Chapter III Transcript map of human chromosome 22q13.31

3.1 Introduction

3.1.1 Gene identification

Genes represent the major biological function of the genome and are therefore a major focus of research interest. Traditionally, experimental approaches such as cDNA selection and exon trapping (see chapter I) have been utilised in positional cloning strategies to produce transcript maps of regions associated with disease. In positional cloning, researchers first map the disease as closely as possible in affected families, then identify genes in the region, before honing in on a candidate gene and showing that patients have mutations in that gene. Genes for important monogenic disorders such as Duchenne's muscular dystrophy (Monaco *et al.*, 1986) and cystic fibrosis (Rommens *et al.*, 1989) have been identified in this way.

However, this kind of approach has several limitations. The experimental strategy is both time-consuming and expensive and does not provide information of the surrounding genomic environment, including other genes, which may influence function. The example of the familial Mediterranean fever locus (FMF) shows that even multiple gene identification methods do not necessarily yield all genes in a specific region. Transcript maps for this region were constructed independently by both Centola *et al.* (1998) and Bernot *et al.* (1998). The maps overlapped by 225 kb and both groups identified genes specific to their approaches (exon trapping, cDNA selection, EST mapping, limited sequencing and computational gene prediction). Each group identified additional genes not annotated by the other, which shows that even a combination of such approaches may not find all the genes.

The availability of genomic sequence for a region of interest significantly alters the gene identification strategy to one of sequence-based analysis. The genome sequence provides the foundation for a systematic approach to gene annotation. The general progress in the human

genome project has had an enormous impact on the smaller scale positional cloning projects, as preliminary transcript maps are now available covering much of the genomic sequence.

Analysis of the genomic sequence may eventually provide a more complete picture of the human transcriptome (the set of expressed genes). However, coding sequences occupy just a small fraction, approximately 3%, of the human genome (Dunham *et al.*, 1999; Duret *et al.*, 1995) and accurate determination of gene structures within the genomic sequence is difficult (see chapter I). Currently, a combination of *ab initio* prediction and similarity searches are utilised to annotate coding sequences.

3.1.2 *Ab initio* prediction packages

Several sophisticated software algorithms have been devised to handle gene prediction in eukaryotic genomes. These algorithms typically consist of one or more 'sensors': a specialised algorithm that tries to detect the presence of a gene feature from motifs or statistical properties of the DNA. Some gene predictors stop with the prediction of a single feature, such as the exon predictor HEXON (Solovyev *et al.*, 1994). Most, however, attempt to use the output of several sensors to generate a whole gene model, in which a gene is defined as a series of exons that are co-ordinately transcribed. Several approaches are typically used (reviewed by Stein, 2001):

a. Neural networks, e.g. Grail (Uberbacher & Mural, 1991), are analytical techniques modelled after the (proposed) processes of learning in cognitive systems and the neurological functions of the brain. Neural networks use a data 'training set' to build rules that can make predictions or classifications on data sets.
b. Rule-based systems, e.g. GeneFinder (Favello *et al.*, 1995) are a type of computer algorithm that uses an explicit set of rules to make decisions.
c. Hidden Markov Models (HMM) represent a system as a set of discrete states and transitions between those states. Each transition has an associated probability. Markov models are

'hidden' when one or more of the states cannot be directly observed. The HMM approach has the advantage of explicitly modelling how the individual probabilities of a sequence of features are combined into a probability estimate for the whole gene. Examples include Genscan (Burge & Karlin, 1997) and Fgenesh (Solovyev *et al.*, 1994).

However, *ab initio* prediction is far from perfect. The performance of the gene prediction programs has been discussed by a number of authors (Burset & Guigo, 1996; Claverie, 1997). An assessment of genome annotation in *Drosophila melanogaster* (Reese *et al.*, 2000) showed that the best algorithms could achieve sensitivities (a measure of the ability to detect true positives) and specificities (a measure of the ability to discriminate against false positives) of ~95% and ~90% respectively when testing if a particular nucleotide is contained within an exon. Accuracy decreased if the criterion was changed to calling the boundaries of an exon correctly and still further if the algorithm was required to predict the entire gene structure correctly. In this case, the best predictor achieved a sensitivity of 40% and a specificity of 30%. To improve the predictions, the use of multiple programs is advocated (Burset & Guigo, 1996; Claverie, 1997; Reese *et al.*, 2000).

Another method to improve the performance of prediction programs is to include similarity searches of the protein and/or EST databases with the gene prediction packages (section 3.1.4).

3.1.3 Sequence similarity

The similarity of a region of the genome to a sequence that is already known to be transcribed provides a powerful prediction of whether or not a sequence is part of a gene. A comparison of a genome sequence with databases of ESTs, cDNAs and proteins (see appendix 2) using programs such as BLAST can identify regions of a contig that correspond to processed mRNA.

However, there are drawbacks to gene finding based purely on similarity searches of expressed sequence databases. Pseudogenes are a common feature of eukaryotic genomes. Many similarity-based gene prediction programs require evidence that the gene is spliced and that the splices maintain an in-phase ORF. However, this criterion biases gene prediction against single exon genes. In addition, ESTs are fragmentary and may suffer from artefacts, including contamination with genomic DNA, chimaerism and lane tracking errors during automated sequencing. cDNA sequences might contain repetitive elements that will cause spurious genomic matches and the method used in generation of EST and cDNA resources (often reverse transcription primed from the poly(A) 3' sequence) can result in 5' incomplete cDNA, as the reverse transcriptase may dissociate at any point from the template. Additionally, similarities to proteins in other species might be altered by evolutionary divergence and the presence of alternative splicing complicates the interpretation of alignments between genomic DNA, cDNAs and ESTs. Finally, even the most comprehensive EST projects will miss low copy number transcripts and those transcripts that are expressed only transiently, or under unusual circumstances.

3.1.4 Combination

The current trend in gene prediction is to combine *ab initio* gene predictions with similarity data into a single model, such as Grail/Exp (Xu *et al.*, 1995), GenieEST (Reese, unpublished) and GenomeScan (Yeh *et al.*, 2001). Reese *et al*., (2000) showed that the algorithms that took similarity data into account generally outdid those that did not. So far, however, most genome-wide annotation systems have run sequence-similarity searches and *ab initio* gene predictors separately, then combined and reconciled the predictions later.

Lander *et al*.,(2001), used a gene identification approach based on the Ensembl gene annotation system (Hubbard & Birney, 2000), which began with *ab initio* Genscan predictions and then

strengthened them with nucleotide and protein similarities. The predicted genes were then merged with Genie (Kulp *et al.*, 1996) output and finally merged with the RefSeq library of well-characterised genes (Maglott *et al.*, 2000). The Celera system took the reverse approach, using firstly sequence similarities found in the RefSeq library, Unigene, of human ESTs (Boguski & Schuler, 1995) and from SwissProt (Bairoch & Apweiler, 2000) before using Genscan to find and refine the splicing pattern of the predicted genes. Both groups gave greater weight to cDNA and EST alignments than to *ab initio* gene predictions. Estimates for the number of genes from both groups were very close: both groups predicted the existence of approximately 30,000 human genes.

However, a comparison of the Celera and Ensembl predicted gene sets (Hogenesch *et al.*, 2001) found little agreement between the two predicted transcriptomes. Collectively, nearly 80% of the 31,098 novel transcripts were predicted by only one of the groups. Using high density oligonucleotide arrays (see chapter I), Hogenesch *et al.* demonstrated that more than 80% of the novel predicted transcripts were detected as expressed in at least one of thirteen human tissues, concluding that the respective transcriptomes are individually incomplete and casting doubt on these estimates of gene numbers. Hogenesch suggests that an integrated approach, combining computational predictions, human curation and experimental validation will be required to complete a finished picture of the human transcriptome.

Another tool for gene identification is becoming more readily available with the completion of the genome projects of several model organisms. In particular, the increasing availability of mouse genomic sequence is expected to have a large impact on annotation of the human genome, through the identification of conserved functional regions (Lander *et al.*, 2001). This aspect of transcript mapping is discussed in more detail in the next chapter.

The availability of intron sequences and surrounding intergenic sequence, allow investigation into several sequence features that are associated with genes. These include analysis of sequence contexts surrounding translation initiation sites (described by Kozak, 1987) and polyadenylation signals (Beaudoing *et al.*, 2000). There is also considerable interest in the prediction of promoter sequences and several programs have been developed which attempt to elucidate the 5' regulatory gene structure (for example Scherf *et al.*, 2000). Investigation of surrounding repetitive sequences and GC content can also be undertaken to give a clearer picture of the genomic environment. Such analysis was most notably carried out on the draft human genome sequence (Lander *et al.*, 2001). This work allows regional comparisons to be made against a broad genomic landscape. Genes in a region of interest can also be compared against the available genomic sequence, to identify paralogous genes and possibly to give an idea of the evolutionary history of the genomic region.

The reference generated by annotation of the human genome sequence will underpin nearly all future genetic research. For this reason it is essential that annotation of genes is as accurate as possible. For example, functional studies using *in silico* analysis programs are heavily dependent on patterns within translated DNA sequences. Errors leading to alteration of the reading frame, or the omission or inclusion of sequences, can have a large affect on experimental outcome. In addition, a huge range of wet-laboratory techniques requires accurate coding sequence information. These include any experiment to express and study the function of proteins encoded within the sequence, as well as investigations of mRNA expression patterns and analysis of potential regulatory sequences (chapter I).

3.1.5 Summary

This chapter discusses the analysis of a 3.4 Mb section of the genomic sequence of chromosome 22 (22q13.31). Availability of 3.2 Mb of genomic sequence from this region

(Dunham *et al*., 1999) enabled study of the genomic environment of genes in the region, through analysis of GC content and density and coverage of repeats.

Computational and experimental data were integrated to aid the assembly of a transcript map of the region. EST, cDNA and protein homologies, as well as Genscan predictions (Burge & Karlin, 1997), were used as a starting point for further experimental investigation to extend and confirm putative gene structures. The specificity and sensitivity of each type of evidence used to identify and annotate genes was calculated by comparison to the final gene annotation.

Northern blot experiments enabled analysis of transcript size and expression pattern of the annotated genes. Additional evidence of expression was provided by the construction and screening of an expression panel representing 32 human tissues from a range of individuals. The availability of the genomic sequence allowed analysis of the intron/exon structure and splice site consensus sequences of all the annotated gene features.

The sequences of fully annotated gene structures were inspected in their genomic context for the presence of poly(A) sites, translation start sites, predicted CpG islands and promoter regions. Availability of the draft genome sequence also allowed a preliminary investigation of gene paralogy and the identification of groups of potentially related genes.

3.2 Gene identification on 22q13.31

Initial analysis was performed on each sequence clone with a standard automated process used by the Sanger Institute annotation group. Figure 3.1 illustrates this analysis process.

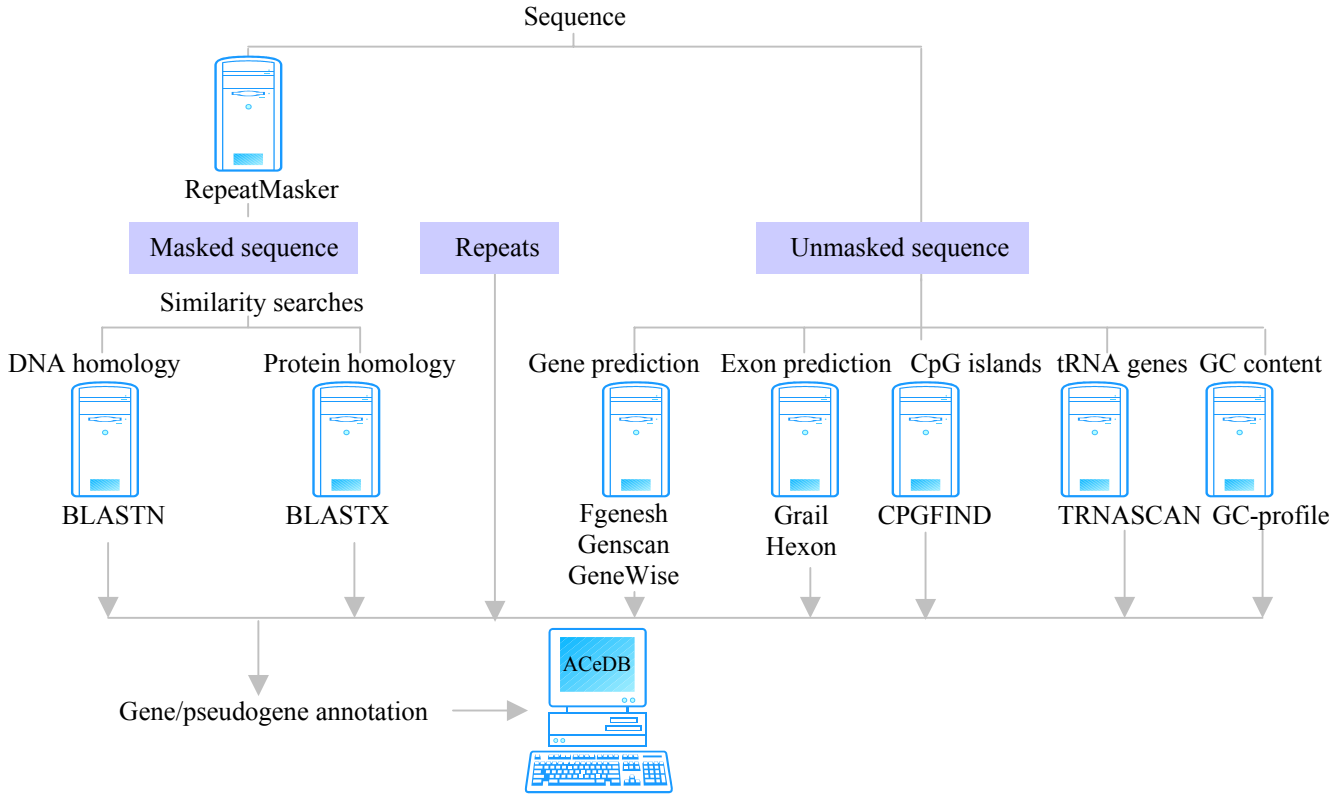


Figure 3.1: Automated analysis strategy. The masked sequence was used in homology searches. The unmasked sequence was used in a number of gene prediction packages and in the prediction of other features such as CpG islands and tRNAs. Both the homology data and the predicted data were integrated with repeat data and displayed by the human chromosome 22 implementation of ACeDB.

The resultant analysis files are read into the HSA22 application of ACeDB (22ace). This data was used for initial gene annotation by a team of annotators and formed the information initially available at the beginning of this project.

The DNA sequence of chromosome 22 is currently contained in 10 contigs. The separate clone sequences that make up these contigs have been linked together and have been reanalysed using the above methods. Additionally, output from relevant novel analysis programs and updated

sequence database searches have been incorporated into 22ace as they became available. All the analysis packages used are described in appendix 2a. Sequence databases, together with the latest version used/release date where applicable, are listed in appendix 2b. The current sequence analysis strategy for human chromosome 22 is illustrated in figure 3.2.

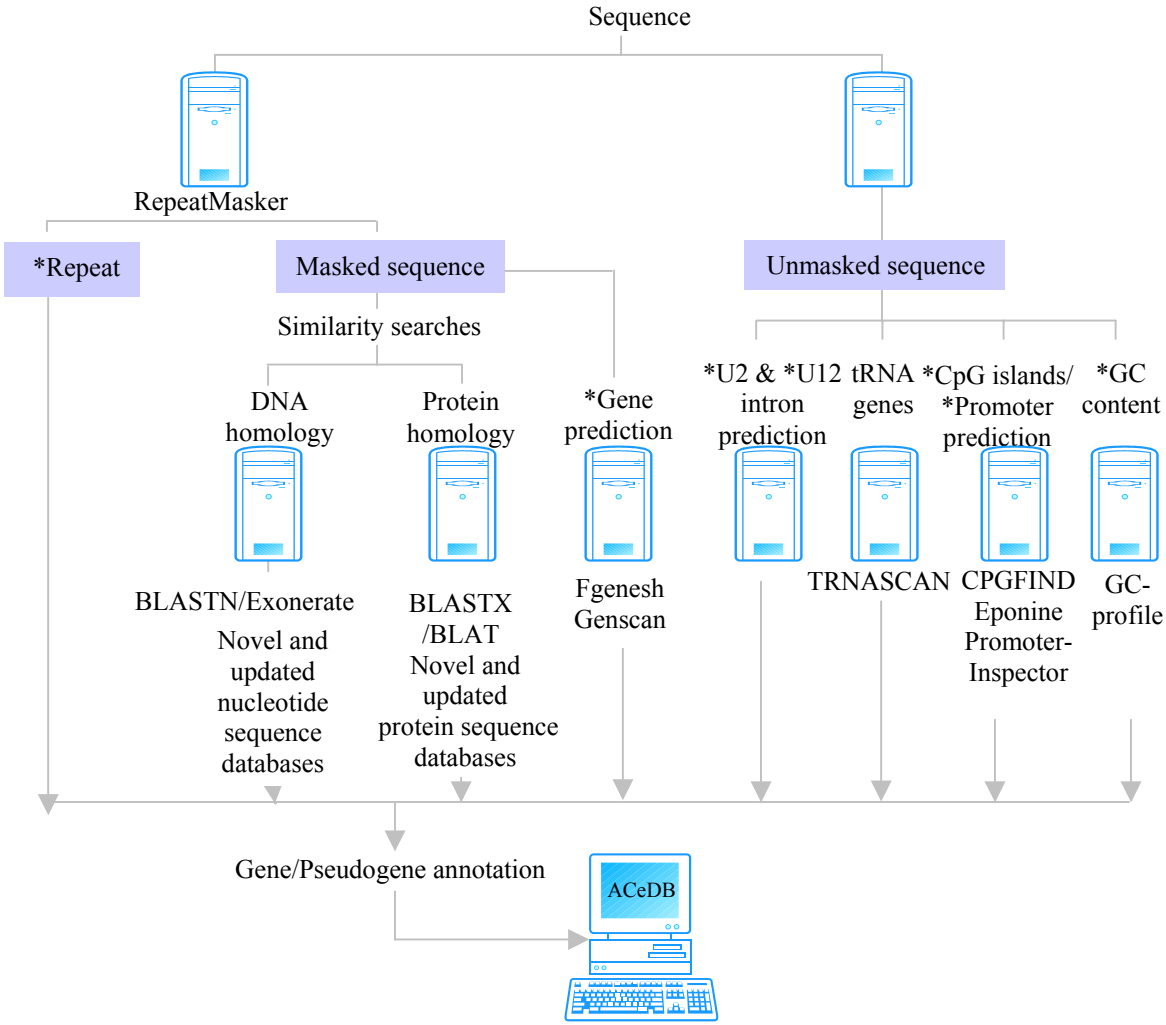


Figure 3.2: Chromosome 22 additional analysis strategy. * denotes analysis performed on linked clone sequences. Masked sequence was used in homology searches against novel and updated sequence databases (see appendix 2a) and in a number of gene prediction packages. Unmasked sequence was used in the prediction of additional features such as CPG islands, promoters, etc. (appendix 2b). Both the homology data and the predicted data were integrated with repeat data and displayed by the human chromosome 22 implementation of ACeDB. This updated information is used in the additional annotation of genes and pseudogenes (section 3.4)

The sequence display of 22ace allows visualisation of these results (figure 3.3). This data has been utilised during the course of the project for annotation of potential genes and regulatory regions (sections 3.4 and 3.8.5), investigation of instances of paralogy (section 3.8.7) as well as

investigation of human-murine sequence conservation (chapter IV) and protein analysis (chapter V).

