Structural, functional and comparative studies of human chromosome 22q13.31

Melanie Elizabeth Anne Goward

A thesis submitted in partial fulfilment of the requirements of Cambridge University for the degree of Doctor of Philosophy

Trinity Hall, Cambridge University

January 2002

This dissertation is the result of my own work and includes nothing that is the outcome of work done in collaboration. The dissertation does not exceed the length limit set by the Biology Degree Committee.

Abstract

As the human genome project nears completion, there is a need to identify and accurately annotate the genes contained within the genomic sequence. The next challenge is the functional analysis of these genes. The aim of this project was to utilise and evaluate different approaches to human gene annotation through analysis of a region of the genomic sequence of human chromosome 22 and then to carry out initial functional studies of the genes identified.

The thesis describes the assembly of a transcript map across a 3.4 Mb region of human chromosome 22 (22q13.31). Candidate gene structures were identified from publicly available expressed sequence evidence and *ab initio* gene predictions, then experimentally verified and extended. This analysis resulted in the annotation of 39 gene and 17 pseudogene structures. Expression of the annotated genes was investigated by Northern blot analysis and RT-PCR screening of RNA isolated from 32 human tissues. The tissue distribution of EST hits to the cDNA sequences were also analysed. The majority of genes demonstrated expression in a wide range of tissues, but the expression of four genes was shown to be limited to reproductive tissues only. Computational analysis of transcription and translation start sites, splice sites and polyadenylation signals showed strong conservation of the sequence contexts necessary for correct transcription and translation. One exception was noted in the gene NUP50, whose features do not correlate with those required by the scanning model of translation initiation.

The contribution that mouse genomic sequence can make, both to human gene annotation and understanding of genome evolution, was evaluated through the construction of bacterial clone maps across a region of mouse chromosome 15, orthologous to human chromosome 22q13.31

and also across a nearby conserved synteny breakpoint between human chromosome 22 and mouse chromosomes 15 and 8. Comparison of available mouse sequence from the mapped clones to the orthologous human regions showed strong conservation of gene order and content, but no conservation of human pseudogenes was noted within the mouse sequence. The analysis of the mouse genomic sequence did not result in extension of the annotation of 22q13.31, but enabled finer mapping of the synteny breakpoint from a 160 kb region on human chromosome 22, to one of 50 kb flanked by adjacent conserved genes.

Functional characterisation was carried out using BLASTP searches to identify protein homologues. The Interpro database was searched to identify protein domains within the amino acid sequences. These results allowed preliminary functional categorisation of the proteins. The localisation of 16 gene products was experimentally determined, by cloning the genes and expressing the encoded proteins in mammalian cells in conjunction with a short peptide tag that conferred antibody specificity. Both N- and C- terminals of each protein were individually tagged. The majority of proteins were distributed in the cytoplasm, with a subset also localised to the cell membrane. An endoplasmic reticular and an unidentified protein localisation pattern were also observed.

Through sequence analysis of regions of human chromosome 22, this project demonstrates and evaluates the contributions that different types of evidence can provide to annotation and analysis of the human genome sequence. It also presents a potential high-throughput approach to determination of protein localisation, which could contribute to the determination of the function of human genes found within the genome.

Acknowledgements

Many people have kindly provided me with their help and advice throughout the course of this project. The most important of these is Ian Dunham, who has given me guidance and encouragement since the beginning of my time at the Sanger Institute. Thank you for all the time you have spent in helping me to complete this project.

I would like especially to thank the past and present members of the chromosome 22 group, Dave Beare, Charlotte Cole, John Collins, Elisabeth Dawson, Carol Edwards, Owen McCann, Andy Mungall, Tamsin Tarling and Charmain Wright, for their encouragement and practical assistance. I would also like to thank Begoña Aguado, Meera Mallya, David Vetrie and Clare East for their help and advice with the protein and RNA work. Many people from the Sanger Institute have provided me with invaluable assistance and I would particularly like to thank Richard Evans, George Stavrides, Rhian Gwilliam, Karen Halls, Elisabeth Huckle, Carol Carder and Paul Hunt and, from informatics, Carol Scott, Sarah Hunt and Kate Rice.

I would like to also acknowledge the helpful discussions and critical reading of this manuscript by Begoña Aguado, John Collins, Ian Dunham, Richard Evans, Brian Goward and Luc Smink. The fold out diagrams were created with the help of Ewan Birney and printed by Richard Summers.

Finally, I would like to thank my family and friends for their support and encouragement throughout this project.

Table of contents

List of tables

List of figures

List of abbreviations

22ace	Chromosome 22 implementation of ACeDB
aa	Amino Acid
ACeDB	A C. elegans DataBase
AUM	Asymmetric Unit Membrane
BAC	Bacterial Artificial Chromosome
BLAST	Basic Local Alignment Search Tool
bp	Base pair(s)
BSA	Bovine Serum Albumin
cDNA	Complementary DNA
CDS	CoDing Sequence
CM	Cytoplasm and cell membrane
Cy	Cytoplasm
DMEM	Dulbecco's Modified Eagle Medium
DNA	DeoxyriboNucleic Acid
EMBL	European Molecular Biology Laboratoriums
ePCR	Electronic PCR
ER	Endoplasmic Reticulum
EST	Expressed Sequence Tag
FBS	Fetal Bovine Serum
FISH	Fluorescent In Situ Hybridisation
FPC	FingerPrinting Contigs
gff	Genome Feature Format
GFP	Green Fluorescent Protein
HSA22	Homo Sapiens chromosome 22
iATG	Translation Initiation site
kb	kilo base pairs
LINE	Long INterspersed repeat Element
LTR	Long Terminal Repeat
Mb	mega base pairs
MGC	Mouse Genome Consortium
MGD	Mouse Genome Database
MGSC	Mouse Genome Sequencing Consortium
Mi	Mitochondria
MIR	Mammalian-wide Interspersed Repeat
MMU8	Mus Musculus chromosome 8
mRNA	Messenger RNA
MS	Mass Spectroscopy
NCBI	National Center for Biotechnology Information
ncRNA	Non Coding RNA
NIH	National Institute of Health
NJ	Neighbour-Joining
nt	Nucleotide
Nu	Nucleus
ORF	Open Reading Frame

PAC	P1 Artificial Chromosome
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PIP	Percentage Identity Plot
R	Purine
RFLP	Restriction Fragment Length Polymorphism
RH	Radiation Hybrid
RNA	Ribonucleic acid
RNAi	RNA Interference
rRNA	Ribosomal RNA
RT-PCR	Reverse Transcription PCR
SINE	Short INterspersed repeat Element
Sn	Sensitivity
snRNA	Small Nuclear RNA
SNP	Single Nucleotide Polymorphism
Sp	Specificity
SSR	Simple Sequence Repeat
STS	Sequence Tagged Site
tRNA	Transfer RNA
upATG	ATG upstream of the iATG
UTR	UnTranslated Region
WGS	Whole Genome Shotgun
WS1	Waardenburg Syndrome type 1
WWW	World Wide Web
Y	Pyrimidine
YAC	Yeast Artificial Chromosome

Publication arising from this work

Dunham I., Shimizu N., Roe B. A., Chissoe S., Hunt A. R., Collins J. E., Bruskiewich R., Beare D. M., Clamp M., Smink L. J., Ainscough R., Almeida J. P., Babbage A., Bagguley C., Bailey J., Barlow K., Bates K. N., Beasley O., Bird C. P., Blakey S., Bridgeman A. M., Buck D., Burgess J., Burrill W. D., O'Brien K. P., and *et al*. (1999). The DNA sequence of human chromosome 22. *Nature* **402:** 489-95.

Chapter I Introduction

1.1 Introduction

The central tenet of molecular biology, first proposed by Francis Crick in 1957, describes how genes encoded by DNA sequences are copied (transcribed) to messenger RNAs (mRNA), which is then translated into functional proteins. This model became the basis of the colinearity theory, which states that the linear arrangement of subunits in the DNA sequence of a gene corresponds to the amino acid sequence of a protein. Determination of the entire genetic code (Khorana *et al.*, 1966; Nirenberg *et al.*, 1966) enabled prediction of protein sequences by translation of DNA sequences. Ten years later, techniques for rapid DNA sequencing were introduced (Maxam & Gilbert, 1977; Sanger *et al.*, 1977), which led to sequencing of large DNA molecules such as the 16.5 kilobase (kb) human mitochondrial genome (Anderson *et al.*, 1981) and the 40 kb genome of the Lambda bacteriophage (Sanger *et al.*, 1982). Since then, further development and high throughput automation of sequencing techniques has been accomplished and complete sequencing of large genomes is now possible, thus enabling researchers to study the fundamental genetic building blocks of life.

The human genome is the largest genome to be extensively sequenced so far. Preliminary analysis has confirmed that knowledge of the genome sequence will provide valuable insights into human biology. An important goal of current research is the generation of accurate annotation of all the genes encoded within the human genome (section 1.5); this gene index is expected to serve as a 'periodic table' for future genetic studies (Lander, 1996). Large-scale studies are being implemented to investigate the function of the genes and proteins identified from this research (section 1.6). Eventual integration of these studies should allow systematic dissection of the circuitry of the human body.

These advances in biological understanding have implications for research into human disease. The human genomic sequence in public databases allows rapid identification *in silico* of potential disease gene candidates and at least thirty disease genes have been identified in research efforts dependent on the genome sequence (Lander *et al.*, 2001). The genome sequence also provides insight into the mechanisms of chromosomal deletion, through homologous recombination and unequal crossing over between large, nearly identical intrachromosomal duplications. Such events are thought to be responsible for several syndromes, including the DiGeorge /velocardiofacial syndrome region on chromosome 22 (Shaikh *et al.*, 2000). Genomic research may lead to the development of new treatments for genetic disease, through the identification of new drug targets and a better understanding of disease mechanisms. Effective approaches to disease prevention may also be developed, as genetic predispositions to disease are recognised.

For the first time, the genomic landscape can be examined from a global perspective. Investigation of the distribution of features such as repetitive elements, GC content, CpG islands and recombination rates, are providing important clues about function and insight into the evolutionary history of the genome (Lander *et al.*, 2001). Comparative genomic data from model organisms also provides a powerful tool for analysis of the human genome, through identification of conserved functional features and novel innovations in different lineages.

This thesis describes the identification and accurate annotation of genes within a 3.4 megabase (Mb) region of human chromosome 22 (HSA22). The availability of the genomic sequence in this region enabled extensive sequence analysis of the gene environment. The utility of genomic sequence from the equivalent region of the mouse genome was explored in comparative

analyses of potentially functionally conserved regions and investigation of chromosomal evolution. Finally, the experimentally verified transcript map was used as a basis for preliminary functional analyses of the protein coding genes encoded in this region, using both *in silico* techniques and an experimental approach to determine subcellular protein localisation. The next sections set out the background to the work reported in this thesis.

1.2 Mapping the human genome

1.2.1 Broad Features of the genome

The complete DNA sequence of a human is approximately 3200 Mb (Lander *et al.*, 2001; Morton, 1991). It is contained in 23 pairs of chromosomes: 22 autosomes and 2 sex chromosomes, X and Y. A basic classification of chromosomes is provided by the position of the centromere. In metacentric chromosomes, the centromere is roughly localised in the middle. Acrocentric chromosomes have the centromere close to one end and submetacentric centromeres are in-between these two positions. Chromosomes can be further distinguished by their banding patterns. A variety of treatments involving denaturation and/or enzymatic digestion of chromatin, followed by incorporation of a DNA specific dye, can cause mitotic chromosomes of complex organisms to appear as a series of transverse light and dark staining bands (Craig & Bickmore, 1993). Banding reflects variations in the longitudinal structure of chromatids, where each band differs from adjacent bands in base composition, time of replication, chromatin conformation and in the density of genes and repetitive sequences (see table 1.1). Such banding permits accurate differentiation of chromosomes – previously the only way of doing this was by examining the sizes of the chromosomes and the positions of the

centromeres. Additionally chromosome banding allows more accurate definition of translocation breakpoints, subchromosomal deletions and other rearrangements.

Table 1.1: Properties of chromosome bands seen with standard Giemsa staining

Dark bands (G bands)	Pale bands (R bands)
Stain strongly with dyes that bind preferentially to AT-rich regions, such as Giemsa and Quinacrine	Stain weakly with Giemsa and Quinacrine
AT-rich	GC-rich
DNase insensitive	DNase sensitive
Condense early during the cell cycle but replicate late	Condense late during cell cycle but replicate early
Gene poor.	Gene rich
LINE rich, but poor in *Alu* repeats	LINE poor, but enriched in *Alu* repeats

(Adapted from Strachan and Read, 1999 and Lander *et al.*, 2001).

1.2.2 Genome maps

A more detailed delineation of the genome has been achieved by the production of various types of genome maps at increasingly fine scales. These maps have provided a framework of marker orders, established along the length of each chromosome. The frameworks, described briefly below, have been used to orientate and anchor the sequence-ready maps of overlapping cloned genomic segments during the human genome project.

1.2.2.1 Genetic maps

The aim of genetic mapping is to discover how often two loci are separated by meiotic recombination. The further apart two loci are on a chromosome, the more likely it is that a crossover will separate them. Thus the recombination fraction is a measure of the genetic distance between the two loci. Human genetic mapping required the development of genetic markers: Mendelian characters, which are sufficiently polymorphic to give a reasonable chance that a randomly selected person will be heterozygous.

The first human linkage map was published in 1987 (Donis-Keller *et al.*, 1987). The markers for this map were restriction fragment length polymorphisms (RFLPs) (Botstein *et al.*, 1980), whose use was soon replaced by the more informative, highly polymorphic microsatellite repeats (Litt & Luty, 1989; Tautz, 1989; Weber *et al.*, 1991). The microsatellite landmarks have been converted to sequence tagged sites (STSs), (Olson *et al.*, 1989), that can be assayed by the polymerase chain reaction (PCR) (Saiki *et al.*, 1988; Saiki *et al.*, 1985). These technical advances aided the construction of genetic maps at increasingly high resolution (Gyapay *et al.*, 1994; Hudson *et al.*, 1992; Weissenbach *et al.*, 1992), culminating in a 1 cM map (Dib *et al.*, 1996; Murray *et al.*, 1994). Efforts are currently focused on the generation of even more dense maps for the mapping of complex traits, using the most common type of DNA sequence variation: single nucleotide polymorphisms (SNPs).

1.2.2.2 Radiation Hybrid (RH) maps

The original approach by Goss and Harris (1975), where chromosome fragments generated by lethal irradiation of donor cells are rescued with suitable recipient cells, was applied to study whole genomes in 1994 (Walter *et al.*, 1994). The presence or absence of markers within a hybrid can be interpreted to produce a linear map order for the DNA clones (Cox, 1992). This is because the nearer two DNA sequences are on a chromosome, the lower the probability of separating them by the chance occurrence of a breakpoint between them (Cox *et al.*, 1990; Gyapay *et al.*, 1996; Walter *et al.*, 1994). RH mapping has been used to produce high-resolution gene maps by assaying the RH panels with genetic markers and RNA-derived expressed sequence tags (ESTs) by PCR. The ESTs are then ordered relative to the genetic markers (Deloukas *et al.*, 1998; Schuler *et al.*, 1996a). RH maps are also used to order and integrate all

chromosome-specific markers to produce a framework map for the construction of bacterial clone maps (Bentley *et al.*, 2001; McPherson *et al.*, 2001; Montgomery *et al.*, 2001; Mungall *et al.*, 1996; Mungall *et al.*, 1997; Tilford *et al.*, 2001).

1.2.2.3 YAC maps

A primary goal of physical mapping is to assemble a comprehensive series of DNA clones with overlapping inserts (clone contigs). This became feasible for larger genomes with the development of yeast artificial chromosomes (YACs) (Burke *et al.*, 1987). The large insert sizes of up to 1500 kb (Chumakov *et al.*, 1995) allow long-range continuity. Many different YAC maps have been published (Bell *et al.*, 1995; Bouffard *et al.*, 1997; Chumakov *et al.*, 1992; Collins *et al.*, 1995; Doggett *et al.*, 1995; Foote *et al.*, 1992; Gemmill *et al.*, 1995; Gianfrancesco *et al.*, 1997; Hudson *et al.*, 1995; Krauter *et al.*, 1995; Nagaraja *et al.*, 1997). However, due to problems with chimaerism and instability (Green *et al.*, 1991; Nagaraja *et al.*, 1994),YACs are not a suitable substrate for sequencing.

1.2.2.4 Bacterial clone maps

Cosmid (Collins & Hohn, 1978) and fosmid (Kim *et al.*, 1992) libraries provided an alternative to YACs, but the disadvantage of these cloning systems is their small insert size (30-45 kb). The development of bacterial clone vectors, which could accommodate larger inserts (up to 200 kb), bacterial and P1 artificial chromosomes (BAC and PAC respectively) (Ioannou *et al.*, 1994; Shizuya *et al.*, 1992), resulted in a number of new clone libraries. These cloning systems are stable, due to the lower copy number replicons and have been shown to contain few rearrangements (Ioannou *et al.*, 1994; Shizuya *et al.*, 1992). For these reasons, these types of library were the chosen resource for sequence ready map construction.

1.2.2.5 Sequence-ready maps and sequencing

Different strategies have been used to construct the sequence ready bacterial clone maps. The Sanger Institute and Washington University Genome Sequencing Center (Lander *et al.*, 2001; McPherson *et al.*, 2001) favour a map-based, hierarchical shotgun method. STSs from previously constructed genetic and physical maps were used to recover BACs and PACs from specific regions. The clones are then assembled into contigs by landmark content mapping (Green *et al.*, 1991) and restriction enzyme fingerprint analysis (Gregory *et al.*, 1997; Marra *et al.*, 1997; Olson *et al.*, 1986) (see chapter IV). A sequence tile path, minimising redundancy from overlapping clones, is then selected for sequencing.

Selected clones were sequenced using a shotgun approach. The cloned genomic insert is fragmented and the 1.4 – 2.2 kb fragments cloned into M13 or plasmid vectors (Bankier *et al.*, 1987). The subclones are then sequenced using the chain termination method (Sanger *et al.*, 1977). This method has been adapted to use two types of fluorescent chemistries: dye labelled primers and terminators (Lee *et al.*, 1992; Prober *et al.*, 1987; Smith *et al.*, 1987). The sequence reads obtained are assembled into contigs, after which a directed approach is used both manually and automatically to edit the sequence. Additional sequence to close any gaps and resolve problems is obtained during 'finishing'.

An alternative whole genome shotgun (WGS) method was utilised by the biotechnology company Celera Genomics, to produce a second version of the human genome (Venter *et al.*, 2001). Human clone libraries of prescribed insert length were produced from the DNA of five individuals. The ends of clone inserts were sequenced (paired end sequences or mate pairs), generating sequence reads amounting to 5.11-fold coverage of the genome. Sequence generated

by the public effort, freely available in public databases, was also used to bring the effective coverage to 8-fold (Venter *et al.*, 2001). Two assembly strategies – a whole-genome assembly and a regional chromosome assembly - were used, each combining sequence data from Celera and the publicly-funded genome effort. Known repeat elements were screened out from the assembly process before sequence overlaps were identified and checked for the presence of repeated elements not removed in the initial screen. Gaps between the assembled contigs could be sized and the contigs orientated, using the mate pair information of sequence reads from opposite ends of the same clone insert. The two assembly strategies yielded very similar results that largely agree with the independent mapping data (Venter *et al.*, 2001).

1.2.2.6 Human genome draft sequence

The public domain sequencing centres published a first draft of the human genome sequence in February 2001. This was generated from a physical map covering more than 96% of the euchromatic part of the human genome and, together with additional sequence in public databases, it covers about 94% of the human genome. A final, accurate draft is promised by 2003 (Lander *et al.*, 2001). On the same day in February, the Celera venture published the second version of the human genome in a rival journal (Venter *et al.*, 2001).

A computational comparison of the two draft sequences (Aach *et al.*, 2001) found that they were overall similar in size, containing comparable numbers of unique sequences and exhibited similar statistics for sample candidate DNA protein-binding motifs. Some differences emerged at a detailed level e.g. contigs in each exhibited different size and gap distributions. However, these differences are expected to diminish as assemblies become more complete and comprehensive.

1.3 Large-scale features of the genome sequence

The availability of the draft genome sequence allows systematic genome wide analysis of the human genome. Analysis has confirmed a variety of large-scale features of the genomic landscape (Lander *et al.*, 2001).

1.3.1 Distribution of GC content

On average the genome is 41% GC but the distribution of base composition varies from 38% to over 55% GC (Lander *et al.*, 2001). Previous studies have indicated that GC-rich and GC-poor regions have different biological properties, such as gene density, composition of repeat sequences and correspondence with cytogenetic bands (Duret *et al.*, 1995; Gardiner, 1996; Hurst & Eyre-Walker, 2000; Saccone *et al.*, 1992; Saccone *et al.*, 1993; Zoubak *et al.*, 1996).

Bernardi and colleagues (1985) proposed that the variation in GC content could reflect that the genome is composed of isochores – local regions of similar GC content. By randomly shearing DNA and fractionating it on CsSO4 gradients, five fractions were identified: two light AT rich fractions L1 and L2 and three increasingly GC rich fractions H1, H2 and H3. The L1 and L2 fractions comprise 62% of the genome, H1 22%, H2 9% and H3 3-4%. The remaining 3-4% consists of satellite and ribosomal DNA. This division was further extended by Saccone *et al.*,(1996) and the H3 isochore was split up into three increasingly GC-rich sub-fractions: $H3^{-}$, H3* and $H3^{+}$. Hybridisation *in situ* of the H3 fractions indicates the positions of the most gene-rich bands.

The draft genome sequence was analysed to see if the existence of strict isochores could be verified. However, the average GC content for a variety of different window sizes showed too

much variation to be consistent with a homogeneous distribution. Although the genome clearly contains large regions of distinctive GC content, Lander *et al*. (2001) concluded that there is substantial variation at many different scales. However, the existence of isochores in the human genome has been supported through the use of the different, window-less approach of recursive segmentation (Li, 2001). A segmentation point is identified that maximises the base composition difference between the left and right subsequences. Each subsequence is then subdivided into two further subsequences in the same manner, until the resulting domains satisfy a previously determined threshold value. Li proposes that a window-approach may not be able to delineate the borders of relative homogeneous domains accurately enough before carrying out a homogeneity test. The alternative recursive segmentation approach, however, supports the existence of isochores in the human genome.

1.3.2 CpG islands

The CpG dinucleotide occurs at about one fifth of the roughly four percent frequency that would be expected by multiplying the typical fractions of Cs and Gs (0.21x0.21) (Matsuo *et al.*, 1993). The shortfall occurs because CpG dinucleotides are often methylated on the cytosine base and spontaneous deamination of methyl-C residues gives rise to T residues (spontaneous deamination of ordinary cytosine residues gives rise to uracil residues that are readily recognised and repaired by the cell) (Coulondre *et al.*, 1978; Sved & Bird, 1990). However, the genome contains many CpG islands in which the CpG dinucleotides are not methylated and occur at a frequency closer to that predicted by local GC content. One feature of these islands is that they are rich in sites for methyl-sensitive restriction enzymes such as *Hpa*II, which recognise unmethylated CpG dinucleotides (Bird, 1986).

Using the definition proposed by Gardiner-Garda and Frommer (1987), a search of the repeat masked draft genome sequence highlighted 28,890 possible CpG islands. CpG islands are of particular interest, because many are associated with the 5' ends of genes (Bird *et al.*, 1985; Bird, 1986; Chan *et al.*, 2000) and may also contain promoter sequences (Cross *et al.*, 2000). Analysis of the draft genome sequence showed that the relative density of CpG islands correlated reasonably well with estimates of relative gene density on chromosomes.

1.4 Coding and non-coding sequence

An important distinction that can be made between the different compartments of the genome is coding versus non-coding sequence. The function of non-coding DNA remains to be fully understood (Gardiner, 1995). There are a number of features that distinguish these two fractions of the genome, which are discussed below.

1.4.1 Non-coding features

Genomes can contain a large quantity of repetitive sequence, far in excess of that devoted to protein-coding genes. Analysis of the draft human genome sequence showed that repeats account for at least 50% of the genome (Lander *et al.*, 2001). Several different classes of repeats have been described.

1.4.1.1 Transposon-derived repeats

About 45% of the genome sequence consists of repeats derived from one of four types of transposable element, of which three transpose through RNA intermediates (LINEs, SINEs and LTR retrotransposons) and one transposes as DNA (DNA transposons).

In humans, full length LINEs are about 6 kb long, contain an internal polymerase II promoter and encode two ORFs. Three LINE families, LINE1, LINE2 and LINE3, are found in the human genome; only LINE1 is still active. The transcribed LINE RNA and translated proteins move to the nucleus, where an encoded endonuclease activity makes a 3' single stranded nick from which the reverse transcriptase is primed. This frequently fails to proceed to the 5' end, resulting in many truncated, non-functional insertions. The LINE machinery is believed to be responsible for most reverse transcription in the genome, including SINE retrotransposition.

SINEs are on average between 100bp and 400bp long, harbour an internal polymerase III promoter but encode no proteins. They are thought to use the LINE machinery for transposition and have been noted to share the 3' end with LINE elements (Okada & Hamada, 1997). The human genome contains three families of SINEs: the active *Alu* and the inactive MIR and Ther2/MIR3.

LTR retrotransposons are flanked by long terminal direct repeats, which contain all of the necessary transcriptional regulatory elements. Mammalian retroviruses fall into three classes (I-III), each comprising of many families. Homologous recombination between flanking LTRs can result in loss of the internal sequence.

DNA transposons have terminal inverted repeats and encode a transposase enzyme. The human genome contains at least seven major classes of DNA transposon, each containing many families. Transposons have been indirectly responsible for many evolutionary innovations in the genome. Over forty human genes have been recognised as probably derived from transposons (Jurka & Kapitonov, 1999; Lander *et al.*, 2001; Smit, 1999).

LINE1 activity can also bring about exon reshuffling by co-transcription of neighbouring DNA. They can also cause reverse transcription of mRNA, which typically results in non-functional processed pseudogenes, but can occasionally give rise to functional processed genes. There are at least eight human genes that may be derived from this origin (Brosius, 1999).

1.4.1.2 Simple Sequence Repeats (SSRs)

Human satellite DNA is comprised of very large arrays of tandemly repeated DNA, often from 100 kb to several megabases in length. The repeat unit can range from 5 base pairs (bp) in length to over 170 bp (centromeric alphoid DNA). Repeated DNA of this type makes up the bulk of the heterochromatic genome regions, approximately 5-10% of the total sequence.

SSRs with a short repeat unit (n = 1-13 bases), often spanning less than 150 bp in total, are termed microsatellites, whilst these with longer repeat units (n = 14 – 500 bases) and spanning within a range of ~0.1 – 2.0 kb, are termed minisatellites. Slippage during DNA replication is thought to result in the production of SSRs (Kruglyak *et al.*, 1998; Toth *et al.*, 2000). SSRs comprise about 3% of the euchromatic human genome, with the greatest single contribution coming from dinucleotide repeats (0.5%) (Lander *et al.*, 2001).

SSRs have been used in human genetic studies (section 1.2.2.1). The microsatellites and especially the expansion of triplet repeats have also been implicated in neurodegenerative disorders. Since the cause of fragile X was shown to be repeat expansion (Yu *et al*., 1991; Verkerk *et al*., 1991; Kremer *et al*., 1991) the list of diseases caused by repeat expansion has continued to grow. Triplet repeat expansions are associated with non-B DNA structures: these structures may account for the expansion and instability and therefore the disease-causing feature of the triplet repeats (Sinden, 1999). Interestingly, one of the genes analysed in this

thesis (E46L) has been implicated in the causation of spinocerebellar ataxia 10, through polymorphism of an unstable pentanucleotide repeat in intron IX (Matsuura *et al.*, 2000).

1.4.1.3 Segmental duplications

Analysis of the draft sequence shows that the human genome seems likely to consist of about 5% segmental duplication. Intrachromosomal duplications occur within a particular chromosome. Interchromosomal duplications are defined as segments that are duplicated among non-homologous chromosomes. Regions near the centromere and telomeres are composed almost entirely of interchromosomal duplicated segments. It is hypothesised that chromosomal breakage products are preferentially inserted here by an unknown mechanism, in order to limit possible damage caused by insertion into more gene-rich regions (Lander *et al.*, 2001).

1.4.2 Coding genome features

1.4.2.1 Non-coding RNA (ncRNA) genes

Less than 5% of the human genome is thought to encode genes (Lander *et al.*, 2001). The majority of human genes ultimately specify polypeptides that carry out numerous diverse functions. However, a smaller minority instead specify a mature RNA product. In addition to the many genes involved in protein synthesis (rRNA genes, tRNA genes), there are other RNA genes that process and modify rRNA in the nucleolus (snoRNAs), spliceosomal RNAs and other ncRNA genes such as telomerase RNA and the 7S signal recognition particle RNAs. ncRNAs do not have translated open reading frames (ORFs), are often small and are not polyadenylated. Accordingly, novel ncRNAs are hard to find by experimental sequencing, but attempts are being made using computational techniques that exploit their secondary structural characteristics (Rivas *et al.*, 2001).

1.4.2.2 Protein coding genes

Human protein coding genes have a complex structure (figure 1.1). Whereas in simple organisms, such as yeast, the genes are simply single ORFs, in complex organisms, such as human, the ORF is segmented with the protein-coding exons being separated by introns. Nuclear pre-mRNA introns are excised from the primary transcript by a large ribonucleoprotein complex, known as the spliceosome (reviewed in Moore & Sharp, 1993), which recognises sites at the 5' and 3' ends of the intron (the donor and acceptor sites respectively) as well as an internal site known as the branch point. With a few exceptions (Sharp & Burge, 1997) nearly all spliceosomal introns begin with GT and end with AG.

Protein coding genes contain a translational start site (usually ATG), often contained in an optimal consensus sequence (Kozak, 1987). Some genes also contain a polyadenylation signal, most commonly an AATAAA hexamer sequence followed by a more complex signal (not yet completely characterised) located 20-30bp downstream (Beaudoing *et al.*, 2000; Gautheret *et al.*, 1998). Less is known about the identity of regulatory sequences that could be present in the 5' and 3' Untranslated regions (UTRs) and introns (section 1.5.2).

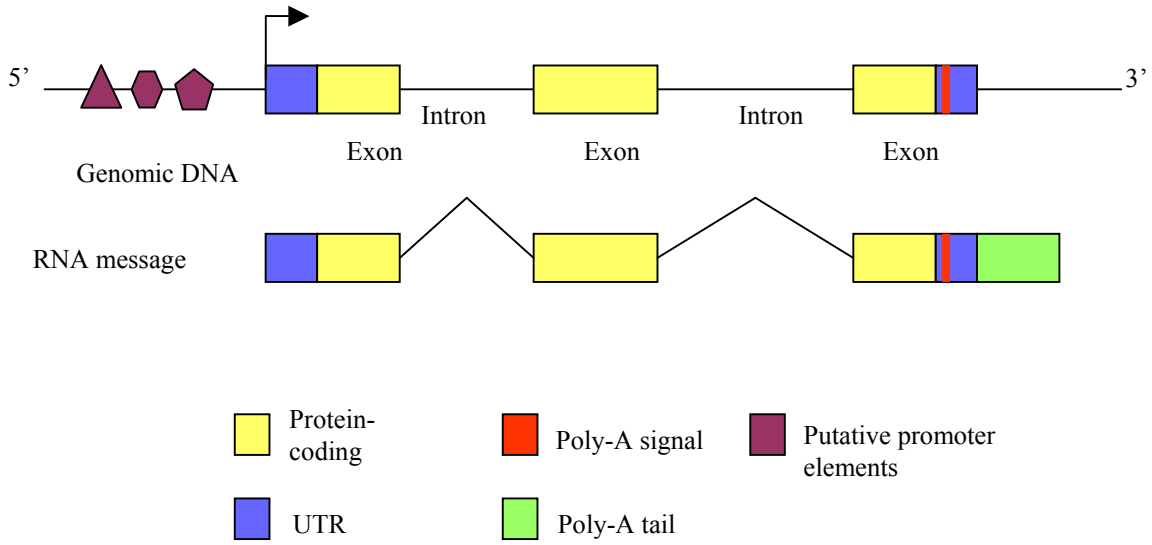


Figure 1.1: Protein-coding gene features. The genomic layout is shown at the top of the figure and the transcribed RNA message below. The colours identify the different features discussed in the text.

1.5 Gene Identification

Genes represent the major functional elements of the genome and are thus the main focus of interest of genome researchers. In principle, three major features permit the DNA of genes to be distinguished from DNA that does not have a coding function:

1. Expression: all active genes are capable of making an RNA product, usually mRNA. Mammalian genes usually contain introns, so the initial RNA transcript undergoes splicing.
2. Sequence conservation: because genes execute important cellular functions, mutations that alter the sequence of the product will often be deleterious and eliminated by natural selection. The sequence of coding DNA and important regulatory sequences is therefore more strongly conserved in evolution than that of non-coding DNA.
3. CpG islands: many vertebrate genes are associated with CpG islands (Bird *et al.*, 1995). Identification of these sites can aid identification of the adjacent gene.

1.5.1 Traditional approaches

Several techniques have been developed that rely on sequence conservation to find genes. For example, a zoo blot (Monaco *et al.*, 1986) involves the hybridisation of a DNA clone to a Southern blot of genomic DNA samples from a variety of animal species. Conserved sequences, which are likely to be genes, are thus identified.

CpG islands usually contain multiple rare-cutter restriction sites (Cross & Bird, 1995). These can be identified by restriction mapping (DNA clones are hybridised against Southern blots of genomic DNA, cut with *Sac*II, *Eag*I or *Bss*HII, to identify clustering of rare cutter sites) (Sargent & Bennett, 1986) or by island-rescue PCR (PCR amplification between islands and neighbouring *Alu* repeats) (Valdes *et al.*, 1994). The identified fragments can be tested for expression by hybridisation to Northern blots containing RNA isolated from a range of different tissues. If a transcript is identified, the corresponding complementary DNA (cDNA) can be isolated from the appropriate library. Alternatively, entire genomic clones can be hybridised against a Northern blot or against appropriate cDNA libraries

More efficient approaches can be used to construct a transcript map of a large region (Gardiner & Mural, 1995). Exon trapping uses a functional assay for splice sites in genomic DNA. The DNA is shotgun cloned into a vector containing a functional splice donor site, an intervening sequence and a selectable marker (Buckler *et al.*, 1991; Duyk *et al.*, 1990). This method has been used to identify the genes for a number of diseases (Trofatter *et al.*, 1993; Vulpe *et al.*, 1993). The technique has also been used to isolate exons from entire chromosomes (Church *et al.*, 1993; Trofatter *et al.*, 1995).

cDNA selection or capture involves repeated purification of a subset of cDNAs that hybridise to a given genomic region. cDNAs that hybridise specifically to genomic fragments immobilised on nylon membranes can be eluted and amplified by PCR. The process is repeated two or three times before the eluted cDNAs are cloned. This results in highly specific and enriched sub libraries for expressed sequences of the genomic region (Lovett *et al.*, 1991; Parimoo *et al.*, 1991). These methods have been improved by using biotin–labelled genomic DNA and streptavidin-coated magnetic beads to capture the genomic DNA-cDNA hybrids (Korn *et al.*, 1992; Morgan *et al.*, 1992). This approach has also been used to generate specific chromosome-enriched libraries (Touchman *et al.*, 1997).

1.5.2 Post-genomic era

The availability of large amounts of genomic sequence, defining the 'post-genomic era', facilitates sequence analysis as a method for gene identification. In small prokaryotic genomes, finding the encoded genes is largely a matter of identifying all the long open reading frames. Ambiguities arise if long ORFs overlap on opposite strands – the true coding region must then be investigated. Genes are found using a computer program that carries out six-frame translation, identifying ORFs longer than a chosen threshold (such as 500 bp (Burge & Karlin, 1998)). However, smaller genes could be missed.

Finding genes in eukaryotes becomes considerably harder as the signal:noise ratio increases. For example, the 8 Mb prokaryotic genome of *H. influenzae* contains 85% coding sequence, whereas more complex eukaryotic genomes, such as those of the fly and worm, are less than 25% coding. The human genome contains an estimated 3% coding sequence (Duret *et al.*, 1995), recently confirmed for chromosome 22 (Dunham *et al.*, 1999). Gene annotation in these

more complex organisms is complicated further by splicing and alternative splicing. The arrangement of genes in genomes is also prone to exceptions. Although usually separated with an intergenic region, there are examples of genes nested within each other (Dunham *et al.*, 1999); that is, one gene located in an intron of another gene or overlapping genes on the same (Ashburner *et al.*, 1999; Schulz & Butler, 1989) or opposite (Cooper *et al.*, 1998) DNA strands. The presence of pseudogenes (non-functional sequences resembling real genes), which are distributed in numerous copies throughout the genome, further complicates the identification of true protein coding genes. Current approaches to gene identification approaches include computer prediction packages and homology searches.

1.5.2.1 *Ab initio* prediction packages

The most natural way to find genes computationally would be to mimic as closely as possible the processes of transcription and RNA processing (e.g. splicing and polyadenylation) that define genes biologically. A number of important signals related to transcription, translation and splicing are now sufficiently well characterised as to be useful in computer predictions of the location and exon-intron organisation of genes. The genomic elements that researchers seek include splice sites, start and stop codons, branching points, promoters and terminators of transcription, polyadenylation sites, ribosomal binding sites, topoisomerase II binding sites, topoisomerase I cleavage sites and various transcription factor binding sites (reviewed by Gelfand, 1995). These conserved elements are used by *ab initio* prediction programs: gene prediction from sequence data without the use or prior knowledge about similarities to other genes. *Ab initio* gene prediction programs are discussed in more detail in chapter III.

