

SECTION 2

**THE IMPACT OF COMPLETE GENOME
SEQUENCES ON GENETICS**

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
 - 5.2 Genomic sequence assembly
 - 5.3 Annotation from a distance: the generalities
 - 5.3.1 Nucleotide level
 - 5.3.2 Protein level
 - 5.3.3 Process level
 - 5.4 Annotation up close and personal: the specifics
 - 5.4.1 Ensembl
 - 5.4.2 UCSC Human Genome Browser (HGB)
 - 5.4.3 NCBI Map Viewer (NMV)
 - 5.4.4 ORNL Genome Channel (GC)
 - 5.5 Annotation: the next generation
 - Acknowledgements
 - References
-

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.nlm.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.embn.net.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://hmmerr.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: 150–800bp 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 $\sim 10\%$ of assembled fragments were assigned the wrong orientation and $\sim 15\%$ 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.emblnet.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 *et 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

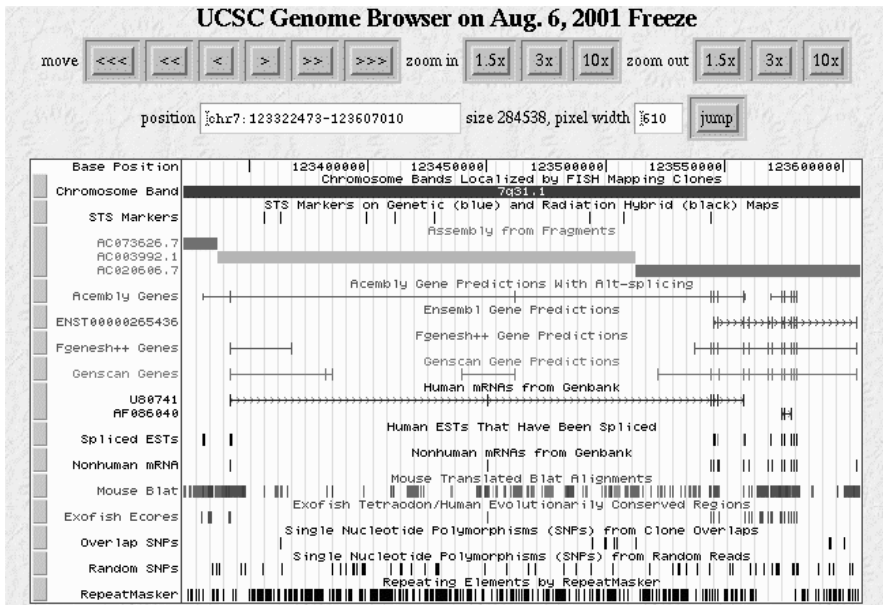


Figure 5.2 The genomic region around the FOXP2 gene according to the UCSC Human Genome Browser.

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

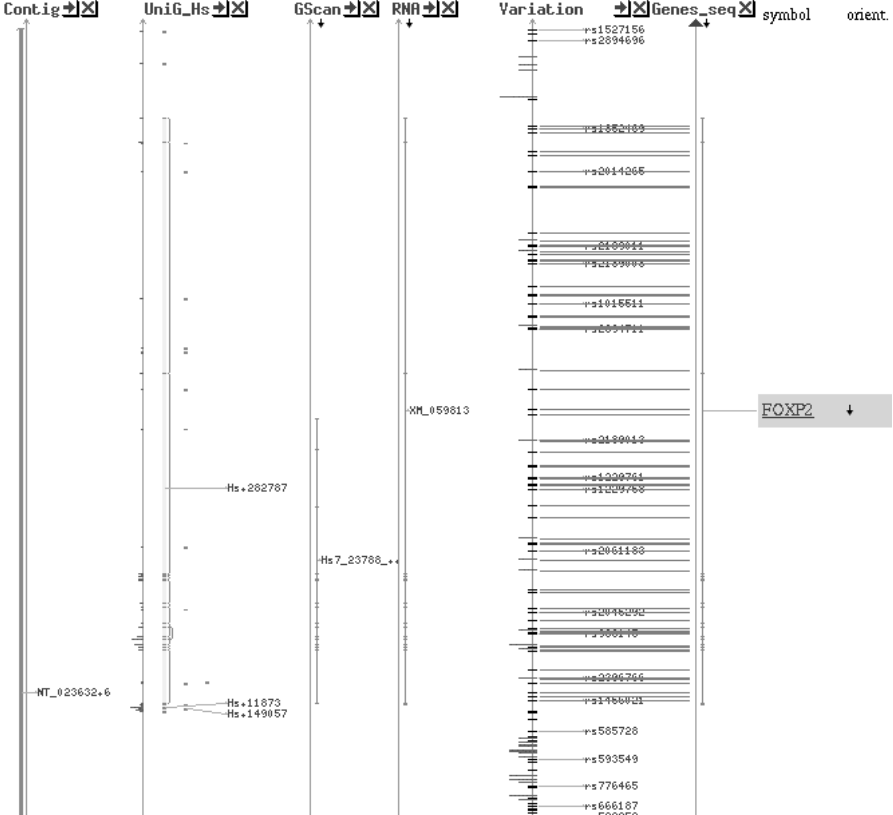


Figure 5.3 The genomic region around the FOXP2 gene according to the NCBI Map Viewer (See Colour Plates).

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

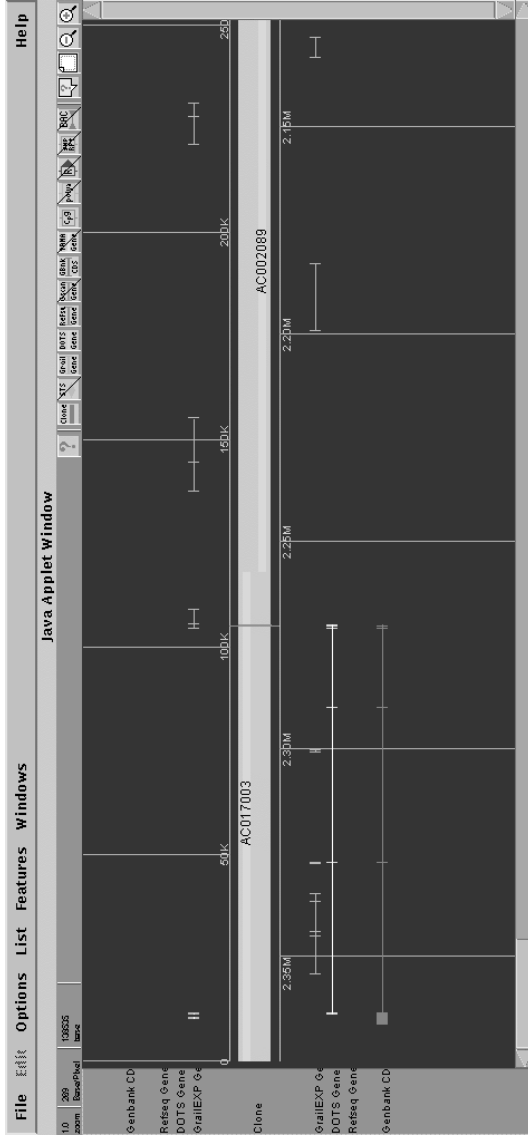


Figure 5.4 The genomic region around the FOXP2 gene according to ORNL Genome Channel (See Colour Plates).

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, Fujiiyama 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.