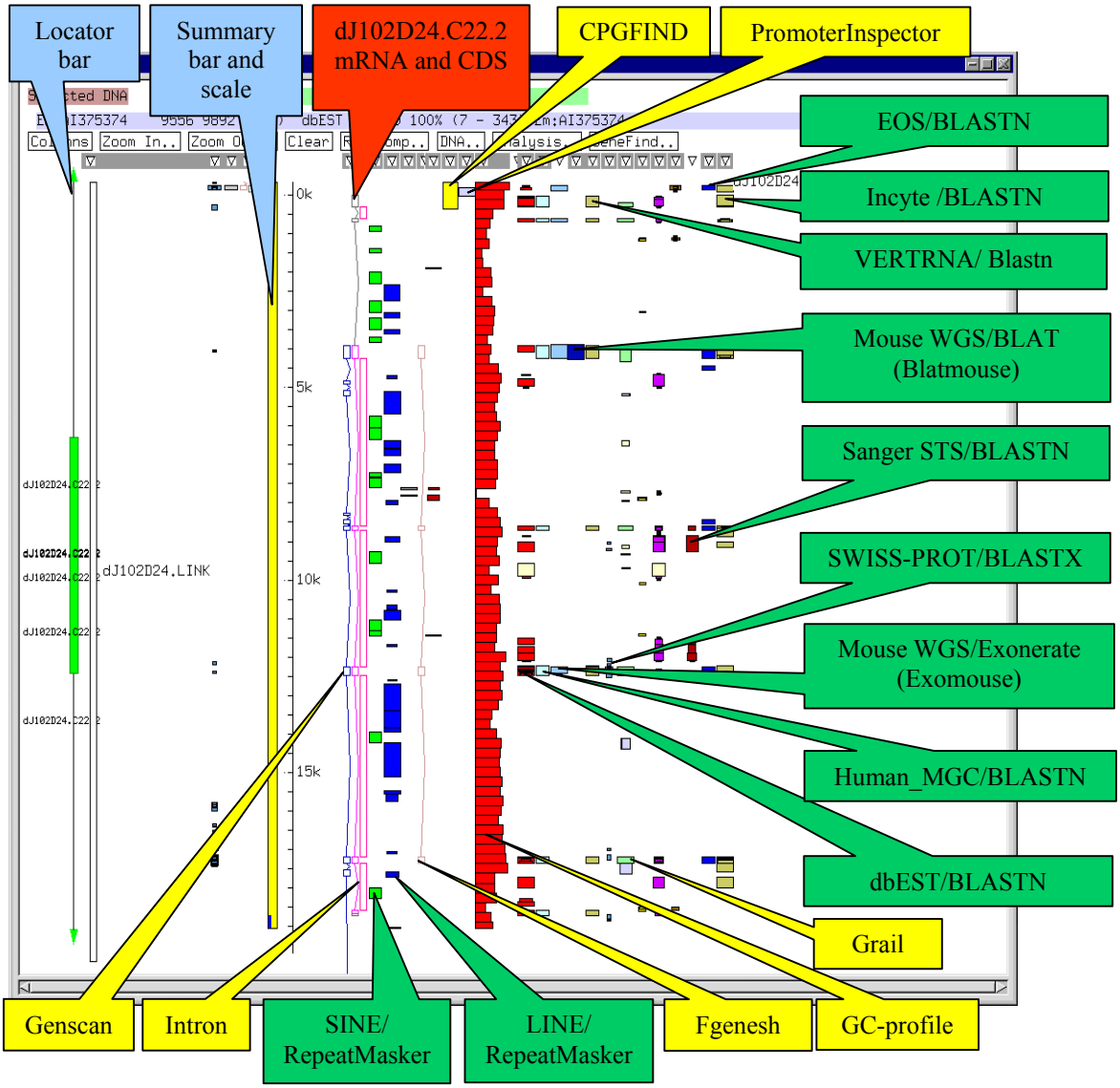


Figure 3.3: An example of the ACeDB display. The blue boxes show ACeDB general features. The green boxes indicate similarities to a variety of sequence databases, listed in appendix 2a. The yellow boxes show the output from a range of prediction programs listed in appendix 2b. Red boxes indicate annotated gene mRNAs and coding sequence (CDS), based on this evidence. The genomic region depicted here surrounds the locus dJ102D24.C22.2.

3.3 Genomic landscape of human chromosome 22q13.31

The region investigated during this project spans approximately 3.4 Mb of chromosome 22. Genomic sequence is available for 3.24 Mb of this region (Dunham *et al.*, 1999). There are two

gaps of approximately 50 kb and 75 kb respectively within this sequence. The region of interest lies within the light band 22q13.31 (Cheung *et al.*, 2001). Some of the sequence differences between chromosomal dark and light bands are noted in the table 1.1, chapter I. In particular, light bands have a high GC content and are expected to be LINE poor, but enriched in *Alu* repeats. The GC and repeat content of the region of interest were therefore investigated, in order to determine of these features agreed with those expected from a chromosomal light band.

3.3.1 Repeat content

The repeat content of the available sequence from the region has been analysed using RepeatMasker (Smit and Green, unpublished). Figure 3.4 shows that approximately 43.1% of all DNA in the region is repetitive. The SINE repeats have the largest coverage at 21.3% followed by the LINE repeat families at 11.53%. The coverage of *Alu* repeats in the region (18.68%) is substantially higher than the equivalent figure generated from the draft genome sequence (13.14%) (Lander *et al.*, 2001). Similarly, LINE coverage in the region is lower than the mean figure from the rest of the available human genomic sequence (20.42%)(Lander *et al.*, 2001). These results are therefore consistent with the characteristics of a light band region.

Table 3.1: The % repeat coverage and density of a 3.4Mb region of chromosome 22q13 and of the draft genome sequence

Repeat	Coverage (%)	Density (repeat/kb)	Coverage (%) Draft genome sequence	Density (repeat/kb) Draft genome sequence
SINE/ALU	18.68	3.69	10.60	3.76
SINE/MIR	2.66	7.67	2.54	6.74
SINE/other	0	0	0	0
LINE/LINE1	7.97	1.55	16.89	1.12
LINE/LINE2	3.34	4.14	3.22	3.57
LINE/other	0.22	3.20	0.31	4.40
LTR/MaLR	2.82	2.46	3.65	2.40
LTR/Other	2.76	2.06	4.64	1.38
DNA/MER1_type	0.80	4.62	1.39	4.78
DNA/MER2_type	0.49	2.69	1.02	2.04
DNA/other	0.10	6.55	0.43	4.78
Other	0.42	2.85	0.14	0.79

The coverage and density of the draft genome sequence (Lander *et al.*, 2001) are included for comparison.

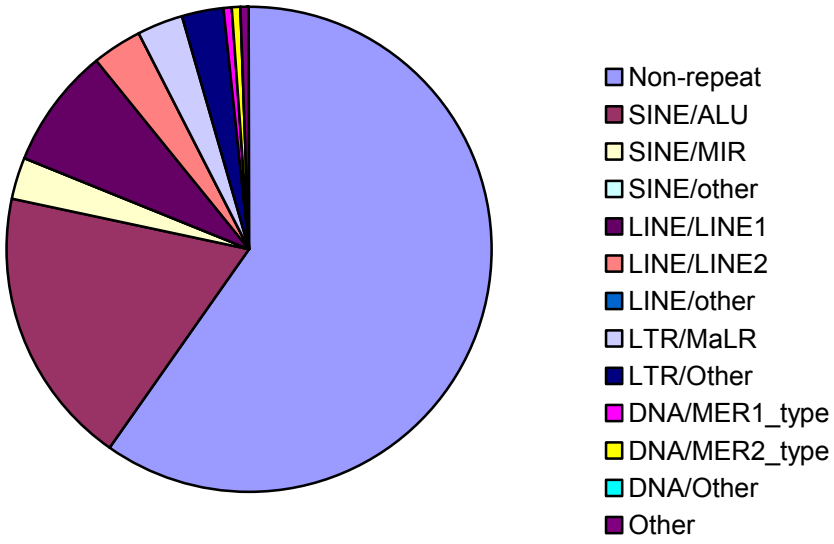


Figure 3.4: Repetitive and non-repetitive DNA coverage (%) for region of interest

3.3.2 GC content

The GC content of the region was calculated using gc-profile, using a window size of 250 bp (Gillian Durham, unpublished). A plot of the GC content over the length of the region is shown in figure 3.5. The mean GC content of the whole region is 50.0%. This is much higher than the

genome-wide value of 41% (Lander *et al.*, 2001) and is again consistent with the characteristics of a chromosomal light band. However, figure 3.5 shows that local GC content can deviate substantially from this average figure. Overall, this region is GC-rich, apart from positions such as 40.65 Mb to 40.82 Mb (denoting the position along the q arm of chromosome 22) where GC content at some points drops below 45%. In addition to the low GC regions, there are some high peaks in GC content. Peaks in GC content also appear to correspond with gaps in the bacterial clone contigs of this region (extrapolated from the sequence immediately adjacent to the gaps) (%GC > 55%). Further analysis of this observation is provided in chapter IV.

Isochores have been discussed in chapter I. The local variations in GC content, seen in figure 3.5 may correspond to different isochores. The amount of DNA corresponding to different GC content fractions was calculated using windows of 250 kb over 22q13.31 (table 3.2). The table shows that 1197.5 kb corresponds to the GC content expected within a H3 isochore (37%) and only 547.5 kb corresponds to L1 isochore (17%).

Table 3.2: GC content, amount of DNA and isochore correspondence.

GC content (%)	Amount of DNA (kb)	Corresponds to isochore (Bernardi, 1993)
>=59	267.5	H3
56<=GC<59	370.0	H3
53<=GC<56	560.0	H3
50<=GC<53	522.5	H2
47<=GC<50	532.5	H2
43<=GC<47	447.5	H1
GC<43	547.5	L1

These results, showing that much of the region consists of H3 isochore, also correlate with the published characteristics of a light band region.

TAKE OUT THIS PAGE FOR FOLDOUT PICTURE

Figure 3.5 (fold-out). Transcript map of 22q13.31. This figure shows the complete transcript map of 22q13.31, with the centromere to the left and telomere to the right. The gene structures are indicated by coloured blocks. Full gene structures are displayed in dark blue, partial structures in light blue and pseudogenes in green (see tables 3.8 and 3.9). The following features are displayed: GC plot of the region (in red) showing deviation from the regional average of 50% GC; transcripts and pseudogenes (those orientated 5' to 3' on the DNA strand from centromere to telomere are designated '+' and those on the opposite strand '-'); predicted CpG islands (yellow); the LINE (pink), SINE (purple) and 'Other ' (blue) repeat distributions; and finally the tiling path of overlapping clones labelled by their GenBank/EMBL/DDBJ accession number.

3.4 Transcript map of a 3.4Mb region of human chromosome 22

3.4.1 Sequence analysis

3.4.1.1 Definition of initial gene features

I used the first-pass annotated data (figure 3.1) and additional analysis data as it became available (figure 3.2) to annotate potential gene features for more in-depth investigation and experimental design. Gene features were initially grouped according to the evidence that was used to identify them as follows:

1. Known genes: identical to known human gene cDNA, ncRNA or protein sequences.
2. Related genes: similar, or containing a region of similarity, to protein sequences from human or other species by BLASTX.
3. Putative genes: similar to only ESTs or exon trap data by BLASTN.
4. Pseudogenes: similar to a known gene or protein, but with a disrupted open reading frame.

In total, 71 features were initially identified for further analysis (see table 3.3).

Table 3.3: Initial feature identification in 22q13.1

Type of Feature	Number
Known genes	10
Related genes	21
Putative genes	23
Pseudogenes	17
Total	71

3.4.1.2 Annotation of known genes

Until November 1999, the Sanger Institute annotation team had annotated most of the genes for which a cDNA was already present in the GenBank/EMBL/DDBJ database. Nine protein-coding genes were identified in this way at the start of the project. Additionally, one non-coding snRNA gene was identified by a subsequent BLASTN search of the EMBL vertebrate RNA database (J. Collins). In total, 10 known genes mapped to the region (see table 3.3). All match the chromosome 22 sequence 100% over the length of the gene, apart from C22orf1, which partially lies in a genomic sequence gap.

3.4.1.3 Annotation of related genes

The BLASTX data that determined the 'related' gene features was used to generate a possible gene structure from the different sequences spanning the gene. Nine related genes were annotated from similarities to other human genes. Three of these genes were annotated from homology to cDNAs sequenced by the Kasuza Institute, found to give partial coverage of the full gene structure. A further 12 genes were annotated based on homology to genes from other organisms. All of these features required further experimental work to confirm the full structure (see below).

3.4.1.4 Annotation of putative genes

In the third category, 23 potential gene features were targeted for the additional investigation in order to annotate and extend a gene structure. These included seven partial gene structures, generated from a composite of splicing EST evidence. Six further features were annotated from non-splicing EST clusters.

Trofatter *et al*. (1995), reported a chromosome 22-specific exon trap study. Twenty-four of the generated exon trap sequences are found in this region. Fourteen of these were already

incorporated into gene structures. The remaining ten exon trap sequences were included as putative genes for further investigation.

3.4.2 Experimental approaches

A summary of the additional experimental work performed to extend and confirm the identified gene features is described below.

3.4.2.1 Vectorette cDNA library production and screening

Production of cDNA Libraries

An adapted version (J. Collins, unpublished) of vectorette PCR (Riley *et al.*, 1990) was used to screen suitably adapted cDNA libraries in order to confirm and extend the predicted gene structures. The vectorette method has the advantage of screening large numbers of clones in pools of a large set of libraries whilst retaining high specificity, due to the use of the vectorette bubble.

Consequently, libraries were prepared from human fetal lung cDNA (Invitrogen) and HL60 peripheral blood cDNA (Invitrogen) (M. Goward) (see chapter II). These two libraries formed part of the Sanger Institute vectorette library resource and have since been extensively used for cDNA PCR amplification and sequencing by a number of research groups. Seven vectorette libraries (see table 2.2, chapter II) were available to screen during this project. An example of a vectorette PCR library screen is shown in figure 3.6, showing PCR amplification of cDNA using primers specific to the putative gene locus ARHGAP8.

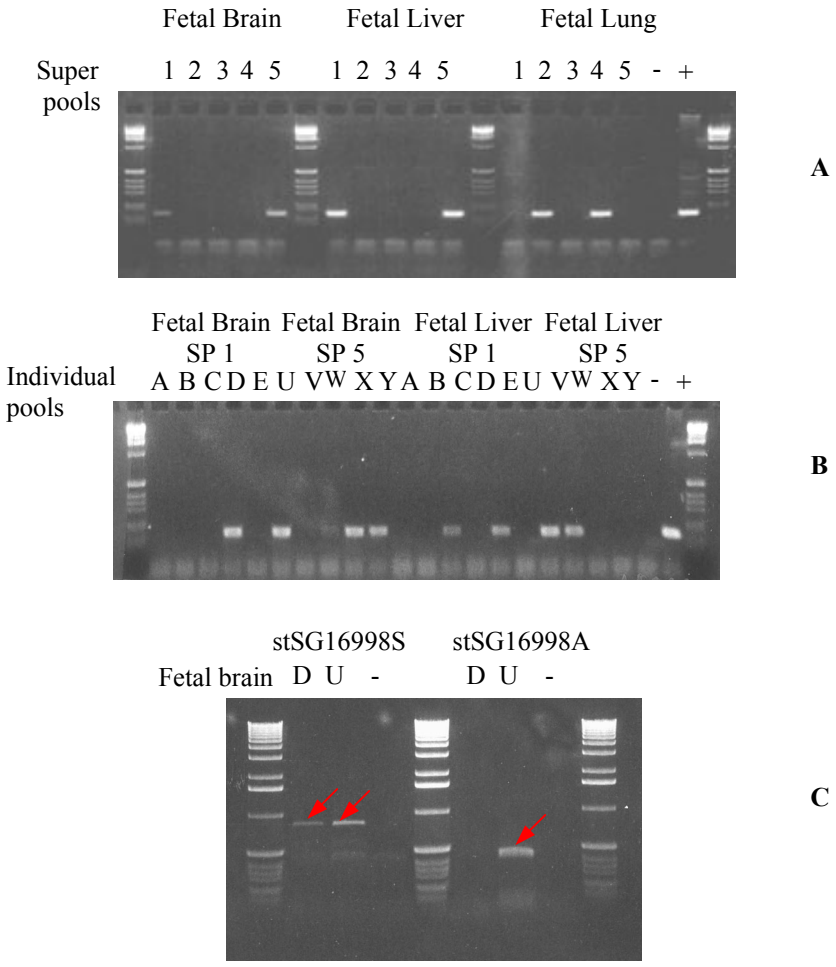


Figure 3.6: Example of vectorette based isolation of PCR fragments from cDNA library using primers stSG16998 (H55372), contained within the locus ARHGAP8. Screening of the super pools (A), is followed by individual pool screening (B). The identified pools are then used as templates in vectorette PCR (C). The marked bands were excised and gel purified prior to sequencing.

3.4.2.2 Screening results

Forty-four potential gene loci were screened (21 related genes + 23 putative genes) against the seven available vectorette libraries. In total, 66 pre-existing and specifically designed primer pairs were used in PCRs to confirm and extend the potential gene structures. This data is summarised in figure 3.7.

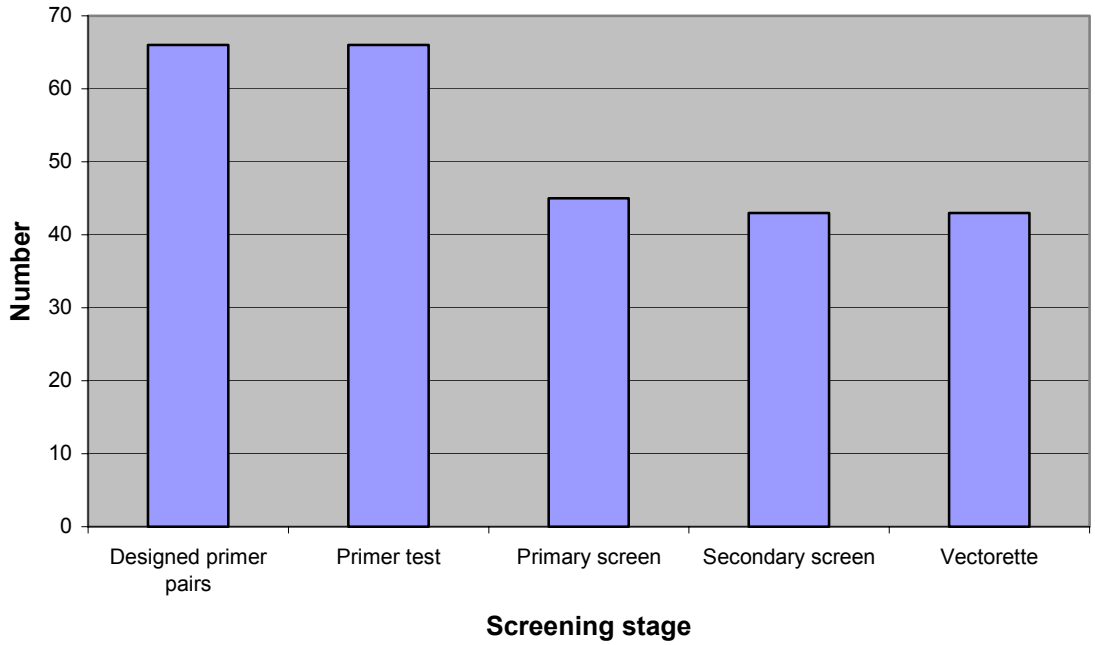


Figure 3.7: Vectorette cDNA library screens. The total number of primer pairs, designed to potential gene features based on similarity evidence, that have been screened across the vectorette cDNA libraries. The bars represent the total number of leads that succeeded at each of the stages.

This data indicates that the largest dropout takes place at the primary screening stage, indicating that either these negative STSs do not correspond to real genes, or they correspond to rare transcripts that occur at very low copy numbers, or are not in the tissues represented by the seven vectorette libraries.

In total 114 sequence reads were generated (E. Huckle) (table 3.4). Of these reads, 69.3% aligned to the chromosome 22 genomic sequence and contributed to the annotation. Twenty-six percent of the sequence reads did not derive from chromosome 22, but demonstrated homology either to other human chromosomes or vector sequences. The remaining sequence reads contained repeat sequence (4.4%). The ability to screen out these false positive results demonstrates a further benefit of having the genomic sequence available.

Table 3.4: Distribution of generated cDNA sequences.

Class	# Sequences
Contributed to annotation	79
Repeat	5
Other homologies	30
Total	114

3.4.2.3 IMAGE clones

In addition to the vectorette approach, a different method was used to obtain additional sequence for the 'related' gene feature E46L. A partial predicted structure was defined from sequence similarity to the mouse brain protein E46 (Em:X61506). A BLASTN search showed that several IMAGE cDNA clones (Lennon *et al.*, 1996) aligned to this region. One of the IMAGE clone inserts (IMAGE I.D. 0035747) was sequenced in order to confirm and extend the E46L gene structure. Subsequently, IMAGE clone resources were not used due to problems of T1 phage contamination.

3.4.2.4 Non-vectorette cDNA libraries

Thirteen gene features did not generate positive results in PCR screens of the seven vectorette cDNA libraries. The remaining 11 cDNA libraries (non-vectorette) available at the Sanger Institute were screened by PCR (table 2.2, chapter II). However, no further positives were found.

3.4.3 Transcript mapping results

3.4.3.1 Library screens

Alignment of the generated cDNA sequence against the genomic DNA allowed the confirmation and extension of 13 putative gene structures. None of the ten remaining exon trap sequences was incorporated into extended gene structures.

Twenty of the 21 related genes were identified in the vectorette and IMAGE cDNA library screens. Incorporation of the generated cDNA sequence into the gene structures allowed seven previously separate features to be incorporated into two extended gene structures. In total, 16 related gene structures were generated.

Eighteen novel mRNA sequences, incorporating an unambiguous ORF and 5' and 3' UTR sequences, were submitted to EMBL/DDBJ/GenBank (Goward and Huckle, unpublished). Accession numbers are listed in table 3.9.

3.4.3.2 Updated BLAST searches

Periodically, BLASTN and BLASTX searches were conducted against novel and updated sequence databases (see appendix 2a), in order to identify new genes and pseudogenes. BLASTN searches of the EMBL vertebrate RNA database identified two human cDNA sequences with 100% identity to human chromosome 22. These were annotated as the loci dJ100N22.C22.4 and dJ753M9.C22.4, but with the note that poly(A) sequence existed in the genomic DNA adjacent to these structures (J. Collins). They were included for further analysis (see below) to check if these structures were true genes, or arisen from spurious reverse transcription from the genomic poly(A) sequences.

Additionally, submission of cDNA sequences by other authors after the start of this project allowed annotation of the full or partial structure of nine of the genes under investigation. These sequences are listed and referenced in table 3.9.

3.5 Investigation of expression

The analysis described above resulted in the annotation of 41 gene structures: 10 initial known genes, 16 generated from the related gene set, 13 confirmed structures from the putative gene set and 2 human cDNAs identified from subsequent BLAST experiments. These loci are listed

in table 3.9. Further investigation of these features was carried out using Northern blot analysis, construction and RT-PCR screening of a human cDNA expression panel and investigation of the tissue origin of EST hits to the cDNA sequences.