1.5.2.2 Sequence similarity

The similarity of a region of the genome to a sequence that is already known to be transcribed, is a powerful predictor of whether or not a sequence is part of a gene. Similarity-based methods rely on matches to DNA and protein databases with the genomic sequence under investigation using, for instance, the Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1997). This type of search has become very powerful due to increased EST availability (Adams *et al.*, 1991; Wilcox *et al.*, 1991) (section 1.8).

Although the ESTs generally cover only a segment of the gene, their utility for gene identification was immediately recognized. In a pilot project, Adams *et al*.(1991) performed automated partial sequencing of more than 600 randomly selected cDNAs from human brain. Of the generated ESTs, 337 represented new genes and 48 had significant similarity to genes from other organisms. Since then, a large number of publications have generated increasing sets of ESTs and their analysis (Adams *et al.*, 1993; Hillier *et al.*, 1996; Houlgatte *et al.*, 1995; Khan *et al.*, 1992; Okubo *et al.*, 1992). All public domain ESTs are deposited in dbEST, a subdivision of GenBank/EMBL/DDBJ (Boguski *et al.*, 1993). An important key development for the widespread use of ESTs, was the formation of the IMAGE consortium (Lennon *et al.*, 1996) (http://bbrp.llnl.gov/bbrp/image/), to ensure that collections of clones as well as sequences would be accessible by the biomedical research community.

A large amount of redundancy exists in the large EST collections, owing to repeated sequencing of the same widely expressed genes in different or the same tissues. Different clustering methods have been devised to address the redundancy. Examples include Unigene (Boguski & Schuler, 1995) and the GeneExpress project (Houlgatte *et al.*, 1995).

There are many applications for which partial sequences are not adequate. For example, accurately predicting the function or structure of a gene product or isolating the protein product, requires a full-length sequence. The Mammalian Gene Collection (MGC) project represents an ongoing effort by the National Institute of Health (NIH) to generate a full-length cDNA resource, eventually representing all human genes (Strausberg *et al.*, 1999).

Protein databases provide a further resource for gene annotation (section 1.8). Notable examples include the SwissProt database (Bairoch & Apweiler, 2000), which is a database of protein sequences derived from translations of DNA sequences from the EMBL nucleotide sequence database, adapted from the Protein Identification Resource (PIR) collection, extracted from the literature and directly submitted by researchers. SwissProt is a curated database, containing high-quality annotation, is non-redundant, and cross-referenced to several other databases. TrEMBL is a computer-annotated protein sequence database, which supplements SwissProt. TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL nucleotide sequence database not yet integrated into SwissProt. TrEMBL can be considered as a preliminary section of SwissProt. Annotation of genes through similarity searches is discussed in chapter III.

1.5.3 Comparative studies

Conserved genetic linkage groups have been documented in a variety of vertebrate species (for a summary, see Jones *et al.*, 1997). Genomic sequence from a range of species is now becoming available (section 1.7) and a wide variety of cross-species comparative studies are being undertaken to identify conserved and novel functional features, to elucidate the mechanisms that have acted during genome evolution and to study gene and protein function in model systems.

1.5.3.1 Identification of conserved functional regions

In order to exploit genomic comparison as an analytical tool for identification of functional regions, species must be selected of great enough evolutionary divergence to permit identification of functionally conserved regions from the rest of the genomic background, yet small enough that comparison of conserved syntenic lineage is meaningful (Lundin, 1993). Such comparisons can allow identification of genes and possible regulatory regions in both genomes with no previous knowledge of the gene content of either.

BLAST (Altschul *et al.*, 1997) is one of a range of alignment algorithms that can also be used to compare genomic sequences with homologous genomic sequences from closely related organisms such as mouse, chicken or pufferfish. For example, the 'Exofish' algorithm (Crollius *et al.*, 2000), utilises a specific implementation of BLAST (TBLASTX) to conduct homology searches of human sequence with available *T. nigroviridis* sequence. Exofish has already proved useful in the identification of human genes and has demonstrated the potential of comparative genomics using the pufferfish genome. Additional algorithms are being developed for the specific purpose of species sequence comparison and gene identification.

The benefits of using mouse genomic sequence for identification of gene and regulatory regions have been illustrated by a large number of small-scale studies and is reviewed in more detail in chapter IV. Additionally, genome-wide alignments of human and mouse sequence are becoming available from the public genome project (Meisler, 2001). As more finished mouse sequence is added, this resource will aid identification of genes and candidate regulatory regions within the human genome.

1.5.3.2 Evolutionary chromosomal rearrangements

Chromosomal rearrangements such as inversions and translocations have played an important role in defining genome organisation in existing mammals. The number of rearrangements that have occurred since divergence from the primordial mammal has been modest and the distribution of these rearrangements among chromosomes appears random (Eppig & Nadeau, 1995). As a result, each mammalian species has a unique arrangement of conserved and disrupted chromosomal segments as compared to other mammalian species. Genes provide excellent markers for these chromosomal segments as their homologies can be detected in highly divergent species (Eppig & Nadeau, 1995; Nadeau & Sankoff, 1998).

The mouse genome represents the most thoroughly studied of all non-human vertebrate genomes. However, rodent chromosomes have undergone an unusually high number of genomic rearrangements per unit of evolutionary time (Graves, 1996). Nevertheless, the degree to which gene content and order is conserved is considerable (Carver & Stubbs, 1997). As the resolution of the physical maps of the human and mouse genomes increases from cytogenetic bands to nucleotide sequence (section 1.2, chapter IV), breakpoints in the comparative map can be mapped more precisely and their characteristics examined. Chromosomal painting by fluorescence *in situ* hybridisation (FISH) with chromosome-specific libraries is an easy way to compare the location of homologous chromosomes in different mammalian species (for example Rettenberger *et al.*, 1995; Scherthan *et al.*, 1994). This method is used to identify the location and approximate size of homologous segments and to estimate the number of rearrangements that have occurred since the divergence of the lineages leading up to the species being compared. Finer mapping of a syntenic breakpoint was provided by the first sequence-level analysis of a conserved synteny breakpoint between mouse chromosome 10 (MMU10)

and HSA21 and HSA22, recently described by Pletcher *et al*.(2000). Examination of the structural features of this segment, and comparison with other breakpoints, should provide insight as to whether particular DNA sequences contain structural features that are predisposed to ancestral chromosomal rearrangements (chapter IV).

Data from comparative maps is also used in functional studies to identify candidate disease genes and to characterise the genetic basis for complex traits (section 1.7).

1.6 Functional genomics

The initial interest in the human genome was precipitated by a desire to identify the cause of observable gene phenotypes: in 1986, Dulbecco stated the wish to 'sequence the whole genome of a whole animal species for the purpose of finding genes involved in cancer'. However, even once the entire complement of genes is established, the function for most of them will remain unknown (Blackstock & Weir, 1999). The emerging field of functional genomics is addressing these problems. This not only involves the determination of gene function, but also the determination of expression patterns and both spatial and temporal analysis of the proteins. Functional genomics is characterised by high throughput or large-scale experimental methodologies combined with statistical and computational analysis of the results. The underlying strategy is to expand the scope of biological investigation from studying single genes or proteins to studying all genes and proteins at once in a systematic fashion. Computational biology will perform a critical and expanding role in this area (Hieter & Boguski, 1997).

1.6.1 Expression studies

The first small-scale approaches to identifying and cloning differentially expressed genes were primarily based on subtractive hybridisation (Hess *et al.*, 1998). Although several genes were cloned using this method, subtractive hybridisation turned out to have some crucial disadvantages: it reveals only a small fraction of the overall changes in gene expression; it requires large amounts of RNA; and it is time-consuming and laborious. In 1992, differential display PCR (DD-PCR) was introduced to compare, identify and isolate differentially expressed transcripts (Liang & Pardee, 1992). In principle, the method utilises reverse transcription (RT)-PCR amplification of two different mRNA populations and separate the resulting fragments side-by-side on a denaturing polyacrylamide gel. Bands expressed at different levels are isolated and cloned.

Large-scale gene expression studies have been revolutionised by microarray technology. This takes advantage of the fact that increasingly complete sets of cDNA clones and sequences representing all human and mouse genes are becoming available for high throughput surveys of gene expression. DNA microarrays consist of genes, gene fragments or oligonucleotides covalently attached in a high-density array on a glass slide. The DNA on the array is selected from databases such as Unigene (Boguski & Schuler, 1995). Arrays can also be produced using photolithography to synthesise specific oligonucleotides *in situ* on the array (Fodor *et al.*, 1991).

The arrays can be used to record differences in expression between a reference and a test RNA population. Each RNA population is used as a template in a reverse transcription PCR reaction, incorporating a distinguishing fluorescently labelled dinucleotide. The fluorescently labelled cDNAs are then hybridised to the array. The relative fluorescence intensity measured for the

two different fluors at each array element provides a reliable measure of the relative abundance of the corresponding mRNA in the two RNA populations. This technique has been used to assay expression in inflammatory disease (Heller *et al.*, 1997), the diauxic shift in *S. cerevisiae* (DeRisi *et al.*, 1997) and tumorigenic versus non-tumorigenic cell lines. Microarrays have also been used in other techniques to assay expression patterns (section 1.6.3.2).

Serial Analysis of Gene Expression (SAGE) also takes advantage of the possibility of a human gene index. SAGE is based on short nucleotide tags of 9 to 10 bp, derived from the complete mRNA pool of a cell population. These tags contain enough information to identify a transcript. Concatenation of short tags allows the simultaneous analysis of multiple transcripts by sequencing of many tags (10-50) within a single clone. The advantage of this method is that it is possible to count the number of distinct mRNA molecules in a given cell population for each condition and, from this accounting, a particular mRNA would be described as differentially expressed if its frequency is significantly greater in one condition versus another.

A current limitation of microarray and SAGE technologies is that there is not yet a comprehensive and accurate index of human genes. cDNAs representing each gene also need to be collected both for the human genome, and for all other species of interest. The second limitation is the need for sufficiently powerful mathematical and visualisation tools for whole-genome expression studies to analyse the mass of new data. This is currently one of the major challenges faced by bioinformatics (section 1.8).

1.6.2 Control of gene expression

Our understanding of how regulatory information is encoded by a DNA sequence is still very fragmented. Large-scale genome sequencing projects currently determine hundreds of

megabases every year. Thus, one of the major challenges that biologists face is to identify the regulatory elements within the bulk genome.

Wet laboratory approaches to the identification of regulatory regions includes the use of reporter genes and deletion analyses to assess how deletion of different segments of DNA upstream of a gene, or occasionally on the first intron, affects gene expression. Gel retardation, DNAse footprinting and methylation interference assays can identify protein-binding sites on a DNA molecule. However, the amount of experimental work that would be required to systematically analyse these non-coding sequences could exceed current research capacity. There is therefore a need for experimental and computational tools that identify potential regulatory elements more quickly to allow focusing of experimental design.

Two main kinds of computational approaches can be distinguished. The first one includes methods that rely on biological knowledge of transcription factor binding sites, promoters, enhancers, locus control regions etc. to set up rules to predict regulator elements. However, the major obstacle to this approach is that the sequence motifs corresponding to these features contain too little diagnostic information for them to be distinguished from chance occurrences (Audic & Claverie, 1998). Experimentalists have shown that transcription factors, facing the same problem in the cell, find their physiological targets mainly via interactions with other factors bound to neighbouring sites. Chromatin structure modulated accessibility may also play a role. Computational biologists now agree that predictive algorithms should do the same thing: using sequence contextual features (e.g. predicted neighbouring elements) in order to distinguish between functional and biologically irrelevant sites.

The second type of approach relies on comparative analysis of homologous sequences. This approach has recently gained considerable interest, thanks to projects intended to sequence large regions of model vertebrate genomes (section 1.7). Tagle *et al.*(1988) proposed the term "phylogenetic footprinting" to describe the phylogenetic comparisons that reveal evolutionary conserved functional elements in homologous genes. However, regulatory elements that have been acquired very recently in evolution may not be detectable by this method. In addition, the conserved feature may not reside in the primary sequence structure, but rather in the spatial structure or a compositional property of the DNA or RNA that is subject to selective pressure.

As well as computational approaches, laboratory protocols are also being developed for the large-scale identification of putative regulatory regions. Frazer *et al*.(2001) recently described the most extensive human/mouse comparison available to date, through hybridisation of orthologous BAC contigs to an oligonucleotide array representing four 25-mers for each nucleotide in 16.6 Mb of non-repetitive DNA from human chromosome 21. The sequence of conserved elements could be determined from the hybridisation pattern. In the human-mouse comparison, 3400 conserved elements ranging in length from 30 bp to > 1 kb were identified, corresponding to 1.6% of the tested sequence. Only 44% of the conserved elements corresponded to known exons. 2.6 Mb of orthologous dog DNA was also hybridised to the oligo arrays, in order to estimate the proportion of conserved sequence resulting solely from common origin in the absence of active selection for function. Only half of the human-mouse noncoding elements were conserved in the dog sequence, indicating that it is worthwhile to extend comparisons beyond two species before initiating functional tests of putative regulatory elements.

1.6.3 Proteomics

The methods described above are critically important for a detailed understanding of the regulation of biological systems; however, such methods provide no information about post-translational control of gene expression. Indeed, experimental evidence suggests that there is no obvious correlation between mRNA expression levels and protein levels either in human liver cells (Anderson & Seilhamer, 1997) or in yeast (Gygi *et al.*, 1999). An emerging field for the analysis of biological systems is therefore the study of the complete protein complement of the genome, the 'proteome'.

1.6.3.1 Two-dimensional gel electrophoresis and mass spectroscopy

One of the most widely used tools for proteome analysis is two-dimensional protein electrophoresis (2.G.E) (O'Farrell, 1975). This technique resolves complex mixtures of proteins first by isoelectric point and then by size. The resulting gel images form a two-dimensional 'barcode' of the proteome of a particular biological system. A comparison of two or more such barcodes might help to identify differences in protein expression that result in particular phenotypes. A protein spot of interest can also be purified and further analysed (by direct amino-acid sequencing, amino acid analysis and mass spectroscopy) to relate the protein to the underlying gene.

The need to characterise spots has fostered an increasing use of mass spectroscopy (MS) for protein characterisation. The two most commonly used approaches for spot characterisation involve peptide mass mapping and tandem MS of a proteolytic digest of a 2.G.E spot. The masses of resulting peptides from a proteolytic digest can be measured using MS. These masses can be compared with *in silico* digests of protein databases or six-frame translations of nucleic

acid databases to help characterise the spot. In a tandem MS experiment, peptide mixtures are studied in an initial MS scan and particular peptides can be fragmented during a second step to generate amino acid sequence information. The sequence information is derived by an attempt to match mass spectra from fragmentation patterns with *in silico* spectra, obtained from databases, or by matching amino acid sequence information with available databases.

1.6.3.2 Chip-based methods for proteome analysis

On array-based methods, protein spots are immobilised onto glass slides. Such arrays can then be used to screen complex protein mixtures for particular binding affinities or other interactions (Walter *et al.*, 2000). Antibodies can be arrayed using bacteria that express recombinant antibodies. These arrays can be probed for specific antibody-antigen binding interactions, using a filter-based enzyme-linked immunosorbent assay (ELISA) technique.

Ziauddin and Sabatini (2001) have produced microarrays of cells expressing defined cDNAs. Arrays are printed with sets of cDNAs cloned in expression vectors. Mammalian cells are cultured on the glass slide and cells growing on the printed areas take up the DNA, creating spots of localised transfection within a lawn of non-transfected cells. Two uses for this approach have been identified so far: as an alternative to protein microarrays for the identification of drug targets and as an expression cloning system for the discovery of gene products that alter cellular physiology.

A commercial device, the ProteinChipSystem (Senior, 1999), combines chip-based techniques with MS to selectively capture proteins from biological systems using surface-enhanced laser desorption and ionisation (SELDI) technology. Protein mixtures are incubated with a variety of available chips that probe Lewis acid/base interactions or hydrophobic, electrostatic and co-

ordinate covalent bonding. The surfaces of these chips are precoded with chromatographic affinity surfaces that extract, structurally modify or amplify a particular protein. ProteinChipArrays have been used to identify disease markers (Xiao *et al.*, 2000). For example, prostate-specific membrane antigen, a protein thought to indicate prostate cancer tumour progression, can be detected in blood sera and quantified, based on a normalised peak, via ProteinChip technology.

The ultimate aim of functional genomics is to integrate information from various 'levels' including DNA sequence, mRNA profiles, protein expression and metabolite concentrations, as well as information about dynamic spatio-temporal changes in these molecules to form effective models of biological system. Attempts have already been made to create whole cell computer simulation models (Tomita, 2001). Rapid accumulation of biological data from genome, proteome, transcriptome and metabolome projects could advance efforts to construct virtual cells in silico. A solid foundation of accurate and complete gene annotation, together with a high quality index of encoded proteins, is a prerequisite for all of this future research.

1.7 Model organisms

1.7.1 Model organism genome projects

The first proposal of the HGP included the study of five model organisms (Watson, 1990). Thus far, the genomes of four of the five initially proposed organisms have been fully sequenced (table 1.2).

Table 1.2: the model organisms initially proposed for genome sequencing (Watson, 1990)

Organism	Genome size	Estimated no. of genes	Sequenced?
Escherichia coli[a]	4.2 x 10^6	4000	Y
Saccharomyces cerevisiae[b]	1.5 x 10^7	6000	Y
Caenorhabditis elegans[c]	1.0 x 10^8	13000	Y
Drosophila melanogaster[d]	1.2 x 10^8	10000	Y
Mus musculus	3.0 x 10^9	30000-100000	Nearly

[a] (Blattner *et al.*, 1997)
[b] (Goffeau , 1996)
[c] (Coulson, 1996)
[d] (Adams , 2000)

In addition, projects are underway to sequence the genome of the rat, zebrafish, and the pufferfish *T nigroviridis* and *T. rubripes*. Plans are also under consideration for sequencing additional primate and other organisms that will help define key developments along the vertebrate and non-vertebrate lineages.

The utility of the genomic sequence of model organisms in comparative sequence analysis is reviewed in section 1.5.3 and chapter IV. Model organisms also play an important role in elucidation of protein function and investigation of human disease (see below).

1.7.2 Functional studies in model organisms

Characterisation of an orthologous gene product through experimental assays in a model organism can offer valuable insights into function. Comparative maps provide the basis for identification of orthologous genes in a variety of organisms. For example, gene mapping placed the murine Pax3 gene on proximal mouse chromosome 1 and made it a candidate for the 'Splotch' mutation (Goulding *et al.*, 1991). When the locus for Waardenburg syndrome type I (WS1) was mapped to the homologous portion of the human genome, 2q37 (Foy *et al.*, 1990), Pax3 became a candidate for WS1 as well. Subsequent molecular studies showed mutations in

the Pax3 gene in 'Splotch' mice (Epstein *et al*., 1991) and individuals with WS1 (Tassabehji *et al.*, 1992).

The nucleotide resolution of the physical maps resulting from the human and model organism genome projects (section 1.7.1) greatly enhances the efficiency of the identification process of candidate disease genes. This may be particularly important in the study of more complex disease traits, such as epilepsy, diabetes and hypertension, which involve more than one locus. These traits are difficult to study in humans because of limited family material, genetic heterogeneity, and complex environmental interactions. In experimental species such as the laboratory mouse and rat, planned crosses can be used, large numbers of progeny can be obtained and environmental factors can be controlled. Experiments can be replicated and gene-gene and gene-environment interactions studied. For example, genetic factors involved in inherited susceptibility to hypertension have been mapped to rat chromosomes 1, 2, 10 and 16 (Deng & Rapp, 1994). In these cases, knowledge of the genomic sequence will provide important clues to identifying homologous susceptibility genes in humans.

The genome projects have resulted in the identification of thousands of novel genes. Many large-scale experiments are underway to systematically study gene function in a more general way using knockouts. In yeast, where each of the ~6000 ORFs is likely to specify a protein product (Goffeau *et al.*, 1996), systematic knockouts of all the ORFs are in progress, either by insertion of transposons (Burns *et al.*, 1994; Ross-Macdonald *et al.*, 1999; Smith *et al.*, 1996), or total deletion of the ORF using a PCR based approach (Baudin *et al.*, 1993). This last method has been refined to include a molecular barcode where each deletion strain is tagged by a unique 20bp sequence (Shoemaker *et al.*, 1996).

A large project is also underway to disrupt all the genes annotated within the *Drosophila* genome using P transposable elements (Deak *et al.*, 1997; Spradling *et al.*, 1995). Similarly, availability of the genome sequence of *C. elegans* has increased the utility of this organism for functional studies. Genes identified within the sequence are easily knocked out using transposons (Collins *et al.*, 1987; Mori *et al.*, 1988; Rushforth *et al.*, 1993; Zwaal *et al.*, 1993) or by double stranded RNA interference (RNAi) (reviewed by Bargmann, 2001).

The increasing availability of mouse genomic sequence is also being exploited by the extensive array of genetic techniques available in mouse. For example, mice can be constructed with pre-determined genetic modifications to the germline by transgenic technology and gene targeting in embryonic stem cells. Recently, RNAi has been demonstrated in mouse cells (Yang *et al.*, 2001). Functional studies in mouse should therefore help to elucidate gene and protein function in humans.

1.8 Bioinformatics

The different large-scale genome related projects have produced a 'tidal wave' of data (Reichhardt, 1999). Bioinformatics has emerged as a science of recent creation, that uses biological data and knowledge stored in computer databases, complemented by computational methods, to aid interpretation of, and derive new, biological knowledge. The development of the World Wide Web (WWW) has accelerated this field as it allows easy access and sharing of data.

Initially, data is collected into databases. Large public domain databases are available for different types of information, for example, EMBL (Baker *et al.*, 2000), TrEMBL for translated

DNA sequences, GenBank for nucleotide and protein sequences and SwissProt for protein sequences (section 1.5.2.2). Individual labs can also use locally maintained databases to store data. For example, ACeDB, a *C. elegans* database, originally developed for the data generated by the *C. elegans* community, has been adapted for data management for many of the human sequencing projects.

A number of tools are available to analyse data. Data retrieval methods can be divided into text based or sequence based retrieval. Examples of text-based retrieval systems are Entrez (Schuler *et al.*, 1996b) which allows access to the data collections at the National Center for Biotechnology (NCBI) and the Sequence Retrieval System (SRS) (Etzold *et al.*, 1996) which allows the exploration of virtually all existing molecular biology databases. Most sequence-based searches are based on pairwise sequence-sequence comparison using algorithms such as BLAST. Similarities and differences are analysed at the nucleotide and/or amino acid level, with the aim to infer structural, functional and evolutionary relationships (Schuler, 1998).

If the sequence of interest contains protein-coding regions, analysis is more sensitive at the protein level as the DNA code is degenerate and, because of selective pressure, protein coding regions are more highly conserved. In general, a set of aligned sequences can be organised into an emerging family to define a ‘profile’. Such profiles aim at capturing the key functionally constrained features of the protein family. As a result, profile-sequence comparisons are a more powerful search tool than sequence-sequence comparisons. Profile-profile comparisons increase the possibility of detecting remotely related family members. In rare cases, discovery of similarities in 3D structure, when it is available, without apparent sequence similarity, can lead to the unification of functional families.

Bioinformatic techniques are also used to find genes. Both comparison-based and predictive methods were discussed previously (section 1.5).

Increasingly, sequence analysis is faced with problems of scale. New sequences become available every day from the various genome projects and processing them by hand is too slow. The flood of new sequence data can be handled only by automation. The genome browser Ensembl (Hubbard & Birney, 2000) provides one example. Ensembl is a software system that produces and maintains automatic annotation on eukaryotic genomes. Currently, *H. Sapiens*, M. *musculus* and *D. melanogaster* Ensembl servers exist. The data can be searched to identify genes, SNPs, proteins and protein families. Currently Ensembl provides identification of 90% of known human genes in the genome sequence and predicts 10,000 additional genes, all with supporting evidence.

Genome sequence provides a static picture. New high-throughput techniques of transcript and protein profiling will soon provide massive amounts of dynamic data, complementing static genome sequences. Database groups are now faced with the challenge of integrating genome sequence data with other data emanating from large-scale molecular biology.

1.9 Chromosome 22

Chromosome 22 is the second smallest of the human autosomes, comprising of 1.6 – 1.8% of the genomic DNA. It is an acrocentric chromosome: the p arm encodes the tandemly repeated rRNA genes and a series of other tandem repeat sequence arrays. There is no evidence to indicate the presence of any protein coding genes on 22p.

There are a number of genetic diseases located to this chromosome, listed in table 1.3.

Table 1.3: Syndromes linked to chromosome 22 genes.

Syndrome	Gene	Position	OMIM
Cat eye syndrome	CECR, CES	22q11	115470
Conotruncal cardiac anomalies	CTHM	22q11	217095
DiGeorge syndrome chromosome region (velocardiofacial syndrome)	DGCR, DGS, VCF	22q11	188400
Thrombophilia due to heparin cofactor II deficiency	HCF2, HC2	22q11	142360
Schindler disease; Kanzaki disease; NAGA deficiency, mild	NAGA	22q11	104170
Rhabdoid tumors; Rhabdoid predisposition syndrome, familial	SMARCB1, SNF5, INI1, RDT	22q11	601607
Breast cancer, t(11-22) associated	?	22q11	600048
Epilepsy, partial, with variable foci	FPEVF	22q11-q12	604364
Epstein syndrome	EPSTS	22q11-q13	153650
Schizophrenia 4	SCZD4	22q11-q13	600850
Glutathioninuria	GGT1, GTG	22q11.1-q11.2	231950
Gamma-glutamyltransferase, familial high serum	GGT2	22q11.1	137181
Bernard-Soulier syndrome, type B; giant platelet disorder, isolated	GP1BB	22q11.2	138720
May-Hegglin anomaly; Fechtner syndrome; Sebastian syndrome; Deafness, autosomal dominant 17	MYH9, MHA, FTNS, DFNA17	22q11.2	160775
Opitz G syndrome, type II	OGS2, BBBG2, GBBB2	22q11.2	145410
Hyperprolinemia, type I	PRODH	22q11.2	239500
Cataract, cerulean, type 2	CRYBB2, CRYB2	22q11.2-q12.2	123620
Agammaglobulinemia, autosomal recessive	IGLL1, IGO, IGL5	22q11.21	146770
Transcobalamin II deficiency	TCN2, TC2	22q11.2-qter	275350
Leukemia, chronic myeloid, Leukemia, acute lymphocytic	BCR, CML, PHL, ALL	22q11.21	151410
Ewing sarcoma; Neuroepithelioma	EWSR1, EWS	22q12	133450
Heme oxygenase-1 deficiency	HMOX1	22q12	141250
Colon cancer (deletions)	?	22q12-qter	
Li-Fraumeni syndrome	CHEK2, RAD53, CHK2, CDS1	22q12.1	604373
Sorsby fundus dystrophy	TIMP3, SFD	22q12.1-q13.2	188826
Neurofibromatosis, type 2; Meningioma, NF2-related, sporadic;Schwannoma, sporadic; Neurolemmomatosis; Malignantmesothelioma, sporadic	NF2	22q12.2	101000
Schwannomatosis	?	22q12.2	162091
Pulmonary alveolar proteinosis	CSF2RB	22q12.2-q13.1	138981
Meningioma	LARGE	22q12.3-q13.1	603590
Meningioma, SIS-related; Dermatofibrosarcoma protuberans; Giant-cell fibroblastoma	PDGFB, SIS	22q12.3-q13.1	190040
Neutrophil immunodeficiency syndrome	RAC2	22q12.3-q13.2	602049
Meningioma	MGCR, MN1	22q12.3-qter	156100
Colorectal cancer	EP300	22q13	602700
Megakaryoblastic leukemia, acute	MKL1, AMKL, MAL	22q13	606078
Spinocerebellar ataxia-10	SCA10	22q13	603516
Cardioencephalomyopathy, fatal infantile, due to cytochrome coxidase deficiency	SCO2	22q13	604272
Waardenburg-Shah syndrome; Yemenite deaf-blindhypopigmentation syndrome	SOX10, WS4	22q13	602229

Male infertility due to acrosin deficiency	ACR	22q13-qter	102480
Ovarian cancer (deletions)	?	22q13.1	
Adenylosuccinase deficiency; Autism, succinylpurinemic	ADSL	22q13.1	103050
Parkinsonism, susceptibility to; Debrisoquine sensitivity	CYP2D, P450C2D	22q13.1	124030
Glucose/galactose malabsorption	SLC5A1, SGLT1	22q13.1	182380
Chromosome 22q13.3 deletion syndrome	PSAP2, PROSAP2, KIAA1650	22q13.3	606230
Metachromatic leukodystrophy	ARSA	22q13.31-qter	250100
Methemoglobinemia, type I; Methemoglobinemia, type II	DIA1	22q13.31-qter	250800
Myoneurogastrointestinal encephalomyopathy syndrome	ECGF1	22q13.32-qter	131222
Megalencephalic leukoencephalopathy with subcortical cysts	MLC1, LVM, VL	22qter	605908

Adapted from OMIM Gene Map (**http://www.ncbi.nlm.gov/htbin-post/Omim**).

Chromosome 22 was the first human chromosome to be completely sequenced by a consortium of labs, providing 33.4Mb of sequence of the euchromatic portion of chromosome 22 in 12 contiguous segments (Dunham *et al.*, 1999). The sequence has been subjected to exhaustive computational analysis (Dunham *et al.*, 1999) and has served as a benchmark for new computational and experimental methods of analysis (for example, de Souza *et al.*, 2000; Mullikin *et al.*, 2000; Roest Crollius *et al.*, 2000; Salamov & Solovyev, 2000; Scherf *et al.*, 2001; Shoemaker *et al.*, 2001).

1.10 This thesis

The huge impact of the human genome project and the availability of genomic sequence can alter approaches to finding and identifying genes. Accurate annotation of genes within the genomic sequence is essential: the genome project will be an important reference for future genetic research and errors in the gene annotation at this early stage could adversely affect future studies of gene and protein function (see section 1.6).

The aim of this project was therefore to generate a highly accurate transcript map of a region of human chromosome 22. The utility of mouse genomic sequence for gene annotation and study of chromosomal evolution was also addressed. The generated human transcript map was then used as a basis for further study of the function of the annotated genes. A variety of bioinformatic and experimental methods were explored to provide a preliminary functional characterisation of the genes encoded within the region of interest.

The thesis will discuss:

1. Sequence analysis of a 3.4 Mb region of chromosome 22 (22q13.31) and the annotation of 39 experimentally verified genes and 17 pseudogenes. Screening of *ab initio* gene predictions in cDNA libraries was carried out to ensure completeness of the transcript map. Northern blotting and creation and screening of a 32 –tissue human cDNA panel, confirmed expression of the annotated gene features. Extensive sequence analysis of the genes and surrounding genomic sequence permitted investigation of translational start sites, polyadenylation sites, splice sites and predicted promoter regions. The correlation of each type of sequence evidence used to generate the transcript map was assessed against the final version, to determine the level of annotation accuracy each approach provided.
2. Comparative study of approximately 3.0 Mb in 22q13.31 and an additional 1.9 Mb region of 22q13.1. This chapter describes the construction of three mouse BAC clone contigs, spanning orthologous regions of mouse chromosomes 15 and 8. The available sequence was used to perform comparative analyses of coding and non-coding regions conserved in both mouse and human genomic sequence. The correlation of the conserved regions with both the existing annotation and with *ab initio* gene predictions

was assessed, but no further genes or exons were identified. Additionally, the study resulted in the refinement of a synteny breakpoint junction on human chromosome 22q13.1 and mouse chromosomes 15 and 8.

3. Preliminary functional characterisation of 27 complete genes identified from the transcript map of 22q13.31. *In silico* analyses were used to identify potential secondary structure and domain features within the predicted protein sequences. Phylogenetic analysis was also utilised to identify orthologous proteins from different species. Additionally, the subcellular localisation of a subset of the proteins was investigated through cloning and expression in a mammalian cell line.

This work was carried out as part of the ongoing work on chromosome 22 at the Sanger Institute.

Chapter II Materials and Methods

All solutions used are listed at the end of this chapter.

2.1 DNA manipulation methods

2.1.1 Polymerase Chain Reaction (PCR)

PCR was performed in 0.5 ml microcentrifuge tubes in a DNA Thermal Cycler (Perkin Elmer) or in a 96 well micro titer plate (Costar Thermowell™ C- or M-type) in an Omnigene (Hybaid) (C-type) or a PTC-225 (MJ Research) (M-type). For most applications, 15 µl reactions were prepared.

1. A premix sufficient for the number of planned reactions was prepared, allowing for a 1X reaction mix once the DNA template (section 2.1.1.2) was added (usually 10 µl of mix and 5 µl of template).
2. The final reaction contained 1X buffer (Buffer 1 unless otherwise specified), 200 µM of each of the four nucleotides (Pharmacia), 40 ng of each primer, and 0.5 units /µl of DNA polymerase (*Taq* (Amplitaq) or *Pfu* (Promega)).
3. The amplifications were performed under the same cycling profile (except where specified): 94°C for 5 minutes, followed by 35 cycles at 94°C for 30 seconds, annealing temperature (specific to each primer) for 30 seconds and 72°C for 30 seconds, and finally followed by 1 cycle at 72°C for 5 minutes.
4. Reaction products were visualised by agarose gel electrophoresis and staining with ethidium bromide (section 2.1.2).

DNA templates

The templates used were

1. Bacterial colonies picked into 100 µl of sterile water and 5 µl used directly.
2. cDNA or bacterial pools.

3. DNA excised from an agarose gel into 100 µl sterile water, left overnight and 5 µl used directly.
4. Human (Sigma D-3035) or mouse genomic DNA at 12.5 ng/µl.

2.1.2 Gel electrophoresis

1. An agarose gel was prepared (2.5% for most PCR amplified products and 1% for fragments over 1 kb) in 1X TBE and ethidium bromide (250 ng/µl).
2. DNA was added to the appropriate amount of 6X loading buffer (e.g. 5 µl of PCR product and 1 µl of 6X loading buffer) and loaded. In the case of Buffer 2, the samples were loaded directly.
3. Size markers (1 kb ladder, Gibco-BRL) were also loaded.
4. Minigels were run at 80 Volts for 10-15 minutes and larger gels were run at 200 Volts for the time required to obtain satisfactory separation.
5. DNA was visualised under UV on a transilluminator and photographed with a Polaroid camera.

2.1.3 Restriction enzyme digests

2.1.3.1 Liquid DNA

1. Up to 10 µg of bacterial clone, or plasmid, DNA was used in a reaction containing the appropriate 1X buffer, 1mM spermidine, 100 µg/ml BSA and 20-50 units of the appropriate enzyme.
2. The DNA was digested for 2 hours or overnight at the appropriate temperature for the enzyme.
3. The DNA was subjected to agarose gel electrophoresis and visualised (section 2.1.2).

2.1.3.2 PCR products

1. After PCR amplification, the required amount (usually 5-10 µl) was transferred to a new 0.5 ml microcentrifuge, and 5 units of the restriction enzyme added.
2. DNA was digested for 1 hour at the recommended temperature and visualised by gel electrophoresis.

2.1.4 DNA purification

2.1.4.1 Ethanol precipitation

1. In a 1.5 ml microcentrifuge tube, 0.1 volumes of 3M sodium acetate and either 1 volume of isopropanol or two and half volumes of ethanol were added to the DNA.
2. The samples were mixed well by vortexing and incubated for 20 minutes at –20°C.
3. The DNA was then pelleted in a microcentrifuge at 13,000 g and washed once with 70% ethanol.
4. The pellet was left to dry and then resuspended in the appropriate amount of $T_{0.1}E$.
5. The recovery was tested by gel electrophoresis (section 2.1.2).

2.1.4.2 Gel purification

The DNA fragment was excised from the agarose gel with a clean scalpel.

1. The gel slice was weighed in a 1.0 ml eppendorf tube.
2. The gel slice was then purified using a Qiaquick Gel Extraction Kit™ (Qiagen) according to the manufacturers instructions.
3. The recovery was tested by gel electrophoresis (section 2.1.2).

2.1.4.3 ExoSAP purification of PCR products

1. A premix, sufficient for the number of planned reactions, was prepared, allowing for a 1X reaction mix once the PCR reaction was added (usually a 15 µl PCR reaction volume).

2. The final reaction contained 1X reaction buffer (SAP Buffer), 1X Dilution buffer, 1 unit/µl of Shrimp Alkaline Phosphatase (USB) and 1unit/µl exonuclease I (USB).

3. The mixture was incubated at 37°C for 30 minutes, followed by 80°C for 15 minutes.

2.2 Clone resources

2.2.1 Libraries used

Different types of clone resources have been used throughout this project. The Sanger Institute clone resource group, who also provided the arrayed filters and PCR pools for screening, maintains the clone resources.

2.2.1.1 Bacterial clone libraries

The RPCI-23 female (C57Bl/6J) mouse BAC library (Osoegawa *et al.*, 2000) was screened in this study. Library details are shown in table 2.1.

