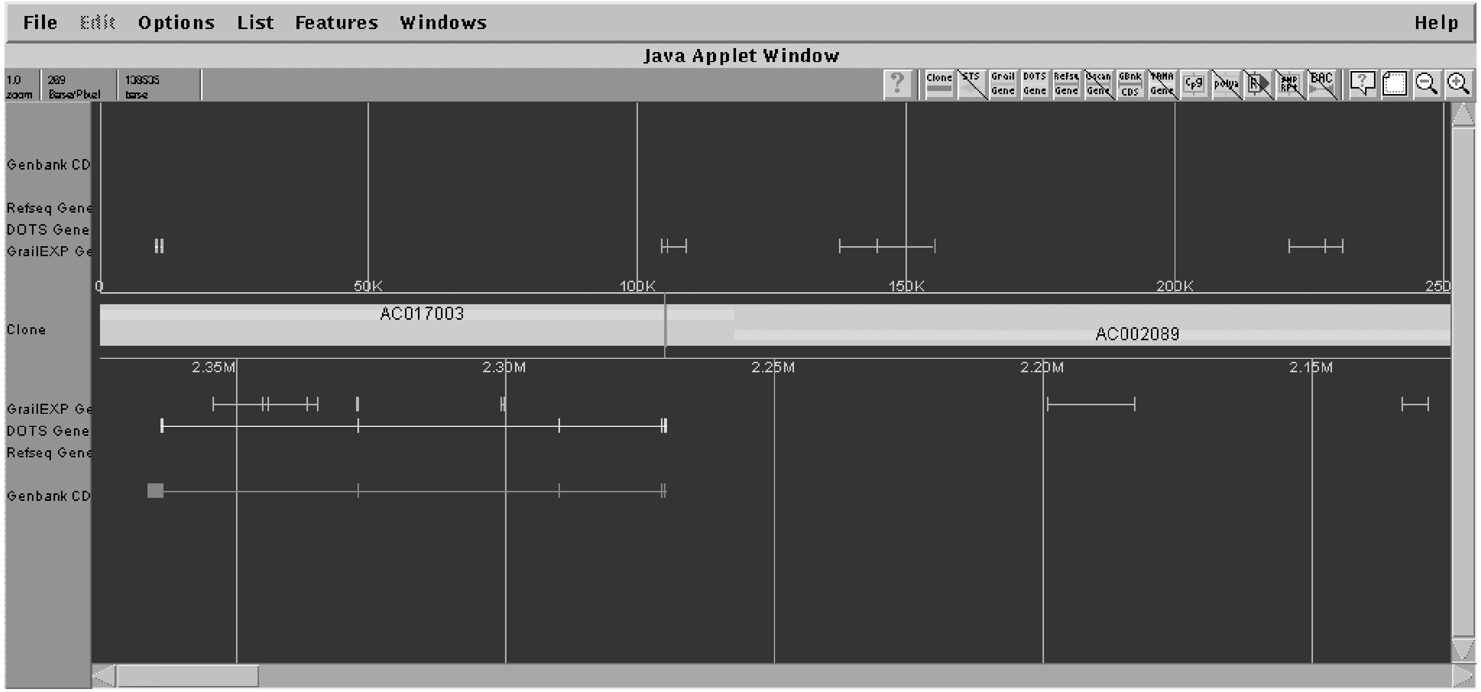


Figure 5.4 The genomic region around the FOXP2 gene according to ORNL Genome Channel (See Colour Plates).

up this NCBI genomic sequence contig, the vertical line intersecting one of the clones represents a CpG island. Repeats and polya sites also appear as lines within this band and gene predictions on either strand are displayed in the panels above and below it. At the time of writing it is not possible to view homologies to other genome sequences or the positions of SNPs. More information on the features that are displayed is available from other windows.

5.5 ANNOTATION: THE NEXT GENERATION

In spite of difficulties with the quality of genomic sequence assemblies and the errors and omissions of computational annotation the browsers discussed above remain extremely useful tools for the cautious biologist. They undoubtedly indicate the presence of most coding sequence in a given fragment of genomic sequence and indicate their location in the genome based on the best genomic sequence available. In addition they have a stab at predicting gene structures for novel genes that should be accurate if the gene in question is known or has a close homologue which is known. Most aspects of the analysis carried out are the subjects of active research, and improvements in performance due to the inclusion of new sequence data and annotation software will be ongoing. The downside of these developments is that all annotation of genomic sequence is potentially in flux and one should not assume that the representation of a region will remain the same between different software or data releases.

At some time in 2003 discussions of draft sequence assembly should be academic for more than 90% of the human genome and large finished contigs, tens of megabases long will be curated at NCBI. The main tasks with regard to the primary sequence data will then relate to data curation rather than assembly. Annotation of these sequences, on the other hand, should still be at a relatively early stage. Even at nucleotide level there is much to be done, particularly in exploiting the data available from model organism genome sequencing projects. There have already been notable successes in using comparative genomics to predict gene structures using the Twinscan program (Korf *et al.*, 2001; http://genes.cs.wustl.edu/). The cutting edge of nucleotide-level annotation is in defining regulatory regions: transcription start sites (TSSs), transcription factor binding sites and promoter modules (Werner, 2001). Here again, comparative genomics is already a rich source of information simply using existing sequence search algorithms such as BLAST (Levy *et al.*, 2001). At a higher level, gene expression is also regulated by the large-scale topology of chromosomes, and annotation may eventually indicate features such as chromosome domains (genomic regions that bind histone-modifying proteins) and matrix attachment sites (regions that facilitate the organization of DNA within a chromosome into loops). However, defining the genes whose transcription is regulated from such features may be an insoluble problem computationally, since they may regulate transcription from a given TSS, from several different TSSs of the same gene or multiple genes in a region.

At the protein and process levels of annotation there is also progress, for instance in our ability to detect more remote homologies and gain clues about function. Homologous proteins, sharing a common three-dimensional structure and function, need not share detectable sequence similarity. There is therefore increasing interest in annotation by similarity at the level of protein structure (Gough and Chothia, 2002). The genome sequence is already changing the way we study biology as we start to fill in the gaps between genetics, cellular function and development. Rather than studying a particular gene or protein we are increasingly able to study all elements in a system of interest, a group of proteins that participate in a complex for example. We might start with a single protein and identify others in the proteome that potentially interact with it, on the basis of the presence of domains known to interact. In the process we may discover previously unknown connections with other complexes or biochemical pathways that can be included in the annotation of the relevant sequences. Studies on this scale are prompting the development of multidisciplinary groups that study the behaviour and perturbation of entire biological systems (Ideker *et al.*, 2001). In the end this should provide a genome sequence with contents which can be browsed at the level of their genomic neighbourhood but also at the level of the interactions, complexes and processes that they participate in and the phenotypes they influence.

This review has only provided a brief introduction to the fields of computational draft genome assembly and annotation but it should be evident that what has already been achieved has involved innovations as great as those in the biotechnology that led to the production of the sequence data itself. At the same time, problems remain at every level and are the subjects of active research. As a result many different groups around the world are working on interpreting the data avalanche that is modern genetics and communication and comparison of results becomes difficult. The Distributed Annotation System (DAS; Dowell *et al.*, 2001; http://biodas.org/) aims to provide a framework for people to exchange data easily using the web. It promises a future without the current confusion of incompatible interfaces and data formats, and an increase in the open exchange of data and ideas.

ACKNOWLEDGEMENTS

Colin Semple enjoys the financial support of the UK Medical Research Council. Martin S. Taylor provided comments on an earlier version of this manuscript.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, *et al.* (2000). InterPro — an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16**: 1145–1150.

Baldock R, Bard J, Brune R, Hill B, Kaufman M, Opstad K, *et al.* (2001). The Edinburgh Mouse Atlas: using the CD. *Brief Bioinform* **2**: 159–169.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, *et al.* (2002). The Pfam protein families database. *Nucleic Acids Res* **30**: 276–280.

Beck S, Sterk P. (1998). Genome-scale DNA sequencing: where are we? *Curr Opin Biotechnol* **9**: 116–120.

Birney E, Durbin R. (2000). Using GeneWise in the Drosophila annotation experiment. *Genome Res* **10**: 547–548.

Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT. (2002). The Mouse Genome Database (MGD): the model organism database for the laboratory mouse. *Nucleic Acids Res* **30**: 113–115.

Burge CB, Karlin S. (1997). Prediction of complete gene structure in human genomic DNA. *J Mol Biol* **268**: 78–94.

Burge CB, Karlin S. (1998). Finding the genes in genomic DNA. *Curr Opin Struct Biol* **8**: 346–354.

Chen R, Bouck JB, Weinstock GM, Gibbs RA. (2001). Comparing vertebrate whole-genome shotgun reads to the human genome. *Genome Res* **11**: 1807–1816.

Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, *et al.* (1998). Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* **282**: 2022–2028.

Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L. (2001). The Distributed Annotation System. *BMC Bioinformatics* **2**: 7.

Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, *et al.* (2002). Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res* **30**: 69–72.

Eddy SR. (1998). Profile hidden Markov models. *Bioinformatics* **14**: 755–763.

Eddy SR. (2001). Non-coding RNA genes and the modern RNA world. *Nature Rev Genet* **2**: 919–929.

Eeckman FH, Durbin R. (1995). ACeDB and macace. *Methods Cell Biol* **48**: 583–605.

Eichler EE. (2001). Segmental duplications: what's missing, misassigned, and misassembled and should we care? *Genome Res* **11**: 653–656.

Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* **8**: 967–974.

FlyBase Consortium. (2002). The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res* **30**: 106–108.

Gavin A-C, Bosche M, Krause R, Grandi P, Marzioch, Bauer A, *et al.* (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**: 141–147.

Gene Ontology Consortium. (2001). Creating the gene ontology resource: design and implementation. *Genome Res* **11**: 1425–1433.

Gough J, Chothia C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res* **30**: 268–272.

Guigo R, Agarwal P, Abril JF, Burset M, Fickett JW. (2000). An assessment of gene prediction accuracy in large DNA sequences. *Genome Res* **10**: 1631–1642.

Hattori M, Taylor TD. (2001). Part three in the book of genes. *Nature* **414**: 854–855.

Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, *et al.* (2000). The DNA sequence of human chromosome 21. *Nature* **405**: 311–319.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, *et al.* (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**: 180–183.

Hogenesch JB, Ching KA, Batalov S, Su AI, Walker JR, Zhou Y, *et al.* (2001). A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106**: 413–415.

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, *et al.* (2002). The Ensembl genome database project. *Nucleic Acids Res* **30**: 38–41.

Huynen MA, Bork P. (1998). Measuring genome evolution. *Proc Natl Acad Sci USA* **95**: 5849–5856.

Ideker T, Galitski T, Hood L. (2001). A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* **2**: 343–372.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**: 860–892.

Iyer LM, Aravind L, Bork P, Hofmann K, Mushegian AR, Zhulin IB, *et al.* (2001). *Quod erat demonstrandum*? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. *Genome Biol* **2**.

Jordan IK, Kondrashov FA, Rogozin IB, Tatusov RL, Wolf YI, Koonin EV. (2001). Constant relative rate of protein evolution and detection of functional diversification among bacterial, archaeal and eukaryotic proteins. *Genome Biol* **2**.

Katsanis N, Worley KC, Lupski JR. (2001). An evaluation of the draft human genome sequence. *Nature Genet* **29**: 88–91.

Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, *et al.* (2001). Functional annotation of a full-length mouse cDNA collection. *Nature* **409**: 685–690.

Kent WJ, Haussler D. (2001). Assembly of the working draft of the human genome with GigAssembler. *Genome Res* **11**: 1541–1548.

Kikuno R, Nagase T, Waki M, Ohara O. (2002). HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res* **30**: 166–168.

Korf I, Flicek P, Duan D, Brent MR. (2001). Integrating genomic homology into gene structure prediction. *Bioinformatics* **17** (Suppl. 1): S140–S148.

Kumar A, Harrison PM, Cheung KH, Lan N, Echols N, Bertone P, *et al.* (2002). An integrated approach for finding overlooked genes in yeast. *Nature Biotechnol* **20**: 58–63.

Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* **413**: 519–523.

Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, *et al.* (2000). SAGEmap: a public gene expression resource. *Genome Res* **10**: 1051–1060.

Levy S, Hannenhalli S, Workman C. (2001). Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics* **17**: 871–877.

Li WH. (1997). *Molecular Evolution*. Sinauer Associates: Sunderland, MA, USA.

Lowe TM, Eddy SR. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 955–964.

Malpertuy A, Tekaia F, Casaregola S, Aigle M, Artiguenave F, Blandin G, *et al.* (2000). Genomic exploration of the hemiascomycetous yeasts: 19. Ascomycetes-specific genes. *FEBS Lett* **487**: 113–121.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, *et al.* (2000). A whole-genome assembly of Drosophila. *Science* **287**: 2196–2204.

Ning Z, Cox AJ, Mullikin JC. (2001). SSAHA: a fast search method for large DNA databases. *Genome Res* **11**: 1725–1729.

Noordewier MO, Warren PV. (2001). Gene expression microarrays and the integration of biological knowledge. *Trends Biotechnol* **19**: 412–415.

Olivier M, Agarwal A, Allen J, Almendras AA, Bajorek ES, Beasley EM, *et al.* (2001). A high-resolution radiation hybrid map of the human genome draft sequence. *Science* **291**: 1298–1302.

Raychaudhuri S, Chang JT, Sutphin PD, Altman RB. (2002). Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res* **12**: 203–214.

Reese MG, Hartzell G, Harris NL, Ohler U, Abril JF, Lewis SE. (2000). Genome annotation assessment in *Drosophila melanogaster*. *Genome Res* **10**: 483–501.

Rice P, Longden I, Bleasby A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276–277.

Rivas E, Klein RJ, Jones TA, Eddy SR. (2001). Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr Biol* **11**: 1369–1373.

Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, *et al.* (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet* **24**: 227–235.

Salamov AA, Solovyev VV. (2000). *Ab initio* gene finding in Drosophila genomic DNA. *Genome Res* **10**: 516–522.

Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. (1982). Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* **162**: 729–773.

Schuler GD. (1997). Sequence mapping by electronic PCR. *Genome Res* **7**: 541–550.

Semple CAM, Morris SW, Porteous DJ, Evans KL. (2002). Computational comparison of human genomic sequence assemblies for a region of chromosome 4. *Genome Res* (in press).

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, *et al.* (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**: 308–311.

Soderlund C, Humphray S, Dunham A, French L. (2000). Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res* **10**: 1772–1787.

Stein L. (2001). Genome annotation: from sequence to biology. *Nature Rev Genet* **2**: 493–503.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**: 33–36.

Todd AE, Orengo CA, Thornton JM. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* **307**: 1113–1143.

Uberbacher EC, Xu Y, Mural RJ. (1996). Discovering and understanding genes in human DNA sequence using GRAIL. *Methods Enzymol* **266**: 259–281.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, *et al.* (2001). The sequence of the human genome. *Science* **291**: 1304–1351.

Watson JD. (1990). The human genome project: past present, and future. *Science* **248**: 44–49.

Werner T. (2001). Cluster analysis and promoter modelling as bioinformatics tools for the identification of target genes from expression array data. *Pharmacogenomics* **2**: 25–36.

Wheelan SJ, Church DM, Ostell JM. (2001). Spidey: a tool for mRNA-to-genomic alignments. *Genome Res* **11**: 1952–1957.

Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, *et al.* (2002). Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res* **30**: 13–16.

Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, *et al.* (2001). The TRANSFAC system on gene expression regulation. *Nucleic Acids Res* **29**: 281–283.

Zhang CT, Wang J. (2000). Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. *Nucleic Acids Res* **28**: 2804–2814.

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)

CHAPTER 6

Mouse and Rat Genome Informatics

JUDITH A. BLAKE[□], JANAN EPPIG and CAROL J. BULT

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

[□] corresponding author

6.1 INTRODUCTION

Mouse and rat genome informatics is grounded in work on mouse and rat genetics and physiology that has been on-going since the early 20th century. The mouse, with its short generation time, small size, and plethora of phenotypic variants excelled as a tool for genetic investigations, especially after the conceptualization and creation of inbred strains, work begun by C. C. Little (Little and Tyzzer, 1916). Genetic crosses between inbred strains led to detailed mapping of genes and phenotypes, the construction of linkage groups, the development of chromosomal mapping techniques and the investigation of genetic components of phenotypes including diseases. Of particular significance was the development of specialized strains for genetic testing and technologies for manipulating the mouse genome. Standard inbred strains, their various derivatives, and 'boutique' mice developed through mutagenesis and genetic engineering have become essential tools. Coupled with advances in micro-technologies that are enabling detailed physiological studies in mice, the rich understanding of mouse genetics is accelerating the studies of genotype–phenotype relationships.

The rat, in contrast, was valued especially for its larger size relative to the mouse, and thus better suitability for physiological studies and experimental interventions. For rat, much is known about diseases, component factors in resistance/susceptibility, and specific networks of disease processes. Areas of research have been broad, including immunology, cancer, diseases of specific organ systems (cardiovascular, urogenital, skeletal, behaviour, growth and metabolism), neurological diseases, haematologic disorders, toxicology, histology, endocrinology, pathophysiology, and pharmacology (Gill *et al.*, 1989; James and Lindpaintner, 1997). The genetics of the rat lagged behind until recently, when genomic tools (expressed sequence tags or ESTs, radiation hybrid and physical maps) for rat have rapidly been created and developed.

Today, rat and mouse are both strong animal models for the investigation of biology particularly with regard to human biology and disease. The availability of two rodent animal models is also fortuitous because it permits the examination of genetic and phenotypic variation between two closely related organisms and the ability, then, to contrast that information with knowledge about the biology of humans.

6.1.1 Bioinformatics for Mouse and Rat Geneticists

The term 'bioinformatics' is used to refer to many aspects of the intersection of computer science, biology, and information science. The term is often equated with the informatics challenges of the genome projects. There are several reasons for this. First, the genome sequencing efforts generate enormous volumes of electronic data that must be organized, stored, and analysed using powerful computers and sophisticated algorithms from the inception of the project. Second, substantial fiscal resources are being devoted to these projects, so the advancement of the informatics component is both absolutely necessary and well funded. Finally, there is the high visibility of the genome projects, with frequent newsflashes about the discovery of new and interesting genes. As a result of these forces, many scientists think of bioinformatics as an endeavour focused solely on the management and analysis of sequence data.