3.5.1.1 Northern hybridisation

Hybridisation of a gene-specific probe to a Northern blot allows investigation of whether the sequence is expressed in the tissues represented on the blot, determination of transcript size and possible indication of the existence of alternative transcripts or gene paralogs. The expression pattern and transcript size results are shown below. Analyses of alternative transcripts and paralogous genes are shown in sections 3.8.6 and 3.8.7 respectively.

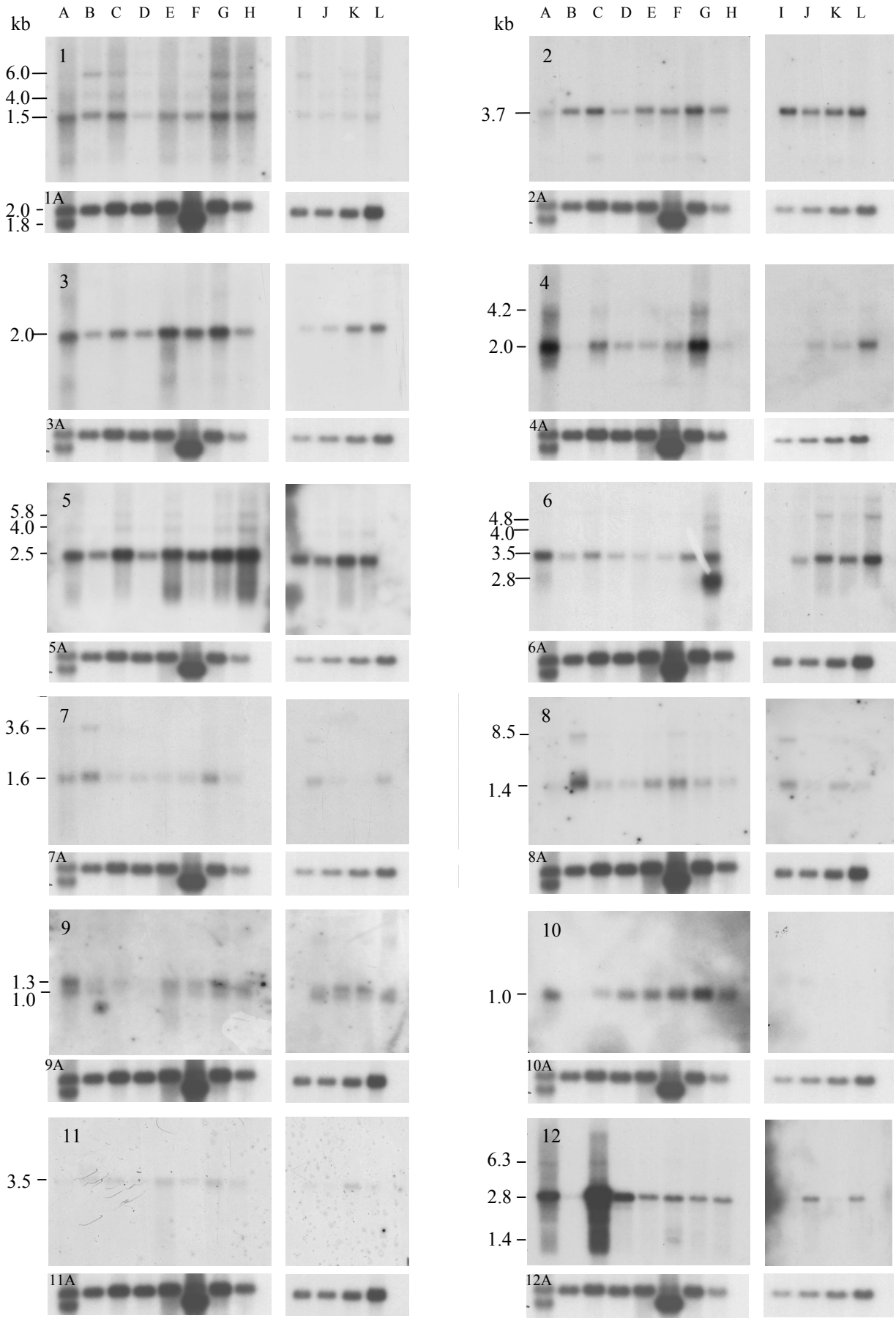
Northern analyses were carried out for the 41 gene loci annotated within the region (see chapter II). Radio-labelled probes were generated by PCR from RNA templates, using primers designed from annotated cDNA sequences and hybridised to Northern blots containing RNA from eight adult and four fetal human tissues (Clontech). Additional hybridisations were perfomed against each Northern blot using a β-actin control probe (Clontech). The results are depicted in figure 3.8. Table 3.5 summarises the obtained sizes and the expected sizes from the current annotation. In cases where the annotated structure is known to be incomplete, the expected transcript size is marked as greater than given by the current annotation. Where available, transcript size estimates from previously published Northern blot data are also shown. Northern results supporting the current gene annotation are highlighted in blue.

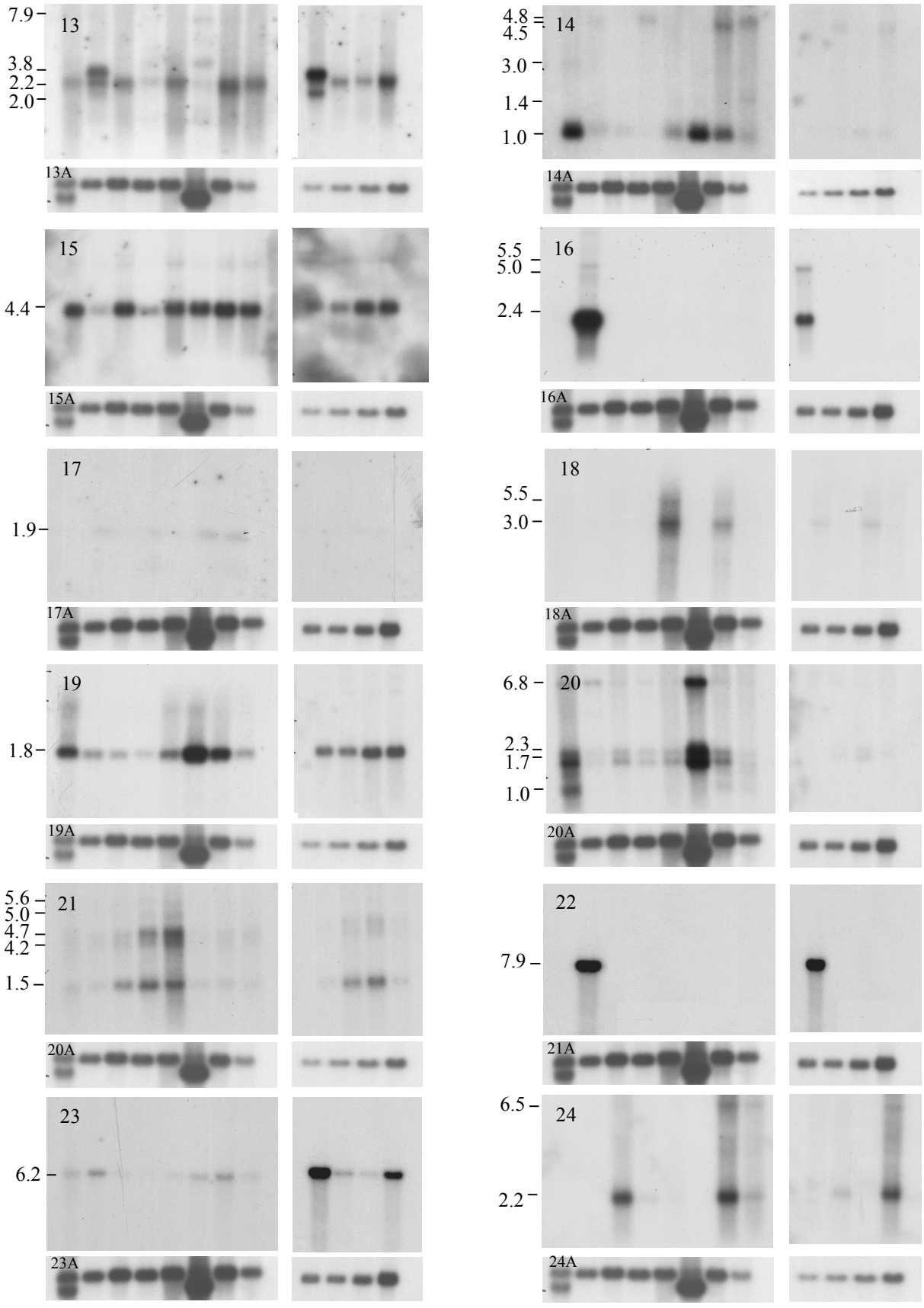
3.5.1.2 Transcript size

All control hybridisations using the β-actin probe generated the expected band intensities of sizes 1.8 and 2.0 kb. Bands were generated from 29 of the 41 blot experiments. Comparison with previously published Northern blot results, where available, showed that the transcript

sizes were generally consistent. Differences may arise through the use of different probes and RNA populations.

In four of the 29 blot experiments that gave a positive result, the annotation was known to be incomplete (dJ526I14.C22.2, dJ345P10.C22.4, HMG17L1 and dJ671O14.C22.6). The larger transcript sizes estimated from the Northern blot evidence may indicate the size of the full transcript and could prove useful in future work to complete the annotation of these genes. However, blots may in fact indicate the existence of larger paralogous gene. This is unlikely for dJ526I14.C22.2, dJ345P10.C22.4 and dJ671O14.C22.6, as BLAST searches of the NCBI human genome sequence database (http://www.ncbi.nlm.nih.gov/genome/seq) do not highlight any potentially paralogous genes that show a high sequence identity to the STS probe used. However, the Northern blot result for HMG17L1 could be explained by hybridisation of the probe to the 7 kb transcript of the human HMG17 gene (Em:X13546). Interestingly, no smaller band sizes were noted that could have originated from the putative HMG17L1 gene.


kb
A B C D E F G H
I J K L
1
6.0
4.0
1.5
1A
2.0
1.8
2
3.7
2A
3
2.0
3A
4
4.2
2.0
4A
5
5.8
4.0
2.5
5A
6
4.8
4.0
3.5
2.8
6A
7
3.6
1.6
7A
8
8.5
1.4
8A
9
1.3
1.0
9A
10
1.0
10A
11
3.5
11A
12
6.3
2.8
1.4
12A


7.9
13
3.8
2.2
2.0
13A
4.8
4.5
14
3.0
1.4
1.0
14A
15
4.4
15A
5.5
5.0
16
2.4
16A
17
1.9
17A
18
5.5
3.0
18A
19
1.8
19A
6.8
20
2.3
1.7
1.0
20A
5.6
5.0
21
4.7
4.2
1.5
20A
22
7.9
21A
23
6.2
23A
6.5
24
2.2
24A

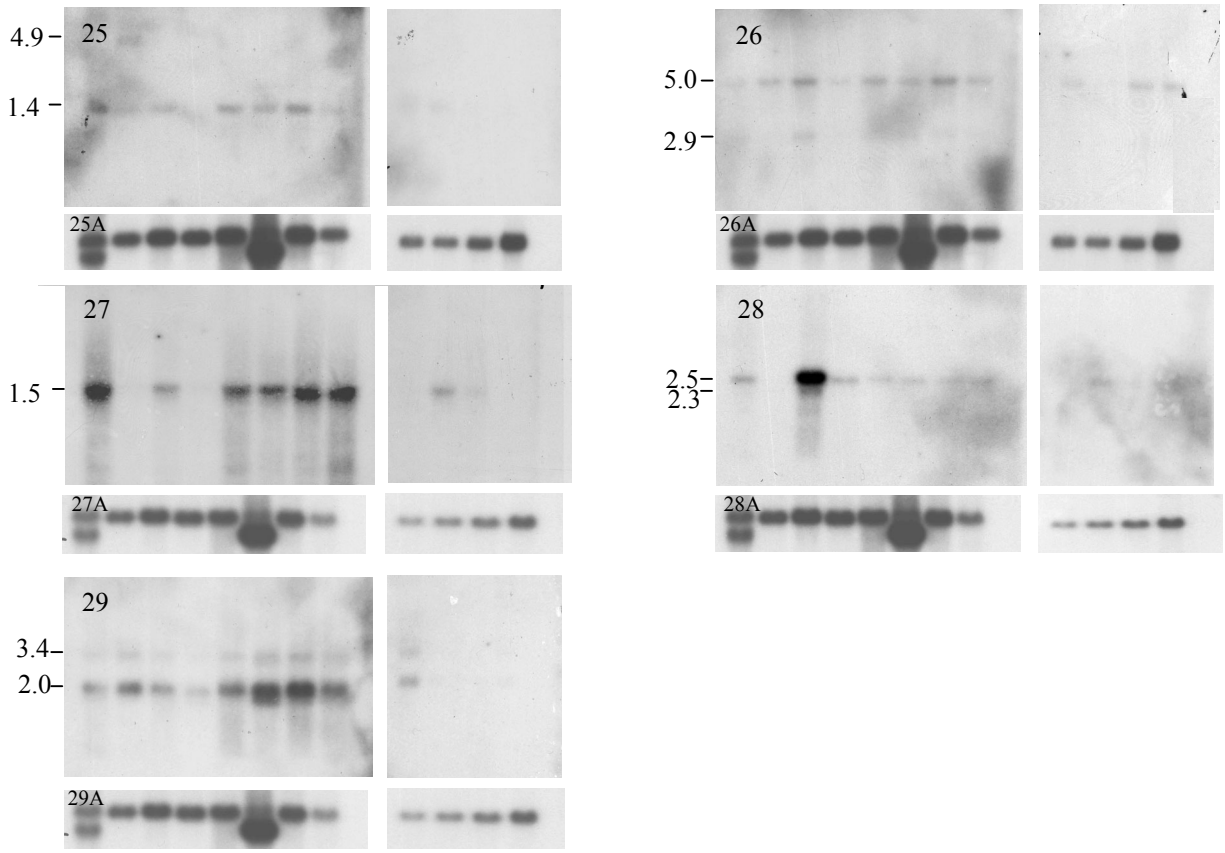


Figure 3.8 Results from 41 Northern blots – only the 29 experiments that gave a positive result are shown. Results generated from hybridisation of each Northern blot to a β-Actin control probe (Clontech) are shown underneath each band (A). Approximate band sizes are shown to the left of each blot (only in the first example in the case of the β-Actin control). The contents of lanes A-L are shown in table 3.6 below.

Table 3.5: Expected and obtained transcript sizes from Northern blot hybridisations from the genes of

Blot #	Locus	Expected transcript size	Approx. Northern blot band size	Previously published size (Northern blot data)
1	dJ222E13.C22.1	1.399, 1.319, 1.271, 1.207	6, 4, 1.5	
2	dJ222E13.C22.3	3.440, 3.272	3.7	
3	DIA1	1.954	2.0	
4	cB33B7.C22.1	2.02	4.2, 2.0	2.02[1]
5	ARFGAP1	2.699, 2.567	5.8, 4.0, 2.5	2.7[2]
6	PACSIN2	3.247, 3.124	3.5	3.5[3]
7	TTLL1	1.684, 1.618, 1.051	3.6, 1.6	8.4, 4.8, 1.8[4]
8	BIK	1.098	1.4, 8.5	1.35[5]
9	bK1191B2.C22.3	1.281, 1.063	1.3, 1.0	
10	BZRP	0.85	1.0	1[6]
11	dJ526I14.C22.2	>3.353, 2.049	3.5	
12	dJ526I14.C22.3	2.805	6.3, 2.8, 1.4	
	dJ100N22.C22.5	2.848		
	dJ754E20A.C22.4	>0.951		
13	C22orf1	2.223	7.9, 3.8, 2.2, 2.0	multiple (<1-4.8)[7]
14	dJ345P10.C22.4	>4.88, >4.746	4.8, 4.5, 3.0. 1.4, 1.0	
15	HMG17L1	>1.159	4.4	
16	SULTX3	2.386, 2.347	5.5, 5.0, 2.4	
17	dJ388M5.C22.4	>1.74	1.9	
18	dJ549K18.C22.1	2.805, 1.177	5.5, 3.0	
19	CGI-51	1.716	1.8	
20	bK414D7.C22.1	1.65	6.8, 2.3, 1.7, 1.0	
21	dJ671O14.C22.2	1.503, 1.43	5.6, 4.7, 4.2, 1.5	
22	dJ671O14.C22.6	>6.332	7.9	
	dJ1033E15.C22.1	>0.618		
23	dJ1033E15.C22.2	2.677	6.2	
	dJ474I12.C22.5	>0.72		
	dJ474I12.C22.2	>0.817		
24	ARHGAP8	2.264	2.2, 6.5	
25	dJ127B20.C22.3	5.17	4.9, 1.4	
	dJ753M9.C22.4	6.412		
26	NUP50	5.172	5.0, 2.9	8, 5, 2.8, 2[8]
	bK268H5.C22.1	6.306		
	UPK3	1.051		
	bK268H5.C22.4	2.879		
	SMC1L2	>4.253		
27	dJ102D24.C22.2	1.392	1.5	
28	FBLN1	2.525, 2.349, 2.156, 1.159	2.5, 2.3	
	bK941F9.C22.6	>0.376		
29	E46L	3.331	3.4, 2.0	

[1] Kojima *et al.* ; [2] Zhang , 2000; [3] Ritter , 1999; [4] Trichet , 2000; [5] Verma , 2000; [6] Chang , 1992; [7] Schwartz & Ota, 1997; [8] Trichet , 1999.

Where available, previously published Northern blot results are included for comparison. Transcript sizes, which may be equivalent at the level of blot resolution, are highlighted in blue.

The expected transcript size agreed with the size of the strongest or most common band established by Northern blotting in a further 22 experiments. A limit of correlation of 500 bp was applied in most cases, extended to 1.5 kb for transcripts larger than 4 kb, due to the limited resolution of the Northern blots. This evidence therefore predominantly supports the current annotation, although differences caused by, for example, missing exons, may not be picked up due to the limited resolution of the blot experiments.

In two more cases (dJ127B20.C22.3 and E46L), the expected transcript size was within the correlation limit of the size of a weaker or less common band established by Northern blotting. These results also support the current annotation. The stronger bands may be generated by more common isoforms or paralogs of the gene, although no potential candidates were identified in TBLASTN searches of the draft human genome sequence (section 3.8.7).

The Northern blot experiment for dJ1033E15.C22.2 (number 23) indicated a much larger transcript, estimated to be six kilobases long from Northern blot evidence, than the one currently annotated. The alignment of the cDNA Em:AL136553 against the genomic sequence indicates that dJ1033E15.C22 has an unspliced structure. This gene may therefore be a processed pseudogene and the transcript indicated by the Northern blot may in fact be the gene from which dJ1033E15.C22.2 is derived. However, BLAST searches of the nucleotide and predicted amino acid sequence of dJ1033E15.C22.2 against the human genome sequence (http://www.ncbi.nlm.nih.gov/genome/seq) failed to identify a candidate for the original gene. This evidence would be required in order to reclassify dJ1033E15.C22.2 as a pseudogene. Alternatively, this evidence may indicate that this gene structure is incomplete.

Overall, the Northern blot evidence supports the transcript size of 24 annotated genes. A further 12 blot experiments gave no result, possibly because these genes are not expressed at high levels in the tissues represented on the blots, or because the annotated structures do not represent true

expressed genes (see below). Four blot experiments provided evidence of the potential transcript size of partial genes. Further experimental work is needed to complete the partial gene structures in this region. This could include screening more cDNA libraries in order to generate further cDNA sequences to complete the annotation. Additionally, 5'RACE experiments could be carried out to extend the annotation of 5' gene sequences.

3.5.1.3 Expression

The Northern blot experiments described above provide evidence of expression patterns. The expression patterns of transcripts of the correct size identified from these experiments (highlighted in blue in table 3.4) are included in figure 3.10.

Twelve Northern blot experiments may have failed because the annotated gene feature was not expressed in any of the tissues represented on the Northern blot. Alternatively, the annotated gene feature may be spurious and not expressed at all. To test this possibility and to further investigate expression patterns of all the gene features of interest, a human tissue mRNA expression panel was constructed and screened.

3.5.2 Construction and screening of expression panel

RNA was extracted from seven different human tissue samples and one human cell line. An additional 24 samples were supplied as RNA (table 2.3, chapter II). In total, RNA from 32 human tissues was reverse transcribed and screened by RT-PCR using primers designed to the 41 gene structures under investigation (chapter II). Although the RNA was treated with DNase during the production protocol, PCR primers were designed across introns where possible, in order to negate the affect of possible genomic DNA contamination. This was not possible for dJ1033E15.C22.1, dJ1033E145.C22.2, dJ100N22.C22.5, dJ753M9.C22.4 and dJ222E15.C22.7, where primers were designed to the single exon. Profiles were obtained for 41 genes in duplicate (figure 3.8). All the expression data from these experiments is summarised in figure 3.10.