Table 2.1: Details of the mouse genomic library used

Library	Library type	Library code	Antibiotic	Vector	Cloning site	Genomic digest
RPCI-23	BAC	bM	Chloramphenicol 12.5 µg/ml	pBACe3.6	*Eco*RI	*Eco*RI

2.2.1.2 cDNA libraries

cDNA libraries used during the course of this project are described in table 2.2.

Table 2.2: cDNA resources used during the course of this project

cDNA library code	cDNA library description	Source/reference	Vector	Vectorette PCR
T	Adult testis	CLONTECH	pCDM8	+
FB	Fetal Brain	Invitrogen	pcDNAI	+
FL	Fetal Liver	Invitrogen	pcDNAI	+
FLu	Fetal Lung	Invitrogen	pcDNAI	+
HL60	Peripheral blood	Invitrogen	pcDNAI	+
AH	Adult Heart	Invitrogen	pcDNA3	+
ALu	Adult Lung	CLONTECH	pcDNAI	+
SK-N-MC	Neuroblastoma cells	Invitrogen	pcDNAI	-
PF	Adult brain	Pfizer	pcDNAI	-
U937+*	(Monocyte -NOT activated- from a patient with promonocytic leukaemia)	Simmons (1993)	pCDM8	-
U937AC*T	(Monocyte -PMA activated- from a patient with promonocytic leukaemia)	Simmons (1993)	pCDM8	-
H9*	Placental, full tem normal pregnancy	Simmons (1993)	pH3M	-
YT*	HTLV-1 +ve adult leukaemia T cell	Simmons (1993)	pH3M	-
NK*	Natural killer cell	Simmons (1993)	pH3M	-
Daudi*	B lymphoma	Simmons (1993)	pH3M	-
HPBall*	T cell from a patient with acute lymphocytic leukaemia	Simmons (1993)	pH3M	-
BM*	Bone marrow	Simmons (1993)	pH3M	-
DX3*	Melanoma	Simmons (1993)	pH3M	-

* Generously provided by David Simmons, Oxford (Simmons, 1993).
+ Screened by vectorette PCR. Remaining libraries are available to screen by single sided specificity PCR (Huang *et al*., 1993).

2.2.2 cDNA clone synthesis

2.2.2.1 Nested PCR

1. A premix sufficient for the number of planned reactions was prepared, allowing for a 1X reaction mix once the DNA template (1 µl cDNA from RT-PCR (section 2.5.5) or 1.5 µl of first round reaction mix) was added.
2. The final reaction volume of 25 µl contained 1X *Pfu* DNA polymerase 10X buffer with $MgSO_4$ (Promega) (PCR buffer 2), 200 µM each of the four nucleotides (Amersham Pharmacia Biotech), 40 ng of each primer and 3 units/µl of *Pfu* DNA polymerase (Promega).
3. First round amplification was performed with the external pair of nested primers, under the cycling profile: 95°C for 2 minutes, followed by 35 cycles at 95°C for 45 seconds, annealing temperature (specific to each primer) for 30 seconds and 72°C for 3 minutes, finally followed by 1 cycle at 72°C for 5 minutes.
4. Two volumes of deionised water were added to 1 volume of the reaction mix. 1.5 µl of the diluted reaction mix was used as a template for second round PCR.
5. Steps 1 to 3 were repeated with the internal set of nested primers.
6. Reaction products were visualised by agarose gel electrophoresis and staining with ethidium bromide (section 2.1.2).
7. Selected bands were cut out and the DNA extracted from the gel (section 2.1.4).

The PCR product was ligated into the pGEM®-T Easy vector (Promega) according to the manufacturers' instructions and used to transform competent cells (section 2.2.2.5).

2.2.2.2 Addition of T7 tag

A schematic of the overall strategy is shown in chapter V.

C-terminus

1. PCR (section 2.1.1) utilising the proofreading enzyme *Pfu* (Promega) was used to introduce suitable restriction sites flanking the open reading frame of the cDNA clone. The designed primers (section 2.4.1) also incorporated a Kozak consensus sequence (Kozak, 1987) prior to the start codon and removed the stop codon.
2. The PCR products were then digested (section 2.1.3) and subcloned (section 2.2.2.4) into pBlue-CT7 (a kind gift from Dr Begoña Aguado (HGMP Resource Centre)).
3. The plasmid was then digested using suitable restriction enzymes and the cDNA plus C-terminal T7.Tag was subcloned (section 2.2.2.4) into the expression vector pCDNA3 (Invitrogen) (a kind gift from B Aguado).
4. The plasmid was then miniprepped (section 2.2.2.6) and the insert sequenced (Elizabeth Huckle, Sanger Institute) using appropriate oligonucleotides.

N-terminus

1. PCR (section 2.1.1), utilising the proofreading enzyme *Pfu* (Promega), was used to introduce suitable restriction sites flanking the open reading frame of the cDNA clone. The designed primers (section 2.4.1) retained the stop codon.
2. The PCR products were digested (Section 2.1.3) and subcloned (section 2.2.2.4) into pCDNA3-NT7 (section 2.2.2.3).
5. The plasmid was then miniprepped (section 2.2.2.6) and the insert sequenced (E. Huckle) using appropriate oligonucleotides.

2.2.2.3 Generation of pCDNA-NT7

A new expression vector containing a *Not*I restriction site and preceded by the T7.Tag sequence was created for used in mammalian cell expression systems. A schematic of the vector is shown in chapter V.

1. The vector pBlue-NT7 (a kind gift from B Aguado) was digested with *Bam*HI.
2. The complementary oligonucleotide 5'-GAT CCA GCG GCC GCT G-3' was heated to 65°C for 10 minutes and then snap cooled on ice. The oligonucleotide was then subcloned into pBlue-NT7 (section 2.2.2.4).
3. PCR (section 2.1.1) utilising the proofreading enzyme *Pfu* (Promega) was used to introduce suitable restriction sites (*Hind*III and *Xba*I) flanking the open reading frame of the T7.Tag. The designed primers (section 2.8.4.2) also incorporated a Kozak consensus sequence (Kozak, 1987) prior to the start codon.
4. The PCR product was subcloned (section 2.2.2.4) into pCDNA3 (Invitrogen) (a kind gift from B. Aguado). As *Xba*I is a methyl-sensitive restriction enzyme, the plasmid was transformed into the *E. coli* strain INV110 (Invitrogen), which has disrupted *dam* and *dcm* genes. Methylation of the plasmid *Xba*I site is thus prevented.
5. The plasmid was then miniprepped (section 2.2.2.F) and the insert sequenced (E. Huckle) using appropriate oligonucleotides to confirm its integrity.

2.2.2.4 Subcloning

1. The vector and insert to be used were digested with the appropriate restriction enzymes (section 2.1.3).
2. The restriction products were gel purified (section 2.1.4). Concentration of the vector and insert was estimated from appearance on the gel against the size markers.
3. An approximate 3:1 molar ratio of the insert and vector (roughly 150 ng of insert: 50 ng of vector) was ligated together in a final reaction volume of 10 µl, including 1 µl of 10X Ligation buffer, and 1 unit of T4 DNA ligase (Roche).
4. The ligation reaction was incubated at 4°C overnight.

2.2.2.5 Transformation

1. 2 µl of the ligation reaction was added to a sterile 1.5 ml microcentrifuge tube on ice.
2. Tube(s) of frozen JM109 High Efficiency Competent Cells (Promega) were removed from –70°C storage and thawed on ice.
3. 50 µl of cells were carefully transferred into each tube and gently flicked to mix. The tubes were incubated on ice for 20 minutes.
4. The cells were heat-shocked for 45 seconds in a water bath at 42°C. The tubes were then immediately returned to ice for 2 minutes.
5. 950 µl of LB broth was added to the tubes, which were then incubated for 1.5 hours at 37°C, 150 rpm.
6. 100 µl of each transformation reaction was plated onto duplicate LB/ 100 µg/ml ampicillin/ 0.5 mM IPTG/ 80 µg/ml X-Gal plates.
7. The plates were incubated overnight at 37°C.
8. White colonies were picked into 100 µl sterile water and used as a PCR template in order confirm identity of the insert using appropriate primers (section 2.1.1).

2.2.2.6 Bacterial clone minipreps

1. A single colony was inoculated into 10 ml of LB broth containing the appropriate antibiotic and grown overnight at 37°C, shaking at 250 rpm.
2. The cells were pelleted at 1500 g and the media fully drained. It was important to ensure that all the media was removed at this stage to prevent inhibition of digestion.
3. The pellet was resuspended in 200 µl of GTE on ice.
4. On ice, 400 µl of fresh 0.2M NaOH and 1% SDS was added with gentle inversion of the tube. The tube was left on ice for 5 minutes.
5. 300 µl of 5M Acetate, 3M K^+ was added. The tube was gently inverted to mix and then incubated on ice for 10 minutes.

6. The precipitate was pelleted in a microcentrifuge at 13,000 g.
7. The supernatant was transferred to a new tube. If the supernatant was still cloudy, step 6 was repeated.
8. 600 µl of cold isopropanol was added to the cleared supernatant.
9. The DNA was pelleted in a microcentrifuge at 13,000g.
10. The DNA pellet was resuspended in 200 µl TE and extracted with 200 µl phenol/isoamyl-alcohol/chloroform (25:1:24).
11. The DNA was ethanol precipitated (section 2.1.4)
12. The DNA was pelleted in a microcentrifuge and washed with 70% ethanol.
13. The DNA was resuspended in 30 µl $T_{0.1}E$ and stored at –20°C.

2.2.2.6 Bacterial clone micropreps

1. A single colony was inoculated into 500 µl of LB broth containing the appropriate antibiotic and grown overnight at 37°C shaking at 300 rpm.
2. 250 µl of culture was aliquoted into a 96 well round-bottom plate (Costar).
3. The cells were pelleted at 2500g for 4 minutes. The plate was inverted to drain the supernatant.
4. The pellet was resuspended in 25 µl.
5. On ice, 25 µl of fresh 0.2M NaOH and 1% SDS was added and mixed by gently tapping. The plate was left on ice for 5 minutes.
6. 25 µl of 5M Acetate, 3M K^+ was added and mixed by tapping gently. The plate was then incubated on ice for 10 minutes.
7. The well contents were transferred to a 96 well filter-bottom plate.
8. The filter plate was taped on top of a 96 well round-bottom plate containing 100 µl of isopropanol.

9. The precipitate was pelleted at 2500 rpm for 2 minutes and the filter plate discarded. The round-bottom plate was incubated at room temperature for 30 minutes and then spun at 3200 rpm for 20 minutes at room temperature.
10. Supernatant was removed by inverting the plate.
11. 100 µl of 70% ethanol was added to each well and the plate tapped gently. The plate was then spun at 3200 rpm, for 20 minutes at room temperature.
12. The supernatant was removed by inversion and the pellet dried at room temperature.
13. The DNA was resuspended in 5 µl of $T_{0.1}E$ with RNase (10 µl of 1 mg/ml RNase per 1 ml of $T_{0.1}E$).

2.2.3 Vectorette Library Synthesis

2.2.3.1 Library titration

1. The cDNA library, consisting of bacteria stored in glycerol, was defrosted on ice. 2 µl of the library was diluted in 198 µl LB. Six tenfold serial dilutions were prepared and 100 µl of each plated onto LB agar plates containing the appropriate antibiotic.
2. The plates were left inverted at 37°C for 4 hours. They were then transferred to 30°C for 16 hours and then back to 37°C for an additional 4 hours to promote growth of separated colonies.
3. The colonies were counted and the library titre estimated.

2.2.3.2 High density liquid pools

1. Ten tubes of 20 ml LB, with the appropriate antibiotic, were prepared.
2. 250,000 clones, diluted in LB, were added to each tube (2,500,000 clones in total).
3. The clones were grown at 37°C for 20 hours at 240 rpm.

4. Meanwhile, dilutions of 1000, 100, 10 and 1 clone(s) were plated onto LB agar plates containing the appropriate antibiotic to check the titration. The plates were inverted and left to grow overnight at 37°C.

2.2.3.3 Low density plated pools

1. 25 X 20,000 clones were plates onto LB agar plates with Hybond N+ filters (500,000 clones in total).
2. The plates were then inverted and the clones left to grow for 4 hours at 37°C, followed by 16 hours at 30°C and a further 4 hours at 37°C.

2.2.3.4 DNA extraction

1. Filters were rolled up and put in a 50 ml falcon tube with 20 ml of SET. The cells were shaken off and the filters removed.
2. DNA was extracted using the bacterial clone miniprep protocol (section 2.2.2.6).
3. 1 µl of RNase (10 ng/ml) was added to the extracted DNA and incubated at 37°C for 1 hour.
4. 0.01 µl of the DNA was visualised by gel electrophoresis to check the extraction outcome.

2.2.3.5 Library preparation

1. 1 mg of DNA was digested with the appropriate enzyme in 30 µl (section 2.1.3).
2. 70 µl of water was added and the DNA extracted with 100 µl phenol/isoamyl-alcohol/chloroform (25:1:24).
3. The DNA was ethanol precipitated (section 2.1.4).
4. The DNA was pelleted in a microcentrifuge and washed with 70% ethanol.
5. The DNA pellet was resuspended with 100 µl of ligation buffer.

6. 10 µl of 1pmol/µl annealed vectorette vectorette bubbles appropriate to the enzyme used, 1.1 µl adenosine 5'-triphosphate and 2.5 units of T4 DNA Ligase (Amersham Pharmacia Biotech) were added.
7. The tubes were incubated at 16°C overnight.
8. The mixes were diluted to 500 µl to generate Stock Pools
9. Equal volumes of sets of five plates Stock Pools were mixed to generate stocks of Super Pools.
10. 1/100 dilutions of stock Super Pools were prepared using $T_{0.1}E$. These pools were screened in the first round of PCR (section 2.3.3).
11. 1/100 dilutions of the plated Stock Pools were prepared using $T_{0.1}E$. These pools were screened in the second round of PCR (section 2.3.3).
12. 1/10 dilutions of the plated and liquid Stock Pools were prepared using $T_{0.1}E$. These pools were used for both PCR pool screening and vectorette PCR (section 2.3.3).

2.3 Screening

2.3.1 Probe labelling

DNA probes were labelled, either by PCR, or by random hexamer labelling (Feinberg & Vogelstein, 1983; Hodgson & Fisk, 1987).

2.3.1.1 PCR labelling of STSs

1. The required fragment was amplified from either genomic DNA or cDNA as appropriate.
2. The fragments were separated on a 2.5% gel (section 2.1.2), cut out and transferred to a microcentrifuge tube containing 100 µl of sterile deionised water. The DNA was allowed to diffuse out of the gel slice (at least one hour).

3. Using PCR buffer 1, 9 µl reactions were set up containing the required primers, 2 µl of the liquid surrounding the gel slice, nucleotides (except dCTP) and DNA polymerase. A single drop of mineral oil was placed on top of the reaction mixture.
4. 1 µl of [α-^{32}P] dCTP (3000 Ci/mol, Amersham Pharmacia Biotech) was added.
5. PCR was performed in a DNA Thermal Cycler (Perkin Elmer) under the following cycling profile, 94°C for 5 minutes, 20 cycles of 93°C for 30 seconds, 55°C for 30 seconds, 72°C for 30 seconds and 1 cycle of 72°C for 5 minutes.
6. After completion, the probes were either denatured in the thermal cycler at 99°C for 5 minutes and then snap chilled, or were added to competitor DNA (section 2.3.1.3) and denatured in a boiling water bath for 5 minutes, before being snap chilled.
7. The probes were then added to the hybridisation mix.

2.3.1.2 Random hexamer labelling for probes

1. 20 ng of DNA was added to a microcentrifuge tube. The volume was made up to 17.5 µl with sterile water.
2. The DNA was denatured in a boiling waterbath for 5 minute, then snap chilled on ice.
3. 5 µl of OLB3, 1 µl of 10 mg/ml BSA, 0.5 units of Klenow DNA polymerase (GibcoBRL) and 0.5 µl of [α-^{32}P] dCTP (3000 Ci/mol, Amersham Pharmacia Biotech) were added to the tube.
4. The reactions were incubated at room temperature for 3 hours and then denatured in a boiling bath for 5 minutes.

2.3.1.3 Competitive reassociation of radio labelled probe

Many of the probes designed from mouse BAC end sequences were of relatively uncharacterised genomic content and were likely to contain repetitive DNA. This was

suppressed with mouse genomic DNA (Sealey *et al.*, 1985). Potential microsatellite sequences were also competed with poly CA/GT (Pharmacia 27-7940).

1. 125 µl of mouse DNA (10 mg/ml, a kind gift from George Stavrides, Sanger Institute), 125 µl of 20X SSC, 5 µl of poly CA/GT (1 mg/ml), was added to a screw cap microfuge tube and the total volume of the mixture made up to 0.5 ml of water minus the volume of the labelling reaction.
2. The labelled DNA was added to the competition mixture and the tube boiled in a water bath for 5 minutes.
3. The tube was snap chilled on ice and the probe was then added to the hybridisation buffer.

2.3.2 Library screening

2.3.2.1 Screening pools by PCR

Using PCR buffer 1, sufficient reaction mix was set up, containing the required primers, nucleotides and 5µl of DNA from the pools making a total of 15 µl, to screen all the pools for the library.

1. PCR amplification was performed under the cycling profile of 94°C for 5 minutes, 35 cycles of 93°C for 30 seconds, annealing temperature specific for each set of primers for 30 seconds, 72°C for 30 seconds followed by 1 cycle of 72°C for 5 minutes.
2. The reaction products were visualised by gel electrophoresis (section 2.1.2).

2.3.2.2 Screening of library filters by hybridisation of PCR-labelled probes

Gridded filters of both the mouse RPCI-23 BAC library and the region-specific subset of bacterial clones were generated by the Sanger Institute clone resource group. These were then screened.

1. STSs were labelled as described (section 2.3.1).

2. Up to 30 filters were sequentially placed in a 15x10x5 cm sandwich box with sufficient hybridisation buffer to cover the filters. A plastic sheet, cut to size, was placed on top to reduce evaporation. The filters were pre-hybridised at 65°C for 2 hours (Innova 4000, New Brunswick Scientific).
3. The filters were removed and the denatured probe was added to hybridisation solution in the box and mixed.
4. The filters were added one by one back into the box and each was carefully submerged under the hybridisation mix. The plastic sheet was replaced on top.
5. After hybridisation overnight at 65°C, the filters were washed by rinsing twice in 2X SSC at room temperature for 5 minutes. The filters were then washed twice in 0.5X SSC and 1% N-lauroyl-sarcosine at 65°C for 30 minutes, before rinsing twice in 0.2X SSC at room temperature for 5 minutes.
6. The washed filters were wrapped in Saran wrap (Dow Chemical Co.) and exposed to pre-flashed Fuji Medical X-ray film (036010) (or equivalent) overnight with two intensifying screens at –70°C in the first instance. This was repeated if longer or shorter exposures were needed.
7. Occasionally filters were re-washed to 0.2X SSC with 1% N-lauroyl-sarcosine at 65°C for 30 minutes if required (i.e. high background).
8. The autoradiographs were developed and labelled with the name of the filter and the data entered into 22ace (section 2.7.1).

2.3.3 Vectorette PCR

This method was adapted from the original (Riley *et al.*, 1990) to isolate cDNA fragments from a cDNA library. The method was developed by Dr. John Collins (Sanger Institute) and libraries were made by J. Collins, G. Stavrides and M. Goward (section 2.2.3).

2.3.3.1 Identification of positive pools

1. For each library, 5 super pools were screened with the appropriate STS.
2. For each positive super pool, the 5 individual constituent pools were screened with the appropriate STS.

2.3.3.2 First round PCR

1. Using PCR buffer 3, 15 µl reactions were set up using 5 µl of a 10X concentrated PCR pool.
2. The reaction was transferred to the Omnigene (Hybaid) and incubated for 1 cycle at 94°C for 5 minutes.
3. After 4 minutes, the program was paused and 1 µl of enzyme mix (0.12 µl *Taq* (Amplitaq), 0.12 µl *Taq*Extender (Promega), 0.12 µl PerfectMatch (Promega) and 0.64 µl sterile water) was added. The cycling was then allowed to continue, for 1 minute at 94°C, 17 cycles of 94°C for 5 minutes, 68°C for 30 seconds and 72°C for 3 minutes, followed by 18 cycles of 94°C for 5 cycles, 60°C for 30 seconds and 72°C for 5 minutes and finally 72°C for 5 minutes.
4. The fragments were then visualised by gel electrophoresis on a 1% agarose gel (section 2.1.2) and the fragments were cut out and stored overnight in 100 µl of sterile water at 4°C.

2.3.3.3 Second round PCR

1. To obtain DNA for sequencing the liquid from around the gel slice was reamplified using standard PCR conditions. Four 15 µl reactions were used to obtain sufficient DNA for sequencing.

2. The amplification was performed under the following cycling conditions, 1 cycle of 95°C for 5 seconds followed by 20 cycles of 94°C for 5 seconds, 60°C for 30 seconds, 72°C for 3 minutes and finally 1 cycle of 72°C for 5 minutes.
3. The pooled reactions were separated by gel electrophoresis on a 1% agarose gel (section 2.1.2). The fragments were cut out and purified using the Qiaquick Gel Extraction Kit™ (Qiagen) according to the manufacturers instructions.
4. The templates and oligonucleotides were used in cycle sequencing (E. Huckle).

2.4 Landmark production

2.4.1 Primer design

The primers were designed using the Primer3 program (Rozen and Skaletsky, 1998; http://www.genome.wi.mit.edu/genome_software/other/primer3.html) from http://www.sanger.ac.uk/cgi-bin/primer3.cgi. Additional primers for amplification and subsequent cloning of full-length cDNA (section 2.2.2) were designed manually.

2.4.2 Primer synthesis

1. Primers were synthesised at the Sanger Institute by David Frazer. A subset of the primers were synthesised by GenSet. Primer concentrations were supplied in both cases.
2. Primers were stored at –20°C and working dilutions for PCR prepared at 100ng/µl for each primer in pairs.
3. The primers were tested at three different annealing temperatures, 55°C, 65°C and 65°C, using the standard cycling on Thermal cyclers to establish optimal PCR conditions.

2.4.3 Fingerprinting

*Hind*III fingerprinting of bacterial clones was performed with the help of Owen McCann (Sanger Institute), using the standard protocol below (Marra *et al.*, 1997).

2.4.3.1. Digestion

1. Bacterial clones were microprepped (section 2.2.2.6) by Carol Carder (Sanger Institute) or M. Goward.
2. 2.6 µl of water, 0.9 µl of the appropriate buffer and 20 units of *Hind*III (Boerhinger) were added to each well, mixed by gentle tapping and then the plate centrifuged up to 1000 g to collect the contents.
3. The plate was incubated at 37°C for 2 hours.
4. The reaction was terminated by addition of 2 µl of 6X Dye Buffer II and the plate centrifuged up to 1000 g to collect the contents.

2.4.3.2 Gel preparation and loading

1. A 1% gel mix was prepared using 450 ml of 1X TAE and 4.5g agarose and poured at 4°C. A 121-well comb was placed in the gel and allowed to set for 45 minutes. The comb was then removed.
2. 3-4 l of 1X TAE was added to the gel tank.
3. 0.8 µl of the marker (Promega, DG1931) was loaded in the first well and then in every fifth well.
4. 1.0 µl of each sample was then loaded into the empty wells.
5. The gel was run at 90 V for 30 minutes at room temperature. Once the dye front had advanced beyond the wells, the gel tanks were transferred to a refrigerated room and run at 4°C for 15 hours at 90 V.

2.4.3.3 Gel staining

1. The gel was trimmed to ~19 cm and stained with vistra green stain for 45 minutes. The gel was covered whilst staining to prevent degradation of the vistra green.

2. The gel was then rinsed with 0.5 l of deionised water. The gel was visualised and the image recorded using a Molecular Dynamics scanner.

2.4.4 SNP verification

1. Candidate SNPs were identified by comparison of cDNA sequence to the genomic DNA. Primers were designed flanking approximately 400 bp of sequence surrounding the candidate variant.
2. Fragments containing the SNP of interest were amplified from the DNA of 24 individuals (Set M24PDR of 24 human DNAs from the Coriell cell repository).
3. Amplification was tested by electrophoresis (section 2.1.2) of 5 µl of the product.
4. The PCR products were purified using the ExoSAP protocol (section 2.1.4).
5. The recovery of the purification was tested by electrophoresis (section 2.1.2).
6. The fragments and correct primers were used in cycle sequencing (E. Huckle).
7. The resultant sequences were aligned in a Gap4 database (Dr. Kate Rice, Sanger Institute).

2.5 RNA manipulation

2.5.1 Steps taken to limit contamination with RNase

Autoclaved plasticware (tubes, pipette tips etc.) was used and bench surfaces, racks etc. were cleaned before use with RNase*Zap*® (Ambion).

All reagents for RNA work were made up with Diethylene Pyrocarbonate (DEPC) water.

1. 0.1% DEPC in deionised water was mixed, and left overnight in a fume hood.

2. The DEPC water was autoclaved before use.

2.5.2 RNA resources

RNA used in this project was obtained from a number of sources.

Table 2.3: RNA resources used during the course of this project

	Tissue	Source	Supplied as		Tissue	Source	Supplied as
A	Heart	Clontech	Human MTN™ Blot (7760-1)	11	Spleen	Stratagene	Total RNA
B	Brain (whole)	Clontech	Human MTN™ Blot (7760-1)	12	Stomach	Stratagene	Total RNA
C	Placenta	Clontech	Human MTN™ Blot (7760-1)	13	Colon I	Stratagene	Total RNA
D	Lung	Clontech	Human MTN™ Blot (7760-1)	14	Colon II	*	Tissue
E	Liver	Clontech	Human MTN™ Blot (7760-1)	15	Rectum	Stratagene	Total RNA
F	Skeletal muscle	Clontech	Human MTN™ Blot (7760-1)	16	Breast	Stratagene	Total RNA
G	Kidney	Clontech	Human MTN™ Blot (7760-1)	17	Ovary	*	Tissue
H	Pancreas	Clontech	Human MTN™ Blot (7760-1)	18	Uterus	Stratagene	Total RNA
I	Fetal brain	Clontech	Human MTN™ Blot (7756-1)	19	Cervix I	Stratagene	Total RNA
J	Fetal lung	Clontech	Human MTN™ Blot (7756-1)	20	Cervix II	*	Tissue
K	Fetal liver	Clontech	Human MTN™ Blot (7756-1)	21	Testis I	Clontech	Total RNA
L	Fetal kidney	Clontech	Human MTN™ Blot (7756-1)	22	Testis II	Invitrogen	Total RNA
1	Kidney I	Invitrogen	Total mRNA	23	Fetal brain I	Stratagene	Total RNA
2	Kidney II	Clontech	Total mRNA	24	Fetal brain II	Clontech	Total RNA
3	Liver I	Stratagene	Total mRNA	25	Fetal heart I	Stratagene	Total RNA
4	Liver II	*	Tissue	26	Fetal heart II	Stratagene	Total RNA
5	Cerebrum	*	Tissue	27	Fetal liver I	Stratagene	Total RNA
6	Skeletal muscle	*	Tissue	28	Fetal liver II	Stratagene	Total RNA
7	Skin	*	Tissue	29	Fetal lung I	Stratagene	Total RNA
8	Tonsil	*	Tissue	30	Fetal lung II	Stratagene	Total RNA
9	Lymphoblast cell line	#	Harvested cells	31	Fetal spleen	Stratagene	Total RNA
10	Thyroid	Stratagene	Total RNA	32	Fetal bladder	Stratagene	Total RNA

* Supplied as tissue, from Tissue Bank, Department of Histopathology, Addenbrookes Hospital, Cambridge.
# Cells supplied as a kind gift from Dr Nigel Carter, Sanger Institute.

2.5.3 RNA isolation

Total RNA was isolated from human tissue samples and cell lines after homogenisation in TRIzol reagent (Chomczynski & Sacchi, 1987).

1. The sample was homogenised in 1 ml of TRIzol reagent per 50-100 mg of tissue, or 10^7 cells.

2. The homogenised sample was incubated at room temperature for 5 minutes.
3. 0.2 ml of chloroform per 1 ml of TRIzol reagent was then added. The tube was shaken vigorously for 15 seconds and incubated at room temperature for 2-3 minutes.
4. The tube was centrifuged at no more than 12000 g for 15 minutes at 4°C.
5. The aqueous upper phase was transferred to a new tube and 0.5 ml of isopropanol per 1 ml of TRIzol reagent used was added.
6. The tube was incubated at room temperature for 10 minutes and then centrifuged at no more than 12000 g for 15 minutes at 4°C.
7. The supernatant was removed and the pellet washed once with 75% ethanol, adding at least 1 ml per 1ml of TRIzol reagent used.
8. The tube was centrifuged at 7500 g for 5 minutes at 4°C.
9. The pellet was dried at room temperature and resuspended in 100 µl of DEPC water. The sample was heated to 55°C for 10 minutes, then stored in 75% ethanol at –70°C.

2.5.4 Ethanol precipitation

1. 0.025 volumes of 3M sodium acetate were added to 1 volume of RNA in 75% ethanol.
2. The samples were mixed well by vortexing and incubated for 20 minutes at –70°C.
3. The RNA was then pelleted in a microcentrifuge at 4°C at 13,000g and washed once with 70% ethanol.
4. The pellet was left to dry at room temperature.

2.5.5 Reverse Transcription PCR (RT-PCR)

2.5.5.1 Preparation of RNA sample prior to RT-PCR

1. 10 µg of RNA was ethanol precipitated (section 2.5.3).
2. The RNA pellet was resuspended in 79 µl of DEPC water.

3. 10 µl of DNase I buffer (GibcoBRL), 1 µl DNase I (GibcoBRL) and 1 µl RNAguard (Amersham Pharmacia Biotech) was added to the tube and incubated at room temperature for 15 minutes.
4. 10 µl of 25 mM EDTA was added to stop the reaction. The tube was then incubated at 65°C for 10 minutes.
5. The tube was briefly chilled on ice and the RNA ethanol precipitated (section 2.5.3).

2.5.5.2 First strand cDNA synthesis

1. To a 10 µg RNA pellet, 2 µl of oligo (dT) (500 ML/ml) was added and the solution was made up to 24 µl with DEPC water.
2. The mixture was heated to 70°C for 10 minutes and then chilled briefly on ice. The contents of the tube were collected by brief centrifugation.
3. 8 µl of First Strand buffer (GibcoBRL), 4 µl of DTT (0.1M) (GibcoBRL), and 2 µl of dNTP mix (10 mM) were added. The tube contents were mixed gently and incubated at 42°C for 2 minutes.
4. 400 units of reverse transcriptase (SuperScript II, GibcoBRL) was added to the reaction and mixed by gentle pipetting.
5. The reaction was incubated at 42°C for 50 minutes.
6. The reaction was then inactivated by heating at 70°C for 15 minutes.
7. RNA complementary to the DNA was removed by addition of 2 units of E. coli RNase H (Promega) and incubating at 37°C for 20 minutes.
8. The cDNA was then used as a template in PCR (section 2.1.1).

2.5.6 Northern blotting

2.5.6.1 Probe generation

1. Probes were generated by PCR from cDNA templates (table 2.2), using primers designed to flank intronic sequence where possible.
2. The STS was radiolabelled by PCR (section 2.3.1).
3. β-actin control probes (Clontech) were radiolabelled by random hexamer labelling (section 2.3.1).

2.5.6.2 Hybridisation

The human Multiple Tissue Northern (MTN) blots (Nos. 7760-1 and 7756-1; Clontech) contain 2 μg of poly (A) mRNAs from different adult and fetal human tissues.

1. The blots were pre-hybridised for 1 hour and then hybridised for 18 hours at 65°C in hybridisation buffer.
2. The blots were washed twice in 2X SSC, 0.05% SDS for 10 minutes at room temperature, then twice in 0.1X SSC, 0.1% SDS for 10 minutes at 50°C.
3. The blots were then subjected to autoradiography at –70°C for an average of 3 days.

2.6 Cell Culture and Protein Manipulation

2.6.1 SDS-PAGE

SDS-PAGE was carried out using a Mini-PROTEAN®Electrophoresis cell (Biorad).

2.6.1.1 Gel preparation

1. The gel unit was assembled according to the manufacturer's instructions
2. A separating gel mix was prepared (12% separating gel was prepared for SDS treated proteins in the approximate molecular weight range of 10-100 k Daltons; lower or higher percentage gels were prepared as required) from a 30% acrylamide/bis stock

(Severn Biotech), containing 0.375M Tris-HCl (pH 8.8), 0.1% SDS, 0.05% ammonium persulfate, 0.05% TEMED and deionised water. The TEMED and ammonium persulfate were added last.

3. The separating gel was poured. A 2mm layer of distilled water was added to the top of the gel. The gel was then allowed to polymerise for 10 minutes.
4. The distilled water was poured off.
5. A 4% stacking gel mix was prepared from a 30% acrylamide/bis stock (Severn Biotech), containing 0.125M Tris-HCl (pH6.8), 0.1% SDS, 0.05% ammonium persulfate, 0.1% TEMED and deionised water. The TEMED and ammonium persulfate were added last. The gel comb(s) were inserted and the stacking gel poured on top of the separating gel.
6. The gel was allowed to polymerise for 30 minutes.

2.6.1.2 Running the gel

1. Cultured cells were harvested in 1X protein sample buffer and boiled at 95°C for 5 minutes, then loaded.
2. Size markers (Benchmark prestained protein ladder, GibcoBRL) were also loaded.
3. Gels were run at 200 Volts for approximately 45 minutes.

2.6.1.3 Electrophoretic transfer

Proteins were transferred to a nitrocellulose membrane using the Mini Trans-Blot® Electrophoretic Transfer cell (Biorad).

1. Nitrocellulose membrane (Hybond ECL, Amersham Pharmacia Biotech) and two pieces of Whatman 3MM were cut to size and soaked in transfer buffer.
2. The gel was equilibrated in transfer buffer.

3. The nitrocellulose membrane was placed on top of the gel. The two were then sandwiched between the Whatman papers. A glass tube was used to remove air bubbles. The sandwich was placed between fibre pads into the electrophoretic transfer cell (Biorad).
4. Electrophoretic transfer was run at 100V for 1 ½ hours.

2.6.2 Western blotting

1. Proteins were transferred to a nitrocellulose membrane (Hybond ECL, Amersham Pharmacia Biotech) (section 2.6.2.3) and blocked in 10% milk powder/0.1% Tween-20/phosphate buffered saline (PBS).
2. The blot was incubated with a mouse anti-T7 monoclonal antibody (stock at ~1 mg/ml) (Novagen #69522-4), at a dilution of 1/2500 in 10% milk powder/0.1% Tween-20/PBS.
3. The blot was washed three times for 10 minutes in 0.1% Tween/PBS at room temperature.
4. The secondary antibody, a sheep-anti-mouse-IgG HRP-conjugate (stock at ~0.32 mg/ml)(Sigma #A67782) was used at a dilution of 1/7500 in 10% milk powder/0.1% Tween-20/PBS.
5. The blot was washed three times for 10 minutes in 0.1% Tween/PBS at room temperature.
6. The signal was detected using ECL (NEN) according to the manufacturer's instructions and visualised by autoradiography.

2.6.3 Cell culture and transfection

1. COS-7 cells (SV40 transformed African Green monkey kidney) were grown in Dulbecco's modified Eagle's medium (DMEM) with 10% foetal bovine serum (FBS) and 100 μg/ml penicillin and 100 μg/ml streptomycin at 37°C in 5% CO_2.

2. Cells were seeded in a 24 well plate at ~40,000 cells/cm^2 in 1 ml of DMEM with FBS.
3. After 24 hours, the cells were then transfected with 0.6 µg DNA using the standard DEAE-dextran protocol (Seed & Aruffo, 1987).
4. Cells were incubated with the DNA for 3 hours, then shocked with 10% DMSO for 2 minutes.
5. The cells were then washed twice with PBS.
6. 600 µl of DMEM with 10% FBS was added and the cells were incubated at 37°C.
7. The cells were harvested after two and three days in 1X protein sample buffer.
8. 25 µl was loaded on a 12% polyacrylamide gel and the proteins were separated by SDS-PAGE (section 2.6.1), before Western blotting (section 2.6.2.4).

2.6.4 Immunofluorescence

1. COS-7 cells were seeded at ~20,000 cells/cm^2 onto coverslips in a 6 well plate.
2. Cells were transfected with 2 µg DNA using the DEAE-dextran method (Seed & Aruffo, 1987).
3. Cells were incubated with the DNA for 3 hours, then shocked with 10% DMSO for 2 minutes.
4. The cells were then washed twice with PBS.
5. 3 ml DMEM/10% FBS was added and the cells were incubated at 37°C for three days.
6. Cells were washed in 250mM Hepes (pH 7.4), and then fixed in 4% paraformaldehyde in 250 mM Hepes (pH 7.4). The reaction was quenched in 50 mM NH_4Cl.
7. The cells were permeabilised with 0.05% w/v saponin/0.2% gelatine in PBS.
8. The cells were then stained with mouse anti-T7. Tag monoclonal antibody (Novagen), at a dilution of 1/100 in 0.05% saponin/0.2% gelatine/PBS at room temperature.
9. The cells were rinsed twice in 0.05% saponin/PBS and then washed three times for 10 minutes in 0.05% saponin/PBS.