However, all aspects of biological investigation benefit from the ordered assembly of the information and from the use of computer technologies to store, query, sort and manage biological data. Prior to a database implementation, many structured datasets about mouse genetics and heritable mutants were maintained manually. The first gene description catalogue for mouse was published in 1941 by Dr George Snell (Snell, 1941). As early as the 1950s Dr Margaret Green began compiling mouse linkage and mapping data on index cards. Linkage maps were drawn by hand and published annually in the *Mouse Newsletter* from 1965–1994. Compilations of mutant genes and polymorphic loci, chromosome atlases, and lists of synteny homologies between mouse and man were irregularly published in journals (cf. Eppig, 1992; Nadeau *et al.*, 1991; Staats, 1985) in addition to books such as *Genetic Variants and Strains of the Laboratory Mouse* (Green, 1981; Lyon and Searles, 1989; Lyon *et al.*, 1996). During the 1980s many of these resources began to be maintained electronically and resulted in an early publicly accessible mouse database GBASE (Doolittle *et al.*, 1991) and the Encyclopedia of the Mouse Genome software tools (Eppig *et al.*, 1994). During the 1990s, this sweep of information about the genetics and biology of the laboratory mouse was integrated and brought fully into electronic form with the construction of the Mouse Genome Database (http://www.informatics.jax.org/; Richardson *et al.*, 1995) and the development of computer programs to manipulate and query the data such as MapManager (Manly, 1993). In addition, large-scale mapping projects redefined the management of genetic data (Dietrich *et al.*, 1992) and led to the construction of additional bioinformatics resources for mouse geneticists.

Compilations for rat information developed in a different way. Billingham and Silvers published the first compilation of rat strain information in 1959 (Billingham and Silvers, 1959). A standard nomenclature for rat strains emerged in 1973 (Festing and Staats, 1973). Rat strain descriptions were catalogued (Greenhouse *et al.*, 1990), and later maintained electronically by M. F. W. Festing and made publicly available in the model organism databases. Gene data was published sporadically and accumulated slowly due to the emphasis of rat researchers on physiology rather than genetics. The pressure for databases and computational tools for rat has been a recent occurrence. Although RatMap, which exclusively curates mapped genes was started in 1993, the need for resources to manage genomic data (simple sequence length polymorphisms or SSLPs, ESTs, comprehensive gene data, genomic sequence, etc.) was not recognized as critical until the joint US–German Rat Genome Project began generating large volumes of data in the mid/late 1990s. This recognition led to the development of the Rat Genome Database (http://rgd.mcw.edu) described more fully below.

6.1.2 Data Integration: The Challenge and the Conundrum

The advent of the Internet and the development of the www permitted the development of multiple sites committed to the presentation of biological data relative to the mouse and rat. Some, such as the sequence repositories GenBank (http://www.ncbi.nlm.nih.gov/; Wheeler *et al.*, 2001) and EMBL (http://www.ebi.ac.uk/embl/; Stroesser *et al.*, 2001), include mouse and rat sequences along with sequences from all other species. Others, such as the Whitehead Institute for Biomedical Research/MIT Center for Genome Research site (http://www-genome.wi.mit.edu/cgi-bin/mouse/index) provide specialized mouse datasets such as the pages for the 'Genetic and Physical Maps of the Mouse Genome'. For investigators, the reality is that information about the genetics and genomics of the laboratory mouse and the rat are found throughout cyberspace. Standards for nomenclature or descriptions of experimental data are not uniformly implemented, and it is often difficult to equate information at one site with information at another. Consequently, the investigator spends much time looking for data, collecting the data, and then manipulating the data before being able to explore and mine the data for knowledge. This has not gone unnoticed by data providers, but efforts to standardize and integrate information are often stymied by the variety of data types, the variability in data annotation, and the diversity of needs of the users. This presents a conundrum for bioinformatics professionals. Scientists do not want to be forced to use standard nomenclature or terminologies in the publication of their own work, but they do want to find a suite of information about a set of genes or sequences without having to do the data integration themselves.

The solution is easy to define, but hard to implement. It is dependent more on the sociology of doing science rather than the need for a technological solution. Data integration requires the implementation of standards and structures across multiple information resources (Bult *et al.*, 2000). Key strategies for data integration are the use of accessioned data entities, the application of nomenclature standards for key objects such as genes and strains, and the use of controlled, structured vocabularies and ontologies for functional annotation of biological information. Most of the larger data providers of interest to mouse and rat geneticists are now working to implement shared standards and to provide curated links between the different resources. Much harder is the integration of the scientific literature. As yet, most authors are unaware of and/or are not required to use standard nomenclature for genes, proteins, anatomy or biochemical reactions in the publication of laboratory research results. The result is that it is more difficult than it needs to be to bring experimental data into electronic form and to integrate it with other information. Hopefully, the use of data and nomenclature standards will become more common as scientists of all types recognize the value of bioinformatics resources and consequently appreciate the necessity and the power of data integration.

6.2 THE MODEL ORGANISM DATABASES FOR MOUSE AND RAT

One approach to integration of information about mouse and rat has been the construction of model organism databases. Several issues swirl around informatics sites devoted to model organisms. On the one hand, better interoperability among large data providers might obviate the need for an organisms-specific site. On the other hand, for model organisms such as *Saccharomyces*, *Caenorhabditis elegans*, *Drosophila* and others, including mouse and rat, there is a need for a central site that integrates all kinds of information about these well-studied species. Various approaches to shared data structures and standards are continually under discussion and have resulted in the increased similarities and links between the model organisms databases. Will there ultimately be one information system for all biology? Or will there continue to be specialized model organism sites loosely connected with other bioinformatics servers? The interconnectivity and transparency between bioinformatics resources continues to evolve, and it is imprudent to envisage bioinformatics systems just a few years hence. Today there exist model organism databases for the mouse and the rat, the Mouse Genome Database and the Rat Genome Database. Both work to provide comprehensive access to experimental and consensus data about these model organisms.

6.2.1 The Mouse Genome Database

The Mouse Genome Database (MGD) (http://www.informatics.jax.org) is the original model organism database for the laboratory mouse (Blake *et al.* 2001). Derived from the merger of several small specialized databases in 1994, MGD now focuses on the integrated representation of genotype (sequence) to phenotype data for the mouse with a particular emphasis on information about genes and gene products. MGD provides official gene nomenclature for the research community and works closely with human and rat genome curators to implement common standards for annotation of genes and other genome features. As part of the Mouse Genome Informatics (MGI)) system (see below), MGD focuses on data integration are through representations of relationships between genes, sequences and phenotypes, the representation of mouse mapping data, the association of genes to the Gene Ontology (GO), the description of targeted mutations and other alleles, and the curation of mammalian orthologies.

6.2.2 Mouse Genome Informatics

MGD is one component of the Mouse Genome Informatics (MGI) consortium based at The Jackson Laboratory. Other components of the MGI consortium include the Gene Expression Database (GXD; Ringwald *et al.*, 2001), the Mouse Tumor Biology Database (MTB; Bult *et al.*, 2001) and the Mouse Genome Sequencing Project (MGS). GXD focuses on the presentation of detailed experimental data about time and place of gene expression during development. MTB provides web-based access to mouse models of human cancers including experimental data and genotype-specific information. MGS works with the public mouse genome sequencing coalition to link the emerging genome with the mouse biological information. Overall, then, the MGI project provides the research community with a canonical set of mouse genes, their official names and genome locations, sequences, mammalian homologies, expression and functional information, phenotypic alleles and variants, associated literature and extensive links to other bioinformatics resources. This highly-integrated system is complemented with many cross-links to genetic and genomic resources for other organisms.

6.2.3 RatMap

RatMap (http://ratpmap.gen.gu.se) focuses on presenting the subset of rat genes, DNA markers, and quantitative trait loci (QTL) that are localized to chromosomes. RatMap maintains a highly-curated set of data, including nomenclature, chromosomal assignment and localization, mapping method statements, human and mouse homologues, references, and links to nucleotide sequences, UniGene and Rat Genome Database (RGD). In addition, RatMap maintains the rat idiograms and current cytogenetic maps. RatMap also provides a ‘gene and position predictor’ (GAPP) report that presents predicted positions for over 6000 rat genes based on conserved syntenic chromosomal segments between mouse and rat (Helou *et al.*, 2001).

6.2.4 The Rat Genome Database (RGD)

The Rat Genome Database (RGD, http://rgd.mcw.edu; Twigger *et al.*, 2002) is a collaborative effort between the Bioinformatics Research Center at the Medical College of

Wisconsin, The Jackson Laboratory and the National Center for Biotechnology Information (NCBI) to gather, integrate and make available data generated from ongoing rat genetic and genomic research efforts. Initially released in 2000, RGD includes curated data on rat genes, QTL, ESTs, sequence tagged sites (STSs) and microsatellite markers as well as details of inbred rat strains. RGD also contains detailed information on nomenclature, genetic and RH maps, mouse and human homologies, Gene Ontology data, and includes key literature citations. Research tools that are provided include 'VCMap', a sequence-based homology tool and gene prediction and RH mapping tools. RGD is introducing disease-based curation for disease processes frequently studied in the rat. Integration of the emerging rat genomic sequence is also planned.

6.3 MOUSE GENETIC AND PHYSICAL MAPS

The genetic map of the mouse has been built over time through the contributions of many research groups, using a variety of methods, including, but not limited to, backcross, intercross and complex cross analyses, congenic strain analysis and recombinant inbred and recombinant congenic strain analyses. Chromosomal rearrangements, somatic cell hybrids and *in situ* hybridization are used to supplement these methods. These diverse methods, utilizing a wide variety of laboratory and wild-derived mouse strains, have been used to develop the consensus linkage map for mouse (MGD, http://www.informatics.jax.org/searches/linkmap_form.shtml). For many purposes, this map is a standard for understanding the overall genomic organization of the mouse and for identifying potential candidate genes for diseases in particular regions.

6.3.1 Mouse DNA Mapping Panels

The development of large interspecific and intersubspecific crosses, for which progeny DNA are stored for cumulative genotyping, provides single-source high-resolution linkage maps containing thousands of markers and with well-defined crossover points (cf. Avner *et al*., 1988; Copeland and Jenkins, 1991; Dietrich *et al*., 1992; European Backcross Collaborative Group, 1994; Rowe *et al*., 1994). Any newly discovered gene for which DNA polymorphism is detectable between the original parental strains can be mapped immediately without setting up a *de novo* cross and the cumulative data can be used to explore questions of recombination distribution across the genome and crossover interference. These DNA backcross panels are, however, not suitable for mapping new genes that are only defined by phenotype.

Genotyping data for individual progeny from many of these DNA mapping panels are available through the Mouse Genome Database (http://www.informatics.jax.org/searches/crossdata_form.shtml). In addition, maps can be generated using these data via the MGD Map Building tool at http://www.informatics.jax.org/searches/linkmap_form.shtml. Two of these DNA mapping panels are also maintained at specific websites: The Jackson Laboratory DNA Mapping Panels (http://www.jax.org/resources/documents/cmdata/bkmap) and the Whitehead Institute for Biomedical Research/MIT DNA Mapping Panels (http://www-genome.wi.mit.edu/cgi-bin/mouse/index#genetic).

6.3.2 Mouse Radiation Hybrid Maps

Recombination maps from DNA mapping panels provide unambiguous placement of gene order. However, for very closely linked genes, these maps may not be able to resolve locus order. For mouse, a radiation hybrid (RH) panel (T31) of 100 cell lines developed from a 3000-rad irradiated primary cell line from mouse embryo fused with hamster fibroblast has been developed (McCarthy *et al.*, 1997). Radiation hybrids can be used for high throughput mapping and high resolution of locus order because each hybrid cell line contains a highly fragmented subset of the mouse genome. The co-retention of mouse genes across the 100-cell panel is indicative of their relative distance apart, assuming random chromosomal breakage and leads to the construction of RH maps (cf. Van Etten *et al.*, 1999). Two complementary databases serve as community resources for gathering, distributing and analysing the T31 RH data.

6.3.3 The Jackson Lab Radiation Hybrid Map

The JAX RHmap provides web-based access to a comprehensive, integrated database that includes all typing data, retention frequency and log of the odds (LOD) scores for markers typed on the T31 panel, as well as RH framework maps for many of the chromosomes (http://www.jax.org/resources/documents/cmdata/rhmap/). All publicly available T31 data from large genome centres at the Whitehead Mouse RH Database (http://www-genome.wi.mit.edu/mouse_rh/index.html), the UK Mouse Genome Centre (http://www.mgc.har.mrc.ac.uk/physical/est_mapping/est.html) and Genoscope–CNS (http://www.genoscope.cns.fr/externe/English/Projets/Projet_ZZZ/rhmap.html), as well as from many individual laboratories are included.

The website includes an electronic submission interface for depositing RH typing data from users, data error checking and quality control, technical support, data analysis and the development of RH maps. All data, with references and experimental notes can be viewed or downloaded. Data are shared with the Mouse Genome Database (MGD) and the EBI data repository (RHdb, below).

6.3.4 The EBI Radiation Hybrid Database

The European Bioinformatics Institute (EBI) Radiation Hybrid Database (RHdb) is a repository for the raw data for constructing radiation hybrid maps, STS data, scores and experimental conditions (Rodriguez-Tomé and Lijnzaad, 2001; http://www.ebi.ac.uk/RHdb/index.html). The EBI RHdb is designed to be a species-neutral database, and currently contains human, mouse, and rat RH data. Data content relies entirely on submissions from data providers and research groups. Maps are not assembled from the accumulating data, but maps may be submitted by data developers.

6.3.5 Mouse Physical Maps

Two genome centres have produced physical maps for mouse that are accessible via the Internet: the Whitehead Institute/MIT (http://www-genome.wi.mit.edu/cgi-bin/mouse/index#phys) and the UK Mouse Genome Centre at Harwell (http://www.mgc.har.mrc.ac.uk/physical/phys.html). Whitehead Institute/MIT data include contigs and STS content mapping across the entire mouse genome and utilizes existing SSLP markers that characterize the MIT genetic map of the mouse to tie the physical and recombination maps together. The UK Mouse Genome Centre data consists of physical maps of selected regions of the genome that are being developed in association with individual research interests, notably regions of chromosomes 13 and X. Data from these sites are integrated into MGD, as well as being available from the originator's site.

A physical map of the genome of the C57BL/6J strain of laboratory mouse has been constructed using Bacterial Artificial Chromosome (BAC) clones (Gregory *et al.*, 2002). This map serves as the framework for the Mouse Genome Sequencing initiative (described below). The current BAC map for the mouse was derived from 305,768 BAC clones from two libraries: RPCI23 (female) and RPCI24 (male) (Osoegawa *et al.*, 2000). These libraries are available for distribution to the scientific community through the BACPAC Resource at the Children's Hospital Oakland Research Institute (http://www.chori.org/bacpac). The RPCI23 library is also available through Research Genetics (http://www.resgen.com/products/RPCI23MBAC.php3).

The clones from the RPCI BAC libraries were fingerprint mapped at the Genome Sequencing Centre in Vancouver, British Columbia (Marra *et al.*, 1997; http://www.bcgsc.bc.ca/projects/mouse_mapping/). The fingerprint data were combined with BAC end sequence data (Zhao *et al.*, 2001; http://www.tigr.org/tdb/bac_ends/mouse/bac_end_intro.html) to produce a mouse physical map that contains 296 contigs and covers an estimated 2,739 Mb (Gregory *et al.*, 2002). The average length of the contigs is 9.3 Mb. Of the 296 contigs, 228 can be localized to a chromosome. Approximately 97% of the total clone coverage for the mouse genome (2,658 Mb in 211 contigs) can be aligned to the human genome sequence.

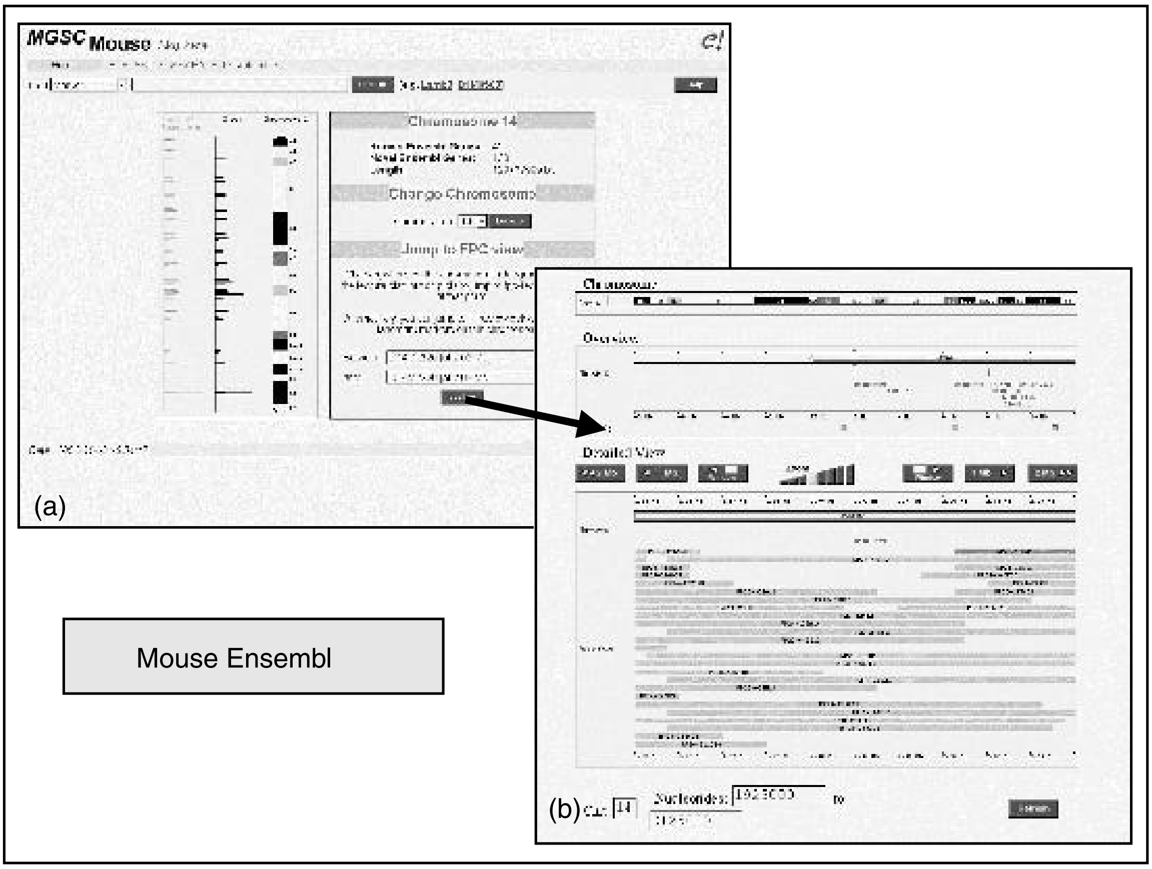


Figure 6.1 Mouse Ensembl. A graphical representation of the clone-based physical map for the proximal end of mouse chromosome 14 from Ensembl. This browser allows users to search for regions of a chromosome between two STS markers and to view the current clone coverage in the selected area. Because the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

There are three ways to view the current status of the mouse BAC physical map. Researchers can download and install a software product called FPC from the Sanger Institute (http://www.sanger.ac.uk/Software/fpc/) (Soderlund *et al.*, 2000) and use this software to graphically display the BAC clone fingerprint data generated by the Genome Sequence Centre in Vancouver. A similar display tool called the internet Contig Explorer (iCE) is available from the Genome Sequence Centre in Vancouver (http://ice.bcgsc.bc.ca/). An option for viewing the map that does not require the installation of software is to view the physical map using the Ensembl mouse browser at the Sanger Institute (http://mouse.ensembl.org/; or the mouse MapViewer of NCBI (http://www.ncbi.nlm.nih.gov).