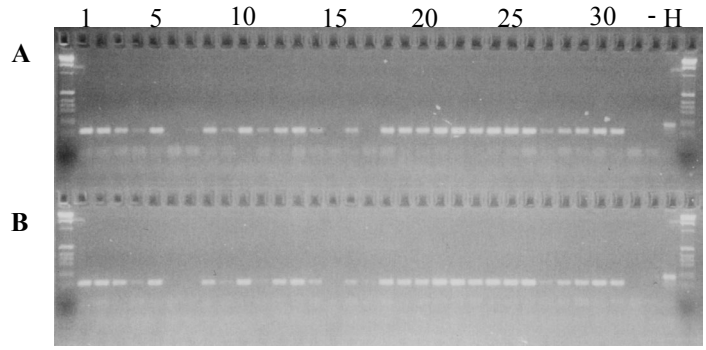


Figure 3.9: Example of a transcription profile for TTLL1. A and B represent duplicate experiments. The experiment was performed in duplicate. - = negative control; H = human genomic DNA. The genomic band is larger as the primers span an intron in TTLL1. The lane designations correspond to the key in table 3.6.

Weak or absent PCR fragments were consistently noted in samples derived from rectum and fetal bladder. This may reflect the true expression profile of the genes tested, but is likely due to experimental error during construction of the cDNA panel. Bands were not always seen from amplification of human genomic DNA; this is because the introns spanned by the primers used were sometimes too large for PCR amplification.

Figure 3.10: Transcription profiles for 41 genes annotated in 22q13.31, tested on 32 tissues by RT-PCR and 12 tissues by Northern blotting. Key to tissue identity is shown in table 3.6.

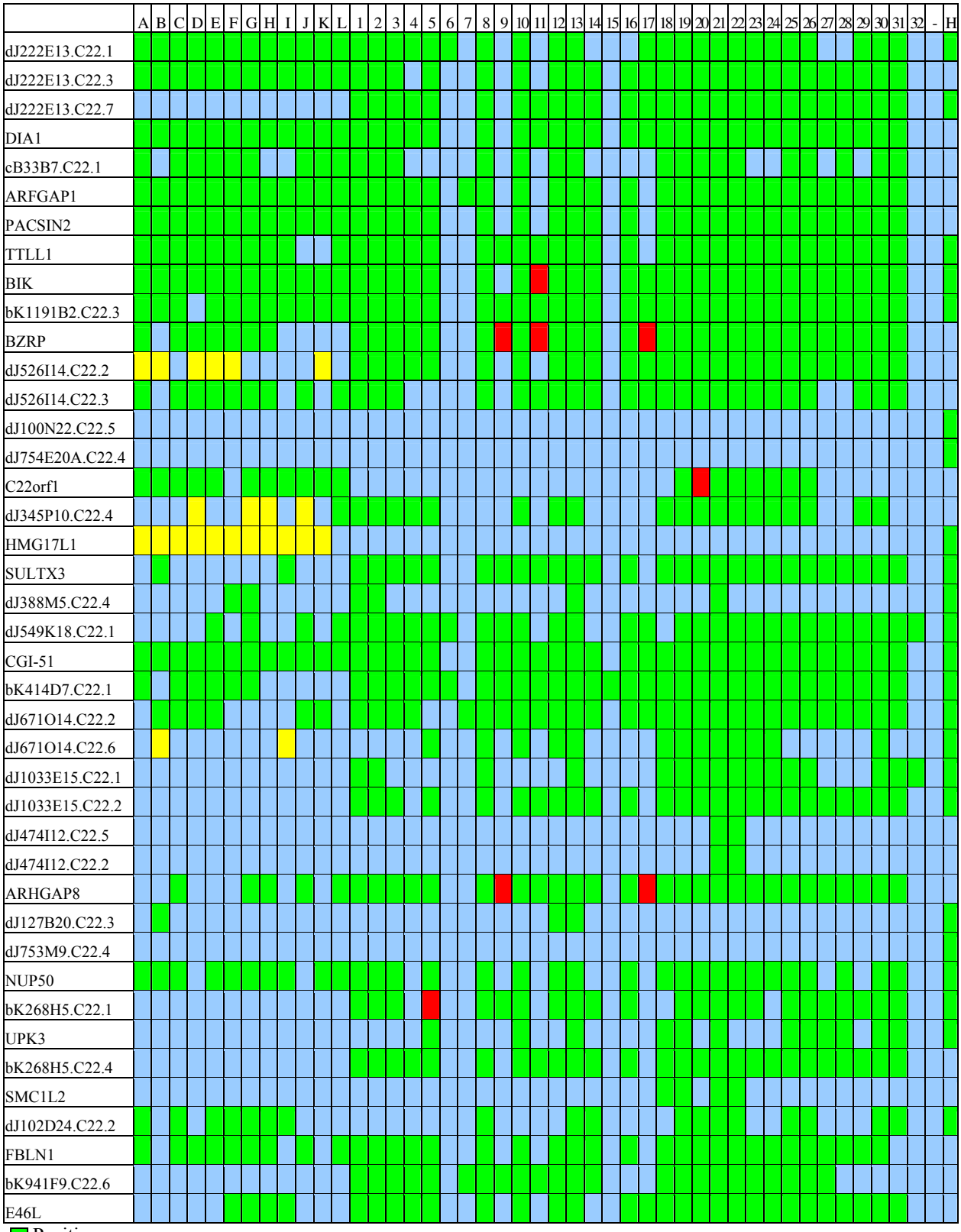


Table 3.6: Key to tissue identity

	Tissue		Tissue
A	Heart	12	Stomach
B	Brain (whole)	13	Colon I
C	Placenta	14	Colon II
D	Lung	15	Rectum
E	Liver	16	Breast
F	Skeletal muscle	17	Ovary
G	Kidney	18	Uterus
H	Pancreas	19	Cervix I
I	Fetal brain	20	Cervix II
J	Fetal lung	21	Testis I
K	Fetal liver	22	Testis II
L	Fetal kidney	23	Fetal brain I
1	Kidney I	24	Fetal brain II
2	Kidney II	25	Fetal heart I
3	Liver I	26	Fetal heart II
4	Liver II	27	Fetal liver I
5	Cerebrum	28	Fetal liver II
6	Skeletal muscle	29	Fetal lung I
7	Skin	30	Fetal lung II
8	Tonsil	31	Fetal spleen
9	Lymphoblast (cell line)	32	Fetal bladder
10	Thyroid	-	water
11	Spleen	H	genomic DNA

3.5.3 EST tissue origin

Additional information about tissue distribution can be derived from the tissue origin of EST sequences that show a high level of similarity to the annotated gene sequences. The script e-profile (Smink and Beare, unpublished) formats the results of a BLASTN search of the dbEST database into an output highlighting the tissue origin of matching EST sequences. An example of e-profile output is shown in figure 3.11. This shows that EST sequences showing 80% or more identity at the nucleotide level to the cDNA sequence of dJ222E13.C22.3a (Em:AL160111) (isoform a) originate from a wide range of tissues. Results from the remaining 40 annotated gene structures in 22q13.31 are shown in appendix 3.

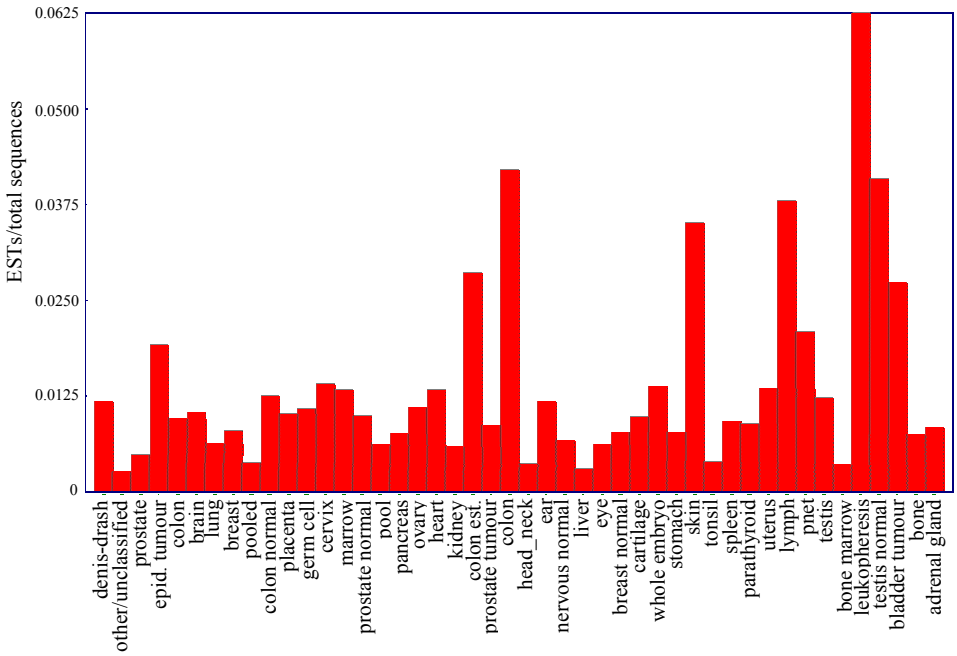


Figure 3.11: Expression profile of dJ222E13.C22.3a (Em:AL160111).
The proportion of ESTs from a range of tissues that show >80% similarity at the nucleotide level to the cDNA sequence of dJ222E13.C22.3 (isoform a). Generated using e-profile (Smink and Beare, unpublished).

3.5.4 Overall expression results

Overall, the Northern blot, cDNA panel and e-profile results show that most of the genes annotated in 22q13.31 show expression in a wide range of tissues. However, SMC1L2 expression appears to be mainly restricted to reproductive tissues (apart from results from e-profile, which also highlight expression in samples of blood from the umbilical cord) and the expression patterns of dJ754E20A.C22.4, dJ474I12.C22.2 and dJ474I12.C22.5 are restricted to testis only.

No evidence of expression was found for dJ100N22.C22.5, or dJ753M9.C22.4. These genes were noted in section 3.4.3.2 as putatively arising from spurious poly(A) priming of genomic DNA during preparation of the cDNA library and the lack of expression data concurs with this possibility.

3.6 Experimental testing of *ab initio* gene predictions

All the gene features investigated above are annotated from expressed sequence evidence, either submitted by other authors or generated as part of this project. It may be that additional genes or exons, without homology to existing expressed sequence evidence, remain undiscovered in the region of interest. *Ab initio* gene prediction programs provide structural information about potential genes that is independent of the spatial and temporal limitations of expression evidence discussed in the introduction. However, studies have shown that these methods have limited accuracy and may have over-prediction rates of over 30% (section 3.9.2). Consequently, *ab initio* gene predictions alone are not considered sufficient for reliable gene annotation, although they may be useful as a starting point for experimental studies (Dunham *et al*., 1999).

Genscan (Burge & Karlin, 1997) and Fgenesh (Solovyev *et al.*, 1994) are *ab initio* gene prediction programs that have been run on the linked clone sequences of chromosome 22. Many predictions coincide with expressed sequence homologies, which combined evidence provides strong evidence for a gene. However, other predicted exons do not align to expressed sequence evidence. These exons could indicate the presence of previously undetected genes, or could be a result of over-prediction by the gene prediction program. Therefore, in order to discover if true genes had escaped previous experimental detection, Genscan exons that had no previous supporting experimental sequence homology were selected for primer design and PCR screening of cDNA libraries.

3.6.1 cDNA library screens

Fifty-nine predicted exons that had no supporting experimental sequence homology were selected for investigation. Primer pairs were designed to each exon and used in PCR screens of vectorette cDNA libraries. This data is summarised in figure 3.12.

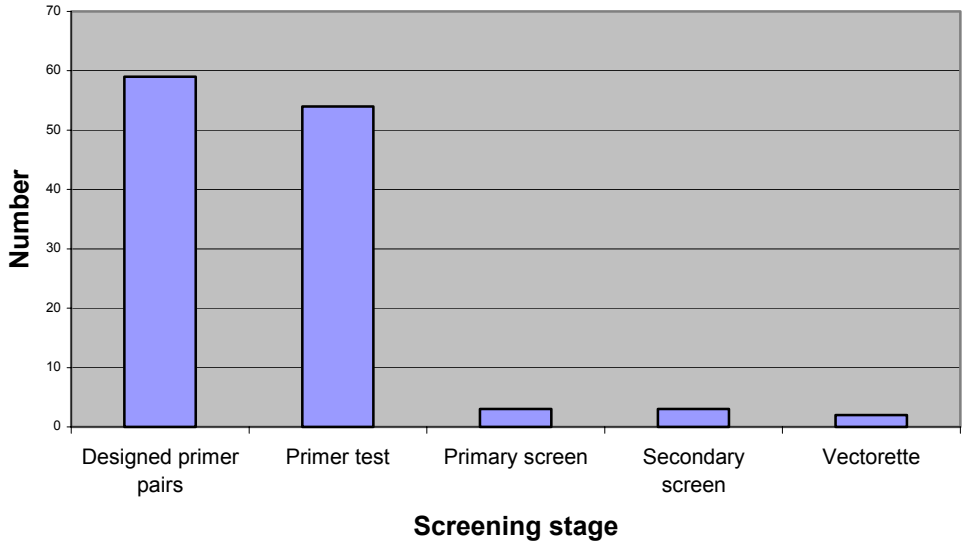


Figure 3.12: The total number of primer pairs, designed to Genscan predicted exons without similarity to expressed sequence evidence, which have been screened across the vectorette cDNA libraries. The bars represent the total number of leads that succeeded at each of the stages.

In total 19 sequence reads were generated (E. Huckle) and 42% of these contributed to the annotation (see table 3.7). Six of the sequence reads defined one partial gene structure from a predicted Genscan exon amplified from three vectorette libraries (fetal brain, fetal liver and fetal lung). Later extension of this structure by vectorette PCR merged this locus with four others previously identified by homology information (dJ345P10.C22.4).

Table 3.7: Number and type of sequence reads obtained from sequencing vectorette cDNA PCR products isolated with primers designed to Genscan predicted exons.

Class	# Sequences
Contributed to annotation	8
Repeat	3
Other homologies	8
Total	19

A second Genscan exon that produced a positive result from the fetal brain vectorette library resulted in generation of two sequence reads with high similarity to a true exon in a gene 6kb upstream (ARHGAP8). The surrounding intron does not appear to be replicated. It could be that

the positive result highlights an alternative 3' end of ARHGAP8, or that this sequence is not truly expressed and the primers amplified DNA from the true exon in ARHGAP8.

A primer pair designed to a third Genscan exon initially gave a positive result in cDNA screens, but failed at the vectorette stage. However, extension of a homology-based gene structure was shown to incorporate this exon (dJ671O14.C22.2).

Overall, only three primer pairs from 59 (5.1%) Genscan predicted exons, which initially had no expressed sequence similarity, were shown to be present in the seven cDNA vectorette libraries screened. None of these identified a novel gene and the three exons were later incorporated into the existing structures as described above.

3.7 Transcription map results

The current annotation of the transcript map is categorised as follows:

1. Full genes: Has a fully defined ORF, including start and stop codon and annotated 5' and 3'UTR sequences. The sequence has been submitted to EMBL/DDBJ/GenBank.
2. Published partial gene: Submitted to EMBL/DDBJ/GenBank, but lacking a fully defined ORF, including start and stop codons.
3. Unpublished partial gene: Not submitted to EMBL/DDBJ/GenBank and lacking a fully defined ORF, and/or start and stop codons.
4. Rejected (Poly(A) in genomic): Annotated from a publicly available cDNA, but probably arisen from spurious genomic poly(A) priming.
5. snRNA: Full gene, submitted to EMBL/DDBJ/GenBank, encoding a snRNA.
6. Pseudogene (R): Homologous to a known gene or protein, but unspliced with a disrupted open reading frame. Possibly derived from retrotransposon (R) activity.
7. Pseudogene (D): Homologous to a known gene or protein, spliced, but with a disrupted open reading frame. Possibly derived from a gene duplication (D) event.

Table 3.8 provides a summary of the results of the work to generate a transcript map of 22q13.31 and includes the EMBL accession numbers of submitted genes and alternative isoforms (designated .a, .b, .c etc. in the text). Table 3.9a lists the annotated pseudogenes, together with the sequence accession number and chromosomal location of the genes from which they were annotated. The annotated genes are listed in table 3.9b. The transcript map of the entire region is shown in figure 3.5 and a table detailing the features of all the genes is in appendix 4. In total, 58 features were annotated.

Table 3.8: Number and type of annotated gene features

Type of feature			Number
Full gene			27
Partial gene			11
	(Published, partial gene)	3)	
	(Unpublished, partial gene	8)	
snRNA gene			1
Rejected (Poly(A))			2
Pseudogene			17
	(Retrotransposon	15)	
	(Duplicate	2)	
Total			58

Table 3.9a: Pseudogenes annotated within 22q13.31. The accession number and chromosomal location of the genes from which they were annotate.

Pseudogene name	Status	Derived from	Chromosomal location
dJ222E13.C22.2	Pseudogene (D)	Em:AF151854	22
dJ222E13.C22.5	Pseudogene (R)	Sw:P36542	10
dJ47A17.C22.1	Pseudogene (R)	Em:U14966	15
dJ47A17.C22.2	Pseudogene (D)	Em:AF035321	9
dJ437M21.C22.4	Pseudogene (R)	Em:AK001665	7
bK1191B2.C22.1	Pseudogene (R)	Gb:AAH4986	11
dJ345P10.C22.1	Pseudogene (R)	Sw:P27348	2
dJ388M5.C22.1	Pseudogene (R)	Sw:P36578	15
dJ796I17.C22.3	Pseudogene (R)	Gb:AAH17093	3
dJ671O14.C22.1	Pseudogene (R)	Em:K02923	19
dJ32I10.C22.9	Pseudogene (R)	Em:U33760	7
bK397C4.C22.1	Pseudogene (R)	Em:AF151892	4
dJ474I12.C22.1	Pseudogene (R)	Em:X12881	X
dJ181C9.C22.1	Pseudogene (R)	Em:Y07569	15
dJ127B20.C22.2	Pseudogene (R)	Em:D17554	18
bK268H5.C22.3	Pseudogene (R)	Em:U14972	11
dJ37M3.C22.5	Pseudogene (R)	Em:AF151805	3

R = possibly derived from retrotransposon activity
D = possibly derived from gene duplication event
Em = EMBL accession no.; Gb = Genbank accession no.; Sw = SwissProt accession no.

Table 3.9b: Genes annotated within 22q13.31. Original status at the beginning of the project, work done and current status is summarised. EMBL accession numbers of the submitted genes are shown.

Gene name	Status at start of project	Work done				Current status	Accession number(s)
		Vec. cDNA library screens	Further cDNA library screens	N blot	RT-PCR		
dJ222E13.C22.1	Related	+		+	+	Full gene	AL589866, AL590120, AL590118
dJ222E13.C22.3	Putative	+		-	+	Full gene	AL160111, AL160112
dJ222E13.C22.7	Known			-	-	snRNA	J04119[1]
DIA1	Known			+	+	Full gene	M16462[2]
cB33B7.C22.1	Putative	+		+	+	Full gene	AB037883[3]
ARFGAP1	Related	+		+	+	Full gene	AL159143, AF111847[4]
PACSIN2	Known			+	+	Full gene	AAD41781[5,] AL136845[6]
TTLL1	Related	+		+	+	Full gene	AL58967, AL096883, AL096886, AF104927[7]
BIK	Known			+	+	Full gene	X89986[8], U34584[9]
bK1191B2.C22.3	Related	+		+	+	Full gene	AL359401, AL359403
BZRP	Known			+	+	Full gene	M36035[10]
dJ526I14.C22.2	Related	+		+	+	Full gene	AL590888, D63487[11]
dJ526I14.C22.3	Related	+		+	+	Unpub. partial gene	
dJ100N22.C22.5	-			-	-	Rejected (Poly(A))	AL442096[12]
dJ754E20A.C22.4	Putative	-	-	-	-	Unpub. partial gene	
C22orf1	Known			+	+	Full gene	U84894[13]
dJ345P10.C22.4	Putative	+		+	+	Pub. partial gene	AB051459[14]
HMG17L-1	Related	+		+	-	Unpub. partial gene	
SULTX3	Related	+		+	+	Full gene	AF188698[15] AF115311[16]
dJ388M5.C22.4	Related	-	-	+	+	Unpub. partial gene	
dJ549K18.C22.1	Related	+		+	+	Full gene	AK025665[17]
CGI-51	Known			+	+	Full gene	AF151809[18]
bK414D7.C22.1	Related	+		+	+	Full gene	AL159142; AF237769[19]
dJ671O14.C22.2	Related	+		+	+	Full gene	AL55092; AF237772[19;] AL590887
dJ671O14.C22.6	Putative	+		+	+	Pub. partial gene	AB051431[20]
dJ1033E15.C22.1	Putative	+		+	+	Pub. partial gene	AF086048[21]
dJ1033E15.C22.2	Putative	+		+	+	Full gene	AL136553[22]
dJ474I12.C22.5	Putative	-	-	-	+	Unpub. partial gene	
dJ474I12.C22.2	Putative	+		-	+	Unpub. partial gene	
ARHGAP8	Related	+		+	+	Full gene	AL355192
dJ127B20.C22.3	Putative	-	-	+	+	Full gene	BC012187[23]

dJ753M9.C22.4	-			-	-	Rejected (Poly(A))	AB051448[24]
NUP50	Known			+	+	Full gene	AF107840[25]
bK268H5.C22.1	Related	+		+	+	Full gene	AB023147[26]
UPK3	Known			-	+	Full gene	AF085808[27]
bK268H5.C22.4	Putative	+		+	+	Full gene	AK000642[28]
SMC1L2	Related	+		-	+	Unpub. partial gene	
dJ102D24.C22.2	Putative	+		+	+	Full gene	AL442116
FBLN1	Known			+	+	Full gene	AF126110[29], U01244[30], X53741[31], X53742[31], X53743[31]
bK941F9.C22.6	Putative	-	-	-	+	Unpub. partial gene	
E46L	Related	+		+	+	Full gene	AF119662

Pu. = published; Unpub. = Unpublished. Unless indicated, all cDNA sequence submitted by Goward and Huckle, unpublished. Additional sequences: [1] Montzka & Steitz, 1988; [2] Yubisui *et al.*, 1987; [3] Kojima , 2000; [4] Zhang , 2000; [5]. Ritter , 1999; [6] Wiemann , 2001; [7] Additional isoform by submitted by Trichet *et al.*, 2000; [8] Pun, unpublished; [9] Boyd , 1995; [10] Riond *et al.*, 1991; [11] Nagase , 1995; [12] Bloecker *et al*., unpublished; [13] Schwartz & Ota, 1997; [14] Hirasawa *et al*., unpublished; [15] Falany , 2000; [16] Sakakibara *et al*., unpublished; [17] Sugano *et al*., unpublished; [18] Lai *et al*., unpublished; [19] Identical submission made subsequently by Olski *et al.*, 2001; [20] Ohara *et al*., unpublished; [21] Woessner *et al*., unpublished; [22] Simpson , 2000; [23] Strausberg, unpublished; [24] Ohara *et al*., unpublished; [25] Trichet *et al.*, 1999; [26] Nagase *et al*., unpublished; [27] Geall *et al*., unpublished; [28] Sugano *et al*., unpublished; [29] Krichevsky , 1999; [30] Tran , 1997; [31] Argraves *et al.*, 1990.