10. The secondary antibody, a goat anti-mouse-IgG FITC-conjugate (stock at ~1.1 mg/ml) (Sigma #F2012) was used at a dilution of 1/100 in 0.05% saponin/0.2% gelatine/PBS at room temperature.
11. The cells were rinsed twice in 0.05% saponin/PBS and then washed twice for 10 minutes in 0.05% saponin/PBS. The cells were finally washed twice for 10 minutes in PBS.
12. Coverslips were mounted onto slides with vectashield (Vector Laboratories) and visualised using a confocal microscope (Nikon 800, using the Microradiance confocal system (BIORAD) and Lasersharp image analysis software (BIORAD)).

2.7 Computational analysis

Details of most of the programs and scripts used can be found on the Sanger Institute WWW pages (http://www.sanger.ac.uk/Software). Some of the main programs and computational protocols are discussed below.

2.7.1 ACeDB

The data produced as part of this project was entered into the lab database 22ace (Dunham *et al*., 1994), the chromosome 22 implementation of ACeDB (Durbin and Thierry-Mieg, 1991). The data entry can either be done by obtaining write access and editing the database in real time, or by importing files prepared previously in a format readable by the database. The database is used for a variety of data, such as sequence, gene annotations and library screen results. The navigation through the database is by clickable links similar to hypertext links, which bring up new windows. There are different graphical representations for sequence data, genetic map data and peptide data. Exampled are shown throughout this thesis (for more detail see http://www.acedb.org).

2.7.2 Sequence analysis

Finished clone sequences were subjected to the standard Sanger Institute computational analysis (Dunham *et al.*, 1999). In brief, the sequences were analysed for repeats and the repeats masked using RepeatMasker (Smit and Green, unpublished). The masked sequences were used in similarity searches against the public domain DNA and protein databases using the BLAST suite of programs. A variety of exon and gene prediction programs, including Genscan (Burge & Karlin, 1997) was used to predict possible gene structures. The unmasked sequence was used in GC content analysis and prediction of CpG islands, tandem repeats, tRNA genes (Fichant & Burks, 1991) and exons. The completed sequences were visualised in the DNA map display in 22ace (section 2.7.1). For more detail, see http://www.sanger.ac.uk/HGP/Humana/human_analysis.shtml.

2.7.3 Gene annotation

1. To align cDNA or EST sequences with genomic DNA, the obtained cDNA sequence or assembled ESTs together with the genomic region to which the gene localised were used with the est2genome program (Mott, 1997).
2. The output file from est2genome was converted to ACeDB format using estg2ace (Dunham, unpublished).
3. The resultant file was imported into the 22ace database.

2.7.4 BLAST

In addition to the above described similarity searches, which were performed as part of the automatic analysis, additional BLAST (Altschul *et al.*, 1997) searches were performed using available websites (section 2.8.5).

2.7.5 Perl scripts

It is often more efficient to analyse large amounts of data using scripts. Perl is a computer language widely used by the bioinformatics community for data management, data format conversion and cgi (common gateway interface) scripts for web forms (Stein & Thierry-Mieg, 1998). Perl scripts were kindly provided by Dave Beare (Sanger Institute) and Dr. Ewan Birney (European Bioinformatics Institute) to aid parts of the analysis described in this project. Additional scripts by Dr. Ian Dunham (Sanger Institute) and Dr. Luc Smink (CIMR, Cambridge) have also been utilised.

Table 2.4: Perl scripts used during the course of this project.

Script	Function	Author
MethComp	Uses GFF file to compare specificity/sensitivity of 'methods' for gene identification/annotation. Compares against the reference set of exons as defined in the keyset of genes (structures) and the GFF	D. Beare
gff2ps	Parses gff format to postscript format	E. Birney
estg2ace	Parses est2genome output to ace format	I. Dunham
e-profile	Classifies results of BLASTn searches of dbEST into tissue origin.	L. Smink and D. Beare
MatchReport	Submits BLAST jobs to multiple databases and processes the output into a variety of formats.	L. Smink, D. Beare and I. Dunham

2.7.6 Calculations of specificity and sensitivity of sequence data

2.7.6.1. Background

The correlation of different types of sequence evidence with the annotated set of genes from 22q13.31 was measured by comparing the alignments of sequence evidence (potential coding value) against the annotated gene features along the test sequence. Analysis of the alignment can take place at the nucleotide, exon and/or gene level as appropriate (see section 2.7.6.3). This has been one of the most widely used approaches in evaluating the accuracy of coding

region identification and gene structure prediction methods. A brief explanation is provided here (for further details see Burset & Guigo, 1996).

A 2x2 contingency table can be used to represent the relationship between the true and putative coding nucleotides on a test sequence (figure 2.1).

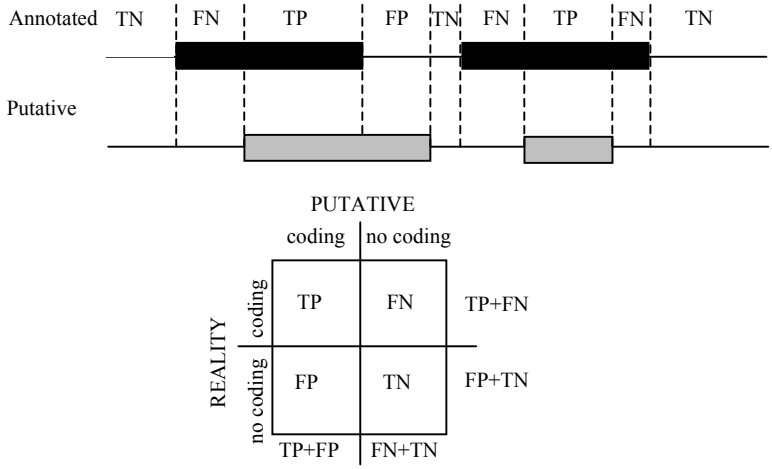


Figure 2.1: Measures of sequence correlation with annotated gene structures

The left upper cell of the table contains the number of coding nucleotides that also aligned with putative coding nucleotides (the true positives, TP), while the right lower cell contains the number of noncoding nucleotides that do not align with putative coding nucleotides (the true negatives, TN). The other two cells register the number of nucleotides in which the annotation and sequence evidence disagree: the number of coding nucleotides that do not align with putative coding nucleotides (the false negatives, FN) and the number of noncoding nucleotides which align with putatively coding sequence (the false positives, FP). Measures of sensitivity (Sn) and specificity (Sp) can be derived from this table and are usually defined as:

$$Sn = \frac{TP}{TP+FN} \qquad Sp = \frac{TP}{TP+FP}$$

Sn is the proportion of coding nucleotides that correctly align with putatively coding nucleotides and Sp the proportion of putative coding nucleotides that are actually coding.

2.7.6.2 MethComp

The correlation of different types of sequence evidence with the annotated sets of genes and pseudogenes from 22q13.31 was calculated using the perl script MethComp (Dave Beare, unpublished). This program analyses the nucleotide alignments of different sequence evidence against annotated exons from genome feature format (gff) files dumped from the chromosome 22 ACeDB database. The alignment of a putative coding feature within the test sequence is described as a 'hit'. In all cases, multiple overlapping hits are counted as one. The following calculations are performed:

Total coverage = No. of base hits/test region size.
Sn (Sensitivity) = No. of bases which hit reference exons/total no. of bases within reference exons = TP/(TP+FN)
Sp (Specificity) = No. of bases which hit reference exons/total coverage = TP/(TP+FP)
Exon hits = No. of reference exons hit/total no. of reference exons.
Gene hits = No. of reference genes hit/ total no. of reference genes.

2.7.6.3 Analysis of the accuracy of Genscan and Fgenesh.

Equivalent calculations of specificity and sensitivity are also calculated at the exonic and genic level in analysis of gene prediction programs. It is assumed that an exon (or gene) has been predicted correctly, only when both its boundaries (and internal exon structure) have been predicted correctly. Predicted exons or genes that only overlap true exons or genes are counted as false predictions. Sn is the proportion of coding exons or genes that correctly align with putatively coding exons or genes and Sp is the proportion of putative coding exons or genes that are actually coding. In this case, only protein coding sequence is taken into account.

Given the stringent criteria used to consider an exon or gene as correctly predicted, two additional measures of specificity and sensitivity are computed. These are the proportion of true exons or genes without overlap to predicted exons or genes – the Missing Exons (ME) or Missing Genes (MG) – and the proportion of predicted exons without overlap to actual exons – the Wrong Exons (WE) or Wrong Genes (WG).

$$\text{ME (or G)} = \frac{\text{number of Missing Exons (or Genes)}}{\text{number of annotated exons (or genes)}}$$

$$\text{WE (or G)} = \frac{\text{number of Wrong Exons (or Genes)}}{\text{number of predicted exons (or genes)}}$$

2.7.6.4 Promoter predictions

The results of the algorithms CPGFIND (Micklem, unpublished) PromoterInspector (Scherf *et al.*, 2000) and Eponine (Down, unpublished) were also correlated with the 38 annotated protein-coding genes within 22q13.31. In this case, correlation limits were set at 6 kb upstream to 0.5 kb downstream of the annotated transcription start site, Unlike CPGFIND and PromoterInspector, Eponine attempts to make strand-specific predictions. Only predictions on the same strand as the annotated gene were counted as a positive correlation. The specificity and sensitivity of each prediction type was calculated as before.

2.7.7 Phylogenetic analysis

1. Each of the 27 full-length protein sequences was used to search the NCBI nonredundant protein sequence database (http://www.ncbi.nlm.nih.gov), using the gapped BLASTP program (section 2.7.4).
2. The BLAST alignments were inspected by eye. Entries which were redundant, contained point mutations with respect to sequences already included in the analysis, or corresponded to sequences known to be previously submitted partial versions of the gene of interest, were excluded. Additionally, entries that demonstrated only partial matches were removed, as the sequences involved shared only some functionally similar parts (e.g. multidomain proteins.)
3. Sequence data were aligned using the default options of clustalw (Thompson *et al.*, 1994)(http://www.ebi.ac.uk/clustalw).

4. Neighbour-joining (NJ) analyses(Saitou & Nei, 1987) of the amino acid alignments were produced using Phylowin (Galtier *et al.*, 1996). Robustness of the NJ trees was tested by bootstrap analyses with 500 pseudo-replications per tree.
5. Potential orthologues, identified from the phylogenetic trees, were then compared against the NCBI nonredundant protein sequence database to ensure that orthologous pairs fulfilled the requirements of being the two most similar proteins between two different organisms (Huynen & Bork, 1998; Tatusov *et al.*, 1997; Tatusov *et al.*, 1996).
6. The chromosomal position of potential mouse orthologues was verified as far as possible by BLASTN comparison of the nucleotide sequence against the available mouse genomic sequence (http://mouse.ensembl.org).

2.8 Materials

2.8.1 Buffers

1X TE

- 10 mM Tris-HCl (pH 8.0)
- 1 mM EDTA

1X $T_{0.1}E$

- 10 mM Tris-HCl (pH 8.0)
- 0.1 mM EDTA

10X PCR buffer 1

- 670 mM Tris-HCl (pH 8.8)
- 166 mM $(NH_4)_2SO_4$ (enzyme grade)
- 67 mM MgCl
- (pH 8.8)

28% Sucrose solution

- 1X TE
- 28% w/v sucrose
- 0.008% w/v cresol red

***Pfu* 10X reaction buffer (PCR buffer 2)**

- 200 mM Tris-HCl (pH 8.8)
- 100 mM KCl
- 100mM $(NH_4)_2SO_4$
- 20 mM $MgSO_4$
- 1.0% Triton®X-100

DNase I reaction buffer

- 200 mM Tris-HCl (pH 8.4)
- 20 mM $MgCl_2$
- 500 mM KCl

First Strand buffer

- 250 mM Tris-HCl (pH 8.3)
- 375mM KCl
- 15 mM $MgCl_2$

10X Ligase buffer (Roche)

- 660 mM Tris-HCl
- 50 mM $MgCl_2$
- 10 mM dithioerythritol
- 10 mM ATP

(pH 7.5)

6X Glycerol loading dyes (I)

- 30% v/v glycerol
- 0.1% w/v bromophenol blue
- 0.1% w/v xylene cyanol
- 5 mM EDTA (pH 7.5)

6X Dye Buffer (II)

- 0.25% bromophenol blue
- 0.25% xylene cyanol
- 15% Ficoll (Type 400: Pharmacia)

Vistra Green stain

For 1 gel:

- 0.01M Tris HCl
- 0.0001M EDTA (pH 7.4)
- 50 µl Vistra green (Amersham RPN5786)

10X TAE

- 890mM Tris base
- 0.05M EDTA
- 5.71% glacial acetic acid (JTBaker)

10X TBE

- 890 mM Tris base
- 890 mM Borate
- 20mM EDTA (pH 8.0)

20X SSC

- 3 M NaCl
- 0.3 M Trisodium citrate

100X Denhardt's

- 20 mg/ml Ficoll 400-DL
- 20 mg/ml polyvinylpyrrolidine 40
- 20 mg/ml BSA (pentax fraction V)

Hybridisation buffer

- 6X SSC
- 2 mg/ml Ficoll 400-DL
- 2 mg/ml polyvinylpyrrolidone 40
- 2 mg/ml BSA (pentax fraction V)
- 1% N-lauroyl-sarcosine
- 50mM Tris-HCl (pH 7.4)
- 10% w/v dextran sulphate

SAP buffer

- 200 mM Tris HCl (pH 8.0)
- 100 mM $MgCl_2$

ExoSAP dilution buffer

- 50mM Tris HCl (pH 8.0)

OLB3

- 240 mM Tris-HCl
- 75 mM β-mercaptoethanol
- 0.1 mM dATP
- 0.1 mM dGTP
- 0.1 mM dTTP
- 1 M HEPES (pH 6.6)
- 0.1 mg/ml hexadeoxyribonucleotides (2.1 OD units/ml)

GTE

- 50 mM glucose
- 1 mM EDTA
- 25 mM Tris-HCl (pH 8.0)

3 M K^+/5 M Ac^-

- 60 ml 5 M potassium acetate
- 11.5 ml glacial acetic acid
- 28.5 ml H_20

Protein transfer buffer

- 10% 10X Protein running buffer
- 25% 100% Ethanol

Protein sample buffer

- 2% w/v SDS
- 10% v/v glycerol
- 60 mM Tris-HCl (pH 6.8)
- 0.01% w/v Bromophenol blue
- 5% v/v β-mercaptoethanol

10X Protein running buffer

- 30 g/l Tris base
- 144 g/l glycine
- 10 g/l SDS

2.8.2 Cell culture

2.8.2.1 Growth media

LB

- 10 mg/ml bacto-tryptone
- 5 mg/ml yeast extract
- 10 mg/ml NaCl
- (pH 7.4)

2.8.2.2 Antibiotic concentrations

Mouse RPCI-23 BAC clones: 12.5 µg/ml chloramphenicol.

Human cDNA clones: 100 µg/ml ampicillin.

Blue/white selection of cDNA clones: 100 µg/ml ampicillin/ 0.5 mM IPTG/ 80 µg/ml X-Gal.

2.8.3 Size markers

2.8.3.1 1 kb ladder (GibcoBRL)

This contains 1 to 12 repeats of a 1018 bp concatenated fragment and vector fragments from 75 to 1636 bp, thus producing the following sized fragments (bp):

Table 2.5: 1 kb ladder (GibcoBRL)

Band no.	Size (bp)	Band no.	Size (bp)
1	12216	12	1635
2	11198	13	1018
3	10180	14	516/506
4	9162	15	394
5	8144	16	344
6	7125	17	298
7	6108	18	220
8	5090	19	200
9	4072	20	154
10	3054	21	142
11	2036	22	75

2.8.3.2 Wide Range Analytical Marker DNA (Promega)

The Analytical Marker DNA, Wide Range, provides an evenly spaced distribution of 32 DNA fragments ranging from 702bp to 29,950bp in size and was used for band sizing in fingerprint experiments. This marker is composed of a mixture of restriction enzyme digests of Lambda DNA and ϕX174 DNA.

2.8.3.3 Benchmark™ Prestained Protein Ladder (GibcoBRL)

This ladder for SDS-PAGE consists of 10 proteins ranging in apparent molecular weight from approximately 10 to 200 kDa. The proteins are rendered blue by a proprietary method that covalently couples dyes to the proteins. The fourth protein band from the top is coupled with a pink dye for easy orientation.

Table 2.6: Benchmark™ Prestained Protein Ladder (GibcoBRL)

Band no.	Apparent molecular weight (kDa)
1	172.6
2	111.4
3	79.6
4	61.3
5	49.0
6	36.4
7	24.7
8	19.2
9	13.1
10	9.3

2.8.4 Primer sequences

2.8.4.1Vectorette primer

244 CGA ATC GTA ACC GTT CGT ACG AGA ATC GCT

T7 TAC GAC TCA CTA TAG GGA GA

SP6 CAT ACG ATT TAG GTG ACA C

2.8.4.2 Production of pCDNA3-NT7

pCDNA3-NT7 cassette	GAT CCA GCG GCC GCT G
stpCDNA3-NT7 S	GGC CAA GCT TGC CAC CAT GGC TAG CAT GAC
stpCDNA3-NT7 A	GGC CTC TAG ATC CAG CGG CCG CAG GAT CCC G

2.8.4.3 Other STSs

The primers of all other STSs used are listed in appendix 1.

2.8.5 URLs and ftp sites

Table 2.7 Useful URL and ftp sites

Title	URL
BLAST services at the NCBI	http://www.ncbi.nlm.nih.gov/BLAST
Clustalw	http://www.ebi.ac.uk/clustalw
Ensembl	http://www.ensembl.org/
Ensembl (Mouse)	http://mouse.ensembl.org
Entrez browser	http://www3.ncbi.nlm.nih.gov/Entrez
GeneCards: human genes, proteins and diseases (Weizmann)	http://bioinfo.weizmann.ac.il/cards
Genomatix PromoterInspector	http://www.genomatix.de/cgi-bin/promoterinspector/promoterinspector.pl
GFP project	http://www.dkfz-heidelberg.de/abt0840/GFP/
Humace home page	http://inteweb.sanger.ac.uk/LocalUsers/humace
Human Gene Nomenclature Database	http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl
Human working draft genome browser	http://www.infobiogen.fr/services/GoldenPath/mirror/goldenPath/gbdDescriptions.html
InterPro	http://www.ebi.ac.uk/interpro/
Maps of Human and Mouse homology	http://www.ncbi.nlm.nih.gov/Homology/
MatInspector	http://transfac.gbf.de/cgi-bin/matSearch/matsearch2.pl
Mouse BAC ends	http://www.tigr.org/tdb/bac_ends/mouse/bac_end_intro.html
Mouse genome database	http://www.informatics.jax.org
PipMaker	http://bio.cse.pse.edu/pipmaker
PIX at the HGMP	http://www.hgmp.mrc.ac.uk/Registered/Webapp/pix/
Primer3	http://www.sanger.ac.uk/cgi-bin/primer3.cgi
REBASE - Restriction Enzymes	http://rebase.neb.com/rebase/rebase.html
RepeatMasker web server	http://ftp.genome.washington.edu/cgi-bin/RepeatMasker
RPCI-23	http://www.chori.org/bacpac/23framefmouse.htm
Sanger Institute	http://www.sanger.ac.uk
Sanger Institute: Human analysis	http://www.sanger.ac.uk/HGP/Humana/human_analysis.shtml
Sanger Institute: SRSWWW	http://www.sanger.ac.uk/srs6
Search Evaluated MEDLINE	http://www.biomednet.com/db/medline
Sequence Logos	http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi
SignalP server	http://www.cbs.dtu.dk/services/SignalP
The HGMP Resource Centre	http://www.hgmp.mrc.ac.uk
The IMAGE Consortium	http://image.llnl.hov
The NCBI BLAST server.	http://www.ncbi.nlm.nih.gov/BLAST
The Sanger Institute: BLAST server	http://intweb.sanger.ac.uk/LocalUsers/humace/BLAST/Internal_blast_server.shtml
Web SequenceLogo main form	http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi
Unigene-Human	http://www.ncbi.nlm.gov/Schuler/UniGene/Hs.Home.html

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.

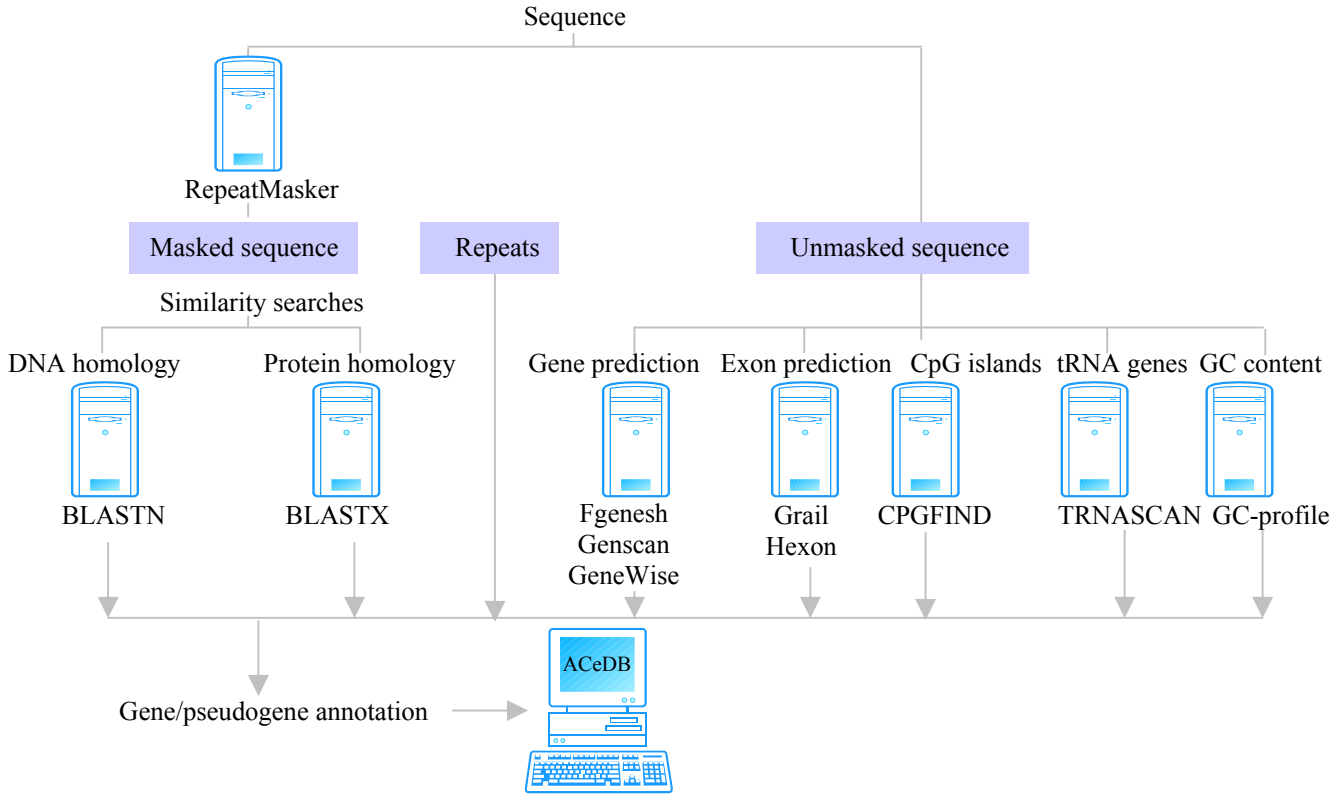


Figure 3.1: Automated analysis strategy. The masked sequence was used in homology searches. The unmasked sequence was used in a number of gene prediction packages and in the prediction of other features such as CpG islands and tRNAs. Both the homology data and the predicted data were integrated with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

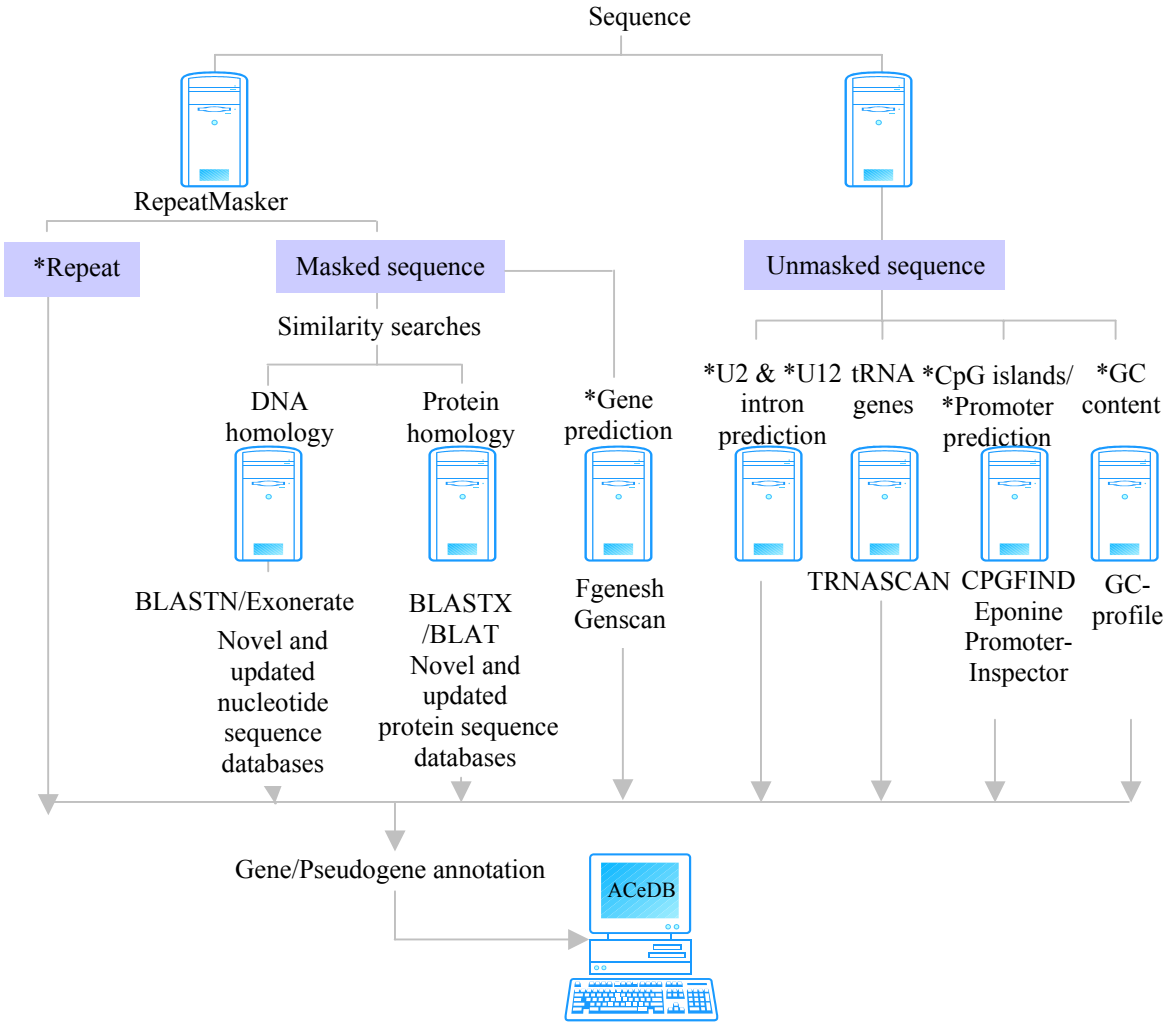


Figure 3.2: Chromosome 22 additional analysis strategy. * denotes analysis performed on linked clone sequences. Masked sequence was used in homology searches against novel and updated sequence databases (see appendix 2a) and in a number of gene prediction packages. Unmasked sequence was used in the prediction 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).

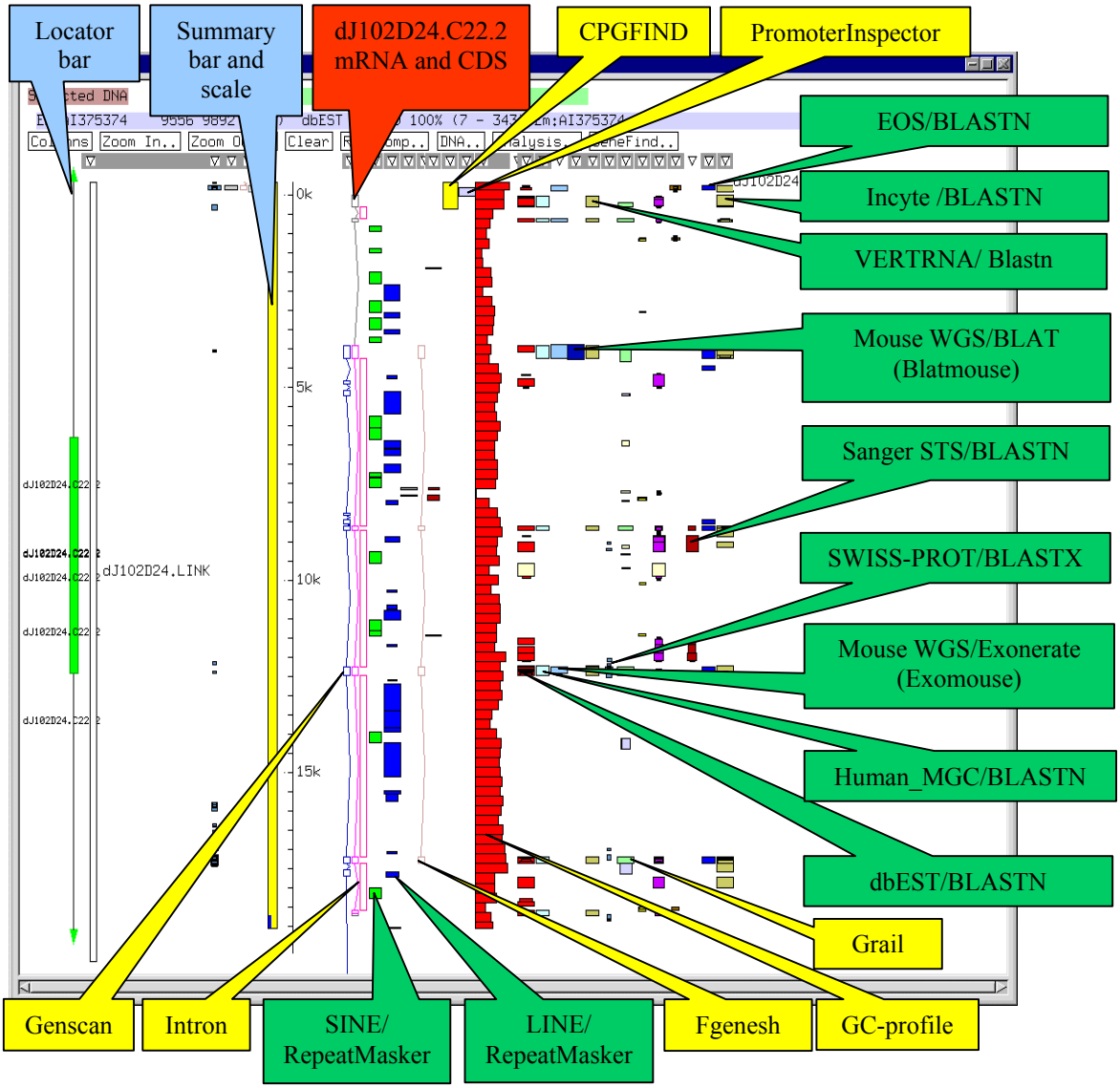


Figure 3.3: An example of the ACeDB display. The blue boxes show ACeDB general features. The green boxes indicate similarities to a variety of sequence databases, listed in appendix 2a. The yellow boxes show the output from a range of prediction programs listed in appendix 2b. Red boxes indicate annotated 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.

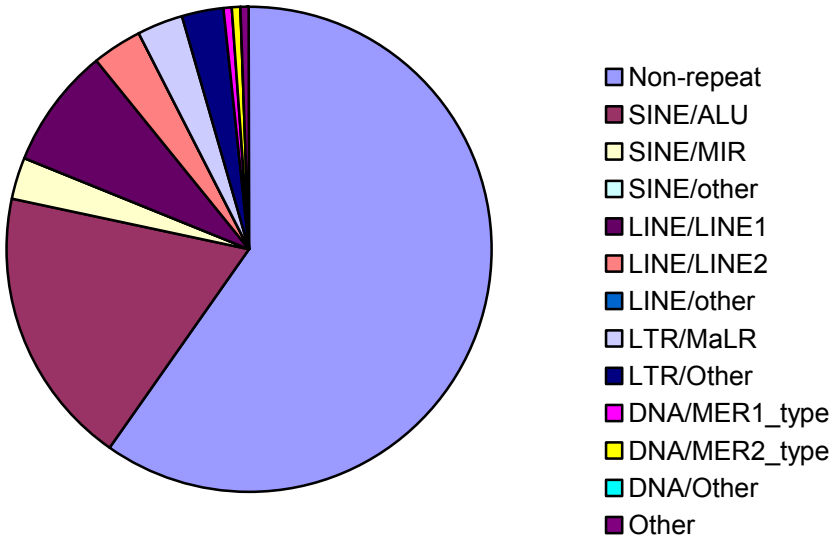


Figure 3.4: Repetitive and non-repetitive DNA coverage (%) for region of interest

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.

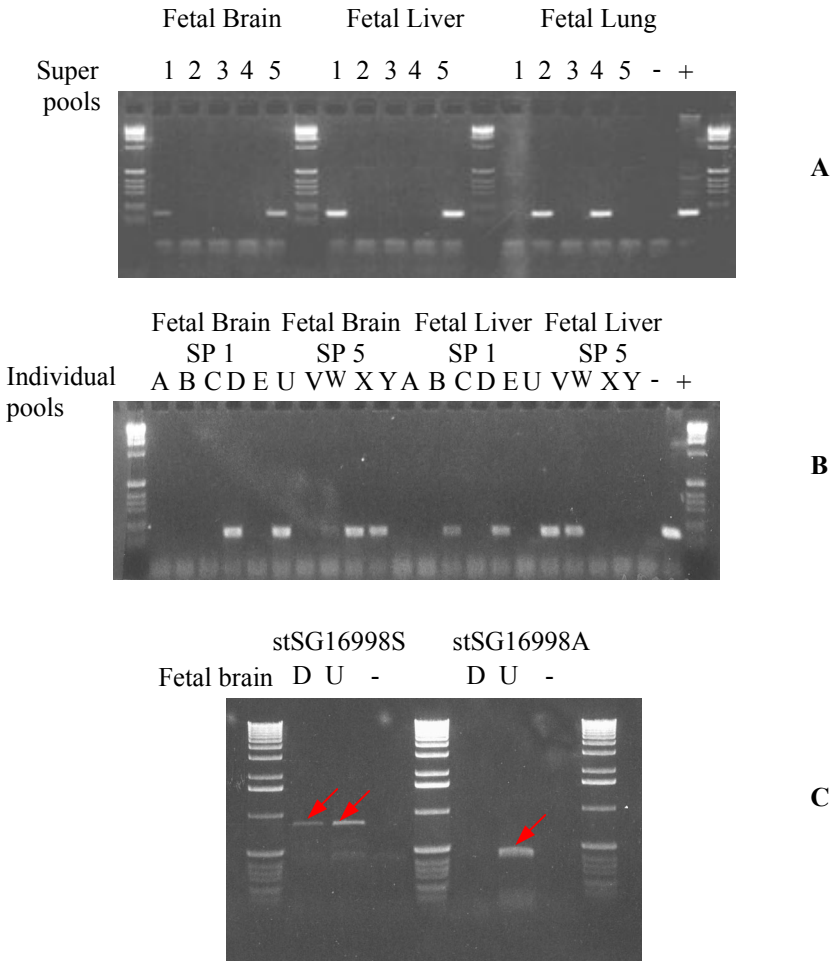


Figure 3.6: Example of vectorette based isolation of PCR fragments from cDNA library using primers stSG16998 (H55372), contained within the locus ARHGAP8. Screening of the super pools (A), is followed by individual pool screening (B). The identified pools are then used as templates in vectorette PCR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

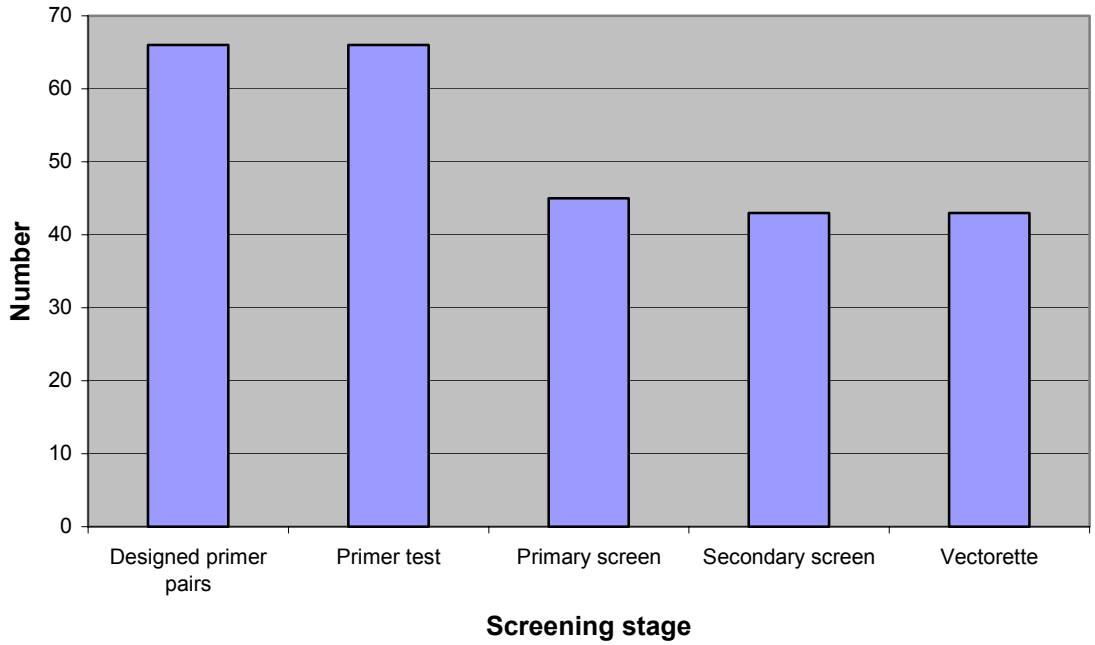


Figure 3.7: Vectorette cDNA library screens. The total number of primer pairs, designed to potential gene features based on similarity evidence, that have been screened across the vectorette cDNA libraries. The bars represent the total number of leads that succeeded at each of the stages.