The ultimate physical map, of course, is the genome sequence itself. Despite the expectation that the mouse genome will be available soon, the need for genetic maps and other physical maps will not disappear. The mouse sequence will continue to be built, reassembled and re-annotated for many years to come, making the physical contig map an important resource for anchoring this new information as it develops. Genetic maps will be needed indefinitely, for the mapping of QTLs, spontaneous mutations and other phenotypes with undetermined molecular defects. In addition, genetic maps are essential for studying chromosome structure and function, and recombination itself.

6.4 RAT GENETIC AND PHYSICAL MAPS

6.4.1 Rat Genetic Maps

The early development of rat genetics paralleled that of the mouse, with the establishment of genetic linkage between albino and pink-eyed dilution in both mouse and rat (Castle and Wachter, 1924; Dunn, 1920). Haldane (Haldane, 1927) recognized that, if these genes were homologues, they represented conserved synteny over evolutionary time. Subsequently, research geneticists focused on mouse and the rat became the major tool for physiologists. Thus, the development of the rat genetic map began to lag behind that of the mouse. As of 1991 there were 214 genes mapped in rat (Levan *et al.*, 1991) in contrast to nearly 3000 genes mapped in mouse (Hillyard *et al.*, 1991). This disparity in the number of genes mapped has continued to this day, with 1576 genes currently mapped in rat (RatMap, 2002) versus 18,983 in mouse (MGD, 2002). Maps of rat genes are largely cytogenetic rather than recombination maps and are maintained by RatMap (http://ratmap.gen.gu.se). After a century of concentrated use of rat by physiologists, rat genetics is now undergoing a revival as genomic tools are developed and its genome is finally being sequenced.

The resurgence of interest in the rat map has paralleled the development of genomic resources for rat. In the 1990s the first rat genome projects were begun to generate ESTs, YAC and BAC libraries, and SSLP maps. There were a number of backcrosses and intercrosses made among rat strains that were used to develop SSLP maps with several hundred to a few thousands markers (cf. Bihoreau *et al.*, 1997; Brown *et al.*, 1998; Dracheva *et al.*, 2000; Watanabe *et al.*, 2000; Wei *et al.*, 1998). Most of these SSLP maps are not yet integrated, although SSLP data and maps for some of the crosses are available through RGD. Data from two F2 intercrosses have been integrated and the resulting map containing 4786 SSLP markers can be found at the Whitehead Institute (http://www-genome.wi.mit.edu/rat/public/). In parallel, a large collaborative Allele Characterization Project was begun to establish allele sizes of 8000 SSLPs among 48 genetically and physiologically important inbred rat strains (http://www.brc.mcw.edu/LGR/research/lgr_acp.

html). Data generated from this project will provide investigators with a means of quickly selecting informative markers for new and existing mapping crosses.

6.4.2 Rat Radiation Hybrid Maps

A rat whole genome radiation hybrid panel (T55) generated by Linda McCarthy in Peter Goodfellow's laboratory has been used to construct high-resolution maps of the rat genome (http://www.well.ox.ac.uk/rat_mapping_resources/rat_radiation_hybrid_maps.html). The first radiation hybrid map was based on 5255 markers and included both microsatellites and known genes (Watanabe *et al.*, 1999). Another map using the same panel was constructed as a framework map using 2000 evenly spaced markers (http://rgd.mcw.edu/RHMAPSERVER/; Steen *et al.*, 1999). Both sites provide RH map web servers for users to map their markers — users submit data to the Rat RH Map Server and a map placement with a summary report is returned.

6.4.3 Rat Physical Maps

In contrast to the mouse, the rat has no genome-wide clone-based physical maps, only a few for specific regions such as the MHC locus (Gunther and Walter, 2001; Ioannidu *et al.*, 2001). Most of the 'physical map' for the rat genome consists of the cytogenetic maps that are maintained in RatMap and include a fair amount of FISH data. A physical BAC map of the rat is in preparation, as part of the NHGRI-sponsored rat genome sequencing initiative. A BAC library (CHORI-230) from the BN/SsNHsd/MCW (Brown Norway) strain of laboratory rat has been prepared using the same methods as were used for the mouse BAC libraries (http://www.chori.org/bacpac/) (Osoegawa *et al.*, 2000). The BAC clones from this library are being fingerprint mapped by the Genome Sequencing Centre in Vancouver, Canada (http://www.bcgsc.bc.ca/projects/rat_mapping/). There are currently (late 2001) 136,195 clones in their database. The BAC ends for this library are being sequenced at The Institute for Genomic Research (TIGR; http://www.tigr.org/tdb/bac_ends/rat/bac_end_intro.html).

6.5 GENOME SEQUENCE RESOURCES

6.5.1 Mouse Genome Sequencing Initiative

The initiative to sequence the genome of the laboratory mouse was announced by the National Human Genome Research Institute (NHGRI) of NIH in September 1999 as part of an overall 'action plan' for mouse genomics (Battey *et al.*, 1999). The goals of the initiative were to have a working draft of the genome of the C57BL/6J strain of mouse completed by 2003 and the finished genome sequence by 2005. The initial strategy for obtaining the mouse genome sequence was to build a physical BAC map of the genome as the BAC clones were sequenced (http://www.nhgri.nih.gov/NEWS/MouseRelease.htm).

In October of 2001 the strategy for obtaining the mouse sequence changed to include a whole genome shotgun approach. Part of the rationale for this change in sequencing strategy was that the shotgun sequences for the mouse genome could be used to assist in the identification of genes in the working draft of the human genome. The sequencing centres of the Sanger Institute, Washington University Medical Centre and the Whitehead Institute for Biomedical Research were funded to generate whole genome shotgun data for the mouse (http://www.nih.gov/science/models/mouse/).

Simultaneously with this shift in sequencing strategy, NIH launched a program to sequence mouse BAC clones that covered genomic regions of high biological interest. Individual investigators were invited to submit applications requesting specific BACs to be sequenced. Several sequencing centres, including the Cold Spring Harbor Laboratory, Harvard University Medical School and the University of Oklahoma were funded to sequence these BACs (http://www.nih.gov/science/models/bacsequencing/). The NIH BAC sequencing program was initially restricted to clones from specific BAC libraries for the mouse. However the program now accepts applications for the sequencing of clones from any BAC library and also from organisms other than mouse.

Several other sequencing centres around the world are using their sequencing capacity for regional and/or comparative sequencing of the mouse genome. For example, the DOE-funded Joint Genome Institute focused on sequencing segments of the mouse genome that are homologous to human chromosome 19 (http://bahama.jgi-psf.org/pub/ch19/; Dehal *et al.*, 2001). The Medical Research Council (MRC) is focusing on sequencing of mouse chromosomes 2, 4, 13 and mouse–human comparative sequencing for chromosome X (http://mrcseq.har.mrc.ac.uk/). Although the primary focus of the Baylor College of Medicine genome centre is now on sequencing the rat genome, it originally focused on sequencing BACs across mouse chromosome 11.

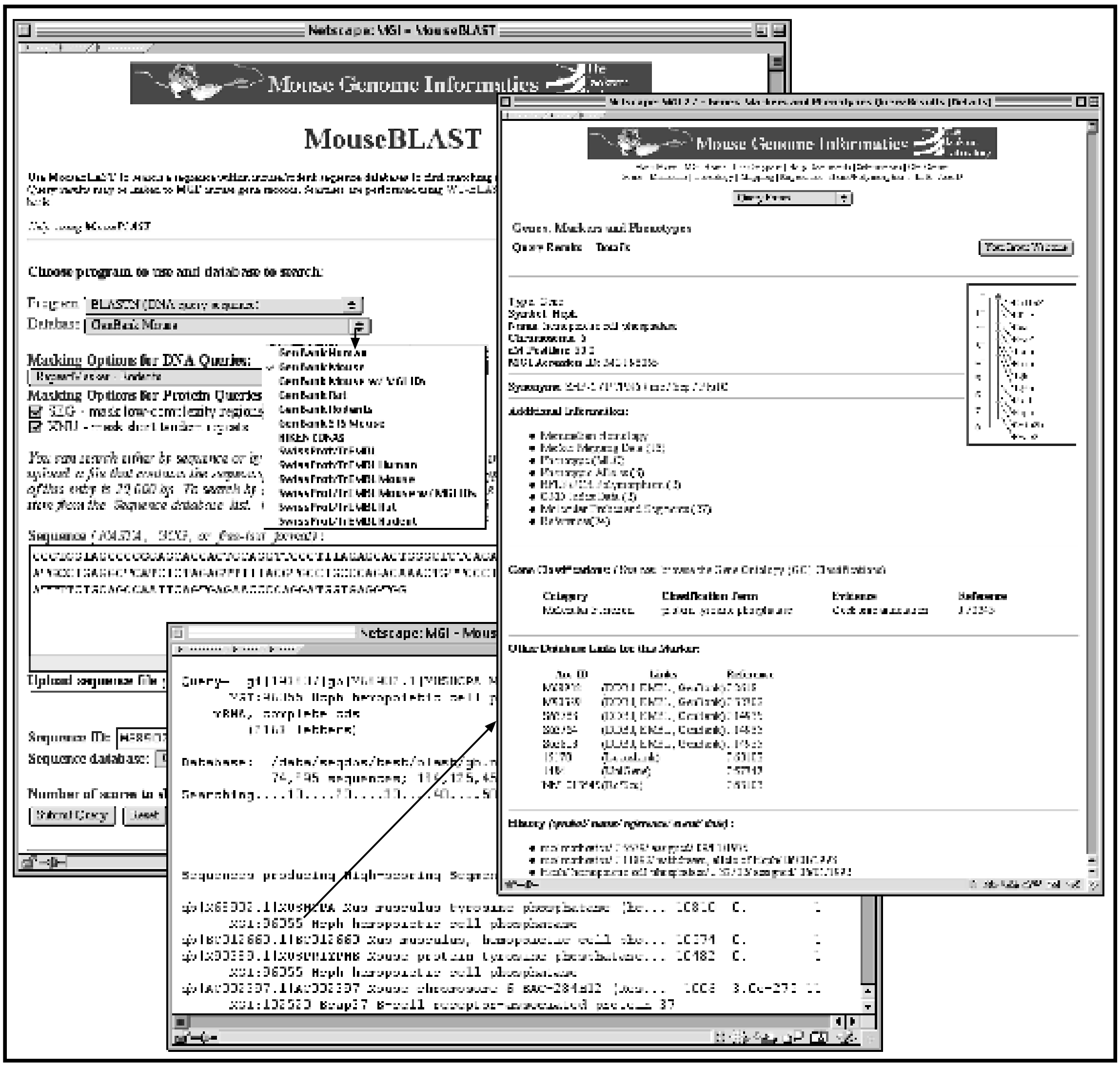
The NCBI maintains a status report of the progress of the mouse genome sequence project (http://www.ncbi.nlm.nih.gov/genome/seq/MmHome.html) as well as a registry of BAC clones that are being sequenced under the auspices of the Trans-NIH BAC Sequencing Program (http://www.ncbi.nlm.nih.gov/genome/clone/cstatus.html).

6.5.2 Mouse Genome Sequence Resources

There are several ways to access mouse genome sequence (here we focus on freely-accessible public resources). The whole genome shotgun data for the mouse can be found in a 'Trace Archive' maintained by the NCBI and can be searched via BLAST (http://www.ncbi.nlm.nih.gov/blast/mmtrace.html). A similar resource is maintained at the European Bioinformatics Institute (EBI; http://www.ebi.ac.uk/blast2). As of December 2001, there were over 31 million sequencing reads available in these archives; greater than six times the coverage of the mouse genome.

The Mouse Genome Sequencing Consortium has released an annotated draft assembly of the mouse genome to the research community (Mouse Genome Sequencing Consortium, 2002). The current draft assembly covers over 96% of the genome; a complete genome sequence for the laboratory mouse is anticipated by 2005. The draft genome and the associated annotations can be accessed using the Ensembl genome browser (http://www.ensembl.org), NCBI's Map Viewer (http://www.ncbi.nlm.nih.gov), and the University of Santa Cruz's genome browser (http://genome.ucsc.edu).

Other genome resources include MouseBLAST (Figure 6.2), a server maintained by the MGS group at The Jackson Laboratory that allows researchers to connect mouse sequence data with the wealth of biological knowledge about the mouse available in the MGI. Finally, the Mouse Genome Resources pages at NCBI (http://www.ncbi.nlm.nih.gov/



Choose program to use and database to search:

Program: BLASTN (DNA query sequence)

Database: GenBank Mouse

Masking Options for DNA Queries:

Masking Options for Protein Queries

☑ SEG - mask low-complexity regions

☑ XNU - mask short tandem repeats

- GenBank Human
- ✓ GenBank Mouse
- GenBank Mouse w/ MGI IDs
- GenBank Rat
- GenBank Rodents
- GenBank EST Mouse
- RIKEN cDNAs
- SwissProt/TrEMBL
- SwissProt/TrEMBL Human
- SwissProt/TrEMBL Mouse
- SwissProt/TrEMBL Mouse w/ MGI IDs
- SwissProt/TrEMBL Rat
- SwissProt/TrEMBL Rodent

Upload sequence file

Sequence ID:

Sequence database:

Number of scores to s

Submit Query | Reset

Type: Gene
Symbol: Hcph
Name: hemopoietic cell phosphatase
Chromosome: 6

Additional Information:

Gene Classifications

Other Database Links for this Marker:

History

```
Sequences producing High-scoring Segment
```

Figure 6.2 MouseBLAST. The MouseBLAST resource available from the Mouse Genome Informatics database website. MouseBLAST returns links to MGI Gene Detail pages as part of a standard BLAST report. The Gene Detail pages in MGI provide a wealth of information about homology, map location, phenotype associations, gene expression data, references and gene function annotation for each gene.

genome/guide/M_musculus.html) provide a compendium of links to various mouse genome resources.

6.5.3 Mouse cDNA Clone Resources

Several groups have undertaken initiatives to obtain full-length cDNA clones and sequences for every mouse gene. The RIKEN Institute from Japan has collected and sequenced over 60,000 cDNA clones for the mouse (http://genome.gsc.riken.go.jp/; The RIKEN Exploration Research Group Phase II Team and the FANTOM Consortium, 2001). The sequences for these clones are publicly available. The Mammalian Gene Collection, an NIH initiative, has a goal to provide a complete set of full-length (open reading frame) sequences and cDNA clones of expressed genes for human and mouse (http://mgc.nci.nih.gov/; Strausberg *et al.*, 2000).

6.5.4 Rat Genome Sequencing Initiative

In February 2001, the National Heart, Lung, and Blood Institute (NHLBI) announced funding support for the sequencing of the rat genome (http://www.nhgri.nih.gov/NEWS/nih_expands_programs.html). Three sequencing centres have been funded to produce enough genome sequence data to have a working draft of the rat genome by 2004: Celera Genomics, Baylor College of Medicine Genome Sequencing Centre and Genome Therapeutics, Inc.

6.6 COMPARATIVE GENOMICS

The sequencing of both the mouse and rat genomes promises to stimulate research based on comparative genome organization and comparative analysis between the human, mouse

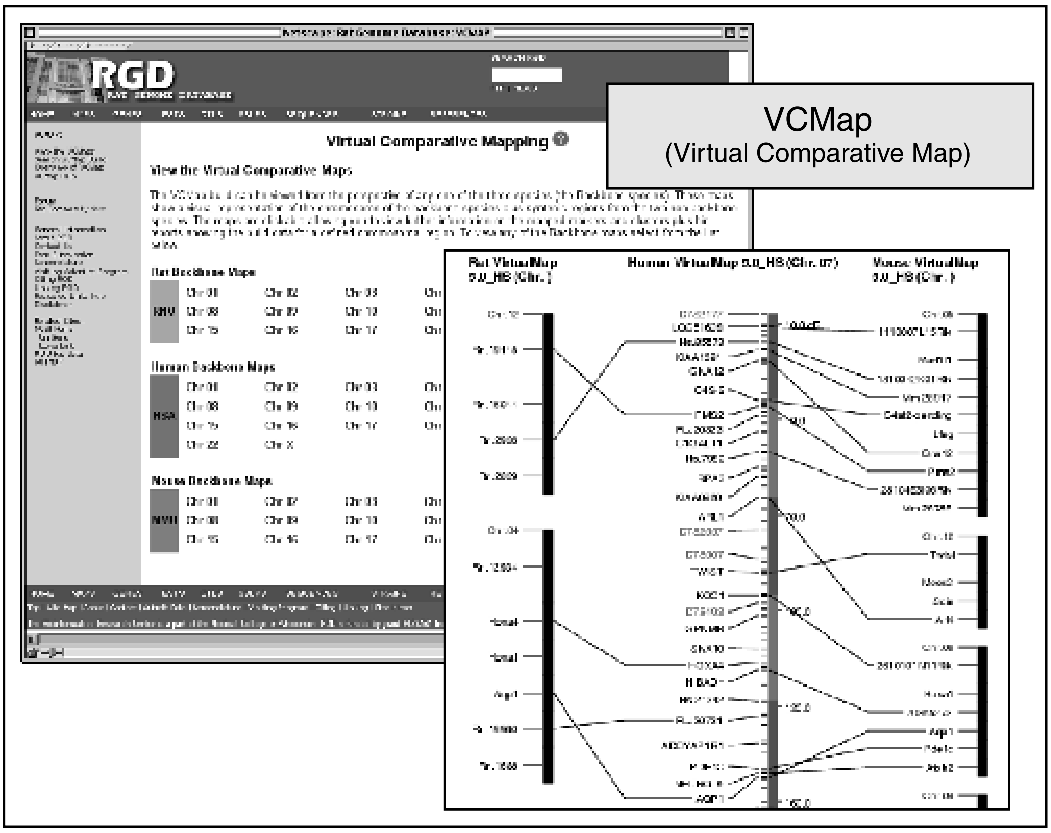


Figure 6.3 Virtual Comparative Map. The Virtual Comparative Map is generated using sequence-based algorithms that predict syntenic regions inferred from homology among mapped sequences. Sequence comparisons between ESTs and cDNAs from human, mouse and rat are combined with Radiation Hybrid map locations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

and rat. Research papers based on comparison of large conserved segments between mouse and human are being published (Dehal *et al.*, 2001; Glusman *et al.*, 2001). Another approach is to use genome comparisons for elucidation of a suite of comparable genome features such as transcription factors (Wasserman *et al.*, 2000). Computational approaches to uncovering conserved regions such as exons or regulatory sites facilitate the discovery of new important genome features (Oeltjen *et al.*, 1997).