3.8 Analysis of annotated genes

3.8.1 General features of annotated genes

Currently, the total length of the sequence occupied by the annotated genes and pseudogenes, including their introns, is 2.07 Mb; 64.6% of the total available sequence of the region. Pseudogenes occupy just over 20 kb and annotated gene exons make up less than 2.8% of the total sequence. This contrasts sharply with the 41.6% occupied by repetitive sequences.

Table 3.10 shows an overview of the characteristics of the 27 full genes contained within 22q13.31. Included in brackets as a comparison are the equivalent figures calculated for 1,804 RefSeq entries aligned to the draft human genomic sequence over their full length, which are purportedly representative of the whole genome (Lander *et al.*, 2001).

Table 3.10: Mean and median values for a range of protein-coding gene properties

Feature	Mean	Median
Internal exon	160 (145)	132(122)
Exon number	9.6(8.8)	25 (7.0)
Introns	6054(3365)	2896(1023)
3'UTR	1181(770)	2085(400)
5'UTR	160(300)	226(240)
Coding sequence	1174(1340)	2718(1100)
(CDS)	391aa(447aa)	906aa(367aa)
Genomic extent	55.4(27)	92(14)

Equivalent values from analysis of 1,804 RefSeq entries aligned to finished human genomic sequence are included in brackets (Lander *et al.*, 2001).

The value of this comparison is limited due to the small gene sample size (27). However, mean coding exon size and number within 22q13.31 are similar to those of the RefSeq set. The 5'UTR sequence annotated in 22q13.31 are smaller than those of the RefSeq set. This may indicate that the full 5'UTR sequences of several genes are incomplete, due to the limitations reviewed in section 3.1.3.

The table also shows that the genomic span and intron size of the genes in 22q13.31 are larger that those of the RefSeq set. The same observation is noted in a comparison of 22q13.31 against the genes annotated in 22q. Although equivalent exon coverage is noted in 22q13.31 and 22q (2.8% and 3.0% respectively), the genomic coverage of the annotated genes is greater in 22q13.31 (64.6%) than 22q (39%). These observations indicate a larger-than-average intron size within 22q13.31.

The sizes of individual genes encoded within the region vary over a wide range. The analysis is incomplete however, as some coding sequences remain partial. However, the smallest complete gene (dJ1033E15.C22.2) is only 1.563 kb in length whereas the largest single gene (dJ345P10.C22.4) stretches over 283.4 kb. dJ1033E15.C22.2 appears to contain only a single exon whilst the largest number of exons within a gene in this region is 33 (dJ345P10.C22.4). The smallest complete exon identified is 20 bp (bK414D7.C22.1) and the largest is 6.0 kb

(dJ671O14.C22.6). The smallest intron spans 86 bp (bK268H5.C22.1) whilst the largest intron stretches over 10.2 kb (dJ323M22.C22.2).

Several pseudogenes are observed to lie within the introns of other functional genes. In addition, the gene HMG17L-1 appears to lie within the 2nd intron of dJ345P10.C22.4. HMG17L-1 lies in the opposite transcriptional direction to the outer gene. This pair of genes seems to be otherwise unrelated (see expression evidence). There are so far few examples of functional genes embedded within introns of higher eukaryotes, although two examples are known to lie within introns elsewhere on chromosome 22 (Dunham *et al.*, 1999). However, HMG-non-histone related proteins show a clear trend to exist as processed pseudogenes (Venter *et al.*, 2001), so it may be that HMG17L-1 belongs to this category. Further evidence is noted from Northern blot and translational start site investigations (sections 3.5.1.2 and 3.8.3). However, the structure of HMG17L-1 does contain an intron, which is not a characteristic of a processed pseudogene.

Interestingly, two members of the same small gene family were found to be adjacent to each other: bK414D7.C22.1 (β-parvin) and dJ671O14.C22.2 (γ-parvin) are 11.7kb apart, in a head to tail orientation. Along with α-parvin, these three proteins make up a family related to the alpha-actinin superfamily, which mediates cell-matrix adhesion (Olski *et al.*, 2001). The two genes have similar expression profiles (section 3.5) so it is possible that they could share regulatory sequences.

A further possible example of shared regulatory sequences is provided by the genes dJ102D24.C22.2 and SMC1L2. These two genes lie only 83 bp apart on opposite strands (head to head). The genes also share a CpG island and both overlap a PromoterInspector prediction (section 3.8.5) suggesting the existence of a possible bi-directional promoter. However, this pair of genes does not share similar expression profiles: dJ102D24.C22.2 is expressed in a wide range of tissues, whereas SMC1L2 is restricted to reproductive tissues (section 3.5).

3.8.2 Splice sites

To examine whether the splice donor and acceptor sites for this region agreed with previous investigations on 1800 introns (Stephens & Schneider, 1992) and 325 chromosome 22q13.3 introns (Smink, 2001), the splice site sequences for 379 introns were extracted from gff (genome feature format) and sequence files and used to generate sequence logos (D. Beare). The sequence logos not only show the frequencies of the nucleotides at each position, but also the importance of each position in the site under investigation. The height of the base reflects the frequency of that base and the height of the stack at each position reflects the contribution of that position to the overall splice consensus. The generated splice site consensus sequences (figure 3.13) agree well with the published splice sites, as expected. There are some minor differences noted between this study and that of Smink, 2001. In sequence logos, the nucleotide on top of the logo at each position is the most frequent nucleotide. In the C/T tract of the splice acceptor consensus from the 379 introns from 22q13.31, thymidine occurs most frequently than cytosine in all positions (except position 5). Stephens and Schneider,(1992) also made this observation, but Smink, 2001, noted that cytosine tended to occur more frequently than thymidine in these regions. Similarly, both this study, and that of Stephens and Schneider, showed that adenine occurred most frequently for position 9 of the splice donor, whereas the study of 325 introns from 22q13.3 showed guanine was most frequent at this position. The frequency of the nucleotides is also reflected in their size. In the cases noted above, the nucleotides involved appear as similar sizes, thus reflecting that these differences may be minimal and unlikely to have biological relevance.

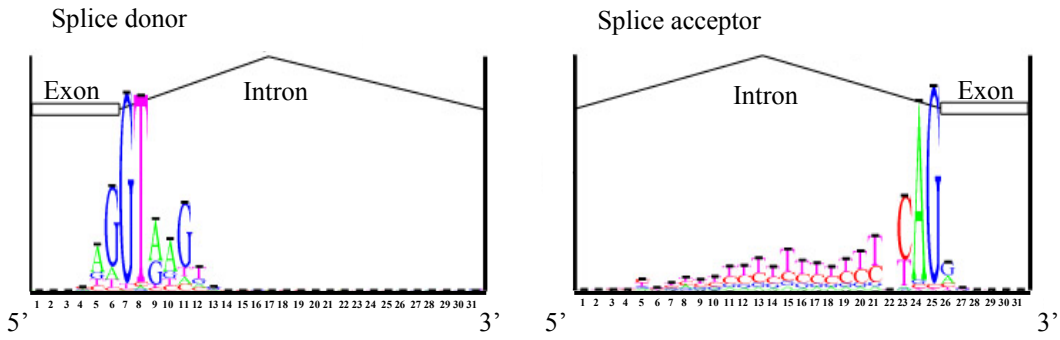


Figure 3.13: Splice donor and acceptor consensus sequences for 379 introns in 22q13.31. The splice site sequences were extracted by D. Beare and visualised using Sequence Logo (Steven Brenner) (http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi).

3.8.3 Investigation of full gene translational start sites

The scanning model of translation initiation (Kozak, 1980) proposes that the majority of translation events initiate at the first ATG codon that is in a particular context. With natural mRNAs, three escape mechanisms – context-dependent leaky scanning, reinitiation and, more controversially, direct internal initiation – are thought to allow access to later ATGs. These mechanisms are reviewed in Kozak(1999). However, recent research (Peri & Pandey, 2001), suggests that translation initiation from downstream ATGs is more common than is generally believed.

3.8.3.1 Translation initiation sites

In this study of the 27 annotated full genes in 22q13.31, putative translation initiation sites were assigned to the first in-frame ATG at the start of the longest ORF (iATG). Alignment of the predicted protein sequence against those of protein orthologues (see chapter V) was possible for 22 of the genes. The alignments supported the choice of reading frame in all cases. Strong conservation was noted at the beginning of the peptide sequences in 16 cases. This provides strong evidence for the choice of initiator codon. In five cases, the sequences at the beginning of the aligned peptides were less conserved, although orthologous proteins were of equivalent lengths. Finally, the alignment of dJ102D24.C22.2 showed that the putatively orthologous mouse

protein extended significantly beyond the chosen translation start site of the human protein. However, no additional evidence can be found to support a longer ORF in dJ102D24.C22.2, so the chosen translation start site was retained.

To examine whether the flanking sequences agreed with the consensus sequence described by Kozak (1987) from an investigation of 640 start sites, the sequences flanking the 27 start sites from –12 (twelve nucleotides upstream from the iATG codon) to +4 were pasted into the Sequence Logo web page (http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi). Figure 3.14 shows the generated Sequence Logo. Kozak's consensus sequence is depicted underneath.

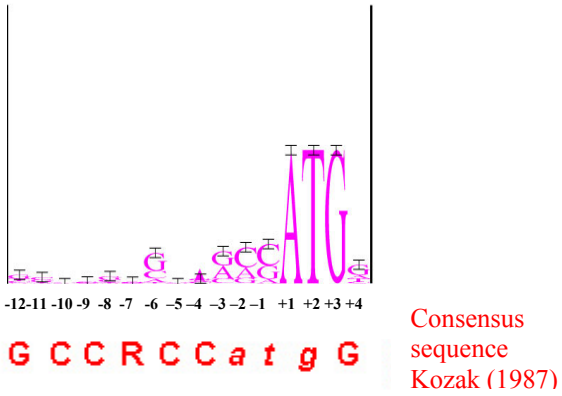


Figure 3.14: Translational start site consensus for 27 full genes on chromosome 22. Kozak's consensus sequence is depicted beneath. Generated from http://www.bio.cam.ac.uk/cgi-bin/seqlogo.cgi (S. Brenner).

Kozak (1987; 1999; 2000) notes that mutations in positions –3 or +4 are most likely to result in leaky scanning and so lead to initiation at a downstream initiator codon. However, flanking sequences lacking only one of the consensus bases at these two positions are still thought to be adequate for translation initiation. The results above show that the consensus sequence is frequently, but not always, found to flank the chosen initiation site. Mismatches are observed at positions –3 and +4 and are commonly found at the remaining positions, particularly in positions –4 and –6.

These findings prompted examination of the 5' UTRs in more detail. The 27 sequences flanking the iATG were categorised according to the degree of mismatch from the motif in the two positions considered optimal; that is, a purine at –3 and a G at +4. If both or one positions were conserved, the site was considered 'strong' or 'adequate' for translation initiation respectively, according to the scanning model of translation initiation. If both positions were mismatched, the site was termed 'weak'. Kozak (2000) suggests that selected initiation sites with the 'weak' characteristic may be inconsistent with the scanning model of initiation.

The results are shown in Figure 3.15.

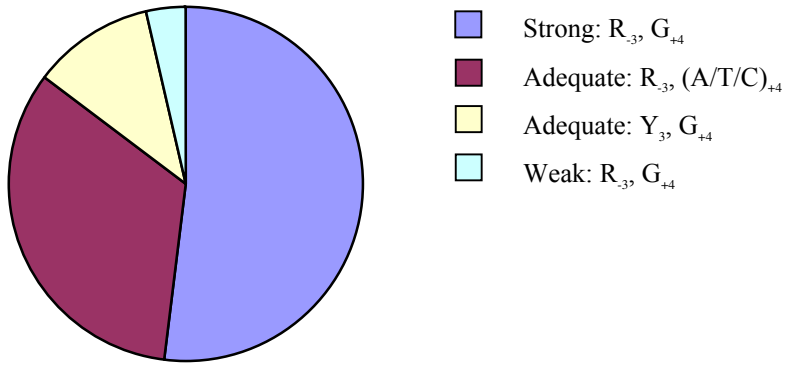


Figure 3.15: Analysis of the sequence contexts surrounding 27 initiator codons from 22q13.31.

Twenty-six sites were at least adequate for translation initiation according to these constraints. However, the gene bK268H5.C22.1 has mismatches at both positions. Inspection of the sequence showed that the first downstream ATG in an at least adequate consensus occurred 120 bp after the original start codon. If this site is the true translation start, the protein produced is shorter by 40 amino acids, or 9.9% of the original predicted protein. Protein features encoded by the original sequence of bK268H5.C22.1 were investigated using Interpro (chapter V). However, no domains or other features were identified within the sequence that might be lost through use of the downstream start site. The available evidence is therefore not sufficient to determine if either (or both) translation start sites are utilised.

3.8.3.2 Upstream ATGs (upATGs)

It has been argued that it is the first ATG with a favourable context that is used for translation initiation. However, under the scanning model, translation initiation may occur at a downstream ATG under the following conditions, which can be inferred from inspection of the mRNA sequence:

1. Leaky scanning. If the downstream ATG is in a stronger context, the upATG may be bypassed by leaky scanning.
2. Reinitiation. If there is an intervening stop codon in frame with the upATG and before the downstream ATG, translation may reinitiate at the downstream ATG.
3. Impaired recognition. Recognition of the upATG by ribosomes may be impaired if the ATG is very near the 5' end (~10 bp).

The 27 transcripts were inspected for the presence of ATGs that were upstream of the putative initiator methionine. Examples were found in nine genes. Additionally, the length of the leader sequence and ORF flanking each ATG was noted so that possible examples of impaired ribosomal recognition, leaky scanning and reinitiation could be identified. The results are shown in table 3.11.

Table 3.11: Possible downstream ATG translation initiation mechanisms.

Gene	No. upATGs		Leaky scanning?	Reinitiation?	Impaired recognition?
cB33B7.C22.1	2	i		•	•
		ii		•	
TTLL1	1		•		
BIK	1		•	•	
C22orf1	1		•	•	
dJ549K18.C22.1	2	i		•	•
		ii		•	
dJ671O14.C22.2	2	i		•	
		ii	•	•	
ARHGAP8	1			•	
NUP50	2	i		•	
		ii			
dJ102D24.C22.2	5	i	•	•	
		ii	•	•	
		iii	•	•	
		iv	•	•	
		v	•	•	

The context, reading frame and leader sequence of ATGs upstream of the annotated translation start site were examined. If the context surrounding the upATG was weaker than the iATG, then leaky scanning was noted as a possible mechanism of downstream initiation. In cases where an intervening stop codon, in-frame with the upATG, was positioned before the iATG, reinitiation may allow downstream translation from the iATG. If the upATG was <10bp from the start of the annotated 5'UTR, impairment of ribosomal recognition may lead to downstream initiation.

The scanning model is consistent with initiation of translation from the annotated downstream ATG (at the start of the longest ORF) in all but one case. This exception is noted in NUP50. The annotated iATG is supported by protein sequence alignments of the orthologous protein in mouse and rat (chapter V) and is in a strong context, with an A at –3 and a G at +4. However, an ATG 190bp upstream is in an equally strong context with G at –3 and +4. The ORF following the upATG is 225 bp (75 amino acids) long, in a different reading frame to the annotated protein, and does not terminate until after the annotated iATG. The 75 amino acid peptide is not similar to any known protein. The mechanism of translation from the downstream iATG is not explained by the scanning model and could be a candidate for internal ribosome entry, or another mechanism of translation initiation.

3.8.4 Polyadenylation signals

The formation of nearly all mature mRNAs in vertebrates involves the cleavage and polyadenylation of the pre-mRNA, 10-30 nucleotides downstream of a conserved hexanucleotide polyadenylation signal. Exceptions include histone transcripts and non-coding RNA genes. The mechanism and regulation of mRNA polyadenylation is reviewed by Colgan & Manley, 1997.

The 3' UTRs of the 27 full genes annotated within the region of interest were examined to see if potential polyadenylation signals could be identified. Putative cleavage sites were recognised by alignment of 3' EST sequences to the mRNA through the graphical BLAST viewer blixem

(Sonnhammer & Durbin, 1994) (figure 3.16). The sequence 10-30 bp upstream of the cleavage/polyadenylation site was then searched for the presence of one or more of the twelve recognised polyadenylation signal sequences (Beaudoing *et al.*, 2000). The results are shown in table 3.12. In cases where more than one polyadenylation hexamer was found, the signal closest to the cleavage site that formed the longest mRNA has been listed.

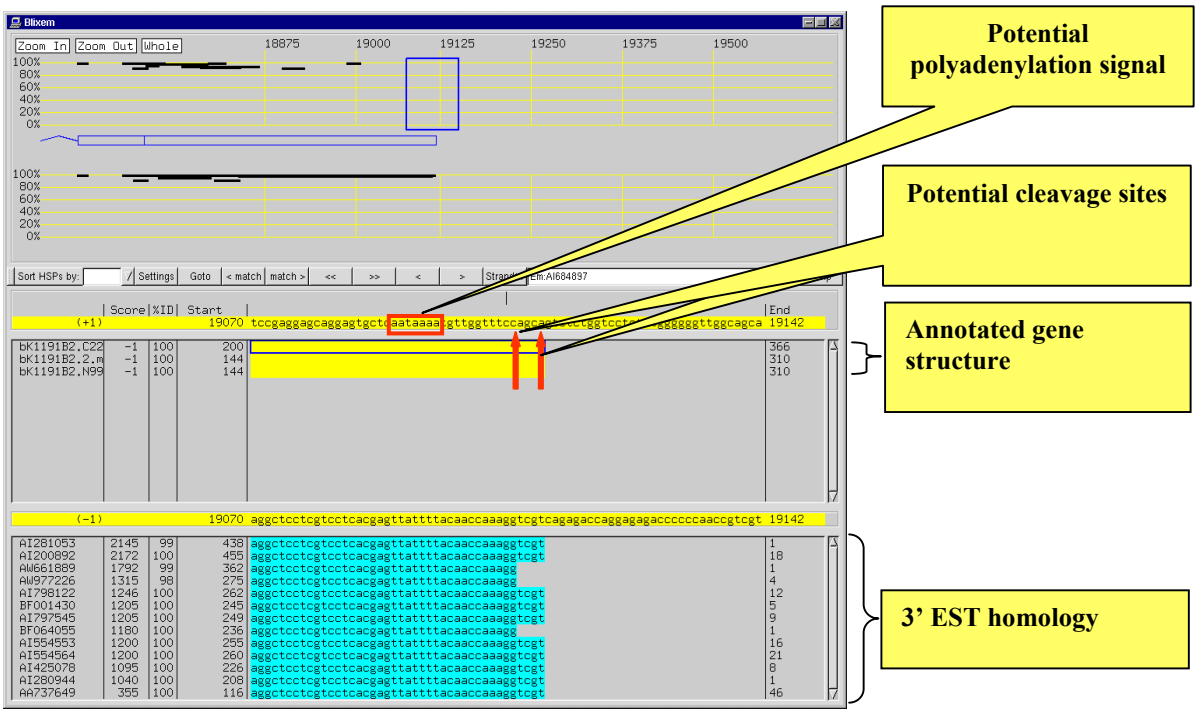


Figure 3.16: An example of Blixem output from ACeDB. EST homology to the 3' end of the BIK gene is shown. Putative polyadenylation signal and cleavage sites are highlighted.

Table 3.12 : The presence/absence of polyadenylation signals and cleavage sites at the 3' end of 27 annotated gene structures from 22q13.31.