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.

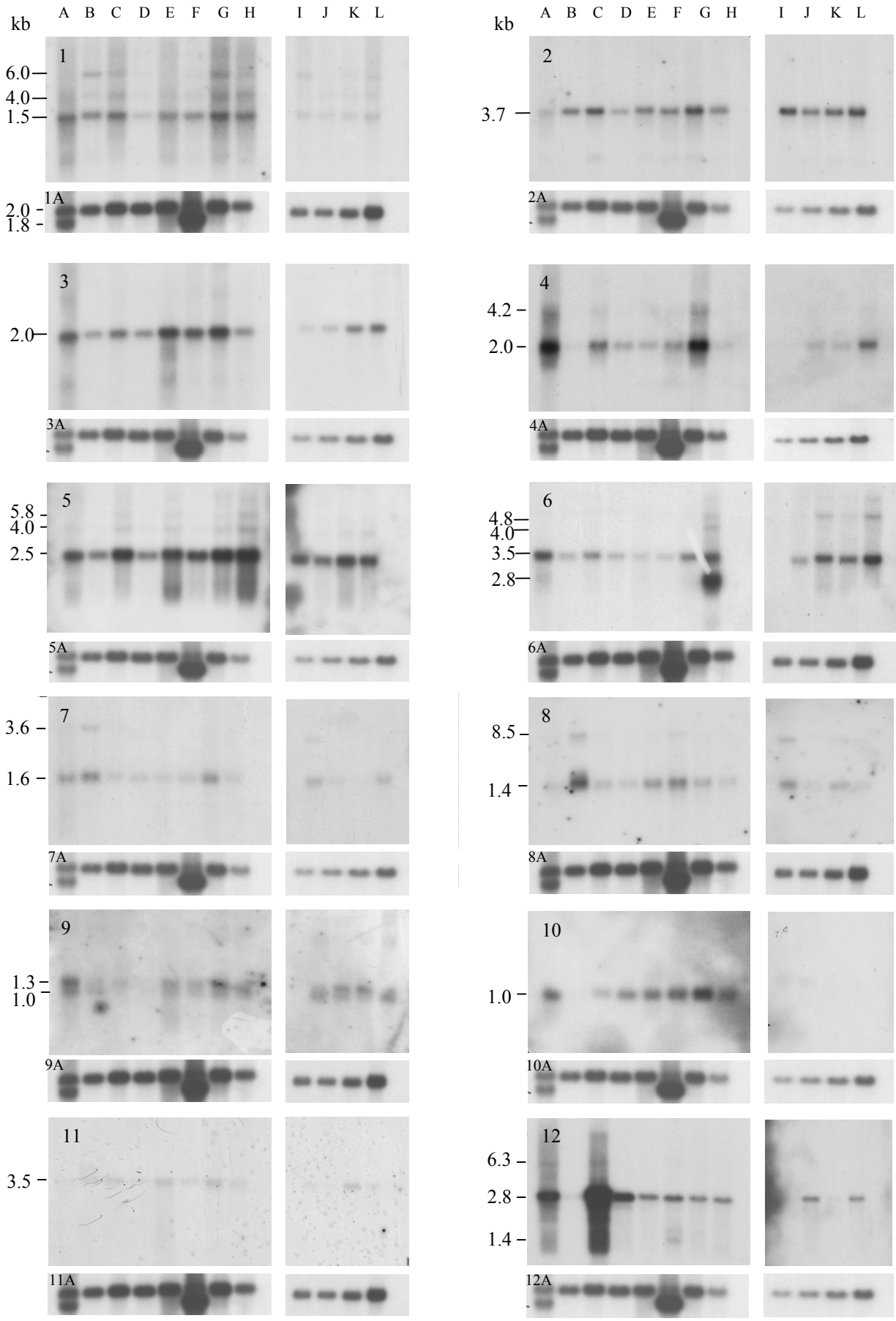
Northern analyses were carried out for the 41 gene loci annotated within the region (see chapter II). Radio-labelled probes were generated by PCR from RNA templates, using primers designed from annotated cDNA sequences and hybridised to Northern blots containing RNA from eight adult and four fetal human 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.

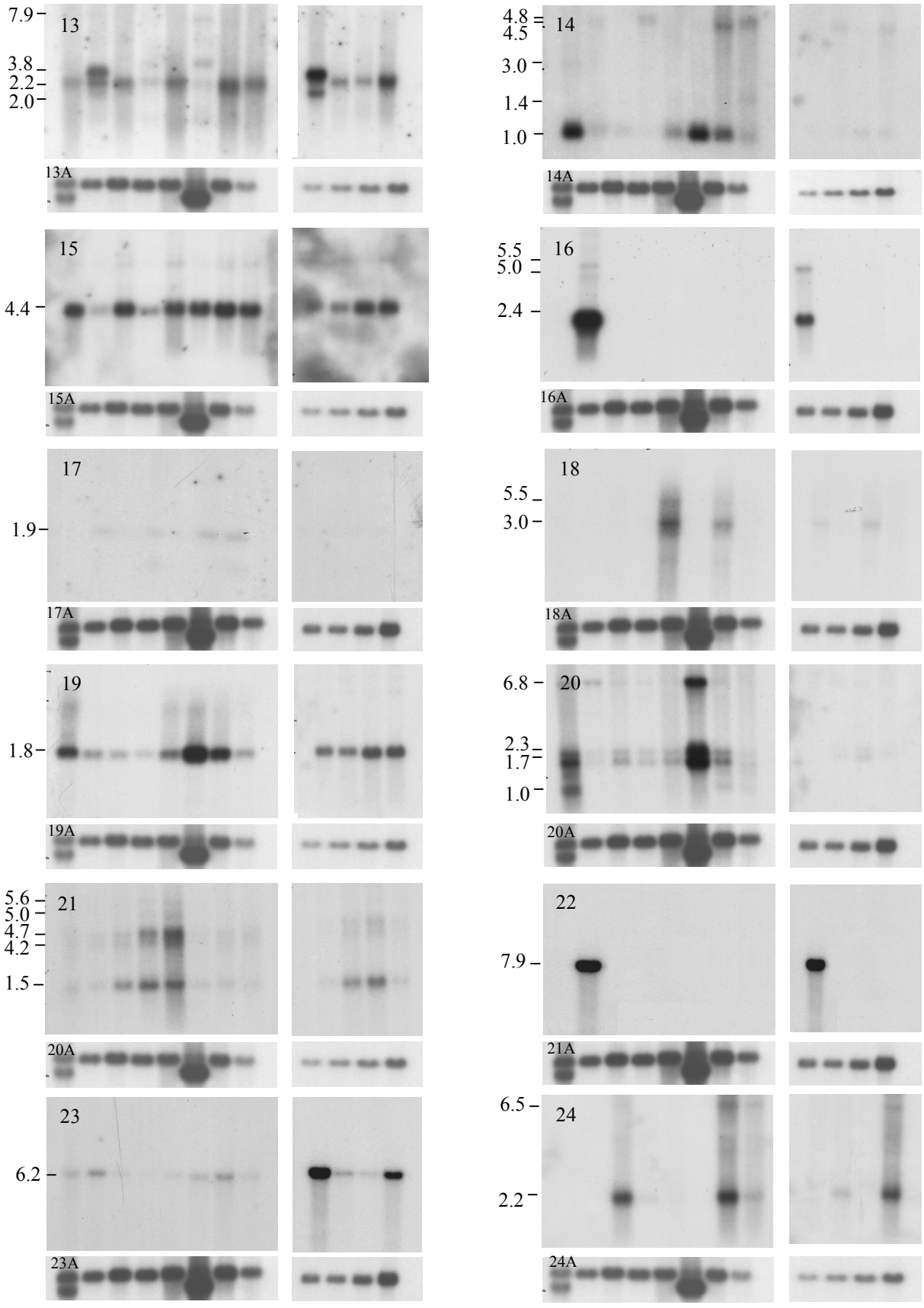
3.5.1.2 Transcript size

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

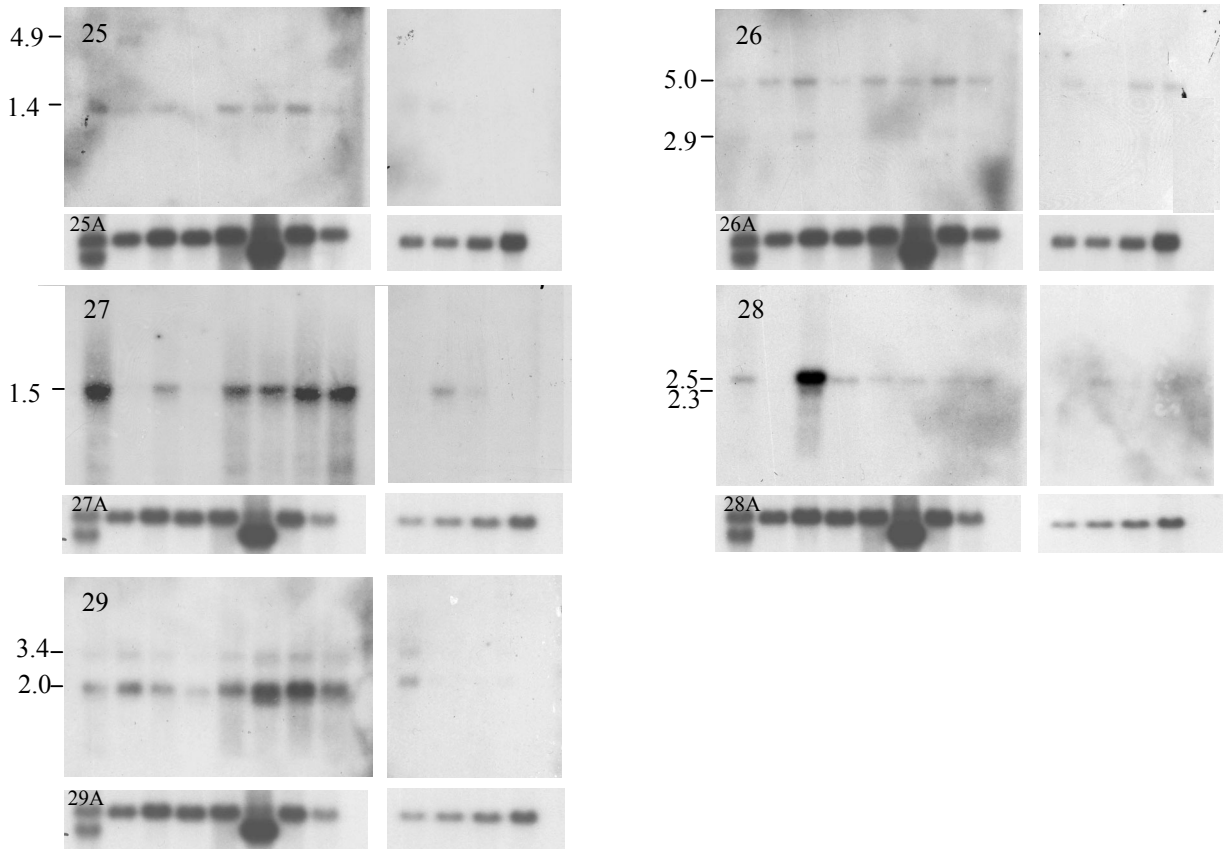


Figure 3.8 Results from 41 Northern blots – only the 29 experiments that gave a positive result are shown. Results generated from hybridisation of each Northern blot to a β-Actin control probe (Clontech) are shown underneath each band (A). Approximate band sizes are shown to the left of each blot (only 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.

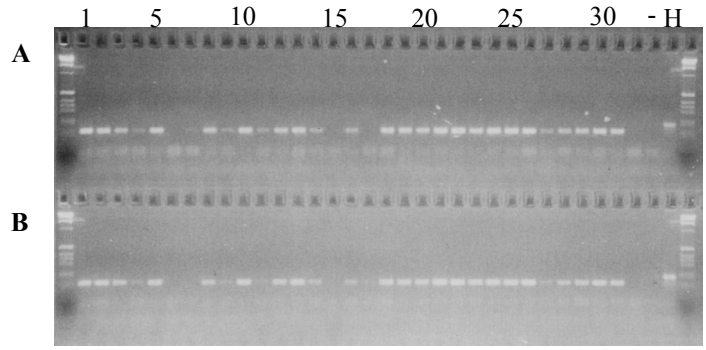


Figure 3.9: Example of a transcription profile for TTLL1. A and B represent duplicate experiments. The experiment was performed in duplicate. - = negative control; H = human genomic DNA. The genomic band is larger as the primers span an intron in TTLL1. The lane designations correspond to the key in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

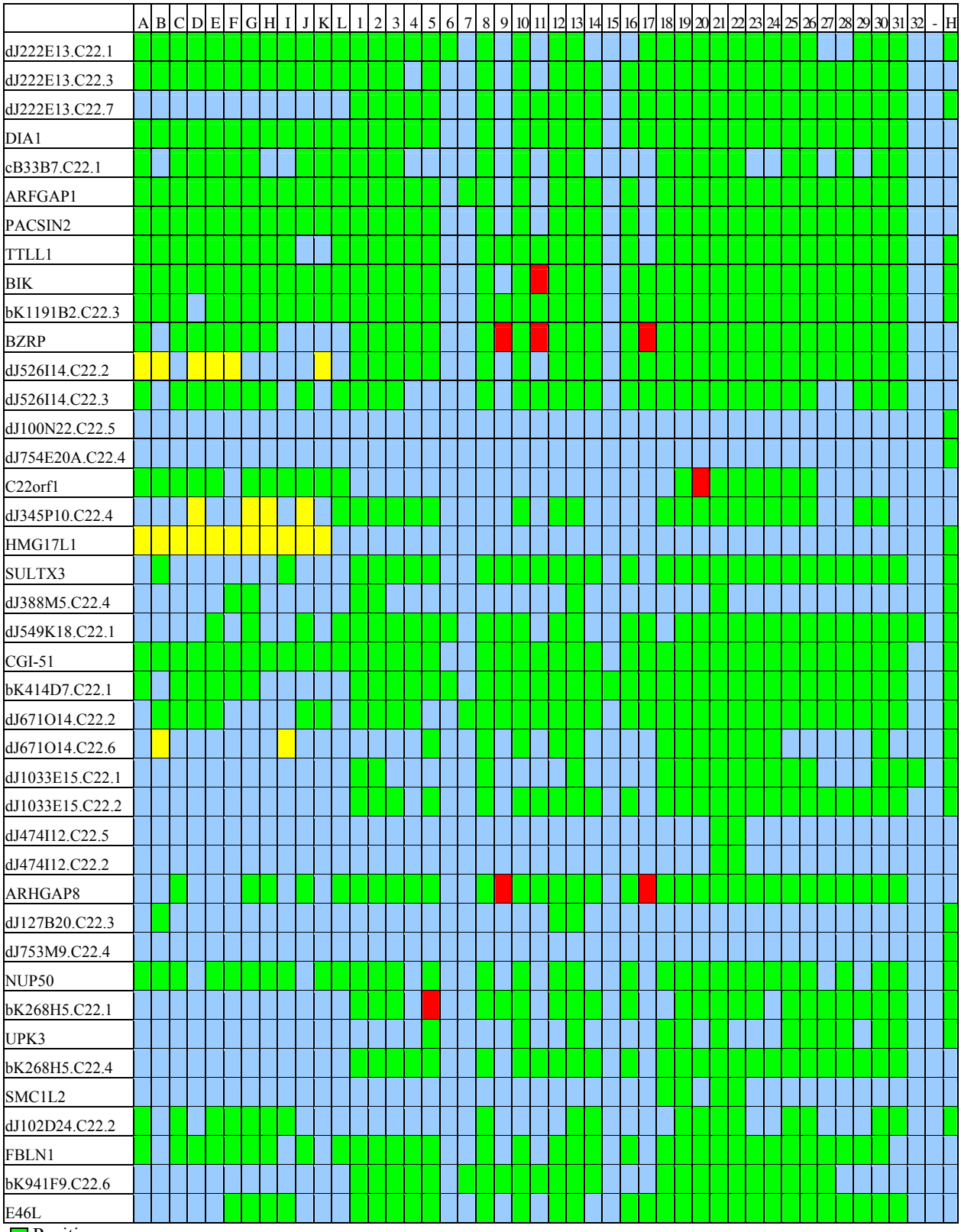


Table 3.6: Key to tissue identity

	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.

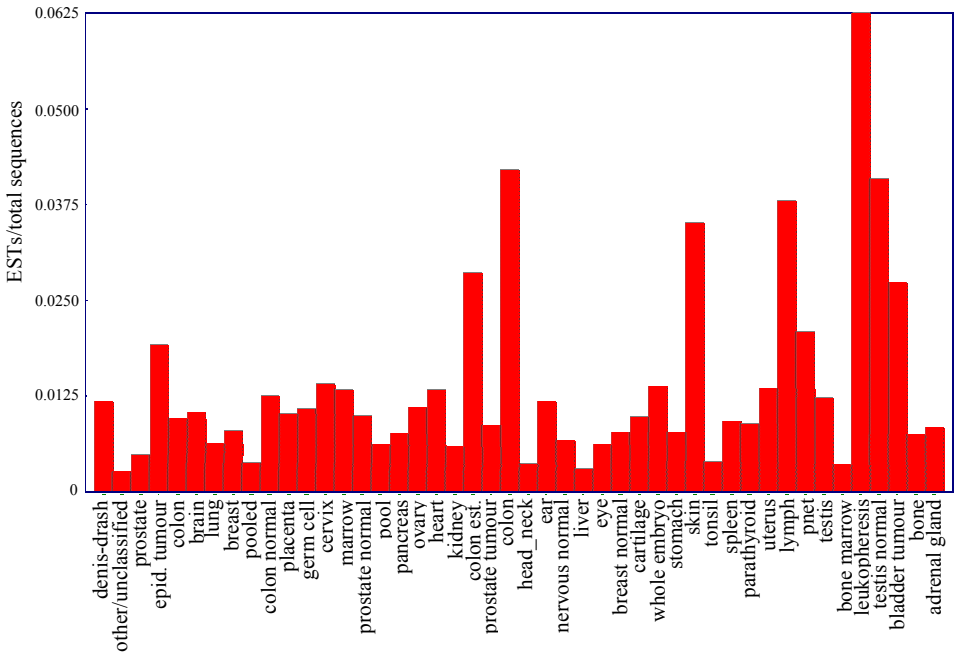


Figure 3.11: Expression profile of dJ222E13.C22.3a (Em:AL160111).
The proportion of ESTs from a range of tissues that show >80% similarity at the nucleotide level to the cDNA sequence of dJ222E13.C22.3 (isoform a). Generated using e-profile (Smink and Beare, unpublished).

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.

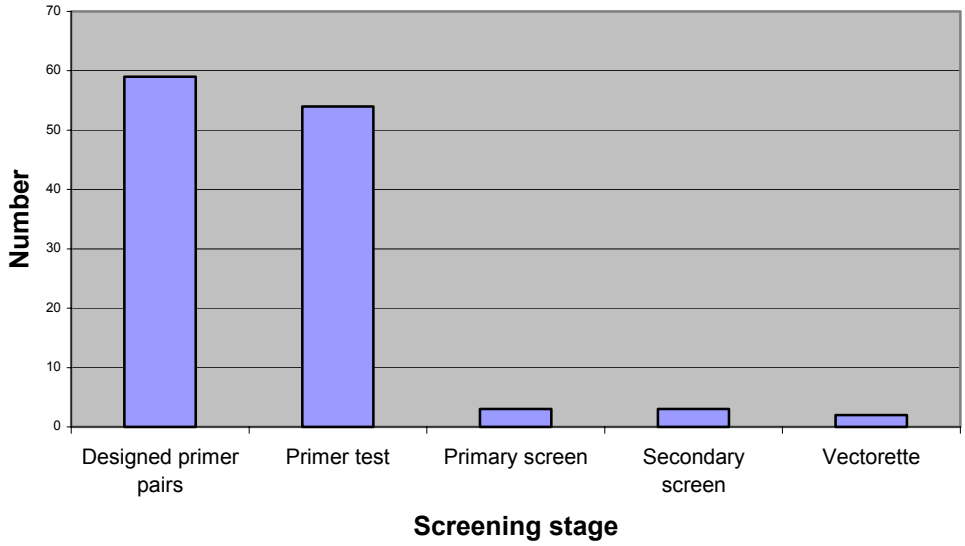


Figure 3.12: The total number of primer pairs, designed to Genscan predicted exons without similarity to expressed sequence evidence, which have been screened across the vectorette cDNA libraries. The bars represent the total number of leads that succeeded at each of the stages.

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.

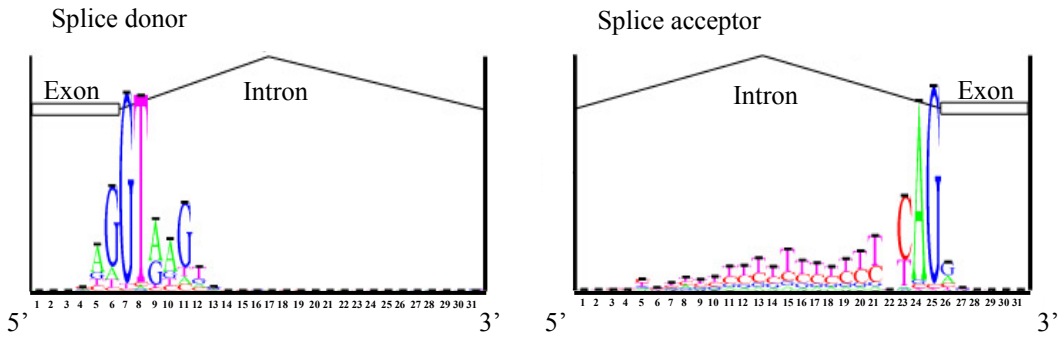


Figure 3.13: Splice donor and acceptor consensus sequences for 379 introns in 22q13.31. The splice site sequences were extracted by D. Beare and visualised using Sequence Logo (Steven Brenner) (http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi).

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.

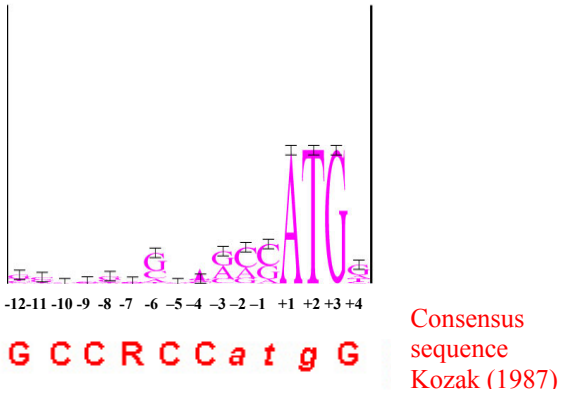


Figure 3.14: Translational start site consensus for 27 full genes on chromosome 22. Kozak's consensus sequence is depicted beneath. Generated from http://www.bio.cam.ac.uk/cgi-bin/seqlogo.cgi (S. Brenner).

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.

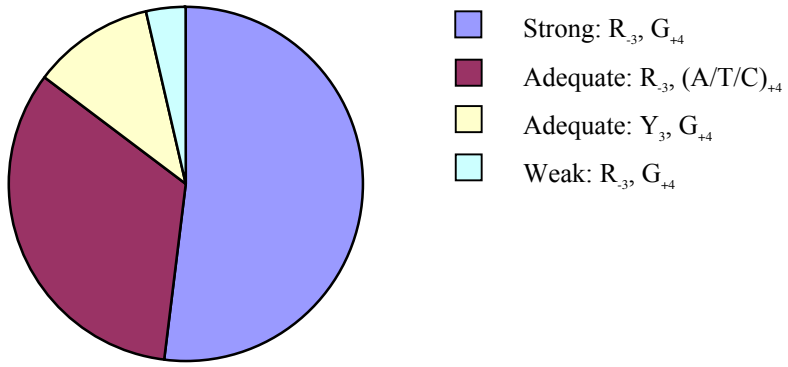


Figure 3.15: Analysis of the sequence contexts surrounding 27 initiator codons from 22q13.31.

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.

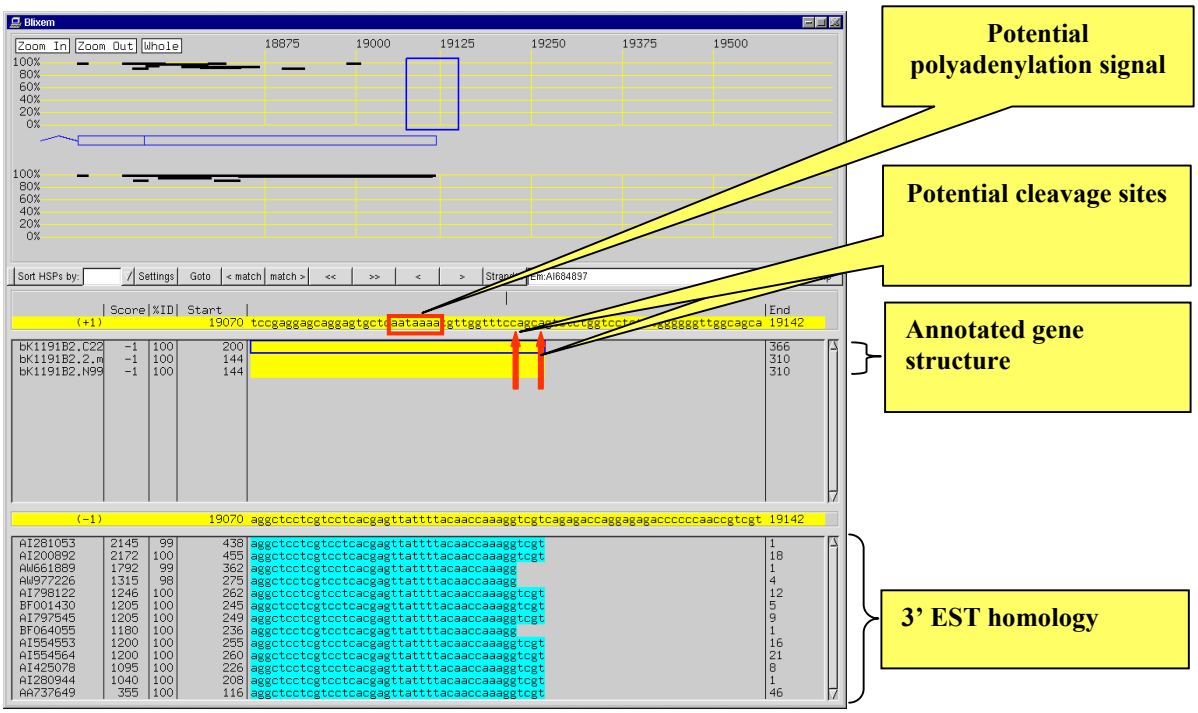


Figure 3.16: An example of Blixem output from ACeDB. EST homology to the 3' end of the BIK gene is shown. Putative polyadenylation signal and cleavage sites are highlighted.

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.

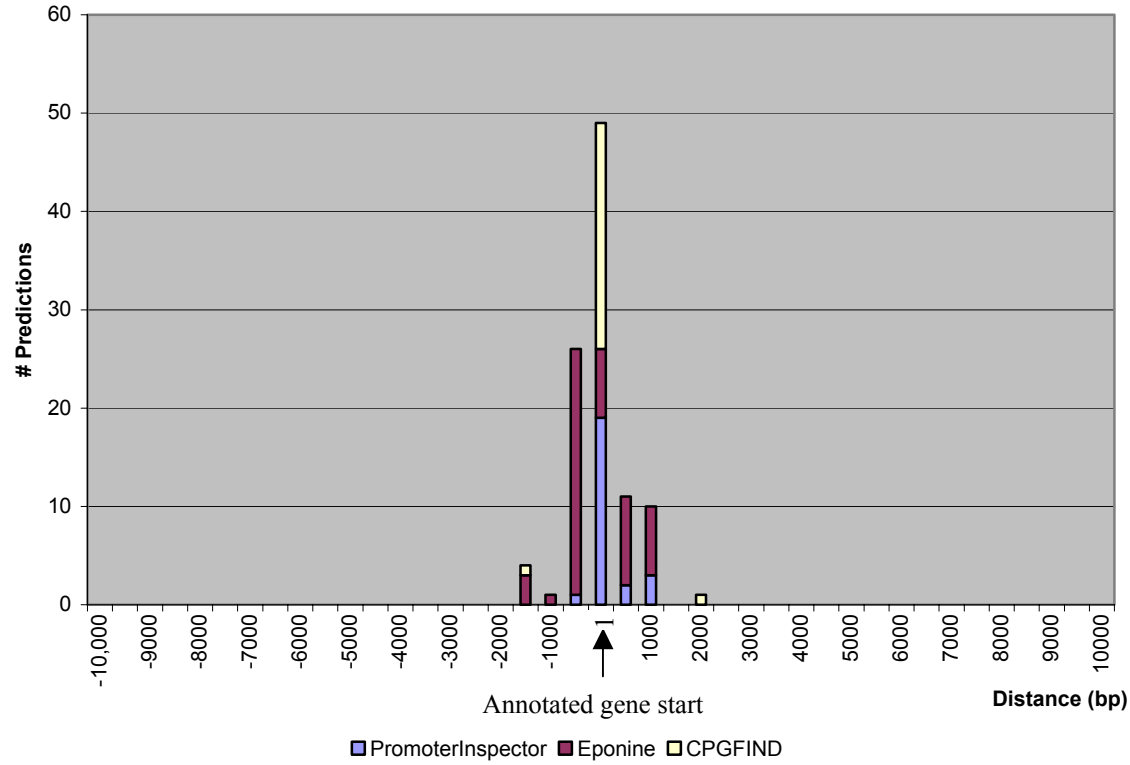


Figure 3.17: Correlation analysis of predicted promoter and transcription start site regions with 27 annotated full gene starts within a 3.4Mb region of chromosome 22. The y-axis indicates the total number of matches found in relative distance to the annotated gene start. Values on the x-axis with a 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.

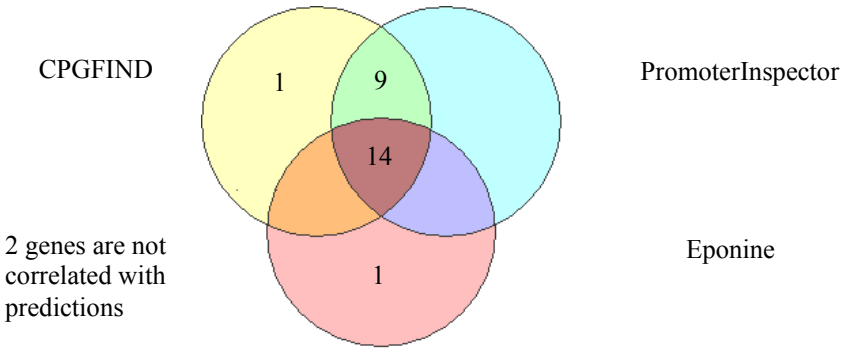


Figure 3.18: Venn diagram shows the number of full gene structures and their correlation with different kinds of promoter prediction algorithms

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.

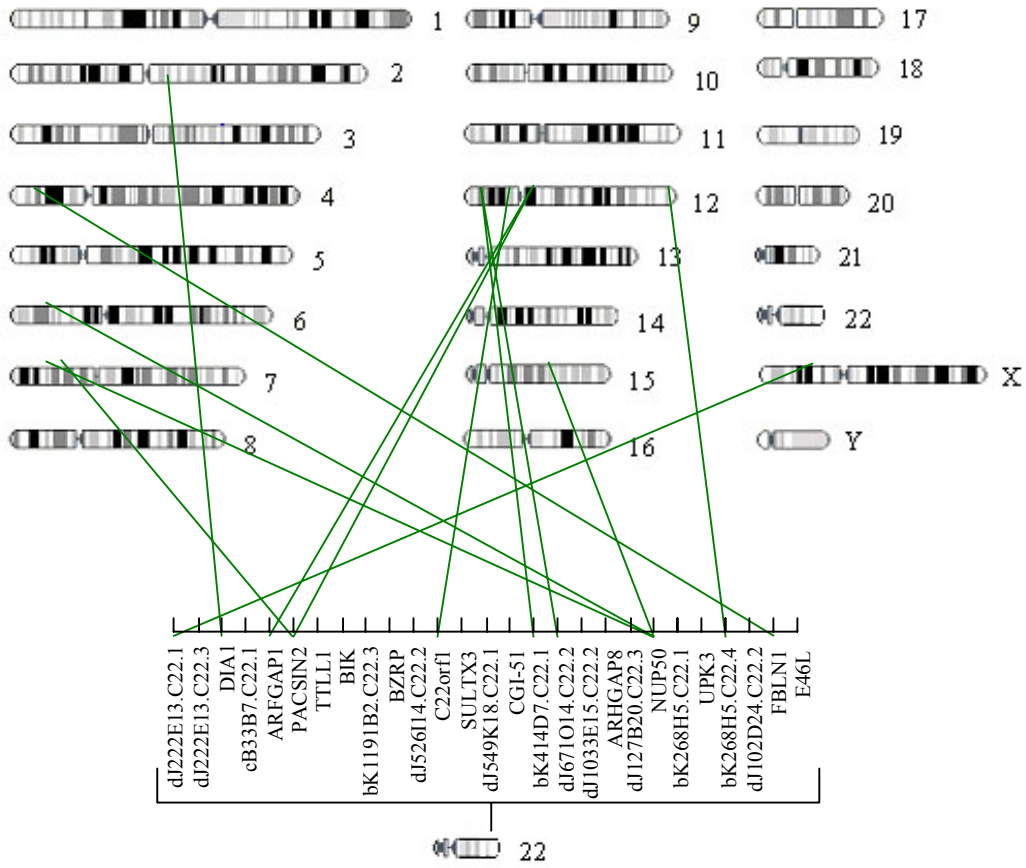


Figure 3.19: Approximate positions of genes putatively paralogous to full genes on 22q13.31. Figure was generated using the Ensembl website (http://www.ensembl.org).

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.

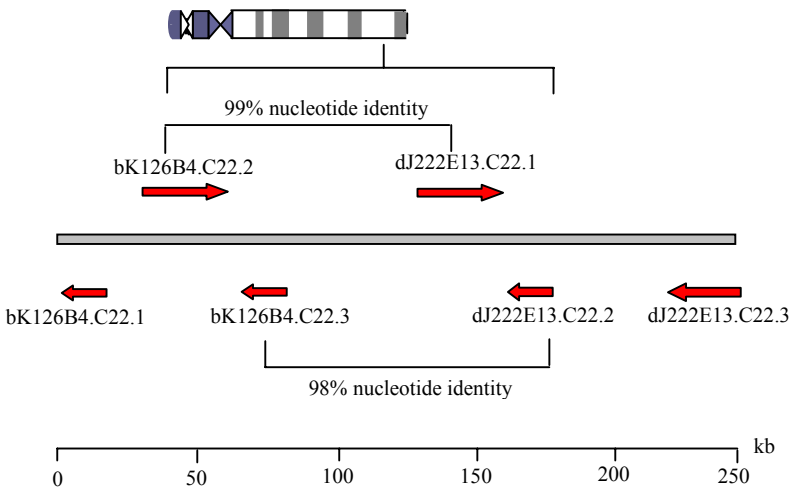


Figure 3.20: Schematic showing a region of interchromosomal duplication on chromosome 22

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.