The direct comparison of genomic sequence from conserved linkage groups between mouse and human (and other organisms) has proven to be an effective strategy for identifying biologically relevant regions (coding and non-coding) in genomes. Two of the most commonly used tools for this effort are VISTA (http://www-gsd.lbl.gov/vista/; Mayor *et al.*, 2001) and PIPMAKER (http://bio.cse.psu.edu/pipmaker/; Schwartz *et al.*, 2000). These resources allow researchers to submit large genomic sequence regions to be aligned and analysed for the presence of conserved sequence elements. The VISTA group provides a set of pre-aligned sequences of mouse and human from finished genomic data in GenBank (http://pipeline.lbl.gov/). Applications include determining all of the protein-coding segments in both species, locating regulatory signals, understanding the mechanisms and history of genome evolution and deducing the similarities and differences in gene organization between the species of interest.

Other comparative map viewers incorporate information about the rat. One resource is the Gene and Position Predictor (GAPP) produced by RatMap which provides predicted comparative maps using known gene orthologues and zoo-FISH data (http://gapp.gen.gu.se/Description.html; Nilsson *et al.*, 2001). A different type of predictive map is the Virtual Comparative Map (VCMap) (http://rgd.mcw.edu/VCMAP/; Figure 6.3). These maps are generated using sequence-based algorithms that predict syntenic regions inferred from homology among mapped sequences. The Otsuka GEN Research Institute posts a genome-wide comparative map of the rat based primarily on extensive RH mapping data (http://ratmap.ims.u-tokyo.ac.jp/cgi-bin/comparative_home.pl). Finally, maps of curated orthologues for mouse/rat/human are available from MGD (http://www.informatics.jax.org/menus/homology_menu.shtml).

6.7 FROM GENOTYPE TO PHENOTYPE

Beyond a generalized representation of the mouse and rat are the intricacies of differences due to differing genetic backgrounds that can be revealed by comparisons between strains, among the rodent species, between rodents and other mammals and even between more distantly related organisms. The publication of the mouse genome sequence and the promise of the rat genome sequence in the near future will facilitate systematic genome-wide approaches to investigate normal and disordered cellular and physiological states. Genome-wide surveys of gene expression or genotype variation will enhance the gene-by-gene approach to the assessment of gene function. Scientists have long known of the importance of genetic background in the analysis of gene function or dysfunction due to the phenotypic variability resulting from epistatic interactions. Now, it may be possible to precisely assess the effect of genotype variability on the expression, function and interaction of gene products. As ever, the challenge for bioinformaticians will be to integrate the data from various experimental approaches into a coherent representation of the model organism. Ideally, one would like to query for a set of gene products expressed at the same time/state, evaluate the effect of genotype on the function and phenotypic presentation of variant gene products or compare ‘snapshots’ of cellular component sets between tissues or strains of rodents.

6.7.1 Genetic Variants

Genetically-engineered strains of mice including mice altered by gene transfer (transgenics), homologous recombination (gene targeting) and chemical mutagenesis provide powerful new tools for biomedical research. The use of these strains has become critical for basic research and for investigating causes of and potential treatments for human disease. The number of genes in mice that have enough characterization to be given descriptive names now exceeds 12,000, perhaps one-third or one-quarter of the estimated total number of genes. Genome manipulation techniques that target specific genes (e.g. knock-outs, knock-ins, and conditional mutations) or that identify sequence variants (e.g. microsatellites and single nucleotide polymorphisms or SNPs), are providing new alleles for biological analysis. Although many factors can contribute to a phenotype, a widely used research approach focuses on the isolated effects of single genes and their mutant alleles on biological systems. An alternative approach is to study quantitative traits where multiple genes contribute to the observed phenotypes. Here a one-to-one relationship between gene and phenotype does not exist and, as in humans, the discovery of the genes underlying complex traits such as obesity and hypertension continues to be challenging, but should become more tractable as new mapping resources are developed.

6.7.2 Mouse Single Nucleotide Polymorphism (SNP) Databases

SNP technologies are being exploited for the investigation of human syndromes and diseases (Schork *et al*., 2000). Human SNP resources such as dbSNP (see Chapter 3) provide access to high-density SNP maps for humans. Large-scale discovery and genotyping of SNPs in mice is underway (Lindblad-Toh *et al*., 2000) and a limited quantity of mouse SNP data is already available in the Roche mouse SNP database (http://mousesnp.roche.com/) and the Whitehead/MIT SNP database (http://www-genome.wi.mit.edu/snp/mouse/). With the sequencing of large genomic regions of multiple mouse inbred strains, further SNP sets for mouse will be defined and could facilitate computer-based identification of QTL loci between inbred strains; one group has already reported some success using this method, but the approach is controversial at present (Chesler *et al*., 2001; Darvasi, 2001; Grupe *et al*., 2001).

6.7.3 Induced Mutant Resources

The rapid generation of many induced mutants of the mouse through the use of technologies such as homologous recombination and targeted knock-outs has created the need for a central facility to collect and distribute them to the scientific community. The Induced Mutant Resource (IMR) (http://www.jax.org/resources/documents/imr/) at The Jackson Laboratory is an example of a national clearing-house for the collection and distribution of a subset of genetically-engineered mice. The IMR maintains an on-line database to provide information about these strains. This information includes a description of the mutant phenotype, husbandry requirements and links to related resources. Another resource providing mouse mutants to the community is the Mutant Mouse Regional Resource Centres (http://www.mmrrc.org/). The MMRRC strive to enhance the availability of genetically-engineered mice for the study of human biology and disease. The European Mouse Mutant Resource (EMMA) (http://emma.rm.cnr.it/) is another repository for mouse mutant stocks.

6.7.4 Resources for Mouse Strain Characterization

Inbred strains in mouse have been specifically generated to facilitate the study of the genetic component of phenotypes including disease phenotypes by being able to isolate the impact of the mutant gene on a standard genetic background. With the advent of many new technologies, molecular information about the whole genome is becoming available for different inbred strains, and the need for standard evaluation of differences between inbred strains is apparent. New initiatives to study strain characteristics in mice and rats are underway with the attendant development of bioinformatics resources.

The Mouse Phenome Database (MPD; http://www.jax.org/phenome/) was established to provide a collection of baseline phenotypic data on commonly used and genetically diverse inbred mouse strains. Many institutions and investigators are involved in this effort to provide standard sets of strain characteristics for the most commonly used strains of mice. The MPD will enable investigators to identify appropriate strains for physiological testing and disease onset and susceptibility.

6.7.5 Phenotypic Variants

In contrast with the reliance on the gene-by-gene approach to discovery of functions and roles for genes and for the investigation of diseases and disorders, a recent development has been the use of systematic large-scale phenotype-driven mutagenesis studies in the mouse. This approach uses chemical or physical disruption of the genome followed by identification of putative mutants using a series of phenotypic screens for particular traits. This phenotype-driven approach to genome characterization has an important role to play in linking gene identification with gene function. This approach will allow researchers to better understand the molecular basis of diseases through the identification of mutants that develop the same or similar phenotypes but that have mutations in different genes. Furthermore, a full appreciation of the genetic basis of a disease requires that the phenotypes associated with multiple alleles of the same gene be studied to identify hypomorphs, alleles that confer gain of function, etc. Although it is unclear how much of the genome can be saturated with this approach, these projects will provide the community with a vast array of new phenotypes for biological analysis.

6.7.6 ENU Mutagenesis Centres

Several public large-scale ENU mutagenesis projects are already underway and are providing new models for the study of disease and gene function to the community (Brown and Nolan, 1998; De Angelis *et al.*, 2000; Justice *et al.*, 1999; Nolan *et al.*, 2000) (Table 6.1). Some of the mutagenesis centres are working in several disease areas to identify new mutants including the ENU Mutagenesis Programme at Harwell (http://www.mgu.har.mrc.ac.uk/mutabase/), the RIKEN Mouse Functional Genomics Group (http://www.gsc.riken.go.jp/Mouse/), and the GSF ENU Mouse Mutagenesis Screen Project (http://www.gsf.de/isg/groups/enu-mouse.html). Several of these mutagenesis centres are focusing on the identification of new mutant mice to serve as models for neurological disorders (Moldin *et al.*, 2001) including the Neuroscience Mutagenesis Facility at The Jackson Laboratory (http://www.jax.org/nmf/), the Neurogenomics Centre at Northwestern University (http://genome.northwestern.edu/), the Tennessee Mouse Genome Consortium (http://www.tnmouse.org/) and the McLaughlin Research Institute (http://www.montana.edu/wwwmri/enump.html). The mutagenesis facility at the Baylor College of Medicine (http://www.mouse-genome.bcm.tmc.edu/ENU/MutagenesisProj.asp) is focusing on developmental defects. The Medical Genome Centre in Australia focuses on cancer-related phenotypes (http://jcsmr.anu.edu.au/group_pages/mgc/CancerGenLab.html). The Mouse Heart, Lung,

TABLE 6.1 Mouse Mutagenesis Centres and Databases

Mutagenesis Centre	Disease Focus	URL
ENU Mutagenesis Programme (Harwell)	General	http://www.mgu.har.mrc.ac.uk/mutabase/
RIKEN Mouse Functional Genomics Group	General	http://www.gsc.riken.go.jp/Mouse/
GSF ENU Mouse Mutagenesis Screen Project	General	http://www.gsf.de/isg/groups/enu-mouse.html
Neuroscience Mutagenesis Facility at The Jackson Laboratory	Neurological	http://www.jax.org/nmf/
Neurogenomics Centre at Northwestern University	Neurological	http://genome.northwestern.edu/
Tennessee Mouse Genome Consortium	Neurological	http://www.tnmouse.org/
McLaughlin Research Institute	Neurological	http://www.montana.edu/wwwmri/enump.html
Baylor College of Medicine	Developmental disorders	http://www.mouse-genome.bcm.tmc.edu/ENU/MutagenesisProj.asp
Medical Genome Centre (Australia)	Cancer	http://jcsmr.anu.edu.au/group_pages/mgc/CancerGenLab.html
The Mouse Heart, Lung, Blood, and Sleep Disorders Centre (JAX)	Cardiovascular	http://www.jax.org/hlbs/index.html

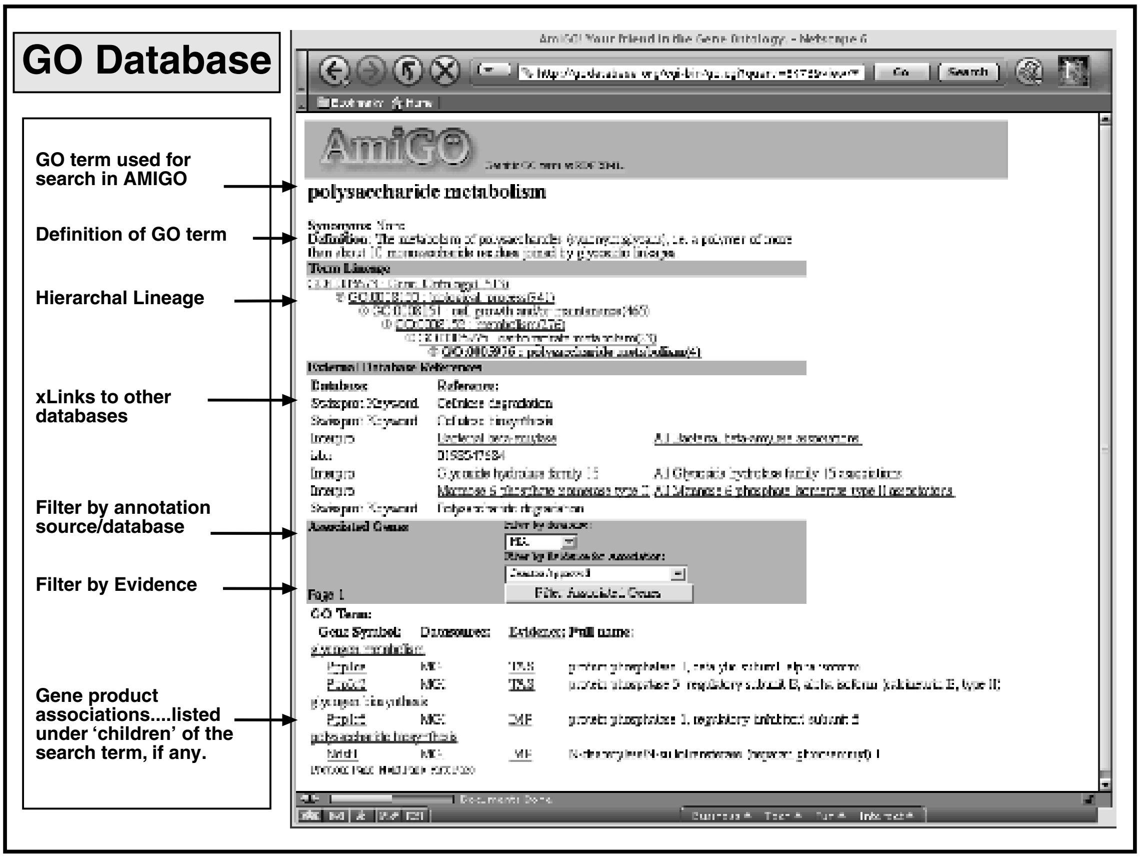
Blood, and Sleep Disorders Centre at The Jackson Laboratory is focusing on the identification of new mutants for cardiovascular diseases (http://www.jax.org/hlbs/index.html).

6.8 FUNCTIONAL GENOMICS

In the post-genome world, mouse and rat models will be heavily used for investigation of gene function and disease pathogenesis (Schimenti and Bucan, 1998; Temple *et al.*, 2001; Zheng *et al.*, 1999). With the completion of the mouse genome, attention can move to genome-wide screens for gene expression and systematic investigation of gene function. The inclusion of functional information with gene annotations first appeared in the sequence data repositories. From the start, issues of quality control for data associations were evident. Evaluation of sequence similarities often led to the transfer of function information from one gene annotation report to another without experimental verification or any statement about the basis for the function assertion. The first detailed functional classification was developed to catalogue the genes of *Escherichia coli* (Riley, 1993). Since then, functional annotation schemes have been developed for single organisms, multi-organism databases, and for pathway-related systems (see Rison *et al.*, 2000 for review).

6.8.1 Gene Ontology

A recent effort initiated by several of the model organism databases has been the development of ontologies describing aspects of biology common to all organisms (GO Consortium, 2000, 2001). These 'controlled structured vocabularies' include defined terms and relationships for the domains of 'molecular function', 'biological process' and 'cellular component'. The Gene Ontology (GO) project (http://www.geneontology.org) is now a consortium of model organism databases, sequence information centres and other genome data providers. In addition to the development of the ontologies, the genome annotation groups contribute gene–GO association files to a central GO repository. MGI and RGD provide detailed GO annotations for mouse and rat genes respectively. A GO database (http://www.godatabase.org/) holds the ontologies, their definitions and relationships, and the contributed sets of gene–GO association files. The AMIGO browser (Figure 6.4) provides access to the data.


GO Database
GO term used for search in AMIGO
Definition of GO term
Hierarchal Lineage
xLinks to other databases
Filter by annotation source/database
Filter by Evidence
Gene product associations....listed under 'children' of the search term, if any.
AmiGO
Go
Search

polysaccharide metabolism

Synonyms: None
Definition: The metabolism of polysaccharides [illegible], i.e. a polymer of more than about 10 monosaccharide residues joined by glycosidic linkages.

Term Lineage

External Database References

Database:	Reference:	
Swissprot: Keyword	Cellulose degradation	
Swissprot: Keyword	Cellulose biosynthesis	
Interpro	[illegible]	[illegible]
[illegible]	[illegible]	
Interpro	Glycoside hydrolase family 15	All Glycoside hydrolase family 15 associations
Interpro	[illegible]	[illegible]
Swissprot: Keyword	Polysaccharide degradation	

Associated Genes

Page 1

Filter Associated Genes

GO Term:

Gene Symbol:	Datasource:	Evidence:	Full name:
[illegible]			
[illegible]	MGI	TAS	[illegible]
[illegible]	MGI	TAS	[illegible]
[illegible]			
[illegible]	MGI	IMP	[illegible]
[illegible]			
[illegible]	MGI	[illegible]	[illegible]

Figure 6.4 GO Database/AMIGO browser. The GO database (http://www.godatabase.org/) and AMIGO browser are recent additions to the tools and resources of the Gene Ontology project (http://www.geneontology.org/). Here a detail page from a query on the controlled GO term 'Polysaccharide metabolism' is displayed. The definition of the term and its relationship to other terms is shown. There are cross-links to other external keyword sets. The detail page has been expanded to show all mouse gene products annotated to this term. The gene product associations detail can be filtered by source of annotation (MGI, RGD, other contributing model organism or genome annotation groups, etc.) and by type of evidence (cf. sequence similarity, mutant phenotype, direct assay, etc).