Locus name	Putative polyadenylation signal and cleavage site
dJ222E13.C22.1	AATAAAAGGTTCTTGTATTCTCA
dJ222E13.C22.3	AATAAACATTTGTTATTCCTA
DIA1	AGTAAAACTTTGCTAATATTAACCCTTC
cB33B7.C22.1	AATAAAAGTGACCGACTGTCA
ARFGAP1	AATAAACACTTGCAGCAGATGGCA
PACSIN2	AATAAACAGTTGATCTCGTGCATATGGAA
TTLL1	AATAAACGAAGGCACTTCTTTGGAA
BIK	AATAAAATGTTGGTTTCCAGCA
bK1191B2.C22.3	AAAAAGCCCTAAAAATGAGTA
BZRP	AATAAAGTTTTTGACTTCCTTTA
dJ526I14.C22.2	AATAAAGGCCATCTTCTCTTA
C22orf1	No signal found in Em:U84894: 3' end in sequence gap
SULTX3	AATAAAGACATGTTCCCGGC
dJ549K18.C22.1	AATAAAGACACAAGACA
CGI-51	AATAAATGTTAAAGACACACTCCGAG
bK414D7.C22.1	AATAAAAGGGTTTTGCAGTTTGAAAAACTTTAAA
dJ671O14.C22.2	AATAAAAGTATTTCTGGGAGGGA
dJ1033E15.C22.2	ATTAAAGATATTAACCTGGTGTGTGTCA
ARHGAP8	No signal found
dJ127B20.C22.3	ATTAAACTCGATCGATGATTT
NUP50	AGTAAACAAAATCCCA
bK268H5.C22.1	AATACAGATATTATAGCAAAGCAATAATT
UPK3	AATAAAATCTTCTGATGAGTTCTA
bK268H5.C22.4	AATAAAATTTTAACTTCAA
dJ102D24.C22.2	TATAAAGAGTGGCTACCTTAAAGAGTCA
FBLN1	AATAAACAACTTTGTGATCCTCCTG
E46L	AATAAAAGGGAGCCTTGTGAGAATACAGA

Potential polyadenylation and cleavage sites were not found for two loci. Further analysis to extend the 3' end of C22orf1 is difficult as it lies within a sequence gap. None of the 12 potential polyadenylation signals described by Beaudoing *et al*.,(2000) could be found at the 3' end of ARHGAP8. A cluster of EST homologies is found 3' to this gene structure and it may be that these represent the remainder of the 3'UTR of this gene. However, not enough evidence is currently available to confirm this.

3.8.5 Promoter Regions

Polymerase II promoters are generally defined as the region of a few hundred base pairs located directly upstream of the site of initiation of transcription. More distal regions and parts of the 5' UTR may also contain regulatory elements and may be part of the promoter. The exact length of a promoter can often only be defined experimentally. So far, no promoters have been experimentally verified for any genes on human chromosome 22 (Scherf *et al.*, 2001). However,

several *in silico* analyses can be carried out to provide initial information that may be useful in subsequent experimental design. Such analysis can also highlight discrepancies between the positions of the annotated gene 5' ends and the program predictions for further investigation.

3.8.5.1 *In silico* promoter predictions

CpG islands are associated with the promoter of ~50% of all mammalian genes (Antequera & Bird, 1993; Larsen *et al.*, 1992) and often contain multiple binding sites for transcription factors (Somma *et al.*, 1991). They are also found within, and at the 3' end, of some gene structures. They are regions of ~1 kb that differ from the rest of the genome, as the unmethylated CpG dinucleotides occurs at a frequency close to that expected from the levels of individual G and C nucleotides (0.21x0.21) (Bird *et al.*, 1985; Bird, 1986; Matsuo *et al.*, 1993). By contrast, bulk genomic DNA is comparatively G+C-poor (40% on average) and heavily methylated at CpG (see chapter I for more details).

The program CPGFIND (Micklem, unpublished) was used to highlight potential CpG islands. This incorporates the definition proposed by Gardiner-Garden and Frommer (1987) (a CpG island is predicted if %GC> 60%, observed CpG frequency/expected CpG frequency > 0.8 and if there is > 200bp of CpG rich DNA). In total, 46 CpG islands were predicted in the 3.2 Mb of available sequencer (CPGFIND, Micklem unpublished) with a mean length of 1016.4 bp, G+C content of 71.73% and an average Obs/Exp CpG of 0.84. The region has approximately 14.3 islands per Mb. This is higher than the mean figure of 10.5 islands per Mb in the draft genome sequence (Lander *et al.*, 2001) but less than the equivalent figure for the whole of chromosome 22 (16.5 islands per Mb) (Dunham *et al.*, 1999; Lander *et al.*, 2001).

PromoterInspector (Scherf *et al.*, 2000) is a program that predicts eukaryotic polymerase II promoter regions in mammalian genomic sequences. Prediction is based on context specific features, which were identified from mammalian training sequences. Details of the algorithm are

published in Scherf *et al*. (2000). PromoterInspector identified 42 possible promoter regions with an average length of 569 bp within 22q13.31.

Eponine (Down, unpublished) is a program that predicts transcription start sites. Eponine models consist of a set of DNA weight matrices, each with a probability distribution over position relative to an 'anchor point'. The model output is the weighted sum of weight-matrix scores that represents an estimate of the probability of the anchor point being a true transcription start site (Down, personal communication). Eponine identified 128 potential transcription start sites in the region.

3.8.5.2 Correlation of predicted promoter regions with 27 full genes from 22q13.31

A correlation analysis of the predicted promoter regions with the annotated genes starts of the 27 full genes within 22q13.31 was performed (figure 3.17). Unlike CPGFIND and PromoterInspector, Eponine attempts to make strand-specific predictions. Only predictions on the same strand as the annotated gene were included in this investigation.

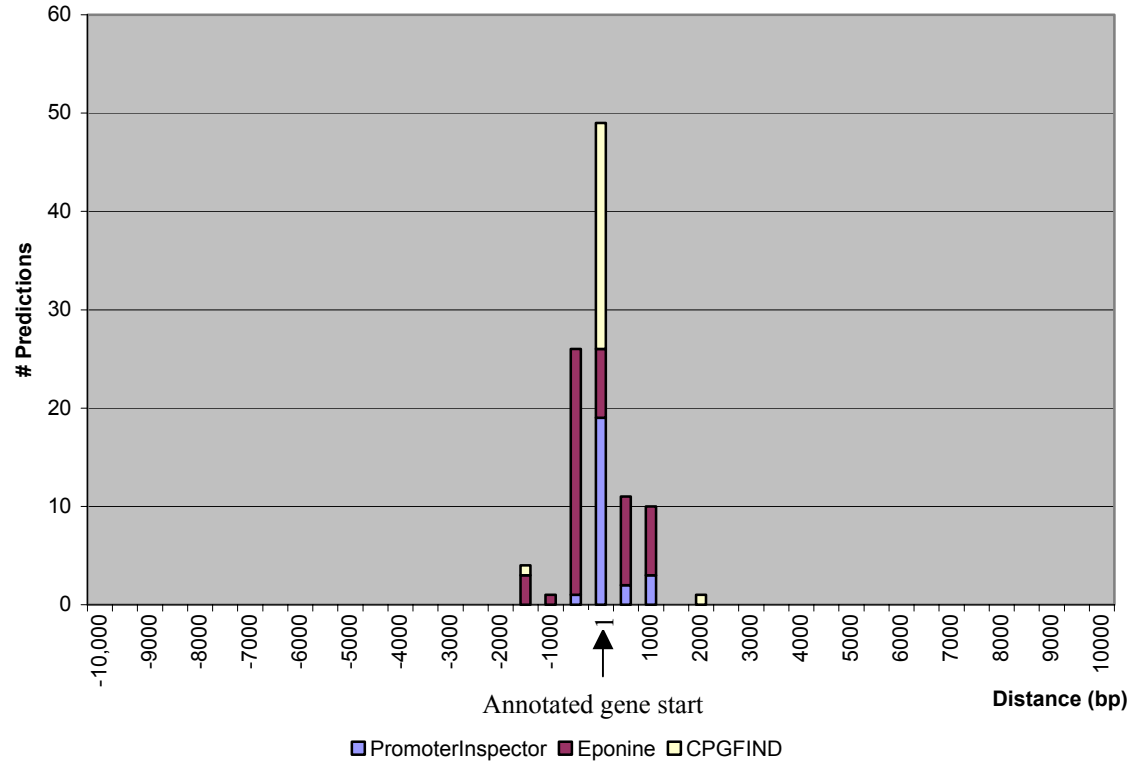


Figure 3.17: Correlation analysis of predicted promoter and transcription start site regions with 27 annotated full gene starts within a 3.4Mb region of chromosome 22. The y-axis indicates the total number of matches found in relative distance to the annotated gene start. Values on the x-axis with a negative sign mark distances to promoter regions, which are located downstream from an annotated gene start. The column at distance value 1 marks the number of promoter regions that overlap an annotated gene start.

Scherf *et al*. (2001) previously denoted PromoterInspector regions as correlated with genes within a region of 2 kb upstream and 0.5kb downstream of the annotated gene starts. From the information provided in figure 3.17, it was decided to maintain this definition for analysis of predicted promoter regions and full genes. (NB. For analysis of the specificity and sensitivity of the promoter prediction packages within this region (see below), this definition was extended to 6kb upstream, to accommodate partial genes structures, (Scherf *et al.*, 2001)).

Figure 3.17 also shows that most Eponine predictions of transcription start site fall within 500 bp upstream (not overlapping) of an annotated start site. Together with the observation that the average 5' UTR length of the full genes in this region was smaller than that of a set of 1,804 RefSeq genes (section 3.8.1), this may indicate that some of the gene annotations analysed here

are foreshortened at the 5' end and are therefore not full-length. However, Northern blot evidence where available (section 3.5.1.2), supports the currently annotated transcript lengths and there is no expressed sequence evidence currently available that extends the 5' UTR regions of these genes.

The fraction of the 27 full genes that correlated with each type of promoter prediction was calculated. Figure 3.18 shows that 89% of the genes correlate with a predicted CpG island, 85% correlate with PromoterInspector predictions and 55% with Eponine predictions. The diagram also shows that 85% of gene structures are correlated with more than one prediction. Just over half (51%) are correlated with all three.

This diagram also highlights two gene structures that are not correlated with promoter predictions. This could indicate that PromoterInspector and Eponine are less accurate when defining the promoters or transcription start sites of genes that are not associated with CpG islands. The sequences 5' of the transcription start sites of dJ671O14.C22.2 and UPK3 were therefore examined in more detail.

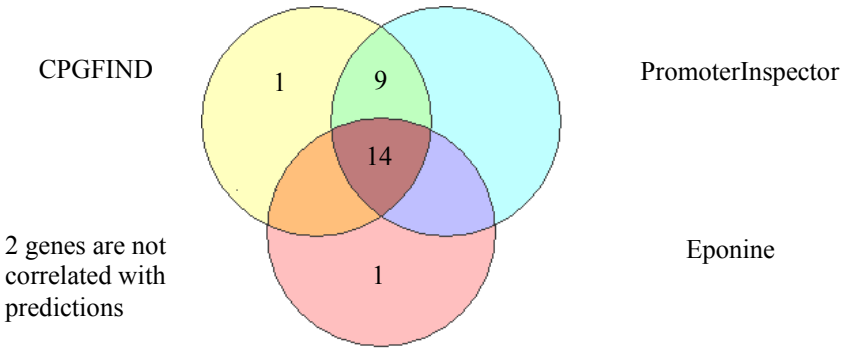


Figure 3.18: Venn diagram shows the number of full gene structures and their correlation with different kinds of promoter prediction algorithms

3.8.5.3 Full gene structures not correlated with a promoter prediction

Promoter Inspector and Eponine identify potential promoter regions independently of the occurrence of specific transcription factor binding site elements such as TATA boxes. However,

many promoters that occur within CpG poor regions contain such elements. TATA boxes are found ~30 bp upstream of the transcription start site. The consensus sequence is $T_{82}A_{97}T_{93}A_{85}(A_{63}/T_{37})A_{83}(A_{50}/T_{37})$ (Lewin, 1994).

One hundred base pairs of sequence upstream of the annotated transcription start site for both dJ671O14.C22.2 and UPK3 was examined for the presence of a potential TATA box, but none were found. It was noted, however, that the 250 bp sequence surrounding the transcription start site of one of these genes, UPK3, was CpG rich: the %GC of 77% and observed/expected GC of 0.77 is only just below the criteria for CpG islands prediction. It may be therefore that the 5' end of UPK3 lies in an unpredicted CpG island.

3.8.5.4 Correlation of predicted promoter regions with 38 protein coding genes

The distribution of predicted promoter regions across the whole region of interest in 22q13.31 was then examined, and the correlation with both full and partial protein-coding gene structures was analysed. The limits of correlation were extended to six kilobases upstream and 500 bp downstream of the annotated 5' end of the gene, in order to accommodate partial gene structures (Scherf *et al.*, 2001). The specificity of each data set (the proportion of predicted promoter regions that correlated with annotated 5' end) and the sensitivity (the proportion of annotated gene 5' ends that correlated with predicted promoter regions) were calculated (chapter II). Table 3.13 summarises these results.

Table 3.13: Correlation of predicted promoter regions and CpG islands with gene annotation on a 3.4 Mb region of chromosome 22.

	A) CPGFIND		B) PromoterInspector		C) Eponine	
	Sn	Sp	Sn	Sp	Sn	Sp
Gene	0.74	0.59	0.71	0.67	0.45	0.38

The correlation boundary was set at 6 kb upstream and 0.5 kb downstream of an annotated transcription start site. Sn (Sensitivity) = No of genes that correlate with prediction/total no. of genes (38) Sp (Specificity) = No of predictions that correlate with a gene/total no. of predictions. Total number of predictions: CPGFIND (46); PromoterInspector(42); Eponine(128). Total number of protein coding genes = 38.

Twenty-eight (74%) of the protein coding genes in this region are correlated with a CpG island. It was noted that all of these islands overlap the annotated transcription start site. Promoter Inspector shows the highest individual specificity with respect to gene correlation with 67% of predictions correlated with annotated gene 5' ends, but Eponine performs less well in terms of both sensitivity and specificity. It was noted, however, that Eponine predictions clustered on both strands around the annotated transcription start sites of several genes, suggesting that Eponine correlation may be greater if strand specificity were ignored.

In total, 113 individual predictions are not currently associated with annotated genes (19 CpG island, 14 PromoterInspector and 80 Eponine predictions). In all, twelve possible promoter 'regions' were identified which had overlapping predictions not associated with gene 5' ends. These regions were examined more closely to determine if these overlapping predictions were likely to indicate the presence of nearby genes. Three were found to lie within introns of annotated genes and three lay within repeat sequence. Six remaining possible promoter regions were identified and all three programs highlighted four of these. One of these regions lies within 20kb upstream of the locus bK941F9.C22.6, which currently has no associated promoter predictions. It may be that further investigation will extend this gene structure and show that this potential promoter is associated with this gene. The three remaining putative promoter regions may be false positives, or may also be associated with existing partial gene structures within 22q13.31. These results could also indicate the presence of regulatory regions of genes that have yet to be identified.

3.8.6 Alternative Splices

Several alternatively spliced exons were identified through the transcript mapping work described in section 3.4 and these results are summarised in table 3.14. Further indications of alternative splicing are provided by the Northern blot analysis described above. However, it may

be that some of the differently sized transcripts identified on the blots derive from paralogous genes (section 3.8.7), rather than from the alternative splicing of a single locus.

Table 3.14: the number of potential alternative splices determined from the transcript mapping of 38 protein-coding genes from 22q13.31.

No. of transcript variants	1	2	3	4	5	6
Number of sequence verified transcripts /gene locus	29	6	1	2	0	0
% sequence verified transcripts /gene locus	76.3	15.8	2.6	5.4	0	0

These results show that 23.8% of gene loci have at least one sequence verified alternative splice form. All of the sequence verified alternative splices found in these genes affect the coding sequence, rather than altering the 5' or 3' UTR. This result could be affected by incomplete 5' UTR sequences, which may be present in the resources used.

The value of 23.8% is probably lower than the real percentage of alternatively spliced transcripts, as a full investigation into identification of alternative splicing in this region has not yet been undertaken. This level of alternative splicing is supported by evidence from three studies (Brett *et al.*, 2000; Mironov *et al.*, 1999; Zhuo *et al.*, 2001), which indicate that, on average, one-third of genes have EST evidence of alternative splicing of any sort. However, these studies may also have underestimated the prevalence of alternative splicing, because they examine EST alignments covering only a portion of a gene.

Investigation of alternative splicing by Lander (2001), using reconstructed mRNA transcripts covering the entire coding regions of genes on chromosome 22, puts this figure much higher at nearly 60%. The true extent of alternative splicing in the genome was expected to be even greater as only a subset of transcripts were sampled in this study.

The percentage of potential alternatively spliced loci detected during this project rises to 74% if Northern blot results are taken into account. Although this figure may more closely represent the true extent of alternative splicing of these genes, the Northern results may be misleading as the

probes used may have hybridised to paralogous genes elsewhere in the genome, and the blots may fail to resolve similarly sized transcripts.

3.8.7 Paralogues

The availability of genomic sequence has already provided insights into genome evolution. Analysis of the duplication landscape of chromosome 22 (Lander *et al.*, 2001) showed that the region of interest contained no inter- or intrachromosomal duplications of more than 90% nucleotide identity and greater than 1kb long when compared to the draft genome sequence. It was decided to extend this investigation to examine paralogy at the exon level, by using a less stringent TBLASTN search to detect shorter stretches of similarity at the amino acid level.

The amino acid translations from the longest ORF from each of the 27 full gene structures were extracted. The sequences were then used in a TBLASTN experiment against the working draft of the human genome, using the NCBI human genome BLAST service (http://www.ncbi.nlm.nih.gov/BLAST). The SwissProt, TrEMBL or NCBI annotation project identities of human peptide sequences that matched along the full length of the chromosome 22 peptides were extracted. The results are listed in table 3.15. Figure 3.19 shows in more detail the approximate chromosomal localisation of the potential paralogues.

These results may still be incomplete as human genome sequencing and annotation is an ongoing project. Apparent duplications may also arise from a failure to merge sequence contigs from overlapping clones in the draft genome assembly.

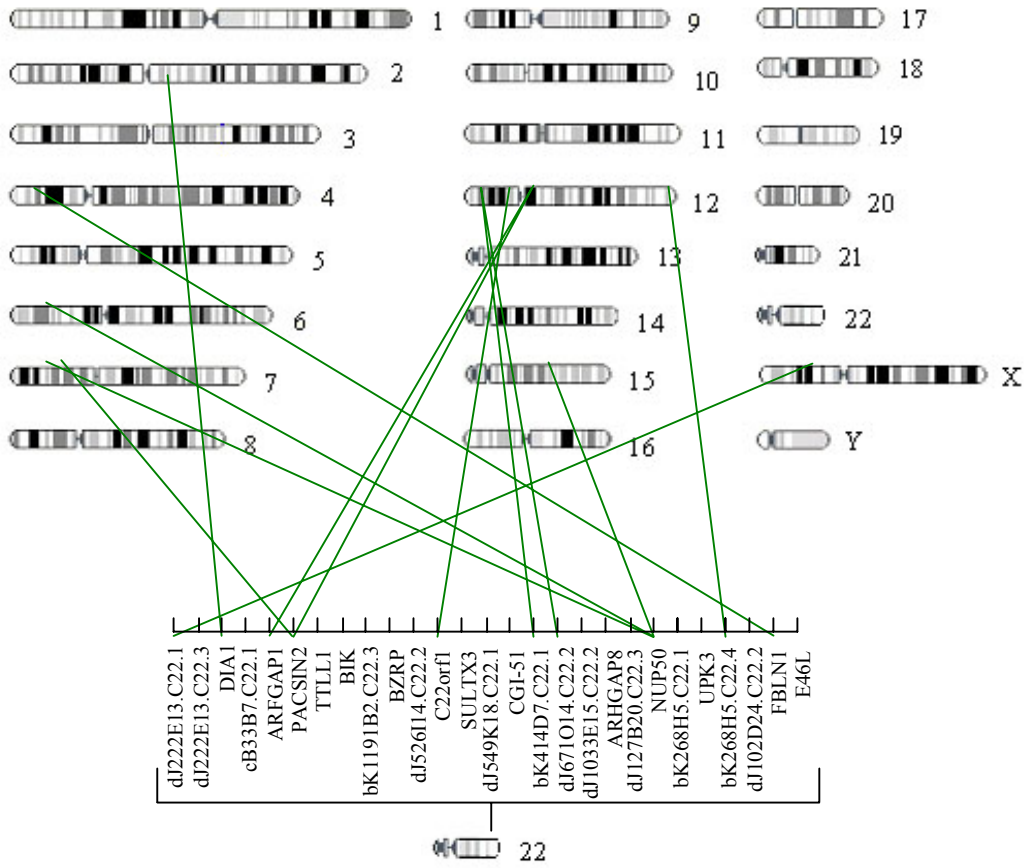


Figure 3.19: Approximate positions of genes putatively paralogous to full genes on 22q13.31. Figure was generated using the Ensembl website (http://www.ensembl.org).