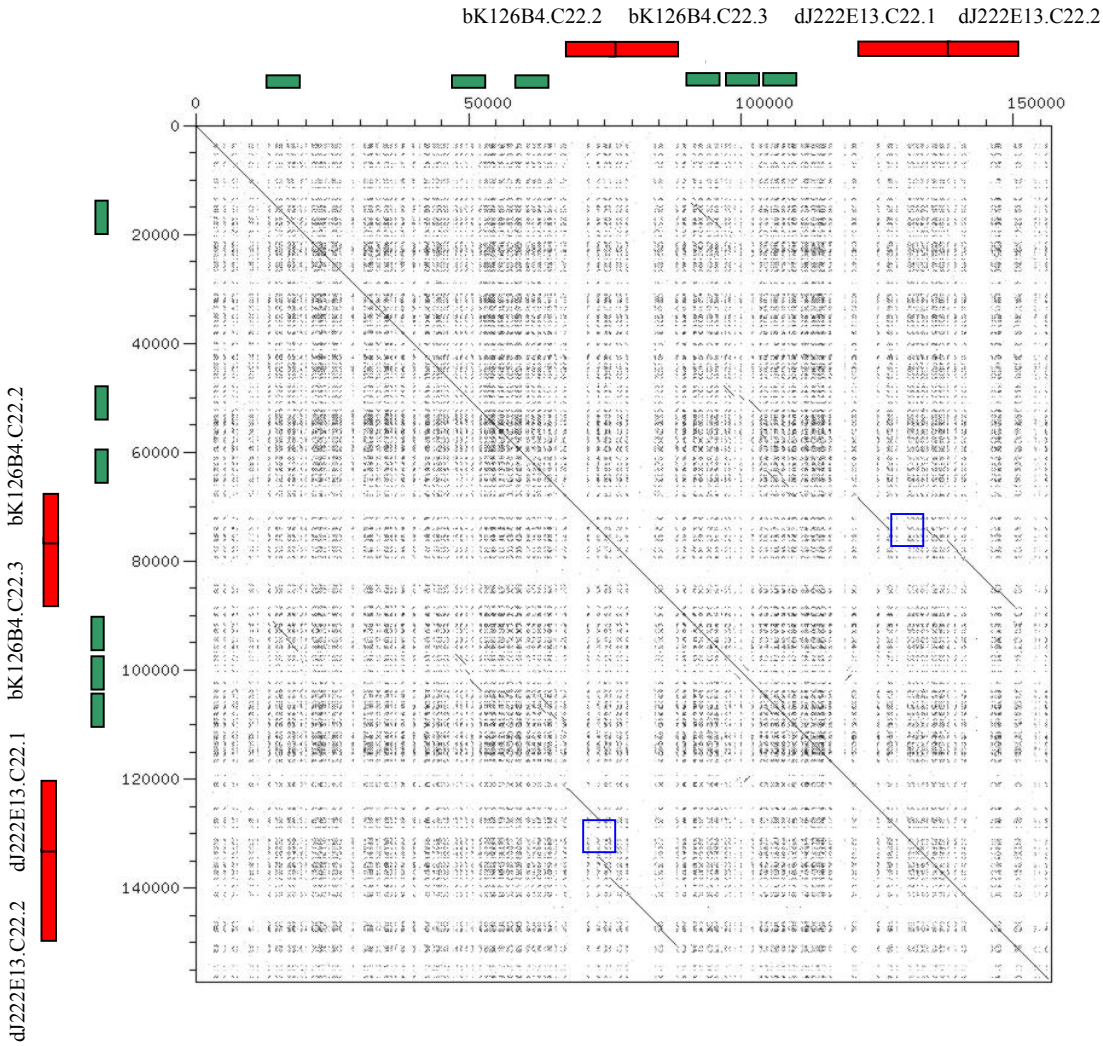


Figure 3.21: Annotated dot plot from identifying an intrachromosomal duplication within chromosome 22. 156366 bp of genomic sequence between genes bK126B4.C22.1 and dJ222E13.C22.3, containing a putatively duplicated region, is plotted against itself. Red boxes along the axes indicate gene structures 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).

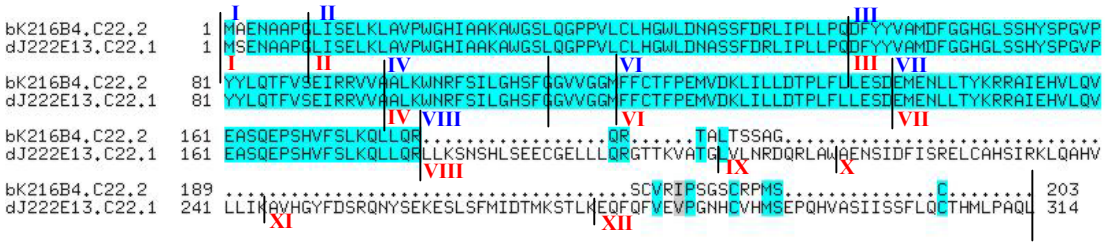


Figure 3.22: Alignment of the amino acid sequences of bK126B4.C22.2 and dJ222E13.C22.1. Exon numbers are marked in blue (bK126B4.C22.2) or red (dJ222E13.C22.1). The alignment was created using clustalw(Thompson *et al.*, 1994) and visualised using belvu (Sonnhammer, unpublished).

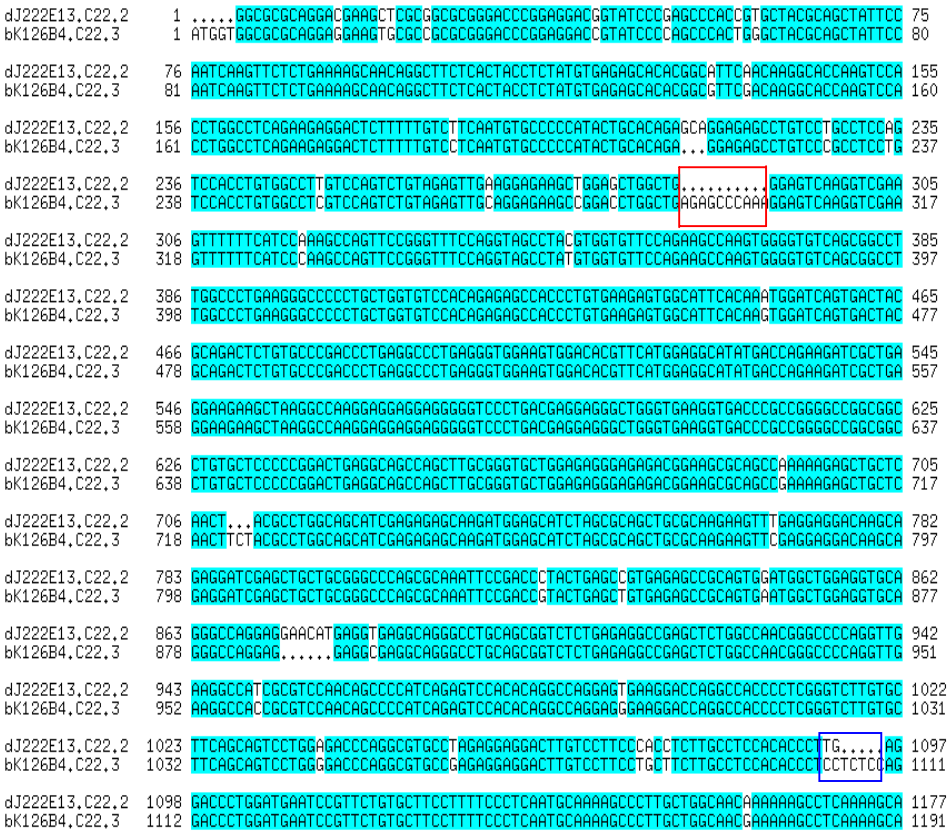


Figure 3.23: Alignment of the nucleotide sequences of bK126B4.C22.3 and dJ222E13.C22.2. A 10 bp insertion/deletion discussed in the text is marked in red and an 8 bp insertion/deletion is marked in blue. The alignment was created using clustalw (Thompson *et al.*, 1994) and visualised using belvu (Sonnhammer,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

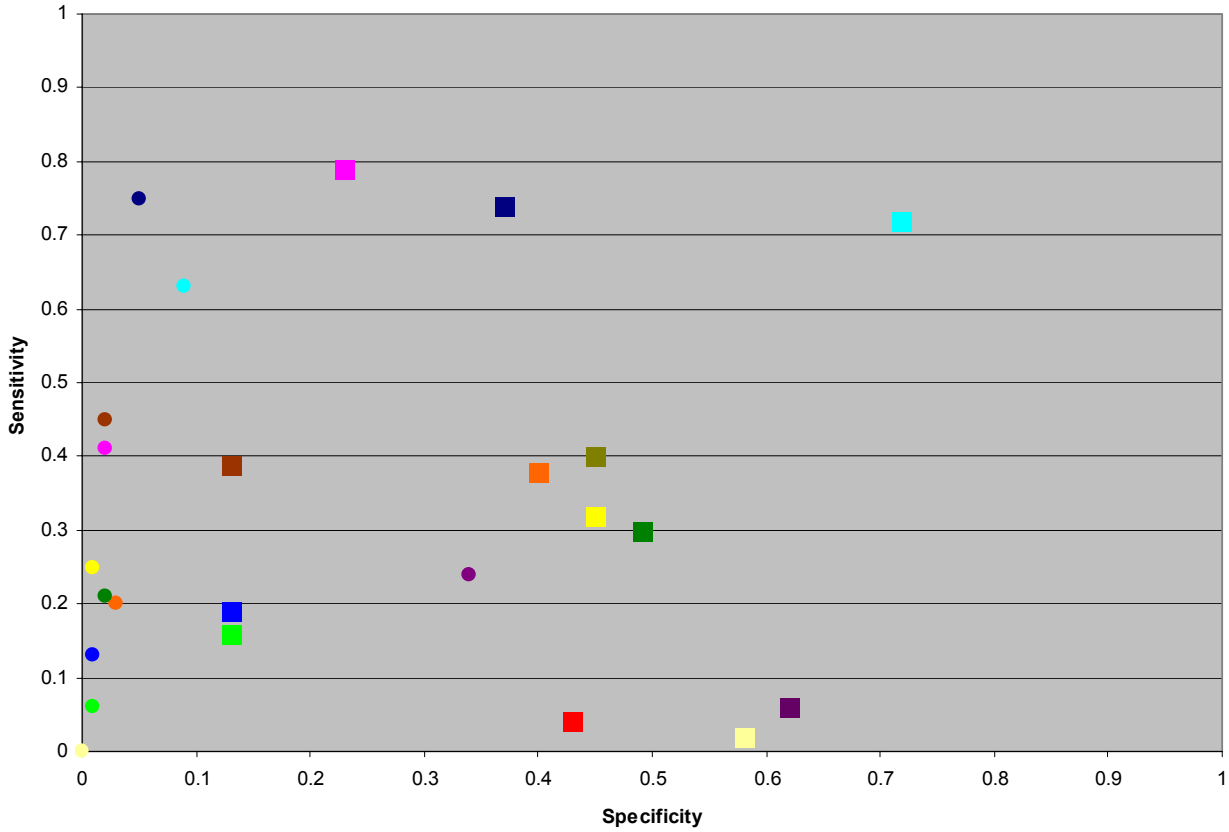


Figure 3.24: Specificity and sensitivity of sequence evidence alignment with the 22q13.31 transcript map. Sensitivity and specificity shown are computed at the nucleotide level.

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.

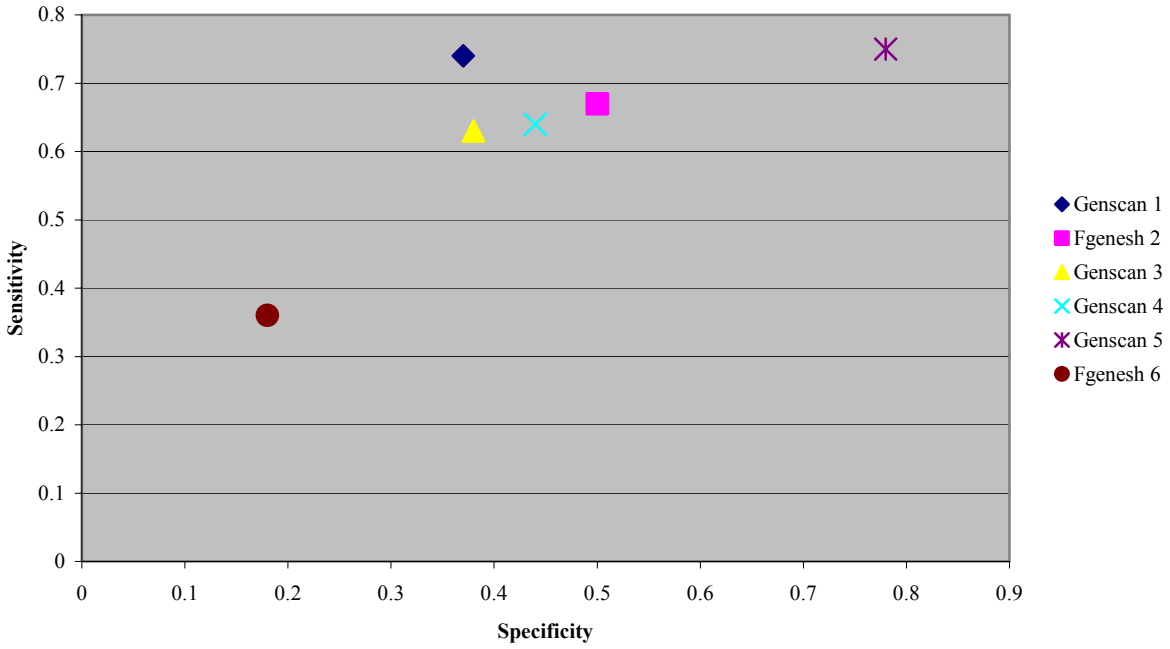


Figure 3.25: Specificity and sensitivity of the alignment of *ab initio* gene prediction programs with a variety of annotated human sequences. Sensitivity and specificity shown are computed at the exon level. The origin of each data set is shown in the legend to table 3.18.

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).

Chapter IV Comparative mapping, sequencing and analysis

4.1 Introduction

4.1.1 Benefits of comparative sequence analysis

The identification of the full complement of human genes as a result of the sequencing and analysis of the human genome in isolation seems unlikely, as discussed in chapter III. Currently, the most efficient approach to gene identification utilises expressed sequence evidence (chapter III). However, some genes with a restricted spatial or temporal expression pattern may not be represented in the available EST and cDNA resources. A second limitation of the EST databases is the paucity of 5' UTR sequences in the entries. Currently the sequence available is mainly limited to the 3'UTR of the mRNA as 5' end information is often scarce due to the method of construction of the resources used (section 3.1.3). In addition, most DNA sequences involving regulation of gene expression are in non-transcribed regions, which cannot be accessed through EST sequence.

Alternative transcript mapping methods discussed in chapter III were also noted to have limitations. For example, *ab initio* gene prediction programs require validation by a second line of evidence, as unsupported gene predictions may have only a limited level of accuracy. Additional expression-independent methods, such as exon trapping, may yield only a few exons of a gene, so an additional strategy is required to confirm the full intron/exon structure.

Comparative mapping and sequencing could aid the identification of conserved genomic regions between model organisms and human which are likely to correspond to exonic or regulatory sequences. The premise for such analyses is that functionally important sequences are conserved, whereas other regions will differ as a result of accumulated mutations since their divergence.

As significant amounts of the mouse genome are now being sequenced, the opportunity to use the mouse sequence as an analytical tool to study the human genome has become increasingly attractive. This chapter therefore focuses on utility of mouse sequence for comparative study. The human and mouse species are estimated to have diverged from a common ancestor 100 million years ago (Burt *et al.*, 1999). The level of evolutionary divergence of the two genomes is, in general, great enough to allow identification of functionally conserved regions from the rest of the genomic background, yet small enough that comparison of syntenic linkage is meaningful (Lundin, 1993).

4.1.2 The Mouse Genome Projects

The mouse genome is roughly 3000Mb in size and a number of genetic maps have been constructed. Dietrich *et al*. (1996) (1996) published an intermediate resolution mouse genetic map based on single sequence polymorphisms. A refined map, based on microsatellite markers, was published in 1998 (Rhodes *et al.*). These genetic maps served as the framework for the construction of a YAC map (Nusbaum *et al.*, 1999). An RH map of the mouse genome, incorporating many markers from the genetic map, was produced in 1999 (Van Etten *et al.*, 1999). RH maps have the benefit of allowing incorporation of all sequence-based markers into an ordered framework. These framework maps provide the resources for the construction of bacterial clone contigs, including the determination of the bacterial clone maps of regions of the mouse genome orthologous to human chromosome 22 (section 4.2).

In 1999, the National Human Genome Research Institute (NHGRI) implemented a program to analyse the mouse genome and sequence areas of biological interest. A parallel approach of restriction enzyme fingerprinting (Coulson, 1996; Gregory *et al.*, 1997; Marra *et al.*, 1997;

Olson *et al.*, 1986) and landmark-content mapping (Green & Olson, 1990) is being taken. The *C.elegans* and human mapping projects (Coulson, 1996; Lander *et al.*, 2001) have demonstrated the utility of restriction enzyme fingerprinting. Fingerprinting has the advantage that the overlap between two clones is assessed over the entire length in shared fingerprint bands, thus providing information on the extent of overlap. Landmark content mapping is based on the detection of the presence or absence of a particular small genomic segment in a clone or clones. This can be done by hybridisation experiments in the laboratory or by electronic PCR (ePCR), a sequence comparison to determine if the STS can be detected in the available genomic sequence (Schuler, 1997). The major advantage of landmark content mapping is that it allows the ordering of clones based on their landmark content by integration with existing framework maps. Together, these methods provide an accurate means to assess the extent of overlap between clones and allow the ordering and anchoring of contigs based on their landmark content (figure 4.1).

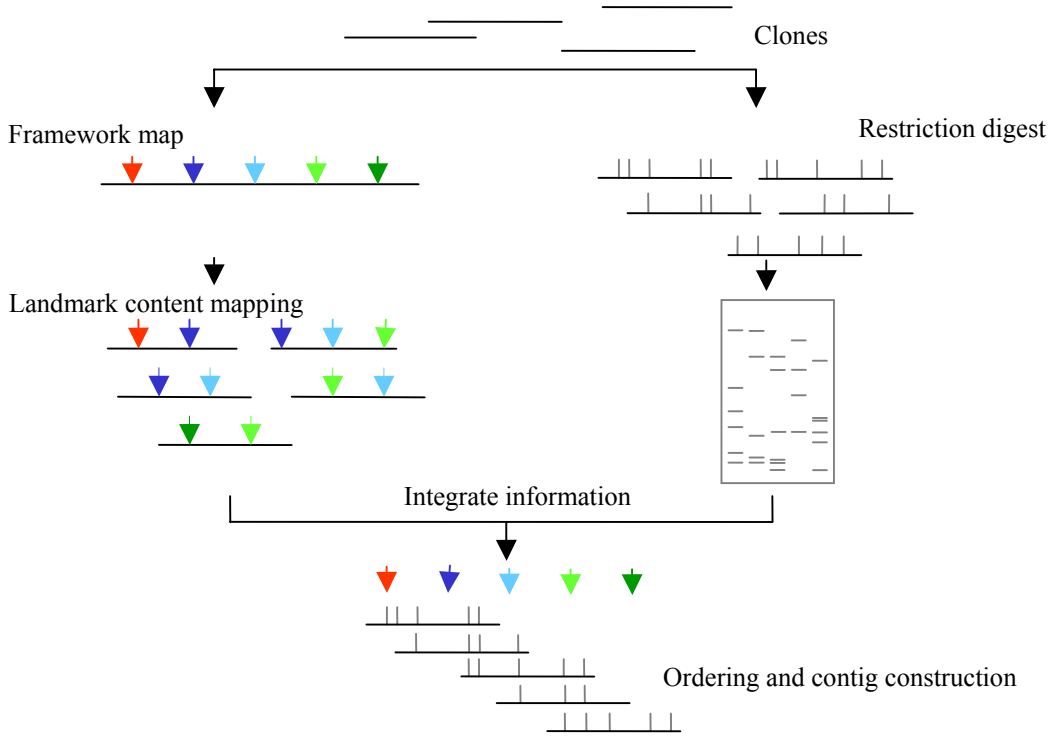


Figure 4.1: Contig construction strategy combining both landmark-content mapping and restriction enzyme fingerprinting (details are explained in the text).

Several different approaches can be used, known collectively as 'walking', to close gaps between contigs. New markers can easily be integrated into the existing framework map, or new markers that localise to the end of existing contigs can be used to isolate new clones. Alternatively, single sequence reads can be generated from clone ends using bacterial vector primers. Those sequences generated from contig ends can be used for STS design.

Resources that are now available for physical mapping projects include a database of over 300,000 fingerprinted clones from two BAC libraries constructed by P. de Jong from C57BL/6J mouse DNA (Marra *et al*., http://www.bcgsc.bc.ca/projects/mouse_mapping). One library, RPCI-23 (Osoegawa *et al.*, 2000) has been constructed from females and the other, RPCI-24, from males of the same strain. A database of sequences from the ends of the cloned genomic fragments has also been produced (Zhao *et al*., http://www.tigr.org/tdb/bac_ends/). These resources have been used to construct both small, regional BAC maps and more recently to assemble a larger physical BAC map of the whole mouse genome, now contained in fewer than 560 contigs. (The Mouse Genome Sequencing Consortium (MGSC), unpublished). The assembly incorporates 1251 framework markers previously placed on genetic and radiation hybrid maps by hybridisation assays or ePCR. A tiling path is currently being selected across the assembled BAC clone contigs, which will be subjected to standard shotgun sequencing, producing a working draft by 2003. The mouse BAC assembly has been imported into the mouse Ensembl database (http://mouse.ensembl.org), which includes predicted transcripts within finished and unfinished mouse sequence clone data.

A parallel effort to sequence the mouse genome was begun in 2000 by a public/private Mouse Sequencing Consortium (MSC). A whole genome shotgun (WGS) strategy has currently

generated over 3-fold coverage of the mouse genome sequence. Initial assembly of these sequences has started. Assembled contigs will be anchored to the mouse BAC end sequences and the available RH and genetic marker data by ePCR. The WGS sequence will then be incorporated with the sequence generated from the MGSC mapping project (Collins, http://www.nih.gov/science/models/mouse/genomics/open_letter.html).

The biotechnology company Celera is also currently engaged in work to sequence the mouse genome, using a strategy similar to that used to sequence the human genome (see chapter I), although, in this case, publicly available sequence has not been included in the assembly process. The Celera assembled and annotated mouse genome is sequenced to over 5-fold coverage representing greater than 98% of the genome, but is only available through subscription (http://www.celera.com).

4.1.3 Comparative Analysis

4.1.3.1 Alignment packages

Human and mouse genomic sequence comparison are being increasingly used to search for evolutionarily conserved regions. A variety of programs are available that allow easy identification of conserved sequences that may correspond to functionally important segments and allow the identification of novel genes and possible regulatory elements.

Percentage Identity Plots (PIPs) (Schwartz *et al.*, 2000) have become a popular method of comparing mouse and human sequence, since they allow the display of conserved regions at a range of identity levels. PIPs use the SIM program (Huang *et al.*, 1990) to identify ungapped blocks longer than 50 bp with an identity > 50%. These blocks are then plotted against the length of one of the sequences. PIPs have been used in a number of studies in regional

comparisons of human and mouse sequence (for example, Footz *et al.*, 2001; Martindale *et al.*, 2000).

The available mouse whole genome shotgun (WGS) sequence has been aligned with the assembled human draft sequence at the translated nucleotide level, using the BLAT alignment package (Kent, unpublished). The alignment can be viewed at http://genome.cse.ucsc.edu and http://www.ensembl.org. A further large-scale nucleotide alignment of the WGS sequence against the human draft sequence has been undertaken using the algorithm Exonerate (Slater, unpublished) (http://www.ensembl.org/Docs/wiki/html/EnsemblDocs/Exonerate.html).

4.1.3.2 Sequence conservation

A number of comparative sequence studies have been published, which demonstrate the conservation of exonic sequence between human and mouse genomes. Comparative sequencing of a number of regions in mouse and human, including the human and mouse β-globin gene cluster (Collins & Weissman, 1984; Shehee *et al.*, 1989); the human and rat γ-crystallin genes (den Dunnen *et al.*, 1989); the human and murine XRRC1 DNA repair gene regions (Lamerdin *et al.*, 1995); the human, mouse and hamster ERCC2 regions (Lamerdin *et al.*, 1996); a gene rich cluster at human chromosome 12p13 and its syntenic region on murine chromosome 6 (Lamerdin *et al.*, 1996); the mouse and human AIRE regions (Blechschmidt *et al.*, 1999); human and mouse T-cell receptor C-δ and C-α regions (Koop & Hood, 1994); human and hamster α - and β-myosin heavy chain genes (Epp *et al.*, 1995); human and murine Bruton's tyrosine kinase loci (Oeltjen *et al.*, 1997); the human and murine ABCA1 regions (Qiu *et al.*, 2001), has underlined the value of comparative sequence for gene annotation.

Conservation of non-coding sequences may, in some cases, arise due to functional constraint, or may be the result of a lack of divergence time. The latter premise suggests that different portions of the human and rodent genomes may evolve at different rates (Hardison *et al.*, 1997; Koop, 1995; Wolfe *et al.*, 1989). This was supported by Makalowski *et al*.(1998), who demonstrated that protein sequence conservation varied from 36% to 100% in a set of 1196 orthologous mouse and human protein sequences.

Many of the regions conserved between the human and mouse genome may correspond to yet unidentified human genes. A recent study, which described the annotation of 21, 076 full-length mouse cDNAs (Kawai *et al.*, 2001), identified 817 mouse transcripts for which no corresponding human gene had been described. The data indicates that comparative sequence analysis could be an important tool in identification of previously unknown genes.

Additionally, conserved non-coding regions may highlight regulatory sequences. Gumucio *et al*. (1988) described such a comparison of potential human and mouse promoter sequences, in order to identify the determinant of tissue specificity of amylase gene expression. The first large-scale study of non-coding sequences compared 100 kb of human and mouse DNA containing the T-cell receptor family (Hood *et al.*, 1995). The non-coding regions of this gene cluster proved to have an unusually high level of sequence conservation. In a more typical 100 kb segment from chromosome 2p13, 1% of the sequence was accounted for by conserved elements of length >80 bp with sequence identity >75% (Jang *et al.*, 1999). Loots *et al*. (2000) demonstrated the function of a conserved non-coding segment from a multi-species sequence comparison of a 1 Mb region containing an interleukin gene cluster. Deletion of a conserved non-coding element was shown to alter interleukin expression in T cells of transgenic mice.

4.1.3.3 Chromosome evolution

Comparative analysis of human genetic and physical maps with those of other organisms, has allowed mapping of the synteny relationships. Chromosome 22, for example, is a recently formed chromosome that is only found in higher primates. In lemurs and most other primates, information from HSA22 is found on at least two different chromosomes, both of which also contain different subsets of HSA12 (Muller *et al.*, 1999). These human chromosomes are posited to have formed from a single reciprocal translocation involving two ancestral chromosomes (Haig, 1999). In contrast, information from HSA22 is found at 21 different sites on eight different mouse chromosomes.

Several studies have suggested that repeated sequences might be associated with genetic instability, possibly leading to evolutionary rearrangement events. For example, the breakpoint of translocations (HSAXp11; HSA1q21) associated with papillary renal cell carcinoma (RCC) were mapped to a small region of HSA1q21 between SPTA1 and a clustered gene family, including CD1C, CD1B, CD1D, CACY and at least four other members (Weterman *et al.*, 1996). Interestingly, the boundary between two segments of HSA1q21 that are related to MMU1 and MMU3 respectively, is located between SPTA1 and CD1C, a region of <200 kb (Oakey *et al.*, 1992). Amadou *et al*. (1995) also reported a syntenic breakpoint in the HSA6p MHC class I gene region, within a tandemly organised family of genes. Related sequences are found on both MMU13 and MMU17.

Sequence analysis permits finer scale mapping of the human-mouse synteny relationships. Pletcher *et al*. (2000), has described the first sequence level analysis of a synteny breakpoint at one of these sites, an 18 kb region of mouse chromosome 10 (MMU 10) containing the junction

of material represented on HSA21 and HSA22. The minimal junction region on MMU10 contains a variety of repeats, including an L32-like ribosomal element and low-copy sequences found on several mouse chromosomes and represented in the mouse EST database. Similar comparative sequence studies could yield further information about the mechanisms of chromosomal evolution.

4.1.4 This chapter

This chapter aims to examine the importance of comparative mapping and sequencing in identifying genes and their control regions. The construction of three mouse clone contigs across the orthologous regions of human chromosome 22 is described. Generated mouse genomic sequences, in both finished and unfinished form, were used in extensive comparative analyses against orthologous human sequences. Dot and percentage identity plots showed extensive conservation of coding regions. The extent of the correlation between the conserved mouse sequence evidence and the annotated transcript map of 22q13.31 was analysed and compared with sequence evidence from other model organisms.

Conserved non-coding sequences were examined for the presence of potential exonic or regulatory features. More detailed analysis of gene structures and sequence content was undertaken on a 0.5 Mb region of finished mouse sequence. This region included sequence from a mouse clone found to span an ‘unclonable’ region in the human chromosome 22 sequence (Dunham *et al.*, 1999).

The utility of mouse genome sequence in the analysis of synteny breakpoints was also examined. A synteny breakpoint junction region between mouse chromosomes 15 and 8 on

human chromosome 22q13.1 was refined through comparative analysis of human and unfinished mouse sequence and the sequence of the junction region was analysed.

4.2 Production of regional mouse BAC maps

4.2.1 Bacterial clone contig construction

The initial framework map used for anchoring bacterial clone contigs was the chromosome 22 transcript map (Dunham *et al.*, 1999). BLAST searches were used to identify mouse cDNA sequences orthologous to cDNAs situated within the 3.4 Mb region of human chromosome 22q13.31 and a 1.9 Mb region of 22q13.1. STSs were designed to the 19 mouse mRNA sequences that were identified by this method. To increase marker density, 39 further STSs were designed from mouse ESTs that demonstrated a level of 100% nucleotide identity to the set of human cDNAs.

In order to isolate mouse clones spanning the three orthologous regions of interest, 11.2X genome equivalents of the female mouse BAC library RPCI-23 (strain C57BL/6J) (Osoegawa *et al.*, 2000) were screened by hybridisation (see figure 4.3).

In initial library screens, four pools of STS PCR products were used. The pools identified 111, 135, 199 and 132 clones respectively (table 4.1). In total, 307 clones were identified (taking redundancy into account). The identified BAC clones were transferred into microtitre plates to form a region-specific library subset. To verify the identified clones, arrayed clone filters (polygrids) were screened with all the markers from the pools individually (figure 4.2). Both the verification and the initial screening data were collated and integrated into 22ace.

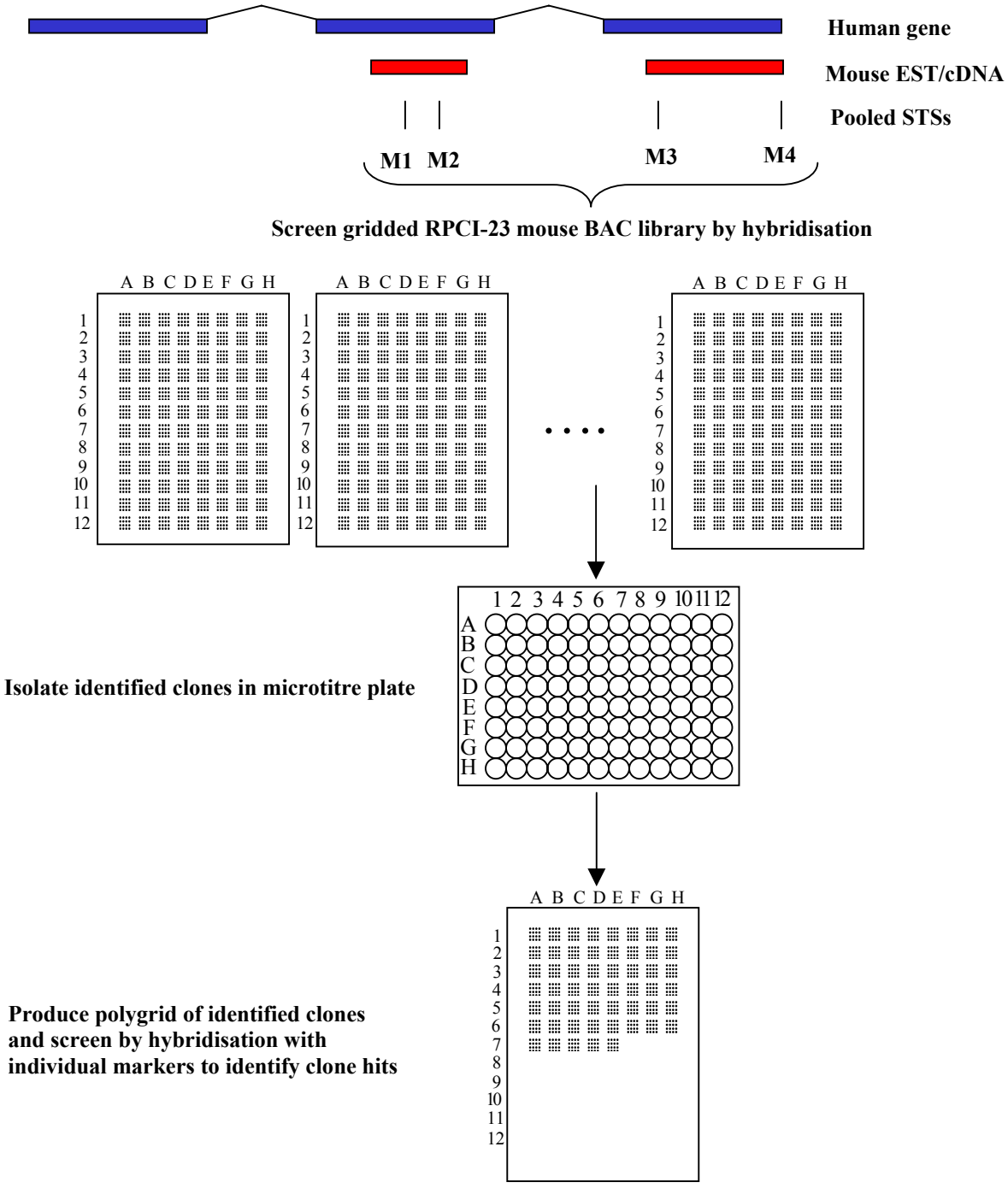


Figure 4.2: Screening strategy. Mouse cDNAs/ESTs homologous to human genes were used to design PCR primers (M1-M4). These were pooled and used to screen arrayed filters of the mouse RPCI-23 bacterial clone library. All identified positive clones were transferred to microtitre plates and gridded onto a specific mouse polygrid. This was then screened with the individual markers to identify specific clone-marker relationships.

Table 4.1: Numbers of pools, markers and isolated clones in the initial library screens

Pool	Contains marker type		BACs
	mRNA	EST	
Pool1	11	0	111
Pool2	1	18	135
Pool3	1	11	199
Pool4	10	10	132
Total	23	39	577
			307*

* Taking into account redundancy between the pools

4.2.2 Fingerprinting

BAC clones from duplicate copies of the microtitre plates were fingerprinted using *Hin*dIII (chapter II). The contigs were built using the program FPC (fingerprinting contig) (Soderlund *et al.*, 1997). FPC automatically clusters fingerprinted clones into contigs using a probability of coincidence score. FPC also allows integration of landmark content data with the fingerprint data, thus providing a workbench for contig assembly, verification and selection of sequence tile path clones.

4.2.3 Landmark content mapping

In addition to fingerprinting, maps were also constructed by landmark-content mapping. Polygrids were screened with each of the markers generated from cDNA information. From the hybridisation results, contigs could be constructed based on shared landmark content using the strategy described in figure 4.1. The initial rounds of screening led to the construction of 33 contigs spanning an estimated 3.8 Mb. (Comparison of sequence and fingerprint data determined that for the mouse library clones, a single fingerprint band corresponded to an average of 5 kb. This figure was used to estimate the size of a region based on the number of fingerprinting bands.)

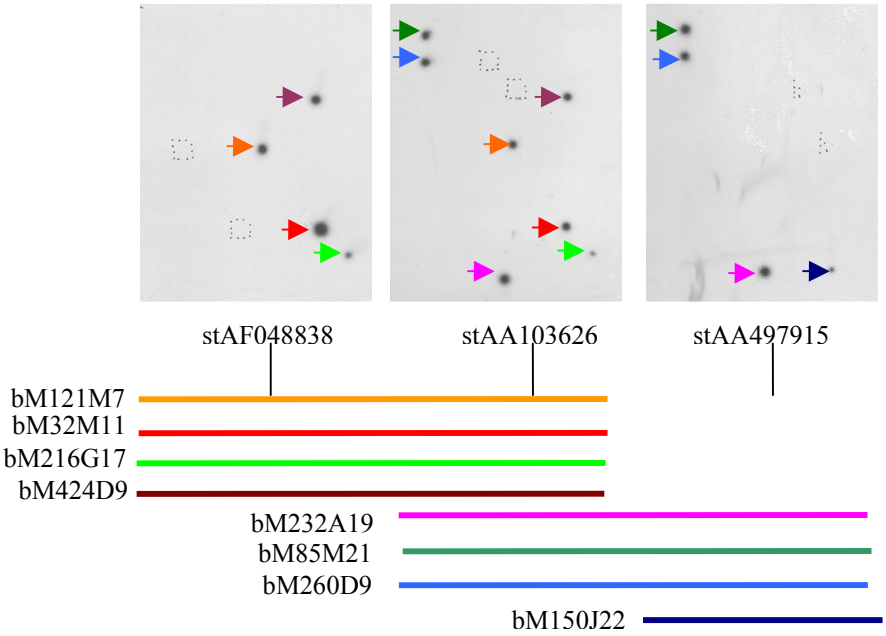


Figure 4.3: Example of landmark-content mapping using three landmarks (stAF048838, stAA103626 and stAA497915). The positives are indicated by coloured arrows, the clones drawn below in corresponding colours.