6.9 RODENT DISEASE MODELS

The experimental manipulation of mice and rats for the purpose of creating animal models for human disease is implicit in the scientific endeavours detailed here. Mouse and rat models will continue to be the best models for experimental manipulation of the mammalian genome for the foreseeable future (Bedell, *et al.*, 1997). The fact that inbred strains exist, providing consistent homogenous genetic backgrounds for experimentation, allows the genesis of diseases characteristic of particular inbred strains to be studied, as well as the development and testing of therapeutic interventions. The occurrence of spontaneous or induced single gene mutations in these strains allows precise detailed studies of the multiple effects of that particular mutation. Targeted mutations that produce knock-out or conditional mutations permit researchers to mimic the molecular defect of human diseases. Comparative studies have uncovered many rodent mutations that reflect their counterpart human disease. Multigenic diseases and quantitative trait loci can be dissected in mice and rats using controlled crosses and through creation of specialized strains, such as congenics and consomics, which place particular parts of the genome from one strain onto the background of another strain (cf. Kwitek-Black and Jacob, 2001; Sugiyama *et al.*, 2001).

In addition to the discovery or creation of models that reflect the underlying genetics of particular disease states, researchers may also find it useful to study animal models that reflect phenotypic similarity alone. That is to say, there is a phenotypic similarity in the animal model to a human disease condition, and the animal model is useful for studying that phenotype, even though we do not know that the underlying genetic dysfunction is exactly the same. For example, many cancers have unknown genetic aetiology, but particular strains or mutants prone to the development of particular cancers can serve as effective animal models (Hann and Balmain, 2001).

The strength of using rat and mouse models clearly lies in the physiological research that has gone before and that is accelerating with micro-technology developments. A fuller understanding of the genetics of these organisms, coupled with the imminent availability of their genome sequences, will enhance our ability to analyse the functions of gene products and to dissect the molecular basis of phenotypes.

6.10 SUMMARY

With new technologies and methods, the pace of data acquisition only quickens. Simultaneously, there are now intense efforts underway to improve data integration and to support rapid access to and interactive use of molecular and related biological information. Biological databases and information resources existed long before the advent of computers and the internet. We are, however, yet developing and realizing the capacity that computers give us to use the databases not just as archives, but also as research tools. The future of computerized scientific databases and information resources will be in their ability to rapidly retrieve and manipulate data in response to complex queries. The full value of the information they contain can then be exploited to address outstanding scientific inquiries.

ACKNOWLEDGEMENTS

This work is supported in part by NIH/NHGRI grant HG00330, DOE grant FG99ER 62850, and HG02273. We thank Greg Cox, Cathy Lutz and Tim O'Brien for their critical reading of this paper; W. John Boddy for preparation of figures and Janice Ormsby for technical review.

REFERENCES

Avner P, Amar L, Dandolo L, Guenet JL. (1988). Genetic analysis of the mouse using interspecific crosses. *Trends Genet* **4**: 18–23.

Battey J, Jordan E, Cox D, Dove W. (1999). An action plan for mouse genomics. *Nature Genet* **21**: 73–75.

Bedell MA, Largaespada DA, Jenkins NA, Copeland NG. (1997). Mouse models of human disease. Part 11: Recent progress and future directions. *Genes Develop* **11**: 11–43.

Billingham RE, Silvers WK. (1959). Inbred animals and tissue transplantation immunity. *Transplant Bull* **6**: 399–406.

Bihoreau M-T, Gaugier D, Kato N, Hyne G, Lindpainter K, Rapp JP, *et al.* (1997). A linkage map of the rat genome derived from three F2 crosses. *Genome Res* **7**: 434–440.

Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT and the Mouse Genome Database Group (2001). The Mouse Genome Database (MGD): the model organism database for the laboratory mouse. *Nucleic Acids Res* **30**: 1–3.

Brown DM, Matise TC, Koike G, Simon JS, Winer ES, Zangen S, *et al.* (1998). An integrated genetic linkage map of the laboratory rat. *Mammal Genome* **9**: 521–530.

Brown SD, Nolan PM. (1998). Mouse mutagenesis — systematic studies of mammalian gene function. *Hum Mol Genet* **7**: 1627–1633.

Bult CJ, Krupke DM, Näf D, Sundberg JP, Eppig JT. (2001). Web-based access to mouse models of human cancers: the Mouse Tumor Biology (MTB) Database. *Nucleic Acids Res* **29**: 95–97.

Bult CJ, Richardson JE, Blake JA, Kadin JA, Ringwald M, Eppig JT and the Mouse Genome Informatics Group (2000). Mouse genome informatics in a new age of biological inquiry. In *Proceedings of the IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, pp. 29–32.

Castle WE, Wachter WL. (1924). Variations of linkage in rats and mice. *Genetics* **9**: 1–12.

Chesler EJ, Rodriguez-Zas SL, Mogil JS. (2001). *In silico* mapping of mouse quantitative trait loci. *Science* **294**: 2423.

Copeland NG, Jenkins NA. (1991). Development and applications of a molecular genetic linkage map of the mouse genome. *Trends Genet* **7**: 113–118.

Darvasi A. (2001). *In silico* mapping of mouse quantitative trait loci. *Science* **294**: 2423.

De Angelis MH, Flaswinkel H, Fuchs H, Rathkolb B, Soewarto B, Marschall S, *et al.* (2000). Genome-wide, large-scale production of mutant mice by ENU Mutagenesis. *Nature Gene* **25**: 444–447.

Dehal P, Predki P, Olsen AS, Kobayashi A, Folta P, Lucas S, *et al.* (2001). Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* **293**: 104–111.

Dietrich W, Katz H, Lincoln SE, Shin H-S, Friedman J, Dracopoli NC, *et al.* (1992). A genetic map of the mouse suitable for typing intraspecific crosses. *Genetics* **131**: 423–447.

Doolittle DP, Hillyard AL, Davisson MT, Roderick TH, Guidi JN. (1991). GBASE — The genomic database of the mouse, In *Fifth International Workshop on Mouse Genome Mapping, Lunteren, Netherlands*, p. 27.

Dracheva SV, Remmers EF, Chen S, Chang L, Gulko PS, Kawahito Y, *et al.* (2000). An integrated genetic linkage map with 1,137 markers constructed from five F2 crosses of autoimmune disease-prone and -resistant inbred rat strains. *Genomics* **63**: 202–226.

Dunn LC. (1920). Linkage in mice and rats. *Genetics* **5**: 325–343.

Eppig JT. (1992). Mouse DNA clones and probes. *Mammal Genome* **3**: 300–330.

Eppig JT, Blackburn RE, Bradt DW, Corbani LE, Davisson MT, Doolittle DP, *et al.* (1994). *The Encyclopedia of the Mouse Genome, an update. Third International Conference on Bioinformatics and Genome Research, Tallahassee*, p. 73.

European Backcross Collaborative Group (1994). Towards high resolution maps of the mouse and human genomes — a facility for ordering markers to 0.1 cM resolution. *Hum Mol Genet* **3**: 621–627.

Festing MFW, Staats J. (1973). Standardized nomenclature for inbred strains of rats, fourth listing. *Transplantation* **16**: 221–245.

Gill TJ, Smith GJ, Wissler RW, Kunz HW. (1989). The rat as an experimental animal. *Science* **245**: 269–276.

Glusman G, Rowen L, Lee I, Boysen C, Roach JC, Smit AF, *et al.* (2001). Comparative genomics of the human and mouse T Cell receptor loci. *Immunity* **15**: 337–349.

GO Consortium. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genet* **25**: 25–29.

GO Consortium. (2001). Creating the gene ontology resources: design and implementation. *Genome Res* **11**: 1425–1433.

Green MC. (Ed.) (1981). *Genetic Variants and Strains of the Laboratory Mouse*, 1st edn. Fischer Verlag: Stuttgart.

Greenhouse DD, Festing MFW, Hasan S, Cohen AL. (1990). Catalogue of inbred strains of rats. In *Genetic Monitoring of Inbred Strains of Rats*, Hedrich HJ (Ed.), Gustav Fischer: Stuttgart, pp. 120–132.

Gregory SG, *et al.* (2002). A Physical map of the mouse genome. *Nature* **418**: 743–50.

Grupe A, Germer S, Usuka J, Aud D, Belknap JK, Klein RF, *et al.* (2001). *In silico* mapping of complex disease-related traits in mice. *Science* **292**: 1915–1918.

Gunther E, Walter L. (2001). The major histocompatibility complex of the rat (*Rattus norvegicus*). *Immunogenetics* **53**: 520–542.

Haldane JBS. (1927). The comparative genetics of colour in rodents and carnivora. *Biol Rev Cambridge Phil Soc (London)* **2**: 199–212.

Hann B, Balmain A. (2001). Building ‘validated’ mouse models of human cancer. *Curr Opin Cell Biol* **13**: 778–784.

Helou K, Walentinsson A, Levan G, Stahl F. (2001). Between rat and mouse zoo-FISH reveals 49 chromosomal segments that have been conserved in evolution. *Mammal Genome* **12**: 765–771.

Hillyard AL, Doolittle DP, Davisson MT, Roderick TH. (1991). Locus map of mouse. *Mouse Genome* **89**: 16–30.

Ioannidu S, Walter L, Dressel R, Gunther E. (2001). Physical map and expression profile of genes of the telomeric class I gene region of the rat MHC. *J Immunol* **166**: 3957–3965.

James MR, Lindpainter K. (1997). Why map the rat? *Trends Genet* **13**: 171–173.

Justice MJ, Noveroske JK, Weber JS, Zheng B, Bradley A. (1999). Mouse ENU mutagenesis. *Hum Mol Genet* **8**: 1955–1963.

Kwitek-Black AE, Jacob HJ. (2001). The use of designer rats in the genetic dissection of hypertension. *Curr Hyperten Rep* **3**: 12–18.

Levan G, Szpirer J, Szpirer C, Klinga K, Hanson C, Islam MQ. (1991). The gene map of the Norway rat (*Rattus norvegicus*) and comparative mapping with mouse and man. *Genomics* **10**: 699–718.

Lindblad-Toh K, Winchester E, Daly MJ, Wang DG, Hirschhorn JN, Laviolette JP, *et al.* (2000). Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genet* **24**: 381–386.

Little CC, Tyzzer EE. (1916). Further studies on inheritance of susceptibility to a transplantable tumor of Japanese waltzing mice. *J Med Res* **33**: 393–398.

Lyon MF, Searle AG. (Eds) (1989). *Genetic Variants and Strains of the Laboratory Mouse*, 2nd edn. Oxford University Press: Oxford.

Lyon MF, Rastan S, Brown SDM. (Eds) (1996). *Genetic Variants and Strains of the Laboratory Mouse*, 3rd edn. Oxford University Press: New York.

Manly KF. (1993). A Macintosh program for storage and analysis of experimental genetic mapping data. *Mammal Genome* **4**: 303–313.

Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, *et al.* (1997). High throughput fingerprint analysis of large-insert clones. *Genome Res* **7**: 1072–1084.

Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, *et al.* (2001). VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**: 1046–1047.

McCarthy LC, Terrett J, Davis ME, Knights CJ, Smith AL, Critcher R, *et al.* (1997). A first-generation whole genome-radiation hybrid map spanning the mouse genome. *Genome Res* **7**: 1153–1161.

MGD. (2002). Statistics for number of localized genes. ftp://ftp.informatics.jax.org/pub/informatics/reports/MGD_Stats.sql.rpt [1 January 2002].

Moldin SO, Farmer ME, Chin HR, Battey JF Jr. (2001). Trans-NIH neuroscience initiatives on mouse phenotyping and mutagenesis. *Mammal Genome* **12**: 575–581.

Mouse Genome Sequencing Consortium (2002). Initial Sequencing and Comparative analysis of the Mouse genome. Nature, in press.

Nadeau JH, Grant P, Kosowsky M. (1991). Mouse on human homology map. *Mouse Genome* **89**: 31–36.

Nilsson S, Helou K, Walentinsson A, Szpirer C, Nerman I, Stahl F. (2001). Rat–mouse and rat–human comparative maps based on gene homology and high-resolution zoo-FISH. *Genomics* **74**: 287–298.

Nolan PM, Peters J, Strivens M, Rogers D, Hagan J, Spurr N, *et al.* (2000). A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nature Genet* **25**: 440–443.

Oeltjen JC, Malley TM, Muzny DM, Miller W, Gibbs RA, Belmont JW. (1997). Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res* **7**: 315–329.

Osoegawa K, Tateno M, Woon PY, Frengen E, Mammoser AG, Catanese JJ, *et al.* (2000). Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. *Genome Res* **10**: 116–128.

RatMap (2002). Statistics for number of localized genes. http://ratmap.gen.gu.se [1 January 2002].

Richardson JE, Eppig JT, Nadeau JH. (1995). Building an integrated mouse genome database. *IEEE Eng Med Biol* **14**: 718–724.

Riley M. (1993). Functions of the gene products of *Escherichia coli*. *Microbiol Rev* **57**: 862–952.

Ringwald M, Eppig JT, Begley DA, Corradi JP, McCright IJ, Hayamizu TF, *et al.* (2001). The Mouse Gene Expression Database (GXD). *Nucleic Acids Res* **29**: 98–101.

Rison SCG, Hodgman TC, Thornton JM. (2000). Comparison of functional annotation schemes for genomes. *Funct Integ Genomics* **1**: 56–69.

Rodriguez-Tomé P, Lijnzaad P. (2001). RHdb: the Radiation Hybrid database. *Nucleic Acids Res* **29**: 165–166.

Rowe LB, Nadeau JH, Turner R, Frankel WN, Letts VA, Eppig JT, *et al.* (1994). Maps from two interspecific backcross DNA panels available as a community genetic mapping resource. *Mammal Genome* **5**: 253–274.

Schimenti J, Bucan M. (1998). Functional genomics in the mouse: phenotype based on mutagenesis screens. *Genome Res* **8**: 698–710.

Schork NJ, Fallin D, Lanchbury JS. (2000). Single nucleotide polymorphisms and the future of genetic epidemiology. *Clin Genet* **58**: 250–264.

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, *et al.* (2000). PipMaker: A web server for aligning two genomic DNA sequences. *Genome Res* **10**: 577–586.

Snell GD. (1941). Genes and chromosome mutation. In *Biology of the Laboratory Mouse*, 1st edn, Snell GD. (Ed.). McGraw-Hill: New York, pp. 234–247.

Soderlund C, Humphrey S, Dunhum A, French L. (2000). Contigs built with fingerprints, markers and FPC V4.7. *Genome Res* **10**: 1772–1787.

Staats J. (1985). Standardized nomenclature for inbred strains of mice: eighth listing. *Cancer Res* **45**: 945–977.

Steen RG, Kwitek-Black AE, Glenn C, Gullings-Handley J, Van Etten W, Atkinson OS, *et al.* (1999). A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat. *Genome Res* **9**, AP1–8, insert.

Strausberg RL, Feingold EA, Klausner RD, Collins FC. (2000). The mammalian gene collection. *Science* **286**: 455–457.

Stroesser G, Baker W, van den Broek A, Camon E, Garcia-Pastor M, Kanz C, *et al.* (2001). The EMBL nucleotide sequence database. *Nucleic Acids Res* **29**: 17–21.

Sugiyama F, Yagami K, Paigen B. (2001). Mouse models of blood pressure regulation and hypertension. *Curr Hyperten Rep* **3**: 41–48.

Temple LKF, McLeod RS, Gallinger S, Wright JG. (2001). Defining disease in the genomics era. *Science* **293**: 807–808.

The RIKEN Exploration Research Group Phase II Team and the FANTOM Consortium. (2001). Functional annotation of a full-length mouse cDNA collection. *Nature* **409**: 685–690.

Twigger S, Lu J, Shimoyama M, Chen D, Pasko D, Long H, *et al.* (2002). Rat Genome Database (RGD): mapping disease onto the genome. *Nucleic Acids Res* **30**: 125–128.

Van Etten WJ, Steen RG, Nguyen H, Castle AB, Slonim DK, Ge B, *et al.* (1999). Radiation hybrid map of the mouse genome. *Nature Genet* **22**: 384–387.

Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE. (2000). Human–mouse genome comparisons to locate regulatory sites. *Nature Genet* **26**: 225–228.

Watanabe TK, Bihoreau MT, McCarthy LC, Kiguwa SL, Hishigaki H, Tsuji A, *et al.* (1999). A radiation hybrid map of the rat genome containing 5,255 markers. *Nature Genet* **22**: 27–36.

Watanabe TK, Ono T, Okuno S, Mizoguchi-Miyakita A, Yamasaki Y, Kanemoto N, *et al.* (2000). Characterization of newly developed SSLP markers for the rat. *Mammal Genome* **11**: 300–305.

Wei S, Wei K, Moralejo DH, Yamada T, Izumi K, Matsumoto K. (1998). An integrated genetic map of the rat with 562 markers from different sources. *Mammal Genome* **9**: 1002–1007.

Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, *et al.* (2001). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **29**: 11–16.

Zhao S, Shatsman S, Ayodeji B, Geer K, Tsegaye G, Krol M, *et al.* (2001). Mouse BAC ends quality assessment and sequence analyses. *Genome Res* **11**: 1736–1745.

Zheng BJ, Mills AA, Bradley A. (1999). A system for rapid generation of coat color-tagged knockouts and defined chromosomal rearrangements in mice. *Nucleic Acids Res* **27**: 2354–2360.

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)

CHAPTER 7

Genetic and Physical Map Resources—an Integrated View

MICHAEL R. BARNES

GlaxoSmithKline Pharmaceuticals, Harlow, Essex, UK

7.1 INTRODUCTION

Not so many years ago, maps of the human genome were restricted to a handful of very low resolution diallelic RFLP marker maps of specific loci. Physical mapping following linkage analysis required a laborious laboratory-based process of contig construction using yeast and bacterial artificial chromosome (YAC and BAC) clones or cosmids. This involved consecutive rounds of library screening and clone characterization to identify overlaps between clones and build contigs. In recent years, as the human genome sequence nears completion, practical approaches to the characterization of genomic loci have changed quite dramatically. Today the process which took many months or even years can be completed in an afternoon using web-based resources. These tools might lead us to believe that the human genome sequence is the only map we need to know, but it actually represents just one dimension of a multifaceted map. Other maps including genetic, cytogenetic and radiation hybrid maps, represent different aspects of the structure, content and behaviour of chromosomes. These properties really need to be integrated with sequence-based maps to fully understand the properties and genomic landmarks that influence genes, mutation and human evolution.