Table 3.15: Genes putatively paralogous to full genes on 22q13.31

Chromosome 22 Locus	Paralogous Locus	Accession number	Chromosome	% identity of amino acid sequences	Result supported by previous publication?
dJ222E13.C22.1			22	99%	
DIA1		O95329 2[2]	1	62%	
ARFGAP1		BAB55144 [2]	11	49%	
PACSIN2	PACSIN1	Q9BY11[2]	6	53%	(Ritter *et al.*, 1999)
	PACSIN3	Q9H331[2], Q9EQP9[2], Q99JB8[2]	11	57%	(Ritter *et al.*, 1999)
C22orf1 (239AB)	239FB	239F_HUMAN 1[1]	11	81%	(Schwartz & Ota, 1997)
bK414D7.C22.1	α-parvin	Q9NVD7 2[2]	11	75%	(Olski *et al.*, 2001)
dJ671O14.C22.2	α-parvin	Q9NVD7 2[2]	11	42%	
NUP50		XP_018531[3]	6	85%	(Trichet *et al.*, 1999)
		XP_017832[3]	5	92%	
		XP_010041[3]	14	70%	
bK268H5.C22.4		Q9H7B0 2[2]	11	48%	
FBLN1	FBLN2	FBL2_HUMAN 1[1]	3	48%	(Zhang *et al.*, 2000)

[1] SwissProt, [2] TrEMBL, [3] NCBI Annotation Project accession number (predicted protein)
Locus name, accession number, chromosomal position and percentage identity to the 22q13.31 gene are shown. Additional evidence of paralogy is provided in the listed references.

Genes from chromosome 22q13.31 were found to have paralogs on chromosomes 1, 3, 5, 6, 11, and 14. Partial gene order from chromosome 22 did not appear to be replicated in cases where more than one paralogue existed on a particular chromosome (6 and 11) and genomic distances between these paralogous genes were at least several megabases. The paralogous regions may be considered to show evidence of ancient intrachromosomal duplications as they are characterised by similarities in the coding regions only. The experiment also highlighted a region of chromosome 22 that appeared to have undergone an interchromosomal duplication. This was examined in more detail.

Comparison of the two regions of chromosome 22, using the 22ace database, identified a direct repeat, occupying ~150 kb of sequence and shown schematically in figure 3.20. The region contained two pairs of paralogous gene structures, bK126B4.C22.2 and dJ222E13.C22.1, and bK126B4.C22.3 and dJ222E13.C22.2, which were duplicated in the region of interest in the same orientation. No other paralogs of these genes were found on any other chromosomes during the TBLASTN experiment above.

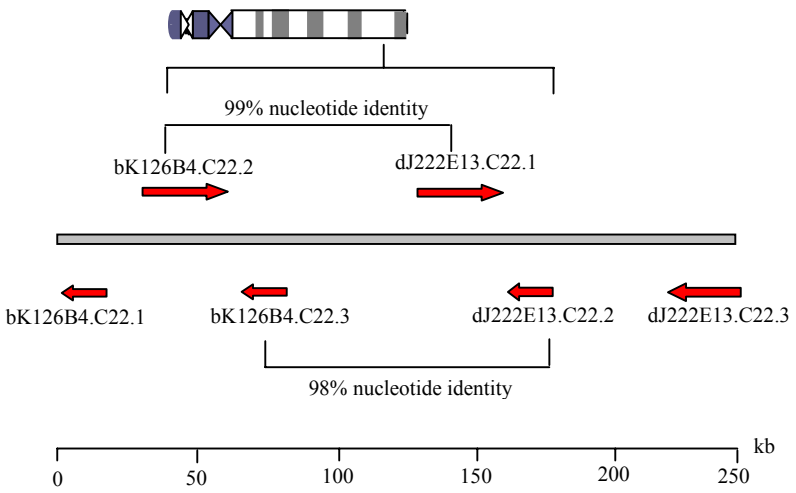


Figure 3.20: Schematic showing a region of interchromosomal duplication on chromosome 22

To investigate this further, the genomic DNA from the region between bK126B4.C22.1 and dJ222E13.C22.3, enclosing the putatively duplicated region, was compared against itself using

the Dotter program (Sonnhammer & Durbin, 1995) (figure 3.21). Dotter is a graphical dotplot program allowing detailed comparison of two sequences. Every residue in the sequence is compared to every other residue in the sequence. Regions of high homology are shown by a row of high scores, which run diagonally across the dot matrix.

This analysis revealed that the two pairs of genes are conserved in both exon and intron sequences, indicating that the duplication could be a fairly recent evolutionary event. Three further groups of homology are noted from repeat regions 5' to the duplicated gene pairs. These regions were found to contain a mixture of repetitive and unique sequences. The remaining sequence in the duplicated is less well conserved, perhaps arising after the duplication event, or diverging more rapidly than the conserved sequences.

There are some important differences between the duplicated gene structures. There is a large insertion or deletion of approximately seven kilobases, highlighted by the blue box in figure 3.21. Exons VIII, I, X and XI of dJ222E13.C22.1 are encoded within this region. Interestingly, the annotated ORF of bK126B4.C22.2 is much shorter than that of its paralogue, dJ222E13.C22.1 (figure 3.22) and the protein sequences diverge after exon VII. Potentially, the coding sequence of bK126B4.C22.2 was truncated by a deletion of this region of genomic sequence and is thus a pseudogene derived from duplication of the ancestral gene.

The nucleotide sequences of dJ222E13.C22.2 and bK126B4.C22.3 were also aligned and a difference of a 10bp deletion or insertion was seen (indicated by a red box in figure 3.23). Interestingly, this difference disrupts the open reading frame of the dJ222E13.C22.2 and thus truncates the protein sequence. dJ222E13.C22.2 could therefore be a pseudogene, which arose after the tandem duplication of the ancestral gene. A second downstream insertion or deletion of 8 bp, that also alters the ORF, is highlighted by the blue box in figure 3.23.

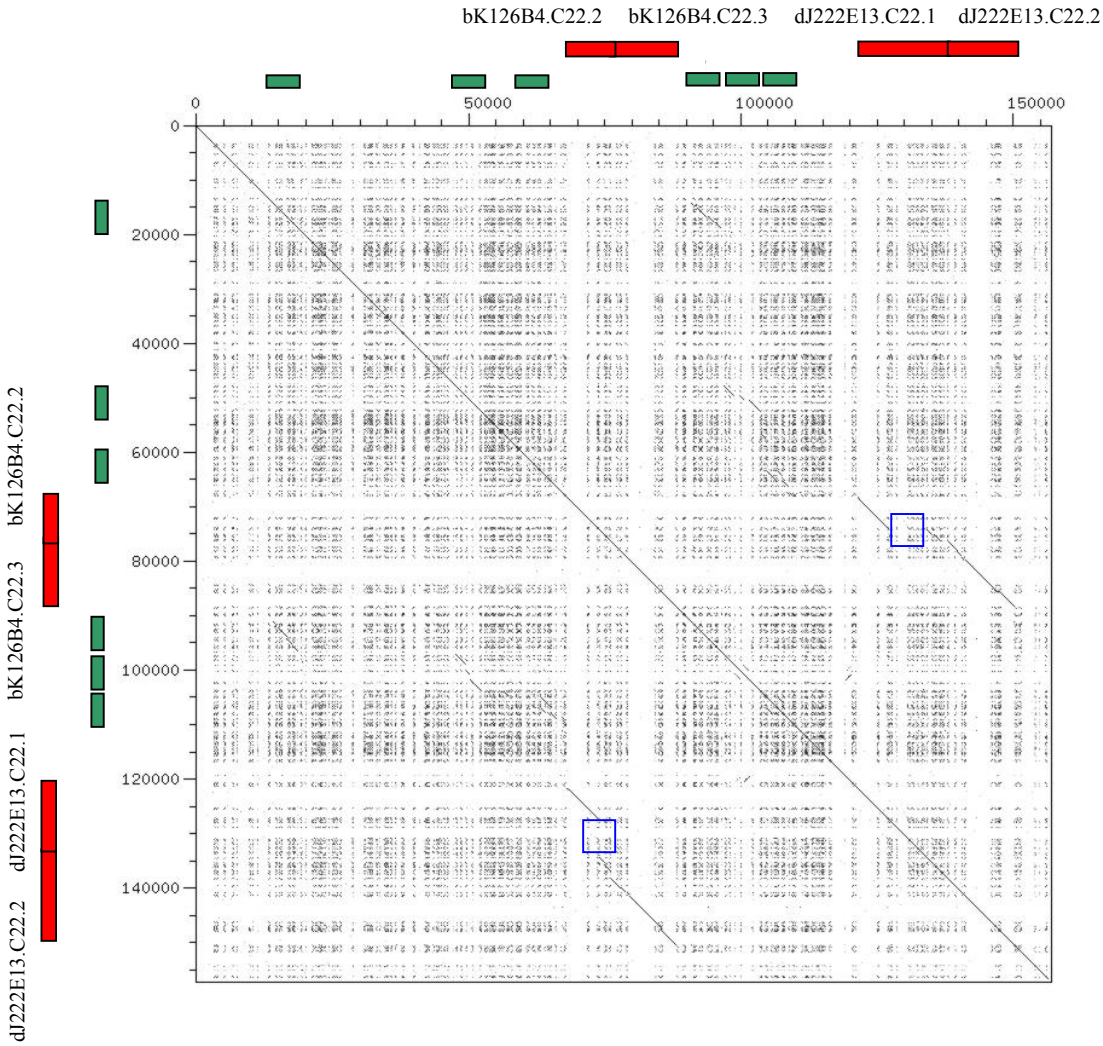


Figure 3.21: Annotated dot plot from identifying an intrachromosomal duplication within chromosome 22. 156366 bp of genomic sequence between genes bK126B4.C22.1 and dJ222E13.C22.3, containing a putatively duplicated region, is plotted against itself. Red boxes along the axes indicate gene structures within the sequence. Further evidence of sequence conservation is also noted in three areas (green boxes). The blue boxes indicate the position of an insertion/deletion of ~7000 bp. The plot was generated using Dotter (Sonnhammer & Durbin, 1995).

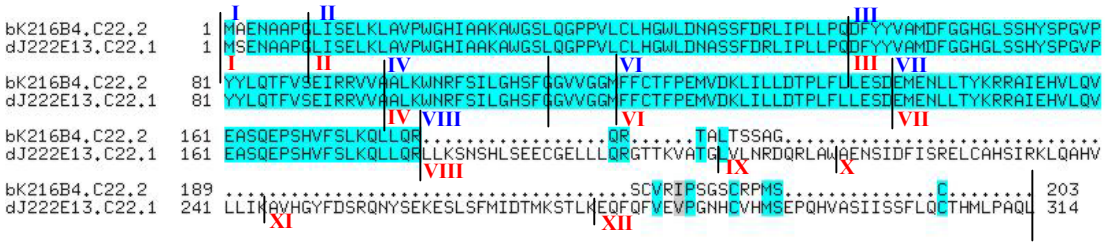


Figure 3.22: Alignment of the amino acid sequences of bK126B4.C22.2 and dJ222E13.C22.1. Exon numbers are marked in blue (bK126B4.C22.2) or red (dJ222E13.C22.1). The alignment was created using clustalw(Thompson *et al.*, 1994) and visualised using belvu (Sonnhammer, unpublished).

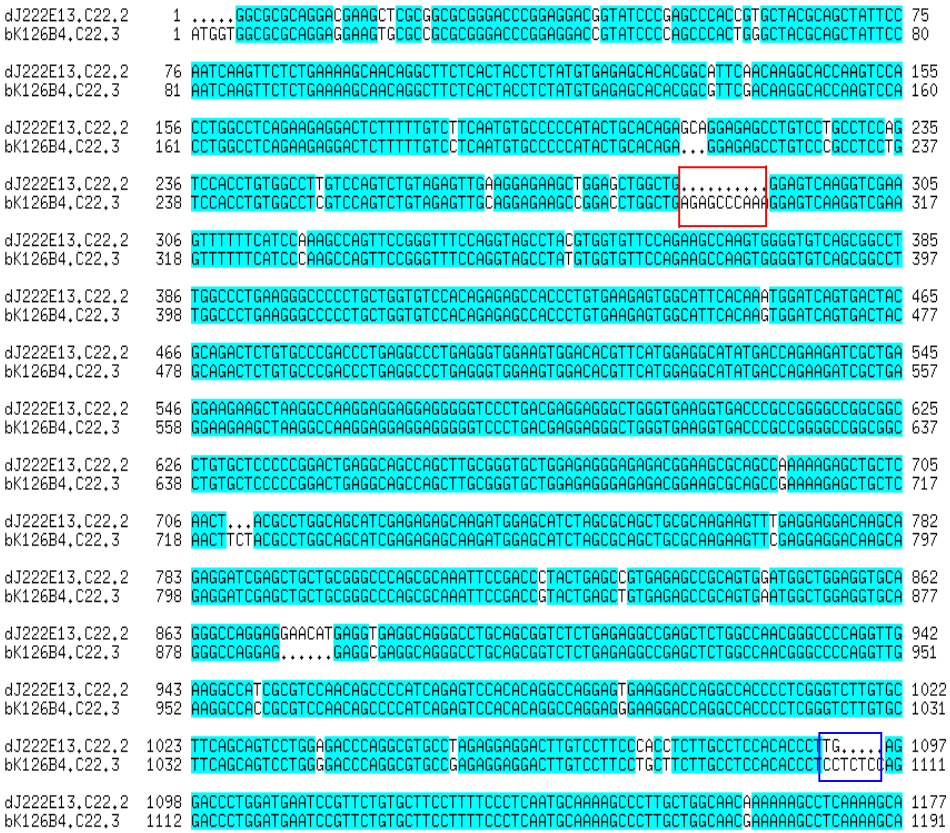


Figure 3.23: Alignment of the nucleotide sequences of bK126B4.C22.3 and dJ222E13.C22.2. A 10 bp insertion/deletion discussed in the text is marked in red and an 8 bp insertion/deletion is marked in blue. The alignment was created using clustalw (Thompson *et al.*, 1994) and visualised using belvu (Sonnhammer, unpublished).

Achaz (2001) describe a study of intrachromosomal duplications of nucleotide sequences in two complete genomes and four partial ones, including *Homo sapiens*. They propose that

intrachromosomal repeats are mostly created in tandem by recombination between sister chromatids or by replication slippage and are turned into distant repeats by later chromosomal rearrangements. The features of this duplicated sequence resemble those most commonly found in the previous study: a direct repeat with the two copies close together with a physical distance, the 'spacer', between them. In this example, the spacer is defined as the 34 kb of sequence separating the genes bK126B4.C22.3 and dJ222E13.C22.1.

To investigate if the vestiges of tandem rearrangement could be determined in the chromosome sequence, NCBI whole genome BLAST server was used to look for paralogs of the spacer within the chromosome 22 sequence. The criteria listed by (Achaz *et al.*) was used to determine matches to spacer sequence paralogs: however, no matches to chromosome 22 or any other genome sequences were found that were at least 80% of the spacer length and identical by more than 80%. This implies that, if the duplication did arise by replication slippage or unequal recombination between sister chromatids, the flanking sequences may have diverged beyond this level of recognition.

3.9 Correlation of expression evidence with annotated gene features

Several different types of evidence have contributed to the generation of a transcript map of 22q13.31 (see appendix 2). Evidence provided by EST sequences has included homologies to the EST database dbEST (Boguski *et al.*, 1993), and a set of EST sequences generated by the biotechnology company Incyte, selected from BLAST matches at 85% nucleotide identity to the genomic sequence of chromosome 22, (J. Seilhamer, Incyte, personal communication). cDNA sequence evidence includes those generated as a result of this project, plus cDNAs identified from the Mammalian Gene Collection (MGC) (Strausberg *et al.*, 1999) and from

vertebrate cDNA sequences submitted to EMBL (Baker *et al.*, 2000). Additionally, protein sequences from the TREMBL and SwissProt databases (Bairoch & Apweiler, 2000) have been used. Chromosome 22-specific exon trap sequences (Trofatter *et al.*, 1995), and a range of exon and gene prediction programs, including Genscan (Burge & Karlin, 1997), provided further evidence. Finally, a database of predicted exon sequences that have been tested for expression by microarray hybridisation was also available (Richard Glynne, Eosbiotech, personal communication).

A region of 22q13.31 sequence that aligns to any piece of such evidence could potentially form part, or all, of a gene. Therefore, it is of interest to investigate the correlation of these data with the annotated gene structures in order to establish the specificity (the proportion of putative coding nucleotides that are actually coding) and sensitivity (proportion of actual coding nucleotides that were identified as putative coding nucleotides) of each method (see chapter II). Such information will be useful in the generation of future transcript maps, by identifying lines of evidence that may lead to more efficient annotation.

Some genes in the transcript map of 22q13.31 remain partial. However, the region has been subjected to extensive experimental analysis. Many potentially coding regions have been screened against cDNA libraries and the negative results produced showed that they were less likely to encode true genes. It is therefore proposed that an investigation of correlation between annotated genes structures and a range of sequence evidence is meaningful and will allow comparison with similar previous studies of *ab initio* gene prediction accuracy (Bruskewich and Hubbard, unpublished; Guigo *et al*., 2000).

3.9.1 Calculation of specificity and sensitivity

The perl script MethComp (D. Beare, unpublished) was used to compare the different methods used for gene identification/annotation against:

(A) The set of 39 annotated 'true' genes within 22q13.31;

(B) The set of 17 annotated pseudogenes within 22q13.31.

Specificity and sensitivity calculations were perfomed at the nucleotide level for all method types. In addition, the fraction of exon hits (the number of reference exons hit/total number of reference exons) and gene hits (the number of reference genes hit/total number of reference genes) were also calculated. In all cases, multiple hits were counted as one hit. These results are shown in table 3.16.a and .b. A plot of the specificity and sensitivity of each type of evidence at the nucleotide level is shown in figure 3.24. Further details of this analysis can be found in chapter II.

Table 3.16: Analysis of the correlation of the evidence types used to annotate genes against:

A: 39 annotated true genes in 22q13.31.

Evidence type	Method	Alignment method	Nucleotide			Exon	Gene
			Total coverage	Sp	Sn		
EST	dbEST[1]	BLASTN	0.060	0.37	0.74	0.81	1.00
EST	Incyte[2]	BLASTN	0.100	0.23	0.79	0.87	0.90
cDNA	ad_hoc[3]	BLASTN	0.005	0.45	0.32	0.69	0.65
cDNA	VERTRNA[4]	BLASTN	0.029	0.72	0.72	0.75	0.82
cDNA	human_MGC[5]	BLASTN	0.003	0.62	0.06	0.08	0.12
Protein	Blastx[6]	BLASTX	0.088	0.13	0.39	0.68	0.92
Exon prediction	Grail1.3[7]	Grail1.3	0.043	0.13	0.19	0.37	0.68
Exon prediction	Xpound[8]	Xpound	0.003	0.43	0.04	0.08	0.17
Exon prediction	fexh[9]	fexh	0.037	0.13	0.16	0.32	0.48
Exon prediction	eos[10]	Genscan	0.026	0.45	0.40	0.75	0.85
Exon prediction	exon trap[11]	BLASTN	0.001	0.58	0.02	0.03	0.31
Gene prediction	Genscan[12]	Genscan	0.028	0.40	0.38	0.58	0.90
Gene prediction	Fgenesh[13]	Fgenesh	0.019	0.49	0.30	0.57	0.90

The test region (22q13.31) contained 3,365,293 bp of genomic sequence. The total number of nucleotides contained within the 39 annotated genes structures is 91,249 bp. The total number of reference exons is 400. For more details, see chapter II.

B: 17 annotated pseudogenes in 22q13.31.

Evidence type	Method	Alignment method	Nucleotide			Exon	Pseudogene
			Total coverage	Sp	Sn		
EST	dbEST[1]	BLASTN	0.060	0.05	0.75	0.86	0.88
EST	Incyte[2]	BLASTN	0.100	0.02	0.41	0.55	0.58
cDNA	ad_hoc[3]	BLASTN	0.005	0.01	0.25	0.37	0.29
cDNA	VERTRNA[4]	BLASTN	0.029	0.09	0.63	0.82	0.88
cDNA	human_MGC[5]	BLASTN	0.003	0.34	0.24	0.24	0.41
Protein	Blastx[6]	BLASTX	0.088	0.02	0.45	0.58	0.76
Exon prediction	Grail1.3[7]	Grail1.3	0.043	0.01	0.13	0.34	0.47
Exon prediction	Xpound[8]	Xpound	0.003	0.00	0.00	0.00	0.00
Exon prediction	fexh[9]	fexh	0.037	0.01	0.06	0.14	0.24
Exon prediction	eos[10]	Genscan	0.026	0.03	0.20	0.45	0.47
Exon prediction	exon trap[11]	BLASTN	0.001	0.00	0.00	0.00	0.00
Gene prediction	Genscan[12]	Genscan	0.028	0.03	0.20	0.45	0.47
Gene prediction	Fgenesh[13]	Fgenesh	0.019	0.02	0.21	0.45	0.41

The test region (22q13.31) contained 3,365,293 bp of genomic sequence. The total number of nucleotides contained within the 17 annotated pseudogenes is 6090 bp. The total number of reference exons is 29. For more details, see chapter II.

[1.] dbEST: dbEST EST database (Boguski *et al.*, 1993); [2.] Incyte: EST database (J. Seilhamer, Incyte, personal communication); [3.] ad_hoc: cDNA sequences generated as a result of this project; [4.] VERTRNA: vertebrate cDNA sequences, EMBL database (Baker *et al.*, 2000); [5.] human_MGC: full-length cDNA sequences (Strausberg *et al.*, 1999); [6.] Blastx: TrEMBL and SwissProt protein sequence databases (Bairoch & Apweiler, 2000); [7.] Grail1.3: (Uberbacher & Mural, 1991); [8.] Xpound: (Kamb *et al.*, 1995); [9.] fexh:(Solovyev & Salamov, 1997); [10.]eos: Genscan predicted exons tested for expression by microarray hybridisation (R. Glynne, personal communication); [11.] exon trap: chromosome 22 specific exon trap sequences (Trofatter *et al.*, 1995); [12.] Genscan: (Burge & Karlin, 1997); [13.] Fgenesh: (Solovyev *et al.*, 1994).