4.2.1.4 Gap closure

Two strategies were utilised to link the contigs. Initially, the publicly available BAC clone end sequences (Zhao *et al.,* unpublished) were used to design PCR primers to those BACs on the ends of the contigs for further library screens. Five pools of clones were screened in two successive rounds of walking which resulted in the identification of 508 clones. Subsequent fingerprinting and mapping of these clones allowed 25 gaps to be filled.

Table 4.2: Numbers of pools, end STSs and isolated clones in gap closure screens

Pool	End STS	BACs
Pool5	17	137
Pool6	23	203
Pool7	23	122
Pool8	23	186
Pool9	17	132
Total	103	880
		508*

* Taking into account redundancy between the pools

As an increasing number of fingerprints (Marra *et al*., unpublished) and end sequences (Zhao *et al*., unpublished) from the mouse BAC library became available, they were anchored by ePCR and hybridisation using publicly available genetic and radiation hybrid markers (Gregory *et al*., unpublished)(section 4.2.5). Incorporation of this data enabled closure of two gaps. Additionally, the information allowed two spurious contigs, containing 261 clones and 31 markers designed to murine genes or EST sequences, that did not map to the correct mouse chromosome and 68 singletons to be discarded.

NB. The three contigs generated during this project have since been incorporated into the large-scale physical mouse mapping effort. Further work has resulted in joining of the two mouse chromosome 15 contigs, creating a contig spanning approximately 6.7 Mb of mouse sequence.

4.2.4 Tile Path Clones

During contig construction, clones with sufficient mapping information (i.e. both landmark and fingerprinting data) were selected for sequencing (Richard Evans, Sanger Institute and M. Goward). Tiling path clones across the three contigs were selected to ensure that minimal overlap of clones reduced redundant sequencing.

4.2.5 Features of the sequence-ready bacterial clone map

The three contigs incorporated 486 BAC clones in total and the final sequence tile paths, containing 34 clones, cover an estimated 3.96 Mb (excluding overlapping sequences). The clone contigs are depicted in figure 4.4. The division of this set of clones is summarised in table 4.3.

Table 4.3: Clone contig data showing the number of clones within the contigs, the number of clones selected for sequencing and the approximate length of the contig

Contig	Mouse chromosome	Orthologous region	Total # clones	# clones in tile path	Approx. length (Mb)
A	15	22q13.31	229	13	2.00
B	15	22q13.1	164	15	1.59
C	8	22q13.1	93	6*	0.37

*including two clones sequenced by the Albert Einstein College of Medicine Human Genome Research Center (AECOM) and the University of Oklahoma Advanced Center for Genome Technology (UOKNOR) respectively.

The maps also incorporate 54 markers from a range of mouse maps listed in the UniSTS (http://www.ncbi.nlm.nih.gov/genome/sts/index.html) database, that have been positioned by ePCR against available mouse sequence (Gregory *et al*, unpublished). Shared markers between different map types allow integration of the sequence-ready map with previously published mouse maps and confirmed the chromosomal location of the mouse contigs. The incorporation of marker types into the contigs is shown in table 4.4.

Table 4.4: Incorporation of marker information into mouse contigs A, B and C

Contig	Mouse chromosome	Orthologous human region	Marker Type				Total no. markers
			mRNA	EST	End STS	UniSTS	
A	15	22q13.31	4	7	27	15	53
B	8	22q13.1	5	2	5	6	19
C	15	22q13.1	6	6	6	33	55

4.2.6 Sequencing

The tile path clones were shotgun sequenced (chapter I) (Sanger Institute sequencing teams). During the project, sequence was released by other groups for several other clones in the contigs. Where possible, these clones were incorporated into the tile path to minimise redundant sequencing.

At the time of writing, finished sequence was available for nine (26%) clones and unfinished shotgun sequence was available for a further 18 (53%) of the 34 tiling path clones. These clones are highlighted in the FPC display shown in figure 4.5. A table of the sequenced mouse clones showing their genomic location, accession number, author, orthologous human region and current sequencing status is shown in appendix 5.

Approximately 85% of 22q13.31 is spanned by mouse clones that have at least unfinished sequence. Approximately 92% of the region of human chromosome 22q13.1 under investigation is spanned by unfinished/finished mouse sequence (see figures 4.2 and 4.6).

Figure 4.4 (foldout): Bacterial clone contigs containing mouse genomic sequence spanning regions of conserved synteny with a) human chromosome 22q13.31 and b) human chromosome 22q13.1. The human transcript map of each region is depicted at the top of each diagram: full genes are shown in dark blue, partial in light blue and pseudogenes in green. Gene structures orientated 5' to 3' on the DNA strand from centromere (left) to telomere (right) are designated '+' and those on the opposite strand '-'. Markers designed from murine sequences orthologous or similar to the named human genes are shown in black. These markers are positioned relative to both the human transcript map and the mouse clone contigs. Mouse chromosome specific markers from the UniSTS database are shown in pink and are positioned relative to the mouse clone contigs only. The .15 or .8 of these marker names refers to the specific murine chromosome. Conserved mouse genes (identified from dot and PIP analyses (section 4.3) are indicated by red arrows. The mouse clone contigs are shown in red below. Figure a shows part of contig A, a region of MMU15 with conserved synteny to 22q13.31. Figure b. shows parts of contigs B and C, from MMU8 and MMU15 respectively. The hashed red blocks denote clones that extend beyond the region of synteny with HSA22q13.1. Only relevant regions of the contigs are shown: clones that extend these contigs further have been mapped (see figure 4.5) but do not yet have sequence.

TAKE THIS PAGE OUT – foldout figure 4.4a

Take this page out too!!! fig 4.4b

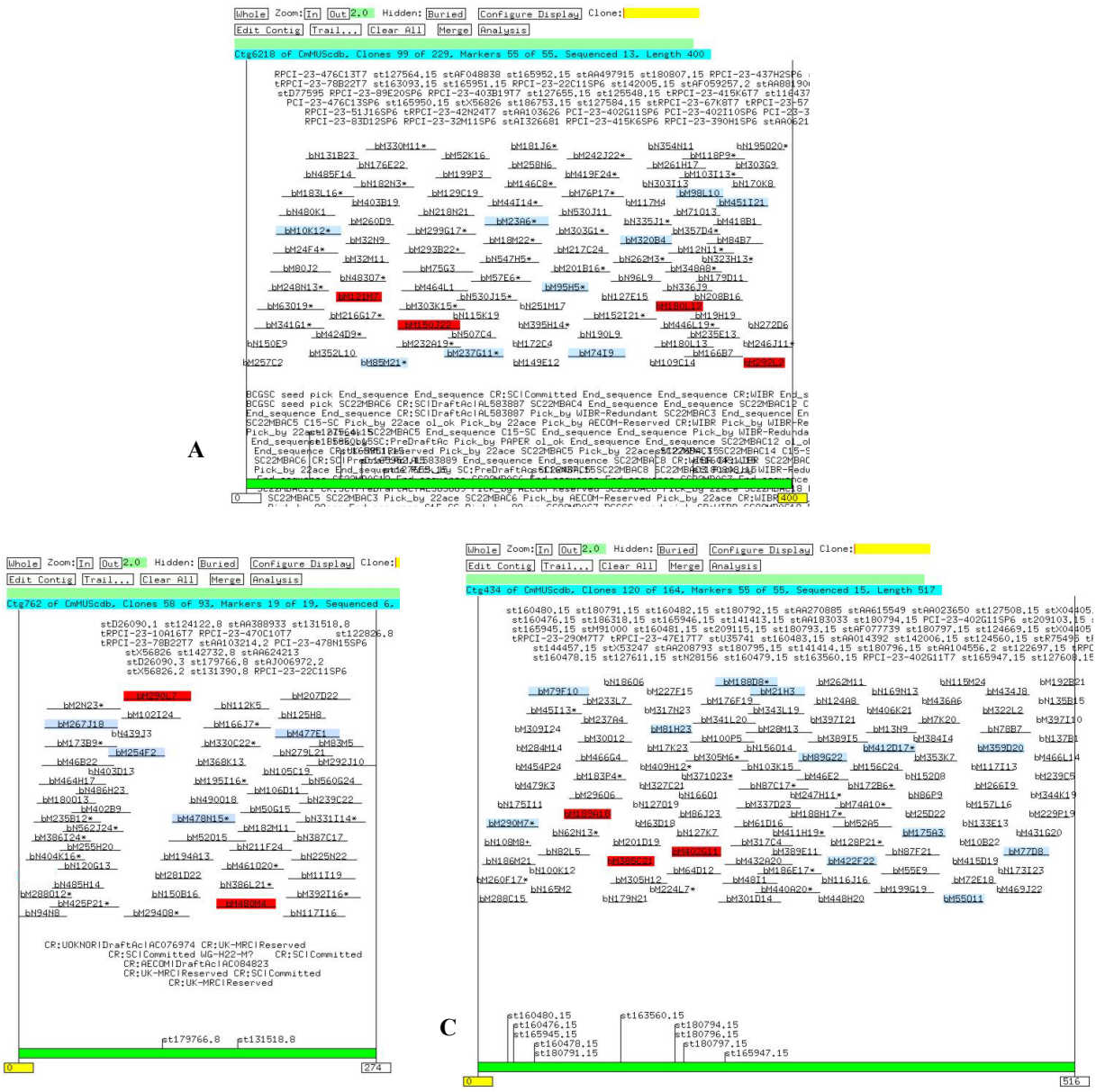


Figure 4.5: FPC display of mouse BAC clone contigs spanning orthologous regions of HSA22.

A) Contig spanning region of mouse chromosome 15, orthologous to human chromosome 22q13.31.

B and C) Contigs spanning regions of mouse chromosomes 8 and 15 respectively, encompassing a synteny breakpoint with human chromosome 22. Contig diagrams extracted from FPC (Soderlund *et al.*, 1997). Tiling paths are indicated in blue and finished sequence clones are highlighted in red.

4.3 Comparative sequence analysis

4.3.1 Dot plot analysis

Available sequence from the three mouse clone contigs (appendix 5) was compared against the orthologous human sequence using the dot plot program from the advanced PipMaker analysis tools available at http://bio.cse.psu.edu/pipmaker (Schwartz *et al.*, 2000). This program is similar to Dotter (Sonnhammer & Durbin, 1995), used in chapter III, but reports only matches contained within a statistically significant alignment. Another feature of this program is that unfinished sequence contigs can be ordered according to their alignment to a second, base sequence. Figures 4.6a and 4.6b show annotated dot plots of the two regions of chromosome 22, aligned against the mouse ordered sequence contigs. Of course, the ordering of the mouse unfinished sequence contigs derived from PipMaker is dependent upon the human reference sequence. The order shown is therefore currently unconfirmed.

The dot plots above show that areas of high similarity correspond to single or multiple genes. In regions of finished sequence, gene order and orientation appear to be conserved between human and mouse. This is supported by the distribution of markers within the contigs, shown in figure 4.4. An apparent inversion of APOL2 exists in AL592187.4, but this is likely due to the unfinished nature of this sequence. Figure 4.6a indicates that two mouse clone sequences, AL513354.14 (finished) and AL603714.4 (unfinished), span sequence gaps in the human sequence of 22q13.31. A more detailed analysis of the finished sequence AL513354.14 is shown in section 4.7. Figure 4.6b confirms the existence of a synteny junction region on human chromosome 22, between genes dJ569D19.C22.1 and MB. This is discussed in more detail in section 4.7.

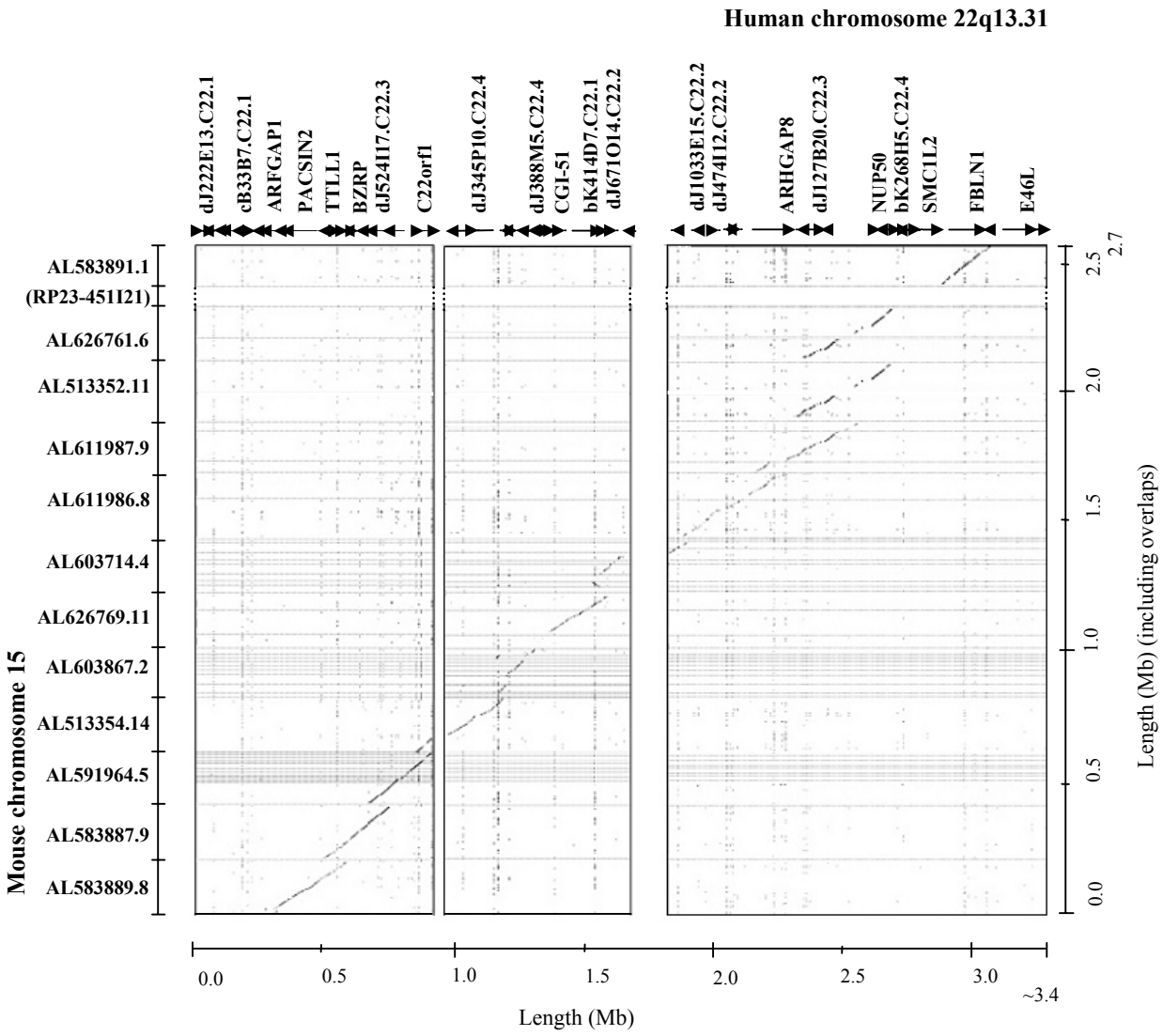
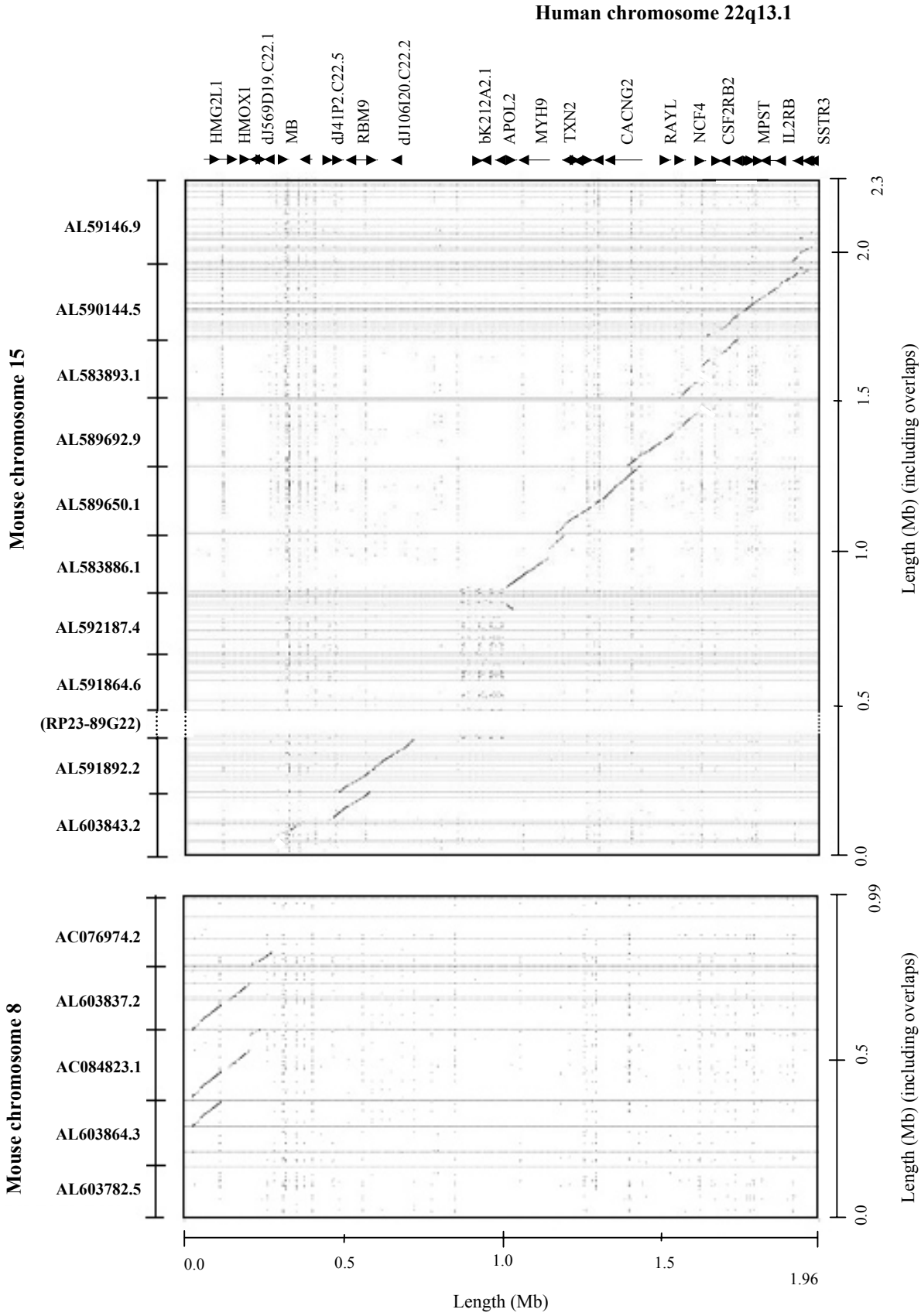


Figure 4.6a: Annotated dot plot of the human sequence of 22q13.31 (X-axis) and orthologous mouse (Y-axis) sequences from MMU 15. Genes present in the human sequence are indicated along the X-axis. Two sequence gaps of approximately ~50kb and ~75kb respectively are shown in the human sequence. The dot plot indicates that these gaps are spanned by the finished mouse sequence AL513354.14 and the unfinished sequence AL603714.4 respectively. Tiling path clone RP23-451I21, for which sequence is not yet available, spans a gap in the mouse sequence.

Figure 4.6b (overleaf): Annotated dot plot of the human sequence of a 1.96 Mb region of 22q13.1 (X-axis) and orthologous mouse (Y-axis) sequences from MMU15 and MMU8. Genes present in the human sequence are indicated along the X-axis. Tiling path clone RP23-89G22, for which sequence is not yet available, spans a gap in the mouse sequence. Further mapped clones have been selected for sequencing, which extend the tiling path along MMU15. However, sequence is not yet available for these clones and these have not been included in the diagram. The dot plot indicates that a MMU8:15 synteny junction exists between genes dJ569D19.C22.1 and MB on 22q13.1 (section 4.8).


Human chromosome 22q13.1
HMG2L1
HMOX1
dJ569D19.C22.1
MB
dJ41P2.C22.5
RBM9
dJ106I20.C22.2
bK212A2.1
APOL2
MYH9
TXN2
CACNG2
RAYL
NCF4
CSF2RB2
MPST
IL2RB
SSTR3
Mouse chromosome 15
AL59146.9
AL590144.5
AL583893.1
AL589692.9
AL589650.1
AL583886.1
AL592187.4
AL591864.6
(RP23-89G22)
AL591892.2
AL603843.2
Length (Mb) (including overlaps)
0.0
0.5
1.0
1.5
2.0
2.3
Mouse chromosome 8
AC076974.2
AL603837.2
AC084823.1
AL603864.3
AL603782.5
Length (Mb) (including overlaps)
0.0
0.5
0.99
0.0
0.5
1.0
1.5
1.96
Length (Mb)

4.3.2 PIP analysis - investigation of exonic conserved sequences

Repeat elements in the human and mouse sequences were identified and masked using RepeatMasker (Smit and Green, unpublished) and the resulting sequences and exon locations were submitted to the PipMaker website (http://bio.cse.pse.edu/pipmaker) (Schwartz *et al.*, 2000) (section 4.1.3.1). An overview of conserved gene structures, derived from the PIP comparisons, is shown in figure 4.4. An example of a PIP, showing in finer detail the alignments made between a region of the human and mouse sequences, is shown in section 4.7.

The coding exons of conserved genes are easily identified by visual inspection of the PIPs. Untranslated regions of exons often show a decrease in percent identity compared to the protein-coding portion of the gene (see the BZRP gene region from ~112K to 124K in figure 4.12). The number of human gene features from each region demonstrating >50% nucleotide identities to gap-free segments of mouse sequence are listed in table 4.5. Overall, over 75% of the annotated human exons, which lay within regions spanned by finished/unfinished mouse sequence, could be aligned with conserved sequences in the mouse.

Interestingly, no pseudogenes showed homology to the mouse sequence outside of repeat regions. The existence of a human pseudogene on human chromosome 22 (CYKB2-ps) that does not have a murine orthologue, has previously been demonstrated by Lund *et al*. (2000) through comparative sequence analysis. A further study has described non-conservation of the human pseudogene EEF1B3 in the mouse genome, although this research was not performed at sequence level (Chambers *et al.*, 2001). These human pseudogenes may have arisen since the divergence of the human and mouse lineages. Alternatively, these non-functional

sequences may have diverged more rapidly in the mouse genome, perhaps because of the shorter murine generation time.

Additionally, no homology was found in the mouse to four human genes: HMG17L1 and dJ1033E15.C22.1 from 22q13.31, and dJ1119A7.C22.4 and dJ1119A7.C22.5 from 22q13.1. This list is not definitive, as analysis of the finished sequence may show further differences. These four gene structures are categorised as partial (see chapter III). It may be that sequence conservation of these genes will be noted when the complete mouse sequence is available. Alternatively, some or all of these features may be pseudogenes (see above) or may be true genes that are not conserved in the mouse sequence.

Table 4.5: Overview of PIP results from comparisons of available mouse genomic sequence to two regions of human chromosome 22.

		No. human gene features spanned by sequenced mouse clones (finished and unfinished sequence)				No. human gene features demonstrating >50% nt. identity to gap-free segments of mouse sequence			
Human Region	Mouse coverage (%)	No. genes	No. exons	No. pseudo-genes	No. pseudo-gene exons	No. genes	No. exons	No. pseudo-genes	No. pseudo-gene exons
22q13.31	85	29	378	12	12	26	243	0	0
22q13.1 (MMU 8)	92	4	55	1	3	4	53	0	0
22q13.1 (MMU15)		29	199	5	5	27	183	0	0
Total	88.5	62	632	18	20	57	479	0	0

Sequence from HSA 22 (6 Mb) was compared against syntenic mouse sequence using the PipMaker website (http://bio.cse.psu.edu/pipmaker) (Schwartz *et al.*, 2000). The resulting PIP was analysed by eye. Coverage shows the estimated amount of the human sequence (%) for which the equivalent orthologous mouse sequence (finished or unfinished) is available. The number of genes and pseudogenes annotated within the human 'covered' region is shown, together with the total number of exons in each category. The numbers of genes, pseudogenes and exons that demonstrate >50% nucleotide identity to gap-free segments of mouse sequence are listed.

4.3.3 Integration of mouse genomic data into 22ace

In order to allow detailed comparison between the mouse genomic data generated during this project, the annotated gene structures described in chapter III and additional data such as

Genscan predictions, it was necessary to generate an alignment of the available mouse genomic sequence with the sequence of 22q13.31 in a format that could be incorporated into the 22ace database.

The program MatchReport (Smink *et al*., unpublished) generates an ace format file from BLAST alignments above a set percentage identity. In order to determine an appropriate value for percentage identity for a local alignment of orthologous mouse unfinished sequence data against human 22q13.31, a preliminary comparison was performed, using three mouse clone sequences against the orthologous human regions using MatchReport at a range of percentage identity values. Repeats in the sequences were masked using RepeatMasker prior to alignment (Smit and Green, unpublished). The compared regions are shown in table 4.6

Table 4.6: Mouse clones and orthologous regions of HSA22q13.31 selected for percentage identity calibration experiment

Mouse clones	Orthologous region of HSA 22q13.31	Size of region (human) (kb)	No. annotated human genes	No. annotated human exons
AL603867, AL513354	dJ345P10.C22.4 – dJ388M5.C22.4	300	3	52
AL583887	TTLL1 – dJ526I14.C22.3	150	6	60
	Total	**450**	**9**	**112**

The generated files were read into 22ace. Values of specificity and sensitivity for each percentage identity value (see chapters II and III) were calculated at a nucleotide level and plotted in figure 4.7.

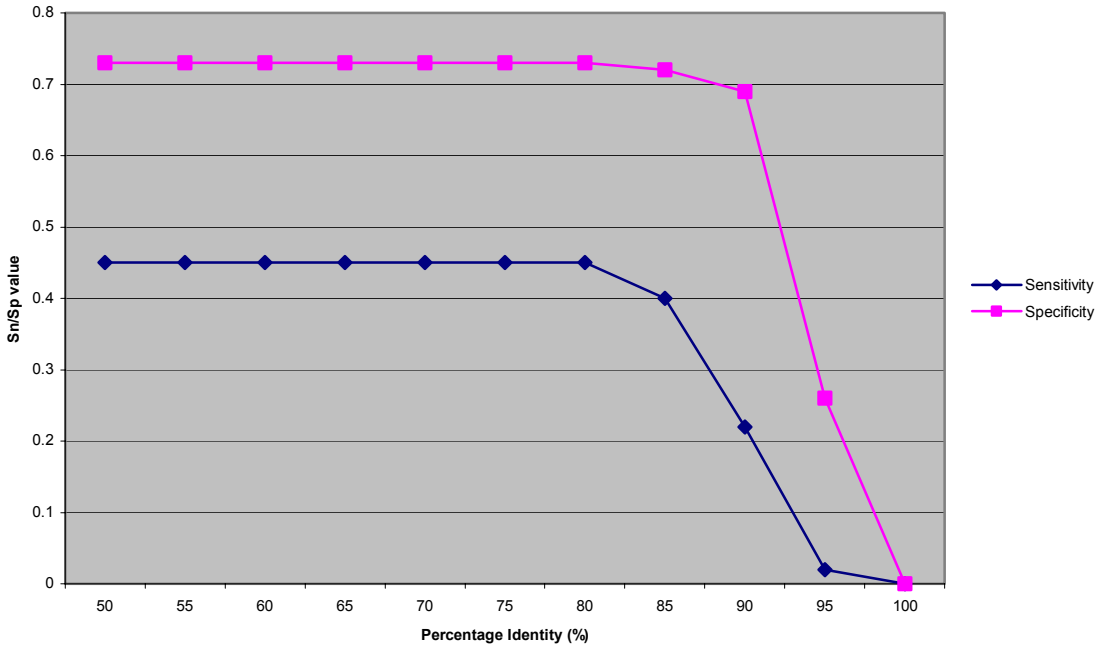


Figure 4.7: Sensitivity and specificity of MatchReport BLAST results from three mouse clone sequences against the equivalent human genomic sequence. The perl script MethComp (D. Beare) was used to calculate specificity and sensitivity of mouse hits to nucleotides contained within exons

These results show that both specificity and sensitivity are compromised if the percentage identity level is raised beyond 80% in this region. Surprisingly, sensitivity did not increase, or specificity decrease, as percentage identity dropped below this level to 50%. A cut-off identity level of 80% was therefore deemed appropriate for a comparative study of this region in order to maximise specificity, without loss of sensitivity. Available mouse sequence from contig A was thus aligned to the human sequence from 22q13.31 using MatchReport at a percentage identity of 80%.

4.4 Correlation of comparative genomic data with 22q13.31 transcript map

The mouse WGS sequence (MSC, unpublished) has been aligned to the draft human genomic sequence using BLAT and Exonerate (section 4.1.3.1). Results specific to HSA22 have been incorporated into 22ace. Additional sequence resources, derived from mouse and other organisms and incorporated into the 22ace database, include sequence from a library of full-length mouse cDNAs (Kawai *et al.*, 2001), output from the ExoFish program (Roest Crollius *et al.*, 2000), which assesses TBLASTX sequence homology to available *T. nigroviridis* genomic sequence, and the translated predicted protein sequences from the *D. melanogaster* (Adams *et al.*, 2000) and *C. elegans* (Coulson *et al*., 1996) sequencing projects. An example of a 22ace display showing alignment of these features to the gene dJ526I15.C22.2 is shown in figure 4.8. The diagram shows that both mouse genomic sequence resulting from this project and mouse cDNA sequence (Kawai *et al.*, 2001) both align to the human sequence along the full length of the gene dJ526I14.C22.2. Output from the Exofish program (Roest Crollius *et al.*, 2000) aligns to only two exons of this gene.

The perl script MethComp (Dave 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. 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, as before (chapter III). In all cases, multiple hits were counted as one hit.

These results are shown in table 4.7. A plot of the specificity and sensitivity of each type of evidence at the nucleotide level is shown in figure 4.9. Further details of this analysis can be found in chapter II.

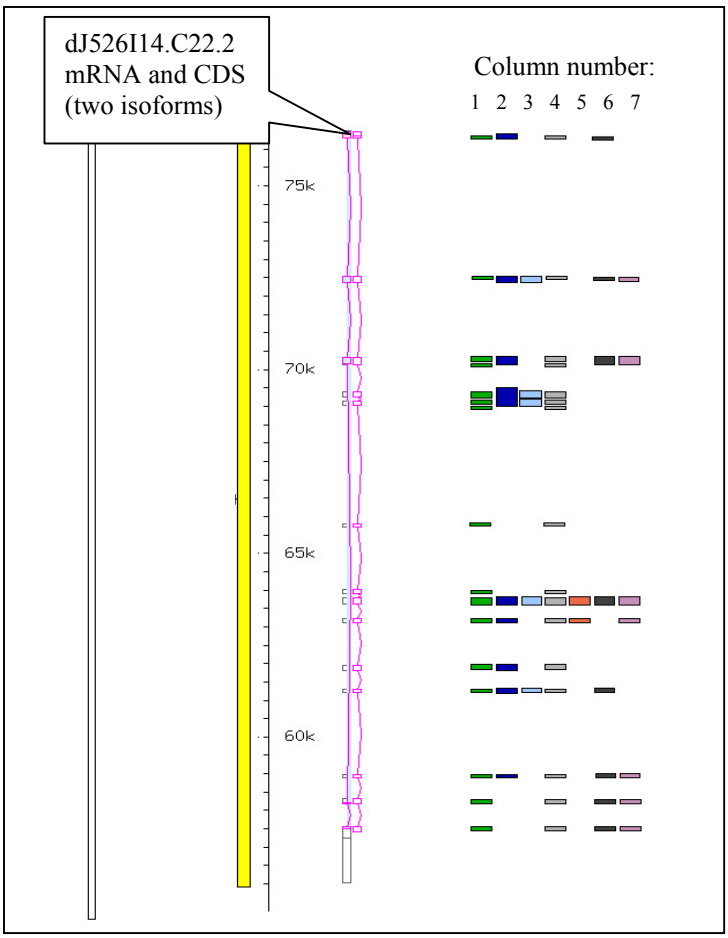


Figure 4.8: 22ace display showing the region surrounding the gene dJ526I14.C22.2. Sequence alignments are shown in columns to the right of the gene structure. Two isoforms of dJ526I14.C22.2 are depicted.
1= Blastn_mus: genomic mouse sequence generated as a result of this project.
2 = Blatmouse: WGS mouse sequence (MSC, unpublished) aligned against the draft human genome sequence with BLAT (Kent, unpublished).
3 = ExoMouse: WGS mouse sequence (MSC, unpublished) aligned against the draft human genome sequence with Exonerate (Slater, unpublished).
4 = fantom: Collection of full-length mouse cDNA sequences (Kawai *et al.*, 2001).
5 = Exofish: Exon prediction program utilising T. nigroviridis genomic sequence (Roest Crollius *et al.*, 2000).
6 = flypep: translated predicted D. melanogaster genes (Adams *et al.*, 2000).
7 = wormpep: translated predicted C. elegans genes (Coulson, 1996).

Additional features have been removed from the display to aid clarity.

Table 4.7 Analysis of the correlation of the evidence types available from different organism genome or gene identification projects used to annotate genes against:

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

Evidence type	Method	Organism	Alignment	Nucleotide			Exon	Gene
				Total Coverage	Sp	Sn		
Genomic	Blastn_mus*	M. musculus	BLASTN	0.016	0.62	0.36	0.60	0.88
Genomic	Blatmouse*	M. musculus	BLAT	0.015	0.51	0.27	0.53	0.78
Genomic	Exomouse*	M. musculus	Exonerate	0.017	0.45	0.26	0.50	0.82
cDNA	fantom*	M. musculus	BLASTN	0.002	0.49	0.03	0.10	0.34
Exon prediction	Exofish*	T. nigroviridis	ExoFish	0.005	0.76	0.12	0.30	0.58
Protein	flypep*	D.melanogaster	BLASTX	0.006	0.69	0.15	0.33	0.56
Protein	wormpep*	C. elegans	BLASTX	0.002	0.58	0.04	0.10	0.17

* Descriptions and references of each method are given in the legend of figure 4.8.

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	Organism	Alignment	Nucleotide			Exon	Pseudogene
				Total Coverage	Sp	Sn		
Genomic	Blastn_mus*	M. musculus	BLASTN	0.016	0.00	0.00	0.00	0.00
Genomic	Blatmouse*	M. musculus	BLAT	0.015	0.12	0.44	0.58	0.76
Genomic	Exomouse*	M. musculus	Exonerate	0.017	0.12	0.45	0.65	0.76
cDNA	fantom*	M. musculus	BLASTN	0.002	0.45	0.18	0.41	0.64
Exon prediction	Exofish*	T. nigroviridis	ExoFish	0.005	0.11	0.11	0.27	0.47
Protein	flypep*	D.melanogaster	BLASTX	0.006	0.13	0.18	0.27	0.47
Protein	wormpep*	C. elegans	BLASTX	0.002	0.24	0.11	0.13	0.23

* Descriptions and references of each method are given in the legend of figure 4.8.

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.

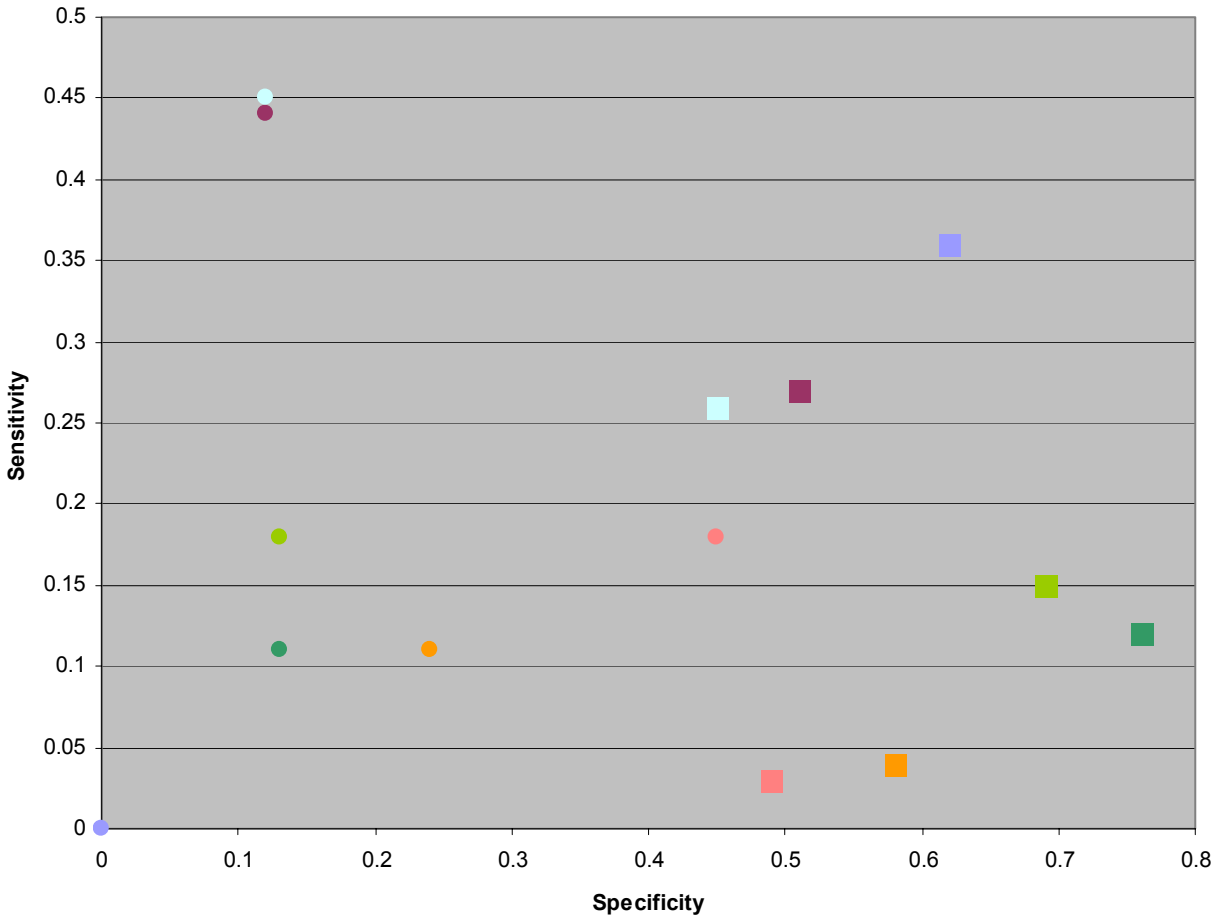


Figure 4.9: Specificity and sensitivity of different comparative sequence data with the 22q13.31 transcript map. Sensitivity and specificity shown are computed at the nucleotide level.