As this book goes to press, the human genome is still unfinished and in the strictest sense it is likely to remain so for several years to come. For example, in April 2002 the human genome draft sequence reached 97.8% coverage, however only 63% of sequence was flagged as finished with 34.8% flagged as draft. The target date for final human sequence completion is 2003. However this may be a moving target, as a combination of contig errors and molecularly intractable regions are likely to continue to keep the genome in at least a partial draft state for many years to come. With this in mind, it is probably pragmatic to assume that the genome will remain unfinished in parts until at least 2005. Mouse genome sequencing is rapidly catching up with human sequencing, with the mouse also projected to finish in 2003. Other mammalian species such as the rat, dog and chimpanzee are further behind, although further genome sequencing will be assisted by existing genomes. The ‘pioneer’ genome sequences (human and mouse) will be used to span gaps and build contigs by comparison with existing contigs. This approach is already being used to accelerate the mouse and human genome sequencing projects, as both assemblies are being used to span gaps in each respective genome assembly (J. Mullikin, personal communication).

As we are becoming more aware of the difficulties of completing whole genome sequences, the role of physical and genetic maps is changing. Generation of new maps continues to be the first line of study for organisms with poorly characterized genomes. But where the genome sequencing of an organism is advanced, emphasis on maps is shifting to a role in the finishing and QC of existing sequencing maps. With this proviso in mind and with a specific focus on human maps, this chapter will review genetic and physical maps as they are being directly applied and integrated with the human genome and other sequenced mammalian genomes. We will not attempt to cover the full complexity of all forms of maps, or attempt to describe the use of these maps to enable the study of unsequenced organisms. Instead we will review the principles and informatics issues that apply to this area, with a focus on the data which is most likely to be useful to the human geneticist. For example we will examine the use of genetic and physical maps to check the order and orientation of marker maps and genomic contigs. For researchers who wish to construct new genetic and physical maps without sequence data we direct the reader to specialist texts in this research area.

7.1.1 What is a Genome Map?

At the most basic level, a genome map is a collective set of markers with known relative positions. A marker could be any genomic element with a uniquely identifiable sequence or property. Markers can exist in many different forms, such as non-polymorphic sequence tagged sites (STS) which act as a unique anchor or SNPs and short tandem repeats (STR), which act as both unique anchors and markers for differentiation between individuals. Genomic maps are divided into two broad categories. Polymorphic markers are used to construct genetic maps and either polymorphic or non-polymorphic markers are used to construct physical maps.

7.2 GENETIC MAPS

The genetic linkage map is a key concept which gives a fundamental insight into the genetic nature of the genome. Genetic linkage maps inform on more than just order of markers, they also give a measure of the underlying genetic recombination that occurs in a particular chromosomal region. Linkage maps show the relative locations of specific DNA markers along the chromosomes of related individuals. Any inherited physical or molecular characteristic that differs among individuals and is easily detectable is a potential genetic marker, for this reason polymorphic markers, such as SNPs and STRs are particularly suited to genetic map construction as they are plentiful, easy to characterize precisely and amenable to laboratory automation (see Chapter 3 for a review of SNPs and STR markers).

Genetic maps are constructed by evaluating the genotypes of a set of markers in groups of related individuals. This raw mapping data is analysed by software packages, such as MapMaker (Lander *et al.*, 1987; reviewed in Chapter 12) which construct genetic maps by observing how frequently the alleles at any two markers are inherited together. The closer the markers are, the less likely it is that a recombination event will separate the alleles, and the more likely it is that they will be inherited together. Thus, unlike physical maps, the distance between markers on a genetic map is not measured in any kind of physical unit; it is a measure of the recombination frequency between those two markers. This genetic map unit is measured in centimorgans (cM). The distance between two markers would be measured as 1 cM if both markers are separated by recombination on 1% of occasions. Genetic distance has an average correlation with the actual physical distance between markers, on average in humans 1 cM is equivalent to 1 Mb (this ratio varies widely between other species). The 1 cM : 1 Mb ratio is often used as a rule of thumb, but it is important to recognize that this is a genome-wide average and can often diverge significantly from this ratio between different regions of the human genome. The genetic/physical ratio also differs considerably between genders, as recombination frequencies vary between males and females. To overcome these differences, genetic maps typically report distances for each sex and a ‘sex-averaged’ distance that integrates male and female recombination frequencies.

7.2.1 Human Genetic Maps

A range of genome-wide human genetic maps has now been published at various resolutions. Most genetic maps are based on STR markers, although a genome-wide SNP linkage map has also been published recently (T. C. Matise *et al.*, unpublished data). Most genome-wide linkage maps are constructed with a marker framework spaced at 2.5–10-cM intervals. Denser marker maps have not been widely used for linkage analysis, as the focus of analysis is on a small number of meiotic events observable within a family. These meiotic events do not require a very dense map of markers to find evidence for possible co-segregation of a disease-influencing gene with marker locus alleles. Higher resolution genetic maps have been described, but they are generally restricted to specific chromosomal regions, such as the long arm of chromosome 21 (Lynn *et al.*, 2000), where they have been used to refine initial linkage analysis. Ideally, to be maximally informative, genetic markers need a relatively high level of heterozygosity (>0.6). This provides a high likelihood that a marker (or cluster of SNPs) will be different between any two copies of a chromosome. Markers with lower heterozygosity, for example, SNPs which range in heterozygosity from □0.1–0.3, need to be used in higher density to give a similar level of information.

The three main genetic maps were developed by Genethon, the Marshfield Institute and the SNP consortium (TSC) (see Table 7.1 for a comparison). The Genethon and Marshfield maps are widely indexed by mapping tools, such as MapViewer and GDB (see below). The newer TSC map is also likely to be available in these tools in the near future.

7.2.2 The Genethon Genetic Linkage Map

The Genethon human linkage map was the first whole genome genetic map to exclusively use STR markers; previous maps were based on less informative RFLPs (which are actually uncharacterized SNPs). The 5264 markers in the Genethon map have a mean heterozygosity of 0.7, which makes it more informative than previous maps. The map was constructed with data from eight CEPH families (comprising 186 meioses) so the fine order of markers is not well resolved, other than by localization within a particular chromosomal region. The map spans a sex-averaged genetic distance of 3699 cM. The average interval size is 1.6 cM, 59% of the map is covered by intervals of 2 cM at most and 1% remains in intervals above 10 cM. The map comprises 2335 positions, of which 2032 could be ordered with an odds ratio of at least 1000 : 1 against alternative orders. This high level of statistical confidence in marker order was subsequently used by DeWan *et al.* (2002), to highlight a number of discrepancies in the order and orientation of clones in the human genome draft assembly. Genethon map data can be accessed at the Genethon website (www.genethon.fr) and the Washington University, St Louis website (www.genlink.wustl.edu/genethon_frame/).

TABLE 7.1 Human Genetic Maps

Map	Genethon	Marshfield	TSC
Marker type	STRs	STRs	SNPs
Marker no.	5264	8325	2679
Av. heterozygosity	0.7	0.68	0.76
Resolution (kb)	1.6 cM	1.3 cM	2.5 cM
Reference	Dib *et al.* (1996)	Broman *et al.* (1998)	Matise *et al.* (unpublished data)

7.2.3 The Marshfield Genetic Linkage Map

The Marshfield genetic linkage map improved on the Genethon map, by offering a larger marker number and a slightly higher resolution. Like the Genethon map, the Marshfield map was constructed with data from eight CEPH families and therefore fine order is still poorly resolved. In particular, markers which are separated by little or no genetic distance generally have no recombination events separating them, and so they are presented in arbitrary order. Accurate ordering information for these markers can be obtained by cross referencing STS marker location with human physical maps, such as RH maps or the human genome sequence itself. The Marshfield database (http://research.marshfieldclinic.org/genetics/), provides a well-documented range of five genome scan marker panels (genome-wide screening sets 6–10), selected from the Marshfield map. These marker panels were initially developed from the first human linkage mapping screening set from the Cooperative Human Linkage Centre (CHLC) (Murray *et al.*, 1994). Each Marshfield marker panel provides a progressively higher density of markers, culminating in set 10 which consists of 405 di, tri and tetra-nucleotide repeat markers with an average spacing of 9 cM. Each marker set is also grouped by allele size so that each panel can be loaded into the same lane or capillary. Primers for marker set 10 are commercially available from Research Genetics, in unlabelled and fluorescent dye-conjugated forms (http://www.resgen.com/).

7.2.4 TSC SNP Linkage Map

Technology developments have brought the cost of SNP genotyping far below the cost of STR genotyping. This has led to calls for the development of a SNP-based linkage map. The only argument against the implementation of such a map is the lower heterozygosity of a single SNP compared to a polymorphic STR (Kruglyak, 1997; see Chapter 8 for a discussion of this issue). Use of single SNPs at similar densities to STRs would essentially be equivalent to the original and less informative RFLP maps. Two related solutions have been proposed to overcome this problem. The first solution is to use a 3–8-fold increase in SNP marker densities to produce an evenly spaced map (Kruglyak, 1997). The second is to use multiple clusters of two to three SNPs in linkage analysis at a similar density to STRs. These SNP clusters provide approximately the same amount of information as an STR in terms of heterozygosity (Goddard and Wijsman, 2002).

Matise *et al.* (unpublished data) used the SNP cluster approach to construct a whole genome SNP linkage map. To do this they selected 666 physically and genetically mapped polymorphic STS anchor loci at 5-cM intervals across the human genome. Ten or more SNPs were then characterized across each STS locus. SNPs were assessed for genotyping success rates, assay quality, allele frequencies (ideally $>20\%$), multi-SNP haplotype heterozygosities (ideally >0.6) and levels of linkage disequilibrium (SNPs in LD with each other were avoided). The three most informative markers per STS locus were then selected to maximize multi-SNP haplotype heterozygosities, to create an informative SNP cluster at each map position. Two thousand SNPs were selected and genotyped in 661 individuals from 48 CEPH reference pedigrees (http://www.cephb.fr/). Linkage maps were constructed without reference to any other mapping or sequence position information. This generated a map with an average resolution of 5 cM; to improve this, a further set of SNPs were identified at half-way points between the SNP clusters loci and were similarly evaluated. The single most informative SNPs at each of these positions were identified ($N = 679$) and genotyped in the CEPH pedigrees. These ‘single’ SNPs were added to the cluster linkage map to produce a final SNP map with a 2.5-cM resolution.

The construction of this map was supported by the SNP Consortium (TSC), all the data and results are available at the TSC website (http://snp.cshl.org).

7.2.5 SNP-based Haplotype and Linkage Disequilibrium (LD) Maps

As new SNPs arise at different loci and at different points in time, groups of neighbouring SNPs may show distinctive patterns of co-inheritance or LD, which are arranged into distinct haplotypes between individuals. The great abundance of SNPs across the genome creates an opportunity to exploit this haplotypic diversity in association studies by identifying SNPs which capture or 'tag' the majority of common human haplotypes. This enables the construction of very efficient maps, which capture maximal diversity with a minimal number of SNPs. Such haplotype tags have already been used to screen candidate genes. For example, Johnson *et al.* (2001) re-sequenced nine genes to identify common SNP haplotypes among 122 SNPs. Once these haplotypes were defined they were able to define just 34 SNPs or 'haplotype tags' which identified all the haplotypes across the genes. Extension of this principle across the genome would enable the construction of powerful haplotype-based maps which could capture most common haplotype diversity with a minimal number of SNP markers. At the time of going to press, such a map does not exist in the public domain, although at least one company has this data. A public domain genome-wide haplotype/LD map is likely to become available early in 2004 if not sooner.

Some data is already available publicly. Public domain LD or haplotype maps are available for three chromosomes, these have been generated by two distinct methods and consequently the exact nature of the data presented differs between the maps. Orchid Biosciences Inc. in collaboration with the TSC have published a SNP-based map of chromosome 19 which will be available from the TSC website before this book goes to press (Michael Phillips, personal communication); Dawson *et al.* (2002) published a SNP-based LD map of chromosome 22 and Perlegen Inc. published a SNP-based haplotype map of chromosome 21 (Patil *et al.*, 2001). We take a closer look at the Perlegen map data in Chapter 9.

7.3 PHYSICAL MAPS

While genetic maps display the linear order of genes or markers and the recombination between them, they do not give reliable information on the physical distance between markers and genes. By contrast a physical map has an absolute and invariant base-pair scale, which defines the physical distance between markers. Two markers may be very close genetically, i.e. very little recombination occurs between them, but very far apart physically. The difference between genetic and physical maps may seem academic, however if a trait or disease is localized on a physical map between two molecular markers it is important to identify the amount of recombination across the region, to select an appropriately dense panel of markers to detect a genetic association. Conversely if a genetic map places a trait or disease between two molecular markers, it is useful to know if that distance represents 1 kb, 1 Mb or further still, to define the likely number of genes or regulatory regions in the locus.

7.3.1 Cytogenetic Maps

There are many different types of physical maps; the first identified and lowest resolution physical map of the human genome is the cytogenetic map. This type of map is based on the distinctive banding patterns of stained chromosomes. Detailed measurements of these patterns were originally used to define the gross physical size of human chromosomes, and led to the size-based sorting of the autosomal chromosomes from chromosome 1, the largest chromosome, to chromosome 22, the smallest. Unsurprisingly these early efforts at physical mapping were quite inaccurate and prone to distortion by differential contraction, which led to the incorrect ordering of chromosome 19 which is actually slightly smaller than chromosome 20 (Morton, 1991). Use of cytogenetic map locations is still remarkably prevalent, perhaps due to the ease of use of the vocabulary of cytobands, e.g. 1q32, 22q11, etc., to describe and cluster groups of genes and loci. Interestingly the cytobanding recognized by early biologists is not just decorative, but in fact the dark cytobands represent regions of higher average GC content, while light cytobands have a lower average GC content (Nimura and Gojobori, 2002). The region where a transition occurs between a dark and light cytoband is known as an isochore, these regions often show a remarkably increased rate of recombination (Eisenbarth *et al.*, 2000). This may make it important to pay special attention to genes and possible regulatory elements in these regions; we specifically address this issue in Chapter 10.

7.3.2 Fluorescence *In Situ* Hybridization (FISH) Mapping

At best a cytogenetic map could be used to locate a DNA fragment to a region of about 10 Mb — the size of a typical chromosome band. Fluorescence *in situ* hybridization (FISH) mapping, is a form of cytogenetic mapping that allows orientation and mapping of DNA sequences to a much higher resolution. Initially FISH resolved markers within 2 Mb, but further development of the FISH method, using chromosomes in interphase when they are less compact, increased map resolution further to around 100 kb. As FISH does not rely on a recombinant map but instead maps a chromosome directly, this has made FISH an important method for the QC of recombinant maps and clone contigs. The level of resolution achieved with interphase FISH, also makes this method directly applicable to the analysis of observable physical traits associated with chromosomal abnormalities, such as prenatal defects or cancer breakpoints. All of these applications are likely to keep the method in regular use well beyond the availability of a complete human genome.

7.3.3 Radiation Hybrid (RH) Mapping

Early physical mapping advanced considerably with the publication of the radiation hybrid (RH) mapping method. Goss and Harris (1975) irradiated human fibroblast chromosomes and fused the resulting fragments with recipient rodent cells. The observed patterns of co-transference of markers in a collection of hybrid cells allowed estimates to be made of linear order and distance between markers by assuming that distant markers are more likely to be separated in different hybrid cell lines than closer markers. The RH mapping technique was refined by Cox *et al.* (1990) who irradiated donor somatic cell hybrids, which contained just a single copy of one human chromosome, and fused the fragments with rodent cells. Several whole genome RH panels were developed in the 1990s which allowed the construction of genome maps containing thousands of STS markers (Gyapay *et al.*, 1996; Stewart *et al.*, 1997). The human RH map finally reached a high-resolution apex, with the development of the TNG panel (Lunetta *et al.*, 1996), which was used to generate an RH map of the human genome consisting of 40,322 STSs (Olivier *et al.*, 2001). From the 40,322 STSs mapped to the TNG radiation hybrid panel, only 3604 (9.8%) were absent from the unassembled draft sequence of the human genome.

7.3.4 Human RH-mapping Panels

Three main radiation hybrid panels have been used for mapping STSs and constructing RH maps, each offers a different level of resolution based on the dose of irradiation. The GB4 RH panel (constructed by using 3000 rad of X-rays) and the G3 RH panel (10,000 rad of X-rays) will resolve markers at 1-Mb and 260-kb intervals respectively, both providing a good long-range continuity for mapping (Deloukas *et al.*, 1998). In contrast, the Stanford TNG panel (50,000 rad of X-rays) allows STS resolution down to 60–100 kb with high confidence (Lunetta *et al.*, 1996). The price of this increased resolution is that a large number of STSs need to be scored to produce good long-range continuity. Olivier *et al.* (2001) found a solution to this by using the TNG panel in conjunction with the Stanford G3 panel to produce an RH map with high-resolution and contiguity. Publication of this map saw a shift in the role of human RH-mapping, from a direct role in mapping new genes to a primarily curatorial role to enable the QC and assembly of the human genome.

RH maps provide a marker order confidence supported by LOD (logarithm of the odds ratio of linkage versus no linkage) scores between adjacent markers, coupled with distance measures between markers. Calculation of distance is based on the frequency of breakage between two markers in the radiation hybrid clones which is measured in centiRays (cR). There is a direct linear correlation between cR units and physical distance in kb, which is fairly constant across any given RH panel. The kilobase equivalent of the centiRay unit differs between RH maps. 1 cR on the TNG map corresponds to an average of 2 kb of physical distance, whereas 1 cR on the G3 map corresponds to a physical distance of 24 kb and a distance of 260 kb on the GB4 map. Table 7.2 illustrates the main features of all three panels and Table 7.3 illustrates the main RH maps generated from these panels. RH panels are available from Research Genetics (http://www.resgen.com/).

TABLE 7.2 Human Radiation Hybrid Panels

Panel	GeneBridge4 (GB4)	Stanford G3	Stanford TNG
X-ray dosage	3000 rad	10,000 rad	50,000 rad
Cell lines	93	83	90
Average retention	30%	18%	16%
Av. Frag. Size	10 Mb	4 Mb	800 kb
Resolution	Low	Medium	High
Resolution (kb)	1000	267	60
Reference	Gyapay *et al.* (1996)	Stewart *et al.* (1997)	Lunetta *et al.* (1996)

TABLE 7.3 Human Radiation Hybrid Maps

Map	GeneMap 99-GB4	GeneMap 99-G3	Stanford TNG	NCBI Integrated
Marker panel	GB4	G3	TNG & G3	G3 & GB4
Marker type	STS	STS	STS	STS
Marker no.	45758	7061	40322	23723
Reference	Schuler *et al.* (1996)	Deloukas *et al.* (1998)	Olivier *et al.* (2001)	Agarwala *et al.* (2000)

Comprehensive RH maps generated from these panels can be viewed and integrated with other maps in GDB, MapViewer and other applications (see below). Novel STS markers can be placed on these existing frameworks by PCR, screening STSs against the three RH panels and submitting the results to a web server, several of which are available. For G3 and TNG RH maps Stanford run a server at http://www-shgc.stanford.edu/RH/index.html. The EBI also runs an RH map server which includes all three human panels and also mouse, rat, pig and zebrafish panels (http://corba.ebi.ac.uk/RHdb/RHdb.html). The Whitehead Institute also maintains a GB4 server (http://www-genome.wi.mit.edu/cgi-bin/contig/rhmapper). Data submissions to all three servers are in a binary format to indicate presence or absence of a PCR product in each hybrid bin, e.g. G3.STS1 11000010000 01000000110011001000001100100010000011101000110000000110.