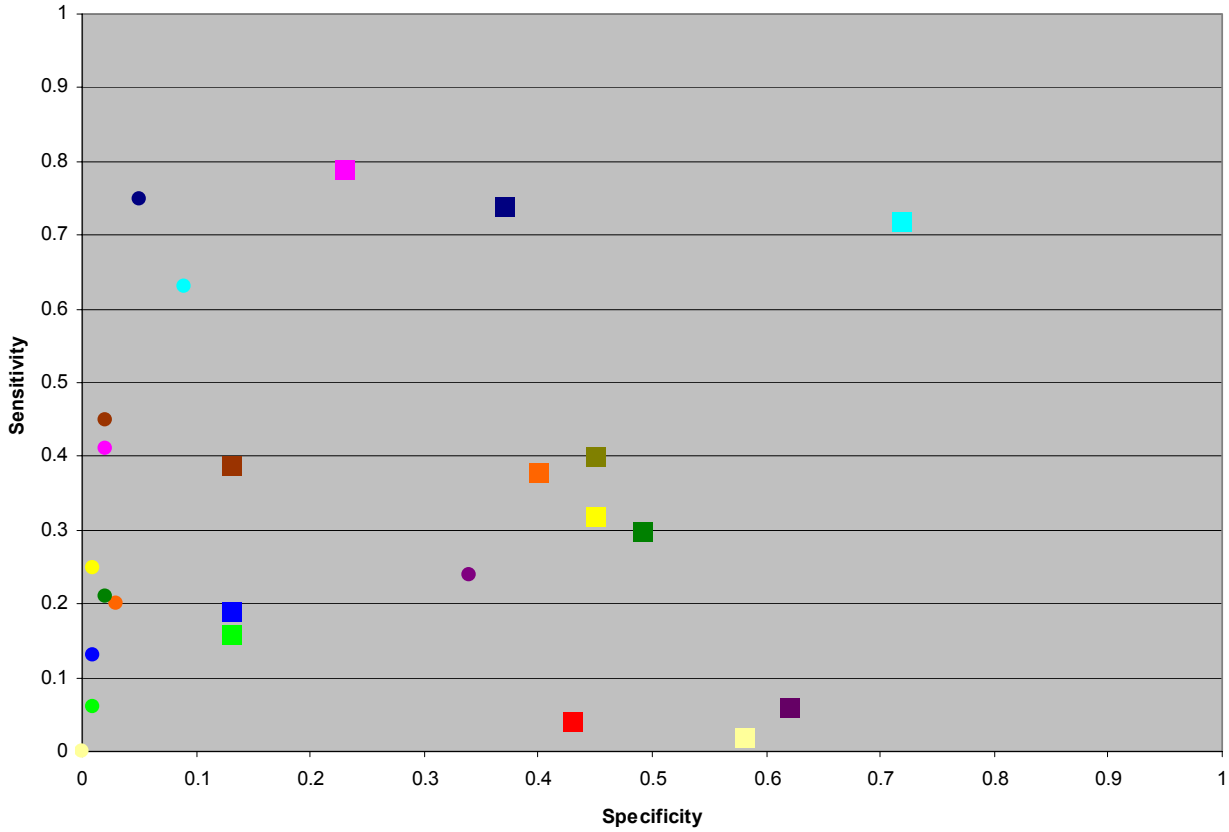


Figure 3.24: Specificity and sensitivity of sequence evidence alignment with the 22q13.31 transcript map. Sensitivity and specificity shown are computed at the nucleotide level.

dbEST[1]
Incyte[2]
ad_hoc[3]
VERTRNA[4]
human_MGC[5]
Blastx[6]
Grail1.3[7]
Xpound[8]
fexh[9]
eos[10]
exon trap[11]
Genscan[12]
Fgenesh[13]

□ = correlation with 39 annotated genes within 22q13.31

○ = correlation with 17 annotated pseudogenes within 22q13.31

Descriptions and references of the sequence evidence are given in the legend to table 3.24.

As expected, the specificity of the correlations with genes structures is much greater than that demonstrated with pseudogenes. The graph shows that most pseudogenes correlate with matches to entries in the dbEST, VERTRNA databases, and to BLASTX matches to known

proteins (Blastx). Most of these pseudogenes were annotated from these sources by the Sanger Institute gene annotation group.

From the analysis of correlation with 39 gene structures, it can be seen that the highest sensitivity is achieved by BLASTN comparison to the VERTRNA mRNA sequences from the EMBL database (Baker *et al.*, 2000). This is not surprising, however, as nearly all of the full and partial gene structures are referenced in this database. The EST databases dbEST (Boguski *et al.*, 1993) and Incyte (J. Seilhamer, Incyte, personal communication) also provide highly sensitive results when aligned by a BLASTN experiment against the annotated sequence of 22q13.31. Similarly, mRNA sequences from the mammalian gene collection (Strausberg *et al.*, 1999) provide the most specific evidence for transcript mapping.
The data derived from the set of exon trap sequences (Trofatter *et al.*, 1995) shows high specificity, but low sensitivity in this comparison against the annotated gene feature set. The table also includes equivalent information for a number of prediction programs. Genscan (Burge & Karlin, 1997) and Fgenesh (Solovyev *et al.*, 1994) achieve the best results. However, this analysis includes UTR and pseudogene sequences within the reference set, which may skew the results against these programs, as they are designed to predict only coding sequences. A more complete investigation of Genscan and Fgenesh accuracy is shown below.

3.9.2 Further analysis of Genscan and Fgenesh predictions

The gene prediction programs Genscan (Burge & Karlin, 1997) and Fgenesh (Solovyev *et al.*, 1994) were taken as a special case, in order to allow comparison between this and previous studies (Bruskewich and Hubbard, unpublished; Guigo *et al.*, 2000). Unlike sequence database

evidence, these data involve *predictions* of gene structures and so specificity and sensitivity at the exon and gene level can also be meaningfully calculated. To compute these measures at exon level, it is assumed that an exon has been predicted correctly only when both its boundaries have been predicted correctly. Annotated pseudogenes are not included in the calculation. Non-coding exons were also excluded, as Genscan and Fgenesh predict coding sequences only. The programs Genscan and Fgenesh were used to generate gene predictions across the linked clone sequences of chromosome 22. The number of predicted gene features within 22q13.31 is shown in table 3.17.

Table 3.17: The number of nucleotides, exons and structures predicted by Genscan and Fgenesh within the region of interest from linked clone sequences.

	Prediction		
Structure Set	# Nucleotides	# Exons	# Gene structures
Genscan	94026	657	83
Fgenesh	63196	449	77
True Genes	44312	334	38

The equivalent figures from the True Genes set of experimentally annotated structures are included for comparison.

The gene predictions were compared at both nucleotide and exon levels against the set of protein coding exons. Sensitivity and specificity calculations were carried out as above. In addition, the fraction of unpredicted missing exons and genes (false negatives) (ME and MG) and wrongly predicted exons and genes (non-overlapping with true exons or genes) (WE and WG) were recorded in table 3.18 (see also chapter II). A plot of specificity and sensitivity values, this time at the exon level, for each data set is shown in figure 3.25.

Table 3.18: Analysis of the correlation Genscan and Fgenesh predictions with 38 currently annotated protein-coding genes 22q13.31.

	Nucleotide		Exon				Gene			
Set	Sp	Sn	Sp	Sn	ME	WE	Sp	Sn	MG	WG
Genscan[1]	0.40	0.85	0.37	0.74	0.18	0.58	0.06	0.13	0.14	0.43
Fgenesh[2]	0.53	0.75	0.50	0.67	0.25	0.44	0.04	0.08	0.14	0.42
Genscan			0.38	0.63						
Genscan	0.64	0.89	0.44	0.64	0.14	0.41			0.03	0.28
Genscan	0.90	0.93	0.75	0.78	0.08	0.10				
Fgenes6			0.18	0.36						

[1] Genscan accuracy in 22q13.31; [2] Fgenesh accuracy in 22q13.31; [3] Genscan accuracy in the BRCA2 region (Hubbard and Bruskewich**,** http://predict.sanger.ac.uk/th/brca2); [4] Genscan accuracy in the set of semi artificial genomic sequences (Guigo *et al.*, 2000); [5] Genscan accuracy in the set of single gene sequences (Guigo , 2000); [6] Fgenesh accuracy in the BRCA2 region (Hubbard and Bruskewich, http://predict.sanger.ac.uk/th/brca2). These previously published results are included for comparison. Calculations of sensitivity and specificity at the nucleotide, exon and gene level are shown. The test region (22q13.31) contained 3,365,293 bp of genomic sequence. The total number of coding nucleotides was 44312 bp. The total number of reference exons was 334, contained within 38 protein-coding genes. For more details, see chapter II.

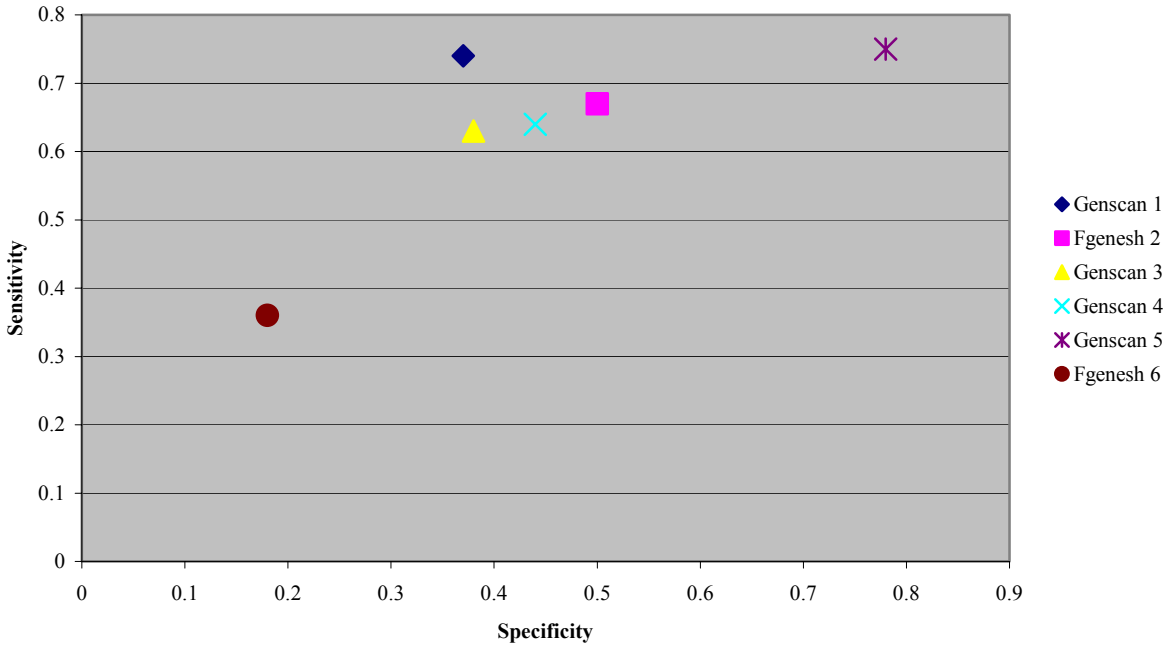


Figure 3.25: Specificity and sensitivity of the alignment of *ab initio* gene prediction programs with a variety of annotated human sequences. Sensitivity and specificity shown are computed at the exon level. The origin of each data set is shown in the legend to table 3.18.

Interestingly, the specificity shown here for Genscan predicted exons is very similar to that reported in the BRCA2 region (Hubbard and Bruskewich) and greater sensitivity is also demonstrated. However, equivalent results for the Fgenesh program were very different and were much lower for the BRCA2 region than those from chromosome 22. As expected, specificity and sensitivity of Genscan performance on this 'real' genomic DNA are generally both lower than in tests conducted on semi artificial and on single gene sequences (Guigo *et al.*, 2000). One exception is that the sensitivity of exon prediction within 22q13.31 was greater (0.74) than that shown by results from the semi artificial test set.

The Genscan results generally agree with the accepted accuracy levels of this program, which have been derived under artificial conditions or on comprehensively annotated DNA. This may imply that this region of chromosome 22 contains a similar level of annotation. Surprisingly Fgenesh did much better on the chromosome 22 DNA than on the BRCA 2 region. The reason for this is unknown, but supports the observation made by Dunham *et al.*(1999) that gene prediction programs show different levels of accuracy in different sequence regions.

3.10 Discussion

This chapter has shown the identification and annotation of 39 genes and 17 pseudogenes in a 3.4 Mb region of chromosome 22 by a combined approach of sequence analysis and experimental work. Integration of the data in a single database has aided the assembly of a transcript map and also enabled further investigation of gene features within their genomic environment.

Publication of the draft genome sequence (Lander *et al.*, 2001) means that comparison can now be made between a specific chromosomal region and the broad genomic environment, in order to identify regional trends or abnormalities. Investigation of the GC and repeat content showed that the region of interest is GC-rich, enriched in *Alu* repeats but LINE-poor. The region contains DNA mainly consistent with the features of the H3 isochore. These characteristics concur with the research of Cheung *et al*.(2001), which mapped the region to the chromosomal light band 22q13.31.

Several different lines of evidence were used as a starting point to identify potential gene features within the sequence of 22q13.31. These included EST, cDNA and protein sequence homologies, exon trap data and *ab initio* gene prediction programs. The use of a wide range of preliminary evidence was followed up by extensive experimental confirmation and manual database inspection to resolve ambiguities and errors.

No single line of evidence was found to be 100% accurate when compared to the current transcript map of 22q13.31. The most sensitive and specific correlations were observed from expressed sequence evidence, such as EST and mRNA databases. However, annotation of genes using multiple ESTs or cDNA sequences from paralogs or orthologs may not be entirely accurate, as data from Wolfsberg and Landsman (1997) suggests. A proportion of these sequences may result from artefacts in generation. This study, for example, disregarded two submitted cDNAs due to the presence of degenerate poly(A)sequence in genomic sequence at the 3' end of the sequence. These cDNAs may have arisen from inaccurate or incomplete splicing, or from oligo-dT primed extension of genomic DNA contamination of the cDNA libraries used in the generation of these sequences. Both of these cDNAs are closely

associated with *Alu* and L1 repeats in the genomic sequence, which contain degenerate poly(A) sequence (Smit, 1996).

Exon traps and *ab initio* gene predictions provided expression-independent information. However, results shown in section 3.9.2 demonstrated that the accuracy of *ab initio* gene programs is insufficient for gene annotation solely on this evidence alone. Similarly, although the results provided by the 'Trofatter' exons demonstrated specificity equivalent to that of EST and mRNA databases, sensitivity of this method was found to be low. Since Trofatter e*t al*.(1995) describes a whole chromosome exon trap, the chance of isolating all exons of a single gene is remote so further evidence is required for full gene annotation.

To assemble a complete gene sequence from preliminary *ab initio* prediction or exon trap evidence, screening of cDNA libraries or whole RNA is required. However, the success of such experiments may depend upon the type or developmental state of tissues tested. Nearly sixty exons predicted by Genscan, but not supported by cDNA or EST evidence, were screened across seven cDNA libraries as part of this study. Only three exons were found to be represented in these resources. The other predicted exons may be incorrect, or may be expressed at low levels, perhaps only in specific tissues or at a specific time. Screening a wider range of cDNA libraries or RNA resources may result in the confirmation of more of these exons. This proposal is supported by a similar recent study by Das *et al*.(2001),involving screens of 230 exons predicted by Genscan from chromosome 22 sequence that were not incorporated in the published gene annotation (Dunham *et al.*, 1999). RT-PCR across 17 tissues and one cell line and sequencing of the resulting PCR products identified spliced

cDNA from 32 (14%) of the Genscan predictions. However, the remaining unsupported predictions can still not be discounted as encoding potential true genes.

Therefore, even a combination of these methods may not yield a complete transcript map as the limitations of expressed sequence resources mean that expression-independent lines of evidence cannot be dismissed. Additionally, eleven genes annotated by the methods described in this chapter are known to be incomplete. This is partly due to the inherent problems described above in generation of the resources used (ESTs, cDNA libraries). Several approaches could be taken in order to complete the transcript map. Screening of further cDNA libraries may identify further sequences to add to the annotation. Additionally, 5' RACE experiments could be undertaken to enable annotation of complete 5' UTR sequences. The increasing availability of genomic sequence from model organism sequencing projects provides another gene annotation tool for the identification of functionally conserved sequences. This approach is examined in more detail in chapter IV.

The availability of the genomic sequence of chromosome 22 allows analysis of the gene structure and surrounding sequence environment. Annotation of known genes onto the genomic sequence has, in some cases, identified the intron/exon arrangement. The gene order and orientation will also be of interest in the study of gene interactions. This thesis identified instances where genes 'shared' a predicted CpG island (SMC1L2 and dJ102D24.C22.2) and related genes are in close proximity (bK414D7.C22.1 and dJ671O14.C22.2), which may indicate the presence of shared regulatory sequences, although preliminary investigations did not indicate similar mRNA expression patterns for the former pair that would be consistent with this theory.

Expression profiles were generated by screening Northern blots, the production and screening of an RT-PCR panel of 32 human tissues and investigation of the tissue origin of EST hits to the cDNA sequences. Each of these approaches demonstrates useful features, but also have disadvantages. Analysis of EST hits allowed investigation of expression in a wide range of tissues. However, inconsistencies may result from different methods used in library preparation, from which the ESTs derive. EST sequences are generally derived from only single-pass reads and therefore represent only part of the full gene sequence and may contain inaccuracies. Additionally, a subsection of the ESTs may derive from spurious priming, mis-splicing, genomic contamination etc. (see section 3.13) leading to further inaccuracies.

In the cases of the RT-PCR panel and Northern blots, information about the origin of each tissue and method of preparation is readily available. The RT-PCR panel represented a wider range of tissues than the Northern blot and screening this panel was quicker and easier than the blot hybridisation approach. However, low levels of genomic contamination were noted in some of the pools, although, where possible, the effects were negated by the design of intron-spanning primers. Northern blots, as well as providing some evidence of expression patterns, also provide information of transcript size, although resolution is limited. Northern blots can also provide evidence of alternative splices and paralogous genes, but this may also lead to confusion as to which band represents the transcript of interest. In the case of the RT-PCR expression panel, generated PCR products could also be sequenced to confirm identity.

Northern blot evidence supported the annotated transcript size of 24 genes and provided evidence of the potential size of the full-length transcript of three partial genes. The hybridisation of probes, designed from the gene features HMG17L1 and dJ1033E15.C22.2, to

particularly large transcripts, may indicate the presence of large paralogous genes (possibly HMG17 in the case of HMG17L1). Additional evidence from this project indicates that HMG17L1 may be a pseudogene, as this feature is situated within an intron of another gene and is a member of a large gene family known to contain a number of pseudogenes (Venter *et al.*, 2001). Further analysis of the coding status of this feature could include an examination of sequence conservation in the conserved syntenic mouse region (see chapter IV) or assays of the encoded protein *in vitro*.

Most of the genes within 22q13.31 demonstrated expression in a wide range of tissues, but the expression of four genes was generally limited to reproductive tissues, suggesting that transcriptional regulation could limit the proteins encoded by these genes to a specific role in these organs. The high quality transcript map described in this chapter provides a foundation for further work to determine the function of the encoded proteins. Preliminary functional characterisation of these proteins is addressed in chapter V, utilising a range of *in silico* and experimental techniques.

Successful identification of additional gene features such as polyadenylation sites and translation start sites can increase confidence that a gene has been annotated correctly. The analysis of translation initiation sites in this project, however, identified a discrepancy between the annotated gene NUP50 and the scanning model of translation initiation. The annotated translation start site is supported by evidence from orthologous genes, but the presence of an upstream ATG in a strong Kozak consensus (Kozak, 1987) with no intervening stop codon precludes translation from this site by the scanning model. This analysis therefore supports the proposal of Peri and Pandey (2001) that additional mechanisms such as leaky

scanning, reinitiation or internal initiation of translation may play a much greater role that previously imagined (Gray & Wickens, 1998; Jackson & Kaminski, 1995; Liu *et al.*, 1984; Slusher *et al.*, 1991). In support of this idea, a growing number of transcripts have recently been reported to undergo internal initiation (Coldwell *et al.*, 2000; Sehgal *et al.*, 2000; Vagner *et al.*, 1995).

With the continuation of large-scale transcript mapping projects, efforts to identify paralogous genes using BLAST experiments become more rewarding. Results in section 3.8.7 supported the previous identification of several small gene families, including the parvin (Olski *et al.*, 2001) and PACSIN (Ritter *et al.*, 1999) families of related proteins, and have identified several more potential groups of related genes. The apparent duplication of two genes on chromosome 22 is of interest in the study of genome evolution. Further investigation of the duplicated region showed that one copy of each gene encodes a full ORF, whilst later mutations in the second copy may have resulted in two unprocessed pseudogenes. The duplication may have arisen as a tandem repeat generated by replication slippage or by recombination between sister chromatids. However, no paralogue of the spacer DNA could be found in nucleotide searches of the chromosome 22 sequence. This may mean that the flanking sequences have diverged as no obvious region where replication slippage or unequal crossing-over occurred could be determined. The increasing availability of annotated human genomic sequence makes the study of evolutionary relationships with the genome easier. Comparison of this data with the genomes of model organisms should further enhance knowledge of chromosomal evolution (chapter IV).