7.4 PHYSICAL CONTIG MAPS

Genetic maps and cytogenetic maps fulfilled many of the short-term goals of the human genome project — to develop low to medium resolution genetic and physical maps of the genome. They have also facilitated longer term goals by assisting in the construction of the more-precise high resolution maps at increasingly finer resolutions needed to organize systematic sequencing efforts (Korenberg *et al*., 1999). FISH and RH mapping in particular have enabled the development of a complex hierarchy of physical YAC and BAC clone contigs at a range of resolutions (Figure 7.1). These physical maps also became an important framework for positional cloning efforts in the years preceding the availability of a draft human genome. Accurate ordering of YAC and BAC clones (and subsequent

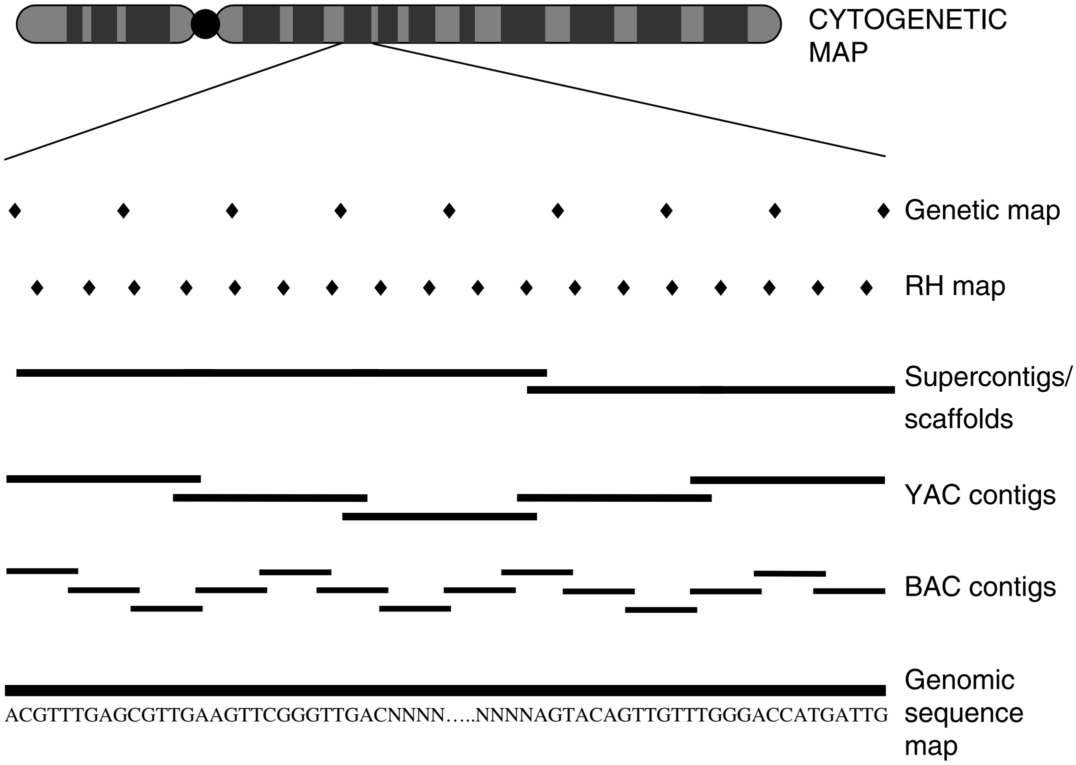


Figure 7.1 Physical and genetic maps used during the sequencing of the human genome. Many different maps were integrated to enable the construction of the framework for human genome sequencing (see Waterston *et al*. (2002) for a review).

shotgun reads) would not have been possible without existing genetic and physical maps which served as a scaffold for orientating, ordering and troubleshooting the human genome sequence assembly.

7.4.1 Yeast Artificial Chromosome (YAC) Maps

Yeast artificial chromosomes (YACs) are the lowest resolution physical clone contig maps, composed of overlapping YAC clones ranging in size from 300 kb–2 Mb. Before YACs were developed, the largest cloning vectors (cosmids) carried inserts of only 20 to 40 kb. YAC methodology drastically reduces the number of clones to be ordered; many YACs span entire human genes, making them a useful resource for further genomic study. The size of YAC inserts can often cause clone instability, which can lead to local rearrangements in the clone, this is the major drawback in the use of YACs for construction of physical contigs and underlines the need to QC YAC contigs with other available genetic and physical maps.

Several whole genome YAC maps are available, including a library of 33,000 YAC clones published by Chumakov et al. (1995). This library and other YAC clones can be obtained from a range of centres which are listed in the CEPH YAC library pages (http://www.cephb.fr/bio/ceph_yac.html).

7.4.2 Bacterial Artificial Chromosome (BAC) Maps

Bacterial artificial chromosomes (BACs), offer a further increase in map resolution, typically ranging in size from 100–300 kb. BAC clones are the primary vehicle of the public human genome sequencing project. Collections of human BACs estimated to represent more than a 10-fold redundancy of the human genome have been used to generate comprehensive BAC maps of the human genome. A minimally redundant set of these BACs have been assembled into physically separate contigs, representing the majority of the human genome. The sequence of these BACs is being determined by shotgun sequencing, where each BAC is digested with restriction enzymes and sub-cloned to generate a library of clones ranging from 0.5–5 kb. These clones are sequenced and assembled to form a complete BAC sequence, which are in turn assembled to form a complete chromosome.

BAC clone data can be accessed in many different ways, either directly from sequencing centres, or alternatively the NCBI have established a Human BAC resource page (http://www.ncbi.nlm.nih.gov/genome/cyto/hbrc.shtml). This page is a useful resource which centralizes information concerning currently available BAC maps and suppliers of BAC clones. Another useful database is GenMapDB (Morley *et al.*, 2001; http://genomics.med.upenn.edu/genmapdb/), which contains over 3000 mapped BAC clones spanning the genome. The database can be searched by map location or accession number. It is also possible to search for BAC clones by using BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) to search the 'HTGS' and 'Genome' divisions of GenBank. BAC sequences can also be accessed indirectly by using tools which show contig information for the draft human genome sequence, e.g. Ensembl, Map Viewer or UCSC human genome browser (see Chapter 5).

7.5 THE ROLE OF PHYSICAL AND GENETIC MAPS IN DRAFT SEQUENCE CURATION

Sequence tagged sites (STSs) are PCR-based anchors used to define a unique genomic sequence in an RH panel, YAC or BAC contig. All that is required to generate a new STS marker is 200–500 bp of unique sequence, this could be a sequence from the 3′ UTR of a transcript or any unique genomic region. Hence STS markers have been extensively identified from characterized genes, expressed sequence tags (ESTs) and random genomic fragments (Schuler *et al*., 1996). STS markers that include polymorphic sequences, such as microsatellites, are the central integrating force between genetic and physical maps. Common sets of such sequence-based markers can be easily screened and therefore can be used to integrate maps constructed by different mapping methods. RH panels and STS markers will play a critical role in the finishing of the human genome by providing a method to obtain markers from regions of the human genome that may be difficult to clone in conventional vector libraries. Hattori *et al*. (2000) found that up to 10% of certain gene-rich regions of human chromosome 21 were composed of such ‘hard-to-clone’ DNA. STSs that fail to hit available sequence can be used to screen different DNA libraries to close existing clone gaps in draft genome contigs. High resolution physical maps, such as the TNG map can also be valuable for curating draft genome contigs. Localization of RH markers to working draft sequences provides an independent measure of order and orientation for the clones underlying the draft sequence. Distances between markers can also be used to estimate the physical length of gaps between non-overlapping clones.

7.5.1 Electronic PCR (e-PCR)

Electronic PCR (e-PCR) is an *in silico* equivalent of the laboratory-based STS mapping process (Schuler, 1997; http://www.ncbi.nlm.nih.gov/cgi-bin/STS/nph-sts). The e-PCR tool at the NCBI maps known STSs from the dbSTS, GDB and RHdb databases to a user-submitted sequence. In a directly analogous process to PCR, e-PCR searches for sub-sequences within a query sequence that match known STS PCR primers and are in the correct order, orientation and spacing to be consistent with the PCR product size. These criteria eliminate the possibility of false positives (e.g. hits to psuedogenes or repeat sequences) that occur with other similarity searching methods such as BLAST. Electronic PCR is a valuable tool to assist in the integration of genomic sequence data with existing maps; this can be useful to assist genomic QC and to correlate genetic distances with physical distances. We offer detailed coverage of the use of this and other tools for genomic contig analysis in Chapter 9.

RH maps are playing a critical role in the QC and finishing of the human genome (see below), but once we have a finished genome, these maps may be of limited further use in humans. However, RH maps will continue to be the physical mapping method of choice for other organisms without extensive genome sequence. Human RH maps may also be of some limited use in the construction of comparative maps with other mammalian genomes (Kwitek *et al*., 2001). But, for the purposes of this chapter, we will focus on the direct integration of genetic and physical maps, with genome sequence as the ultimate integration framework. For consideration of non-human maps we refer the reader to Chapter 6 and other specialist texts.

7.6 THE HUMAN GENOME SEQUENCE — THE ULTIMATE PHYSICAL MAP?

The complete DNA sequence of the human genome will be an accurate physical map resolved down to a single base pair resolution, but we do not have this map as yet. Geneticists will need to work with a draft assembly of the human genome for a somewhat indeterminate number of years, until this task is truly finished. However, the draft genome assembly is still a very valuable asset for genetics, particularly if data are treated with care. With this in mind it is very important to be aware of some of the issues relating to the curation of draft sequences. Genetic and physical maps are one aid in this process.

For example, Olivier *et al.* (2001) used a 40,000-marker RH map to provide an estimate of the size and location of missing sequence in the human genome draft in relation to the existing sequence, and to provide order information for the 15,000 + clones that constitute the human genome working draft. They found that 9.8% of STS markers were absent from the October 2000 draft of the human genome. They suggest that these are likely to represent the 'hard-to-clone' regions of the human genome. Other studies have made similar observations (Hattori *et al.*, 2000) which suggests that a small intractable percentage of the human genome may remain in an unfinished state for longer than we may have anticipated.

Genetic maps are also playing an important role in the QC of human genomic sequence. DeWan *et al.* (2002) compared the genetic order of the Marshfield genome-scan markers (set 9 and 10) with their physical order in the April 2001 public golden path contig and the February 2001 Celera genome assembly. They found inconsistencies in 5 and 2% of the markers in the Celera assembly and the golden path assembly, respectively. The genetic order of these markers was supported with high confidence by a LOD of >3 and most discrepancies were not observed in both contigs, which suggests errors in the physical map order of both genome assemblies. Chromosome-by-chromosome breakdown of this data are available on a website: http://linkage.rockefeller.edu/maps/.

7.7 QC OF GENOMIC DNA — RESOLUTION OF MARKER ORDER AND GAP SIZES

The studies by DeWan *et al.* (2002) and Olivier *et al.* (2001) demonstrate the value of genetic and physical maps in the curation and QC of human genomic sequence contigs. As discussed previously the relationship between genetic distance (cM) and physical distance (Mb) is not uniform, however both genetic and physical mapping methods can resolve marker order to varying degrees of confidence, depending on the map characteristics. Marker or contig order across a locus can be validated by integrating information from different maps. The value and accuracy of different maps is not necessarily hierarchical or directly related to the density of the map. For example, one might assume that a dense RH map, or even a finished sequence map might be more accurate than a less dense genetic map, however as the studies above have shown, this does not always hold true. Using maps in an hierarchical manner may avoid the inevitable discordances between different maps, but this is not necessarily the best order for integration. In some cases for example, YAC STS content data may be more accurate than RH data, or a genetic map may be more reliable than a BAC contig. Both genetic maps and RH maps show a relative confidence in marker order by LOD scores using appropriate maximum likelihood statistical methods (Boehnke *et al.*, 1991). A LOD of 3.0 (odds 1000 : 1) or more is generally accepted as a strong indication of a contiguous relationship between markers. Comparison of LOD scores can help to integrate different data sources in an attempt to reach a consensus. But sometimes, all that can be done *in silico* is to flag up an unresolvable discrepancy between maps to receive special attention in the laboratory. Bioinformatic tools and databases can be a great help in the integration and evaluation of genetic and physical maps. MapViewer, GDB and UDB allow the user to compare and integrate maps; these are described below.

The UCSC human genome browser also provides graphical data on the positions of STS markers on the golden path versus their positions in other maps, including radiation hybrid, sex-averaged genetic (Marshfield), cytogenetic and YAC STS maps at the following URL: http://genome.ucsc.edu/goldenPath/mapPlots/.

It is also possible to view and integrate genetic and physical maps on an *ad hoc* basis using bioinformatics tools, such as Map View at the NCBI (reviewed below).

7.8 TOOLS AND DATABASES FOR MAP ANALYSIS AND INTEGRATION

There are some excellent tools which have recently become available for viewing the human genome sequence. Ensembl and the UCSC human genome browser are shining examples of the kind of biological data integration that geneticists need for their studies. But unfortunately they are lacking in functionality to enable map integration. Other more specialized tools, such as GDB and UDB exist, which allow a user to view and integrate different maps, but unfortunately these have not generally been integrated with the human genome sequence. Fortunately Entrez Map View at the NCBI, is one tool which straddles both the human genome and human genetic maps.

7.8.1 Entrez Map View (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map_search)

In Chapter 5 we reviewed Map View alongside Ensembl and the UCSC human genome browser as a tool for annotation of the human genome sequence. Map View would probably appear lower in most researchers' preferences for this purpose, although it does provide some unique gene annotation information. However, as the name suggests, Map View truly excels in its integration of a wide range of cytogenetic, genetic and physical maps with the NCBI draft and finished sequence contigs. Although this tool is sometimes a little difficult to navigate, once these idiosyncrasies are overcome, Map View becomes a complex and powerful tool.

Map View is an integrated component of the NCBI Entrez system, in the Entrez Genomes division. This division presents a unified graphical view of genetic and physical maps (including sequence maps) for over four vertebrates, including human and mouse. The tools present different genomes at four levels of detail:

- Organism home page — summarizing the resources available for that organism
- Genome View — graphical display of chromosome ideograms and search page
- Map View — presents one or more maps aligned against a master map
- Sequence View — graphically annotates the biological features in a region

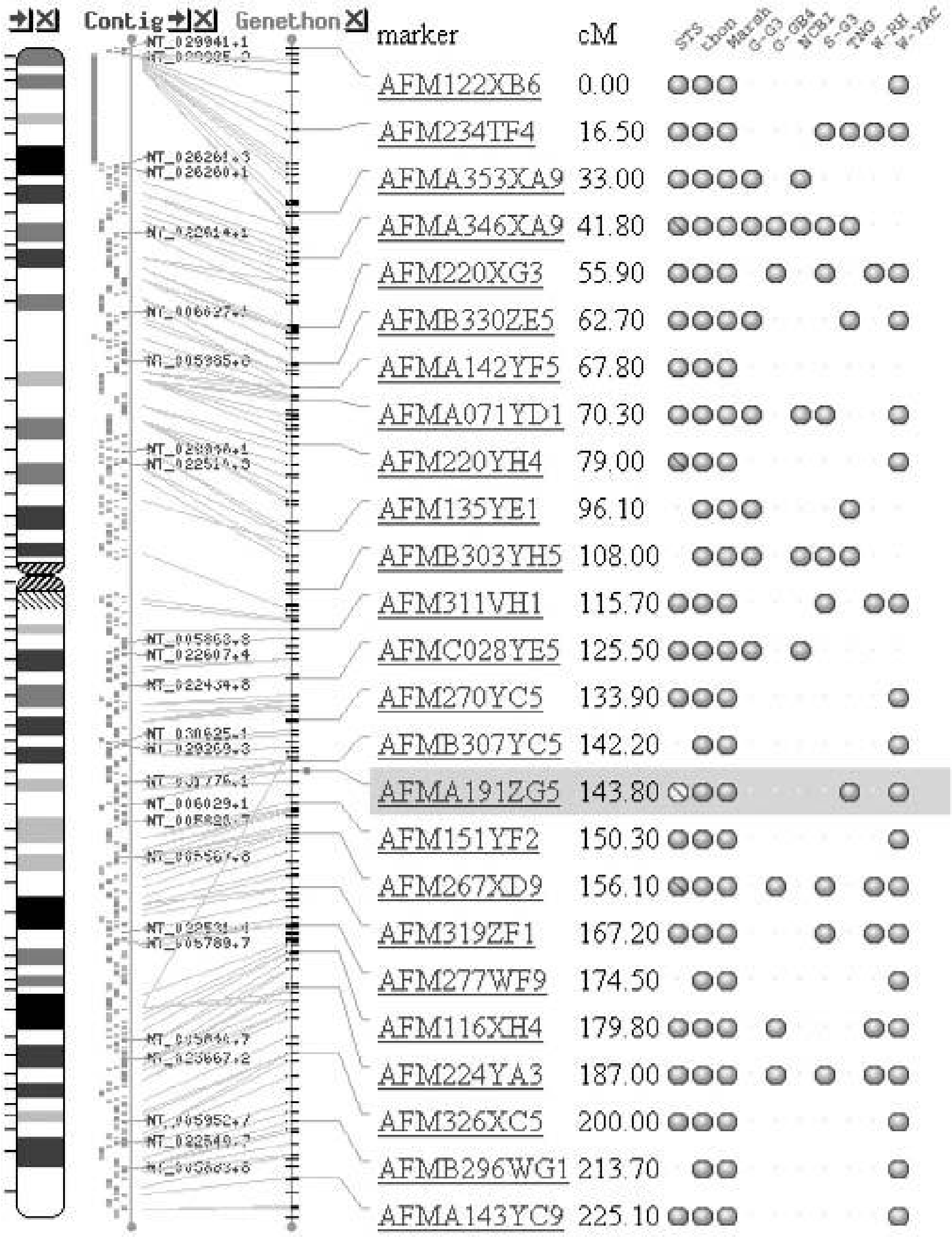
7.8.1.1 Searching and Browsing Map View

Map View can be searched with almost any marker, SNP, gene or genomic element either targeted at a chromosome or genome level. Searches at the genome level return a graphic view of the location of the hit with red marks on the chromosome ideograms, this will quickly identify if a query hits multiple regions or chromosomes. A summary of the maps in which the query exists is returned in tabular format at the bottom of the page. This is the essence of the Map View tool — selection of a map from the tabular summary links to a detailed Map View of the corresponding genomic region, with the selected map as the 'master' map. The master map is presented in detail with supporting information, such as LOD scores, cM locations or gene information. To view and integrate the master map with other maps, select the 'maps & options' link at the top of the page. This will summon a pop-up window for Map View configuration. It is possible to select up to eight maps to view alongside the master map, each is presented in a compact view alongside the master map. The alignment between maps is based on common or corresponding objects. Markers or objects shared between maps are indicated by lines connecting the maps. Map View allows the user to zoom and pan into progressively more detailed views.

It is also possible to search and browse Map View by map position or cytoband. This can be achieved from the Map View of a chromosome, by entering a range of interest in the boxes in the side window. A range can be specified in base pairs, cytogenetic bands or between two gene symbols. General chromosomal browsing is possible by clicking on the region of interest in the chromosome thumbnail graphic in the sidebar, or by clicking on a region of interest on the ideograms in the genome view.

Map View is very effective for integration of genetic and physical maps on an *ad hoc* basis. Figure 7.2 shows an integrated view of the Genethon genetic map and the human genome contig for chromosome 3. In this map, the Genethon markers are mapped to sequence and a line is drawn between the marker positions on the two maps. This clearly illustrates some key map integration issues. Firstly several markers in the genetic map are seen to conflict with the order of markers on the sequence (or physical) map. This may be due to an error in either map, so further maps need to be compared to support either order and the LOD scores on the genetic map need to be examined. Figure 7.3 shows such a comparison. The red line traces the Genethon marker, AFMA121WD5, through the Marshfield, GB4, G3 and TNG maps through to the genomic contig level. In this case the marker order is confirmed by each map. Sometimes it may not be possible to conclusively determine which map is 'right', instead further laboratory work may be necessary to resolve marker order. Figure 7.2 also clearly shows the variable relationship between genetic and physical distance. In particular it highlights some of the physical properties of chromosomes, for example the genetic physical distance ratio at the telomere of the P arm of chromosome 3 is very low; the marker AFM234TF4, for example, has a genetic location of 22 cM and physical location of 8 Mb. This illustrates the higher rates of recombination that are often observed in telomere regions (Riethman, 1997). Both figures indicate the presence or absence of each marker in available maps by an array of symbolic green circles at the far right of each marker. This helps to indicate non-specific markers. For example, some markers map to multiple locations in the same chromosome, these are indicated by green circles with a strike through. Other markers map to more than one chromosome, these are indicated by yellow circles and finally some markers, map to multiple chromosomes and multiple locations, indicated by a yellow struck through circles (e.g. AFMA191ZG5 in Figure 7.2).

There are a number of somewhat idiosyncratic features in Map View which might confuse the user. Firstly if a map is viewed in low resolution, it seems to display a somewhat arbitrary selection of markers, the full marker set only becomes visible when the user zooms in. Secondly, if the locus is too large to view in one window it is broken up into pages indicated at the top of the window. This pagination feature can make it slow and difficult to assess a whole locus, but this can be overridden by altering the page size in the configuration window. Setting a page size of 100–200 will allow a very large map to be viewed in a single window, this may take some time to load but it is worth it in the end.



marker	cM	STS	Gbon	Marsh	G-G3	G-GB4	NCBI	S-G3	TNG	W-RH	W-YAC
AFM122XB6	0.00	●	●	●							●
AFM234TF4	16.50	●	●	●				●	●	●	●
AFMA353XA9	33.00	●	●	●	●		●				
AFMA346XA9	41.80	⊘	●	●	●	●	●	●	●		
AFM220XG3	55.90	●	●	●		●		●		●	●
AFMB330ZE5	62.70	●	●	●	●				●		●
AFMA142YF5	67.80	●	●	●							
AFMA071YD1	70.30	●	●	●	●		●	●			●
AFM220YH4	79.00	⊘	●	●							●
AFM135YE1	96.10		●	●	●				●		
AFMB303YH5	108.00		●	●	●		●	●	●		
AFM311VH1	115.70	●	●	●				●		●	●
AFMC028YE5	125.50	●	●	●	●		●				
AFM270YC5	133.90	●	●	●							●
AFMB307YC5	142.20		●	●							●
AFMA191ZG5	143.80	⊘	●	●					●		●
AFM151YF2	150.30	●	●	●							●
AFM267XD9	156.10	⊘	●	●		●		●		●	●
AFM319ZF1	167.20	●	●	●				●		●	●
AFM277WF9	174.50		●	●							●
AFM116XH4	179.80	●	●	●		●				●	●
AFM224YA3	187.00	●	●	●		●		●		●	●
AFM326XC5	200.00	●	●	●							●
AFMB296WG1	213.70		●	●							●
AFMA143YC9	225.10	●	●	●							●

Figure 7.2 Integration of genetic maps and genome contigs. This figure shows an integrated view of the Genethon genetic map and the human genome contig for chromosome 3 generated by the NCBI Map View tool. The Genethon markers are mapped to sequence with a line drawn between the marker positions on the two maps. Lines which cross over show markers which conflict in order between the genetic map and the physical sequence map.

7.8.2 The Genome Database (GDB) (www.gdb.org)

The Genome Database (GDB) was the first web-based graphical interface to the human genome, as such it was a pioneering bioinformatic tool. Now Ensembl, UCSC and Map View present effortless graphical genome views and the GDB graphical interface is starting

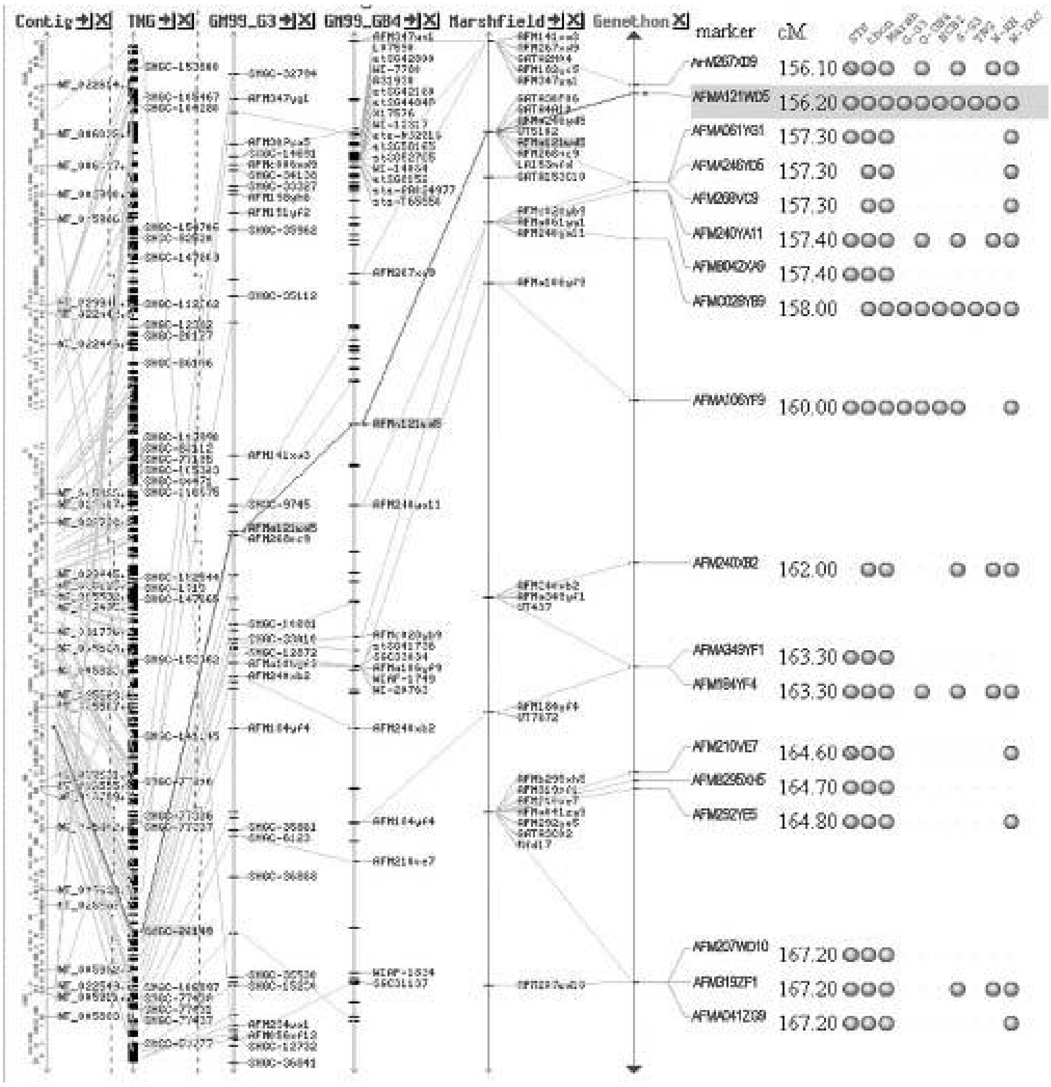


Figure 7.3 Integration of genetic and physical maps with the human genome contig on chromosome 3 using the NCBI Map View tool. The grey line traces the Genethon marker, AFMA121WD5, through though the Marshfield, GB4, G3 and TNG maps through to the genomic contig level. In this case the marker order of this marker is confirmed by each map.

to look a little tired and most of the graphical functionality is covered by Map View. But GDB does have a productive text/table-based search interface which is an improvement on Map View's limited text-based capability. GDB is also a comprehensive source for some forms of genetic data, particularly tandem repeat polymorphisms (it contains over 18,000), and an eclectic range of information on fragile sites, deletions, disease genes and mutations, collected by a mixture of curation and direct submission. This makes GDB a valuable tool for text-based data mining to assist in the construction of marker lists and the identification of marker variables, such as primer and marker sequences.

The text-based search interface is accessible on the front page of the GDB database by following the 'advanced search' link. This interface allows complex queries, for example, it is possible to retrieve all known polymorphic or non-polymorphic markers between two markers or genes. Results are retrieved and ordered based on the genetic distances of the markers, along with a very roughly estimated Mb location (unfortunately actual integration with the human genome draft is currently lacking). As the markers are ordered by genetic distance, the distances are very approximate, with no fine measure of distance or order. It may be necessary to clarify the order with another tool such as Map View.

7.8.3 The Unified Database for Human Genome Mapping (UDB) (http://bioinformatics.weizmann.ac.il/udb/)

The Unified Database for Human Genome Mapping (UDB) is maintained by the Weizmann Institute of Science, Israel. UDB has attempted to create an integrated map based on a diverse range of human genome mapping data retrieved from a number of public databases. The map consists of an integrated hierarchy of genetic, RH, cDNA and YAC maps down to a kilobase resolution, on a scale converted from centiRays (cR) to megabases (Mb). UDB generates its maps using data from the Whitehead/MIT STS map, GeneMap'98, the Stanford TNG map and Genethon maps. The database can be searched in several different ways. An initial search by chromosome number can be narrowed by specification of cytogenetic band, position (in Mb) or marker interval. It is also possible to search by gene or marker name. This gives the estimated location of the gene as well as links to GeneCards and the Genome Database (GDB). The database also displays the estimated boundaries (in Mb) of the cytogenetic bands of any chromosome.

The UDB database is a good starting point for constructing physical or transcript maps across a genomic region. The main benefit of the database is that it eliminates the need to look at a number of different websites and integrates markers from several different maps with genomic contigs from NCBI. Unfortunately UDB is somewhat over zealous in its map integration, sometimes this might cause problems. It assumes an hierarchical value of RH maps over genetic maps and genetic maps over YAC maps which is not necessarily the best order for integration, it may have been better to flag conflicting marker orders for laboratory-based resolution. However as the human genome map solidifies around finished sequence this approach will begin to represent the simplest and most effective use of time and resources.

7.9 CONCLUSIONS

As this chapter has described, there are many tools available to give an integrated view of genetic and physical maps across a defined chromosomal locus. Comparison of the physical and genetic distances between markers can provide a great deal of information about the underlying nature of a locus. Yu *et al.* (2001) compared the genetic and physical distances across the whole genome and found that the genetic/physical distance ratio ranged widely between 0 and 9 cM per Mb. They used this ratio to infer recombination rates and identified several chromosomal regions up to 6 Mb in length with very low or high recombination rates, which they termed recombination ‘deserts’ and ‘jungles’, respectively. Linkage disequilibrium (LD) was much more extended in the deserts than in the jungles as higher rates of recombination are likely to reduce the extent of LD.

When sequencing of the human genome is truly complete genetics will become technically much easier. Human map QC may become a distant memory, but presently we are still struggling to study complex phenotypes with draft contigs and incomplete datasets. Every piece of data and data curation may count in this struggle — in Section III of this book we review how physical and genetic map data can come together with literature data, marker data, gene data and comparative organism data to assist genetic studies in the laboratory.

REFERENCES

Agarwala R, Applegate DL, Maglott D, Schuler GD, Schaffer AA. (2000). A fast and scalable radiation hybrid map construction and integration strategy. *Genome Res* **10**: 350–364.

Boehnke M, Lange K, Cox DR. (1991). Statistical methods for multipoint radiation hybrid mapping. *Am J Hum Genet* **49**: 1174–1188.

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet* **63**: 861–869.

Chumakov IM, Rigault P, Le Gall I, Bellanne-Chantelot C, Billault A, Guillou S, *et al.* (1995). A YAC contig map of the human genome. *Nature* **377**(Suppl.): 175–297.

Cox DR, Burmeister M, Price ER, Kim S, Myers RM. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* **250**: 245–250.

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, *et al.* (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature* (in press).

Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, *et al.* (1998). A physical map of 30,000 human genes. *Science* **282**: 744–746.

DeWan AT, Parrado AR, Matise TC, Leal SM. (2002). The map problem: a comparison of genetic and sequence-based physical maps. *Am J Hum Genet* **70**: 101–107.

Dib C, Faure S, Fizames C, Sampson D, Drouot N, Vignal A, *et al.* (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**: 152–154.

Eisenbarth I, Vogel G, Krone W, Vogel W, Assum G. (2000). An isochore transition in the NF1 gene region coincides with a switch in the extent of linkage disequilibrium. *Am J Hum Genet* **67**: 873–880.

Goddard KA, Wijsman EM. (2002). Characteristics of genetic markers and maps for cost-effective genome screens using diallelic markers. *Genet Epidemiol* **22**: 205–220.

Goss SJ, Harris H. (1975). New method for mapping genes in human chromosomes. *Nature* **255**: 680–684.

Gyapay G, Schmitt K, Fizames C, Jones H, Vega-Czarny N, Spillett D, *et al.* (1996). A radiation hybrid map of the human genome. *Hum Mol Genet* **5**: 339–346.

Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, *et al.* (2000). The DNA sequence of human chromosome 21. *Nature* **405**, 311–319.

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, *et al.* (2001). Haplotype tagging for the identification of common disease genes. *Nature Genet* **29**: 233–237.

Korenberg JR, Chen XN, Sun Z, Shi ZY, Ma S, Vataru E, *et al.* (1999). Human genome anatomy: BACs integrating the genetic and cytogenetic maps for bridging genome and biomedicine. *Genome Res* **9**: 994–1001.

Kruglyak L. (1997). The use of a genetic map of biallelic markers in linkage studies. *Nature Genet* **17**: 21–24.

Kwitek AE, Tonellato PJ, Chen D, Gullings-Handley J, Cheng YS, Twigger S, *et al.* (2001). Automated construction of high-density comparative maps between rat, human, and mouse. *Genome Res* **11**: 1935–1943.

Lander ES, Green P, Abrahamson P, Barlow A, Daly MJ, Lincoln SE, *et al.* (1987). MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1**: 174–181.

Lunetta KL, Boehnke M, Lange K, Cox DR. (1996). Selected locus and multiple panel models for radiation hybrid mapping. *Am J Hum Genet* **59**: 717–725.

Lynn A, Kashuk C, Petersen MB, Bailey JA, Cox DR, Antonarakis SE, *et al.* (2000). Patterns of meiotic recombination on the long arm of human chromosome 21. *Genome Res* **10**: 1319–1332.

Morley M, Arcaro M, Burdick J, Yonescu R, Reid T, Kirsch I, *et al.* (2001). GenMapDB: a database of mapped human BAC clones. *Nucleic Acids Res* **29**: 144–147.

Morton NE. (1991). Parameters of the human genome. *Proc Natl Acad Sci USA* **88**: 7474–7476.

Murray JC, Buetow KH, Weber JL, Ludwigsen S, Scherpbier-Heddema T, Manion F, *et al.* (1994). A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC). *Science* **265**: 2049–2054.

Niimura Y, Gojobori T. (2002). *In silico* chromosome staining: reconstruction of Giemsa bands from the whole human genome sequence. *Proc Natl Acad Sci USA* **99**: 797–802.

Olivier M, Aggarwal A, Allen J, Almendras AA, Bajorek ES, Beasley EM, *et al.* (2001). A high-resolution radiation hybrid map of the human genome draft sequence, *Science* **291**: 1298–1302.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, *et al.* (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719–1723.

Riethman H. (1997). Closing in on Telomeric Closure. *Genome Res* **7**: 853–855.

Schuler GD. (1997). Sequence mapping by electronic PCR. *Genome Res* **7**: 541–550.

Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, *et al.* (1996). A gene map of the human genome. *Science* **274**: 540–546.

Stewart EA, McKusick KB, Aggarwal A, Bajorek E, Brady S, Chu A, *et al.* (1997). An STS-based radiation hybrid map of the human genome. *Genome Res* **7**: 422–433.

Waterston RH, Lander ES, Sulston JE. (2002). On the sequencing of the human genome. *Proc Natl Acad Sci USA* **99**: 3712–3716.

Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, *et al.* (2001). Comparison of human genetic and sequence-based physical maps. *Nature* **409**: 951–953.