Blastn_mus
Blatmouse
Exomouse
fantom
Exofish
flypep
wormpep

□ = 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 figure 4.8.

Once again, the sensitivity and specificity of matches to annotated pseudogenes are, in general, lower than the correlation to annotated genes. In the case of Blastn_mus (mapped mouse genomic sequence derived from this project), no alignment to pseudogenes was noted. In comparison, BLAT and Exonerate alignments of the WGS mouse sequence demonstrated

relatively high sensitivity of correlation to pseudogene structures: this is because the WGS sequence resource is not limited to the sequence from one particular region. These matches to human pseudogenes may be from sequence of the true mouse gene, orthologous to the true human gene from which the pseudogene is derived.

This analysis shows that the highest sensitivity of correlation with the annotated genes is currently demonstrated by the mapped mouse genomic sequence resulting from this project. However, as the large-scale murine genome project is completed, and gene identification in this and in other genomes advances, values of sensitivity and specificity will alter. The highest values of specificity here originate from the Exofish gene prediction program, followed by matches to DNA and protein sequence databases. These values are comparable to those derived from human cDNA collections (chapter III) and indicates that comparison to known, or predicted, genes in other species is a powerful tool for accurate gene annotation. However, this high level of specificity is, in general, linked with lower sensitivities than those shown in chapter III and may therefore enable identification of only a subset of genes present in the region of interest.

4.5 Investigation of intronic and intergenic conserved sequences

The results shown in table 4.7 indicate that there are areas where high similarity is observed outside of the annotated human genes. These regions may just be non-functional sequences that have not diverged or could indicate the presence of regulatory element. Some of these conserved features may also be unidentified human exons. This latter possibility was initially investigated through a comparison of the conserved human-mouse sequences and Genscan predicted exons.

4.5.1 Correlation of Genscan predictions with human-mouse conserved sequences

A correlation analysis of Genscan predictions with the gene annotation of 22q13.31 is described in chapter III. From this study, 384 (58%) of the 657 Genscan predicted exons are identified as 'wrong' i.e. do not overlap an annotated true coding exon. Eighteen of the 'wrong' predictions overlap annotated pseudogenes and are therefore discounted from this analysis.

The correlation of the remaining 366 Genscan predicted exons with the Blastn_mus, Blatmouse and Exomouse alignments were manually assessed by eye from the visual display of the 22ace database. Genscan predicted only six exons outside of the annotation, which contained sequence that aligned to mouse genomic DNA. The results of this analysis are shown in detail in table 4.8

Table 4.8: The position of exons predicted by Genscan, which do not overlap annotated true exons, but overlap aligned mouse genomic sequences

Genscan exon no.	Position on human transcript map	Correlates with Human-Mouse genomic alignment:		
		ExoMouse	Blatmouse	Blastn_mus
1	intergenic			•
2	within dJ345P10.C22.4	•	•	•
3	intergenic			•
4	within dJ474I12.C22.2	•	•	•
5	within ARHGAP8	•	•	•

4.5.2 Test of expression

The three intergenic Genscan predictions had previously tested negative for expression in seven cDNA libraries by PCR (see chapter III). In a similar experiment, primers were designed to the remaining three Genscan predictions, as well as to an additional twenty-five

exon candidates identified from the Blastn_mus alignment, which were over 30 bp long and contained an ORF. Altogether, six exon candidate regions were not associated with any annotated gene structures, whilst 22, including those supported by Genscan predictions, lay within introns of annotated genes.

The twenty-eight primer pairs were used in PCR screens of seven cDNA vectorette libraries (see chapter II). Only one positive result was obtained from a candidate exon (not supported by a Genscan prediction) within the gene E46L. cDNA sequence from the resulting vectorette PCR product partially matched the existing exon structure, but appeared to result from spurious poly(dT) priming within a repeat. No new human exons or genes were therefore experimentally confirmed in this test.

4.6 Finished mouse sequence analysis

Two finished mouse clone sequences, AL583887.9 (bM121M7) (220050bp) and AL513354.14 (bM150J22) (22703bp) were selected for more detailed analysis. These clones map in close proximity to each other (see figure 4.5) but do not overlap, as a gap of ~60kb (estimated from fingerprint data) exists between them. This gap is spanned by clone bM85M21, which is currently being sequenced.

4.6.1 Mouse gene annotation

Initial annotation of the finished mouse clones was performed by Dr. Laurens Wilming (Sanger Institute) by similarity comparison to:

1. EMBL vertebrate cDNA sequences (see appendix 2)
2. Publicly available EST sequences (see appendix 2)

3. Human annotated gene sequences from 22q13.31.

This initial annotation was extended by similarity comparison to non-publicly available ESTs (appendix 2) and partial, but not submitted, cDNA sequences from 22q13.31 (chapter III) (M. Goward). The approach is similar to the human sequence analysis discussed in chapter III. In total, eight genes were annotated in the mouse clones. The longest isoforms of these genes are summarised in table 4.9. Figure 4.10 shows the genomic distribution of the mouse genes in comparison with the syntenic human region.

Table 4.9: The annotated mouse genes and their exon number, genomic span, transcript size and ORF size.

Mouse gene	Human orthologue	No. of exons	Genomic size (bp)	Transcript size (bp)	ORF size (bp)
bM121M7.1	TTLL1	12(12)	26956(49751)	2003(1684)	1272(1272)
Biklk	BIK	5(5)	17795(19110)	1370(1099)	453(483)
bM121M7.3	bK1191B2.C22.3	4(4)	15727(11180)	1679(2048)	1146(1173)
Bzrp	BZRP	4(4)	10623(11697)	849(850)	510(510)
bM121M7.5	dJ526I14.C22.2	14(14)	19502(20479)	3209(3353)	1920(1935)
Scube1*	dJ526I14.C22.3	>19(22)	>72041(139476)	>4914(5741#)	2886#(2967)
bM150J22.1	C22ORF1*	6(>4)	66530(>63349)	3180(2323#)	981(909#)
bM150J22.2*	dJ345P10.C22.4	>26(33)	>121975(283449)	>4032(4878)	>3965(4575)

*Gene structure extends beyond available genomic sequence
Size calculated from EMBL cDNA entry
The equivalent values for the orthologous human genes are shown in brackets.

Figure 4.10 (foldout): Alignment of the human and mouse annotated genes. The figure depicts the human clones (blue boxes) with sequence accession numbers, the human and mouse CpG islands (yellow), the human gene features (genes with orthologues shown in the mouse sequence are shown in dark blue, genes for which equivalent mouse sequence is not yet available in light blue and pseudogenes in green), mouse genes (red) and mouse sequence clones (red boxes) with accession numbers. Similar exons are indicated by the grey lines.

take out this page for figure 4.10

Additionally, five alternative splice forms were annotated based on mouse EST evidence (L. Wilming). Three isoforms of bM121M7.3 have been annotated. Two of these are orthologous to alternative splices verified in human: bK1191B2.C22.3a (Em:AL359401) and bK1191B2.C22.3b (Em:AL359403). The remaining isoform of bM121M7.3 shows a possible alternative 5' end. Additionally, alternative 3' ends are indicated from EST evidence for bM121M7.5 and Scube1. However, there is currently no evidence to support the existence of these isoforms in the orthologous human genes. EST evidence can be unreliable (chapter III) so further experimental evidence is required to confirm these structures.

4.6.2 Human-mouse finished sequence alignment

4.6.2.1 Dot plot

The annotated mouse and human sequences were compared using the PipMaker dot plot program (http://bio.cse.psu.edu/pipmaker) (Schwartz *et al.*, 2000). Figure 4.11 shows the mouse sequence displayed on the x-axis and the human sequence on the y-axis. Drawn along both of the axes are boxes corresponding to each of the annotated genes. Regions of high similarity correspond with gene structures. Gene order and orientation are conserved. The human gene dJ754E20A.C22.4 lies within the mouse sequence gap. The genomic span of the human sequence is approximately 1.6X greater than the equivalent genomic mouse sequence (see sections 4.6.4 and 4.6.5). The mouse clone bM150J22 spans a gap in the human sequence. This is discussed in more detail in section 4.7.

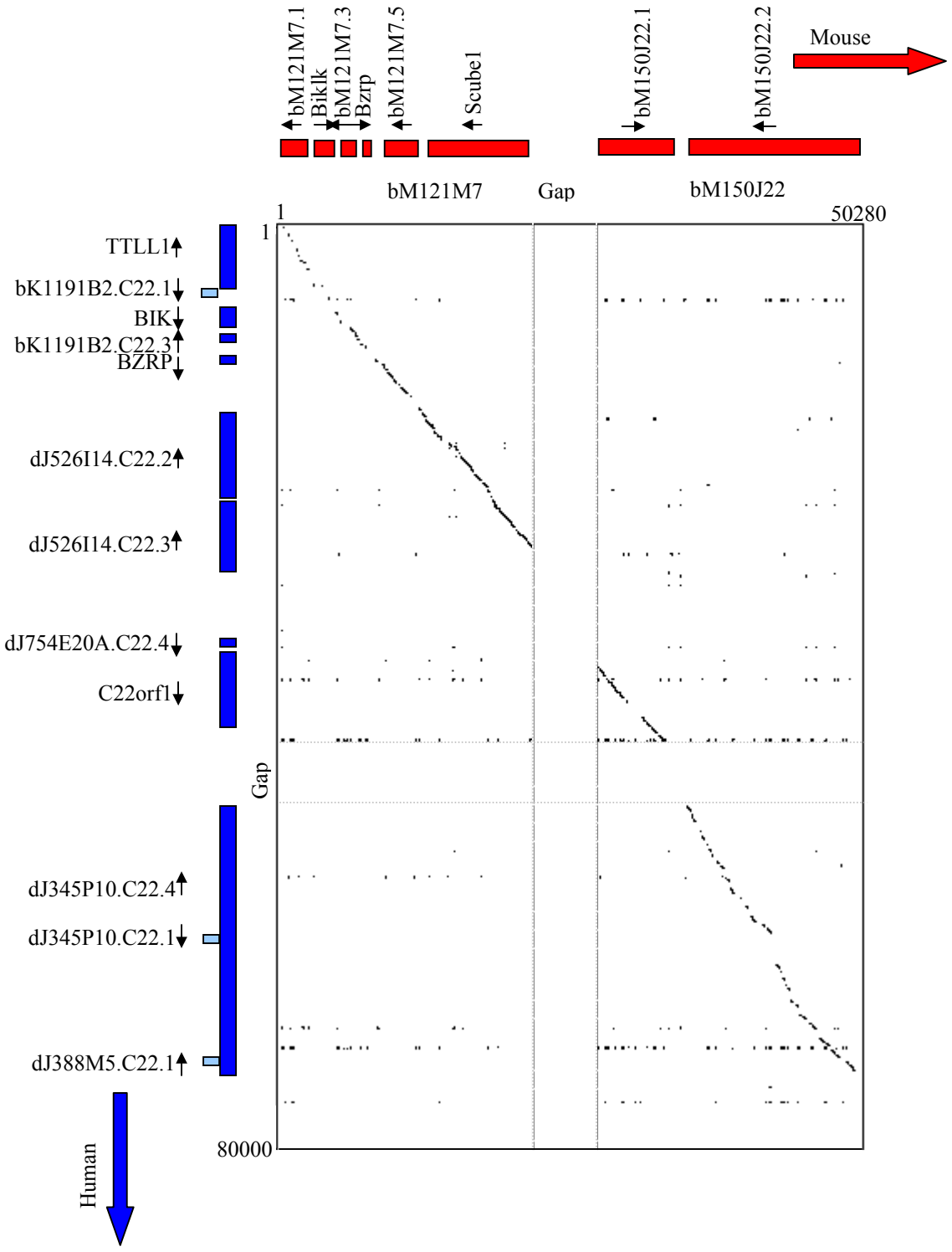


Figure 4.11: Annotated dot plot of the mouse (x-axis) and human (y-axis) sequences. The plot was generated using the PipMaker suite of analysis tools (Schwartz *et al.*, 2000). The boxes along the axes indicate the positions of human (blue) and mouse (red) genes. Light blue boxes depict possible human pseudogenes, which are not conserved in the mouse sequence.

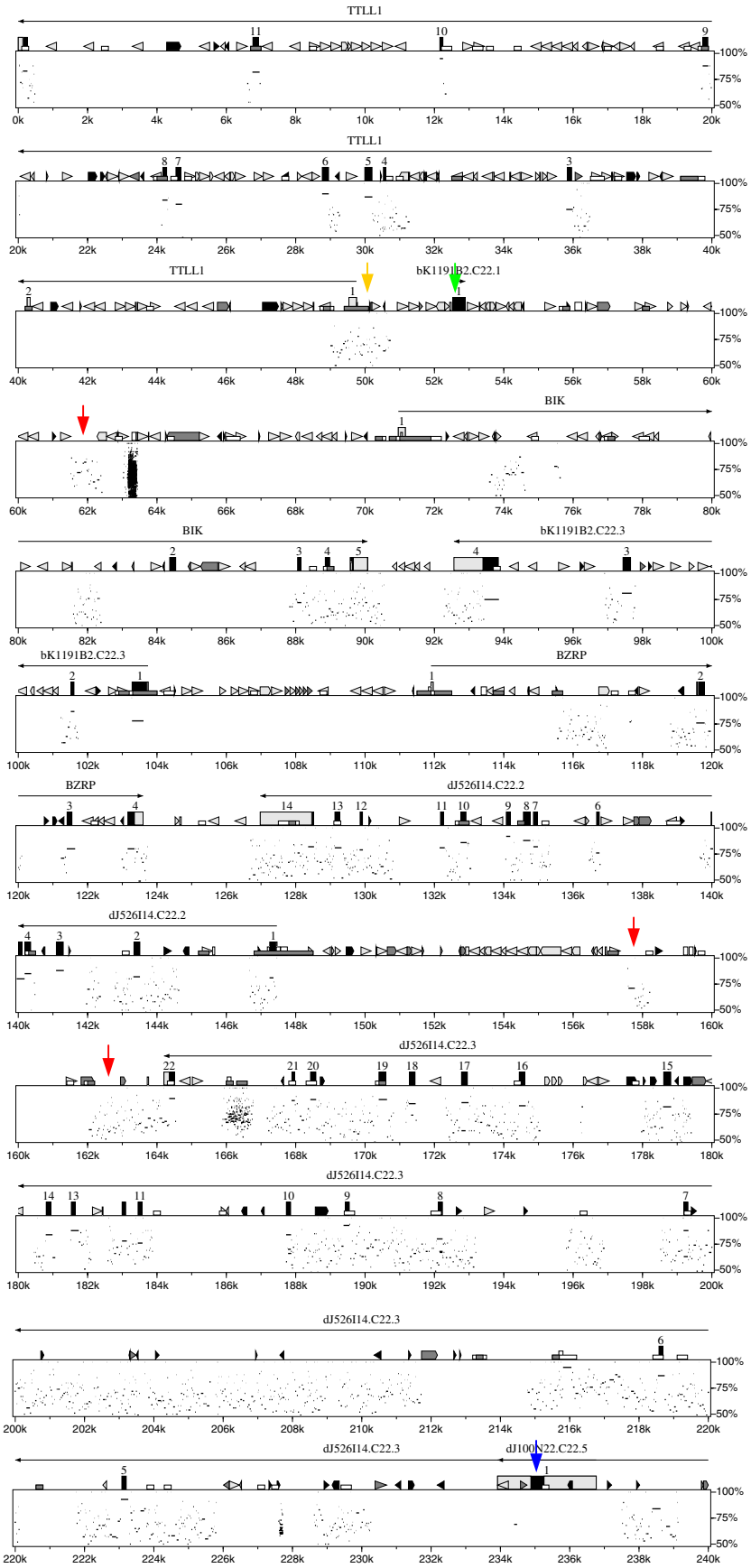
4.6.2.2 PIP analysis

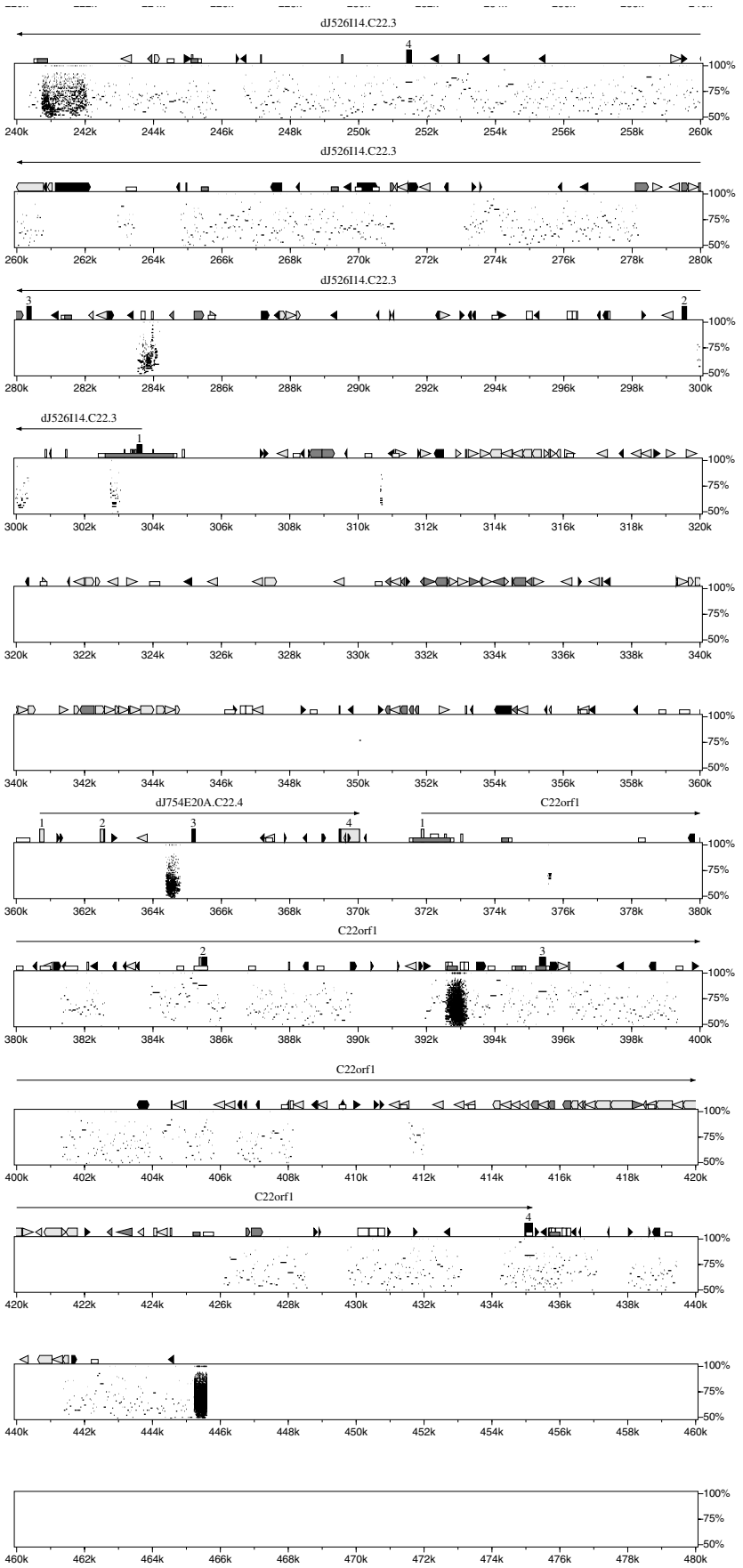
A PIP (Schwartz *et al.*, 2000) was generated to show the conservation of this region between finished human and mouse sequences in more detail. The plot displays the human sequence along the x-axis, incorporating features such as genes, repeats (generated from RepeatMasker output) etc. The y-axis displays the percent identity of the mouse sequence. Figure 4.12 shows that overall the areas of high similarity correspond well with the annotated human genes. There are a few exceptions:

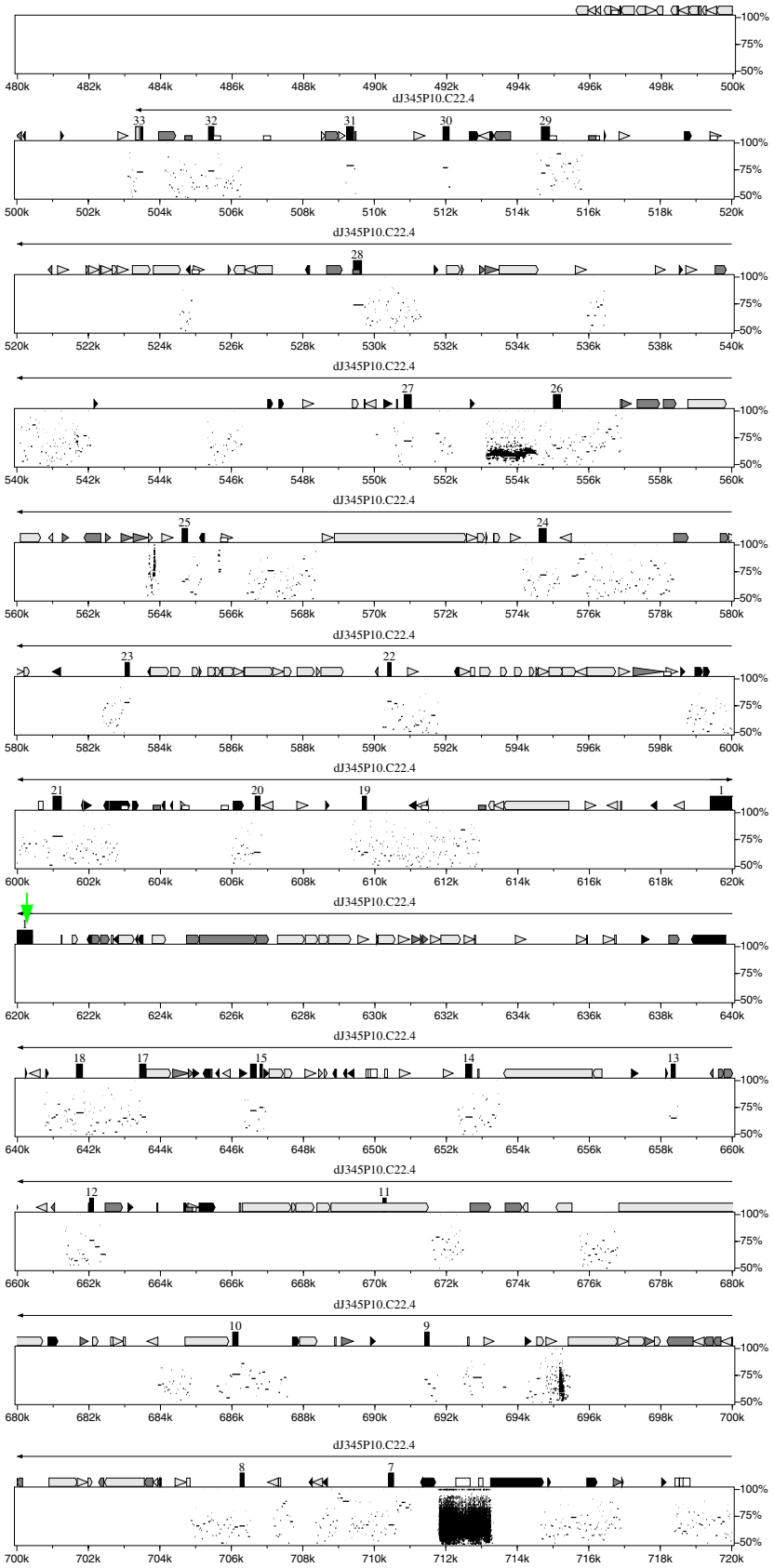
- Conserved sequences are located in an intergenic region around 62K (between TTLL1 and bK1191B2.C22.3) and between 157.5K and 164K (between dJ526I14.C22.2 and dJ526I14.C22.3) (indicated by red arrows).
- Conserved sequences are also found in the 5'UTR of TTLL1 (yellow arrow) and in the introns of most genes.

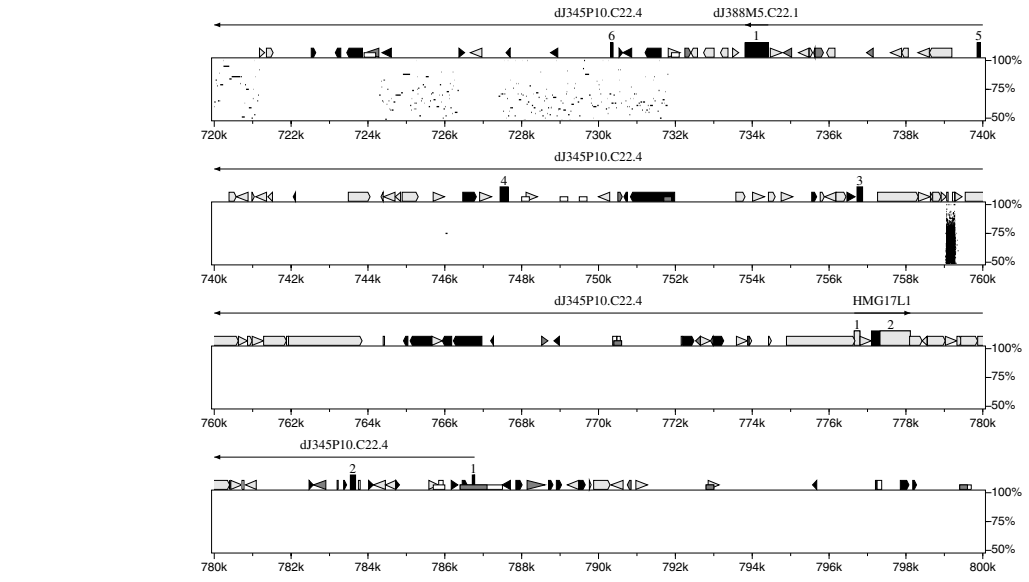
These sequences may highlight additional exons that have not been annotated in the human sequence, or may indicate the presence of regulatory regions.

- The cDNA sequence Em:AL442096 (Bloecker et al., unpublished), was previously noted as possibly resulting from spurious priming of an adjacent genomic poly(A) tract (chapter III). The sequence is not conserved in mouse (blue arrow), which supports the premise that this cDNA does not originate from a true gene.
- Similarly, the human pseudogenes bK1191B2.C22.1 and dJ345P10.C22.1 were not conserved in the mouse sequence (green arrows).


TTLL1
TTLL1
TTLL1
bK1191B2.C22.1
BIK
BIK
bK1191B2.C22.3
bK1191B2.C22.3
BZRP
BZRP
dJ526I14.C22.2
dJ526I14.C22.2
dJ526I14.C22.3
dJ526I14.C22.3
dJ526I14.C22.3
dJ526I14.C22.3
dJ100N22.C22.5
100%
75%
50%


dJ526I14.C22.3
dJ526I14.C22.3
dJ526I14.C22.3
dJ526I14.C22.3
dJ754E20A.C22.4
C22orf1
C22orf1
C22orf1
C22orf1
100%
75%
50%


100%
75%
50%
480k 482k 484k 486k 488k 490k 492k 494k 496k 498k 500k
dJ345P10.C22.4
33 32 31 30 29
500k 502k 504k 506k 508k 510k 512k 514k 516k 518k 520k
dJ345P10.C22.4
28
520k 522k 524k 526k 528k 530k 532k 534k 536k 538k 540k
dJ345P10.C22.4
27 26
540k 542k 544k 546k 548k 550k 552k 554k 556k 558k 560k
dJ345P10.C22.4
25 24
560k 562k 564k 566k 568k 570k 572k 574k 576k 578k 580k
dJ345P10.C22.4
23 22
580k 582k 584k 586k 588k 590k 592k 594k 596k 598k 600k
dJ345P10.C22.4
21 20 19 1
600k 602k 604k 606k 608k 610k 612k 614k 616k 618k 620k
dJ345P10.C22.4
620k 622k 624k 626k 628k 630k 632k 634k 636k 638k 640k
dJ345P10.C22.4
18 17 15 14 13
640k 642k 644k 646k 648k 650k 652k 654k 656k 658k 660k
dJ345P10.C22.4
12 11
660k 662k 664k 666k 668k 670k 672k 674k 676k 678k 680k
dJ345P10.C22.4
10 9
680k 682k 684k 686k 688k 690k 692k 694k 696k 698k 700k
dJ345P10.C22.4
8 7
700k 702k 704k 706k 708k 710k 712k 714k 716k 718k 720k



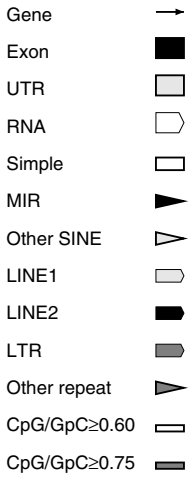


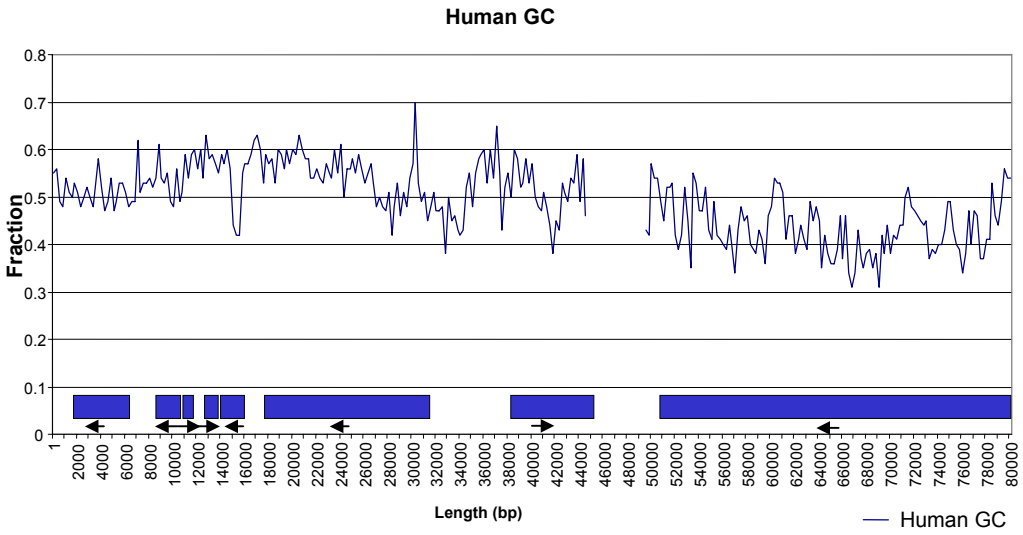
Figure 4.12: Percentage identity plot calculated by PipMaker for the human interval TTLL1 to dJ345P10.C22.4, compared with sequence from the region of conserved synteny on mouse chromosome 15. Black horizontal bars beneath the graphical depictions of interspersed repeats and gene structures indicate gap-free segments demonstrating> 50% nucleotide identities. Exons are numbered from the 5'-most annotated exon. A single gap-free alignment underneath a protein-coding exon indicates the mouse exon is conserved, and thus the mouse locus maintains a homologous ORF.

4.6.3 GC content

4.6.3.1 Comparison of human and mouse GC content

The fraction GC content in 1kb intervals was calculated by GC profile (Gillian Durham) and the GC content profiles plotted (Figure 4.13). The two GC profiles are similar, although direct comparison is complicated by the expansion of the human sequence to 1.6X the length of the equivalent mouse sequence. The 5' ends of genes align well with peaks in GC content. The human sequence has a higher overall GC content of 51% compared with the mouse sequence value of 49%.

A



B

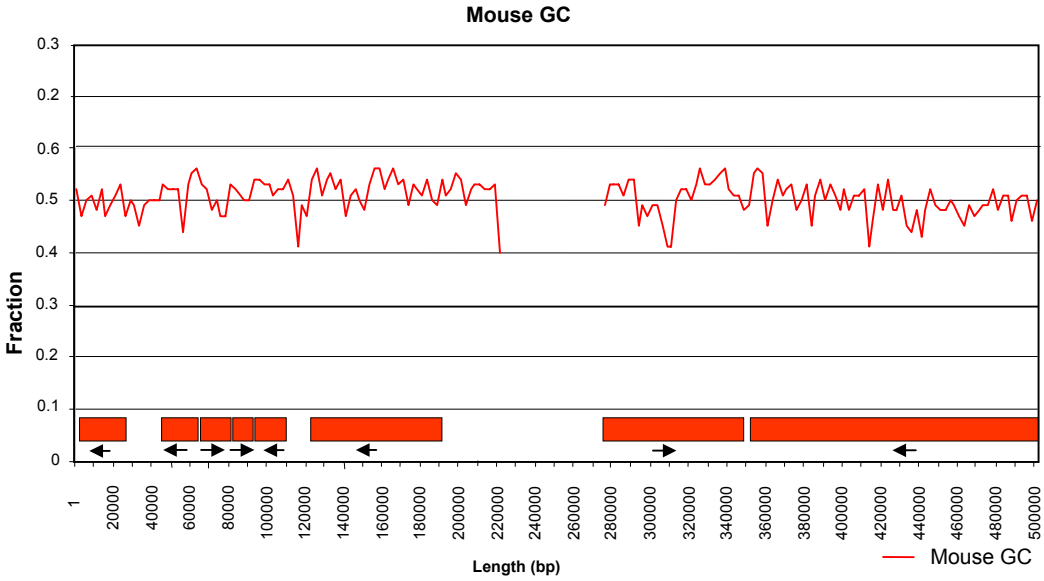


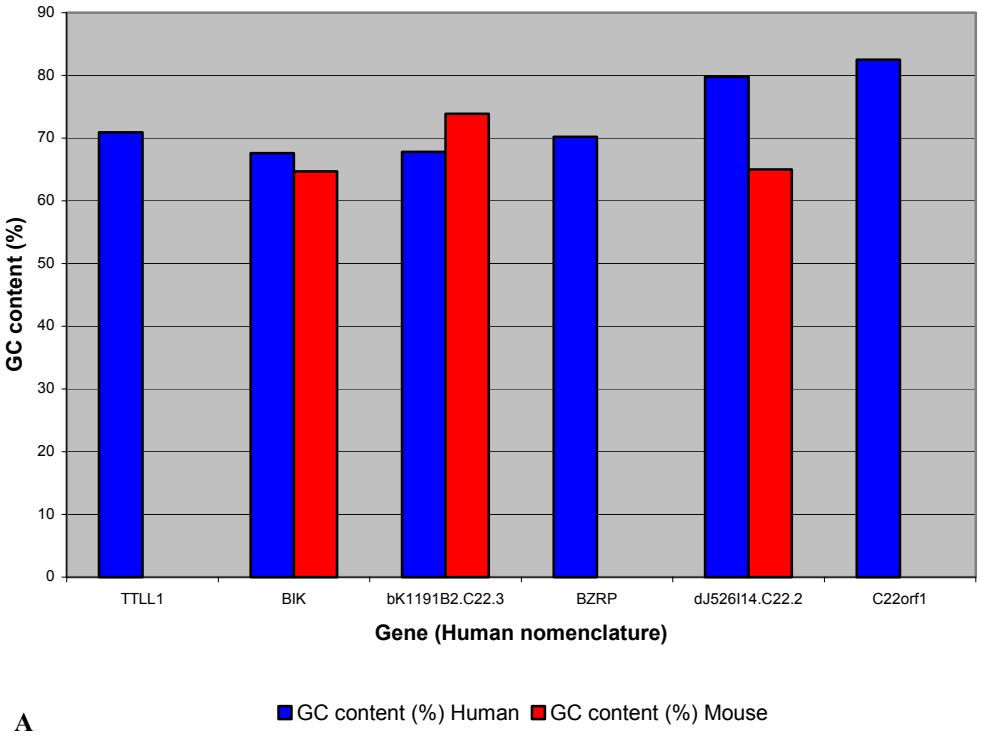
Figure 4.13: Human and mouse GC distribution, calculated using GC profile (G. Durham), with a window size of 1 kb. Human and mouse genes are depicted by blue and red boxes respectively, along the x-axes.

4.6.3.2 CpG islands

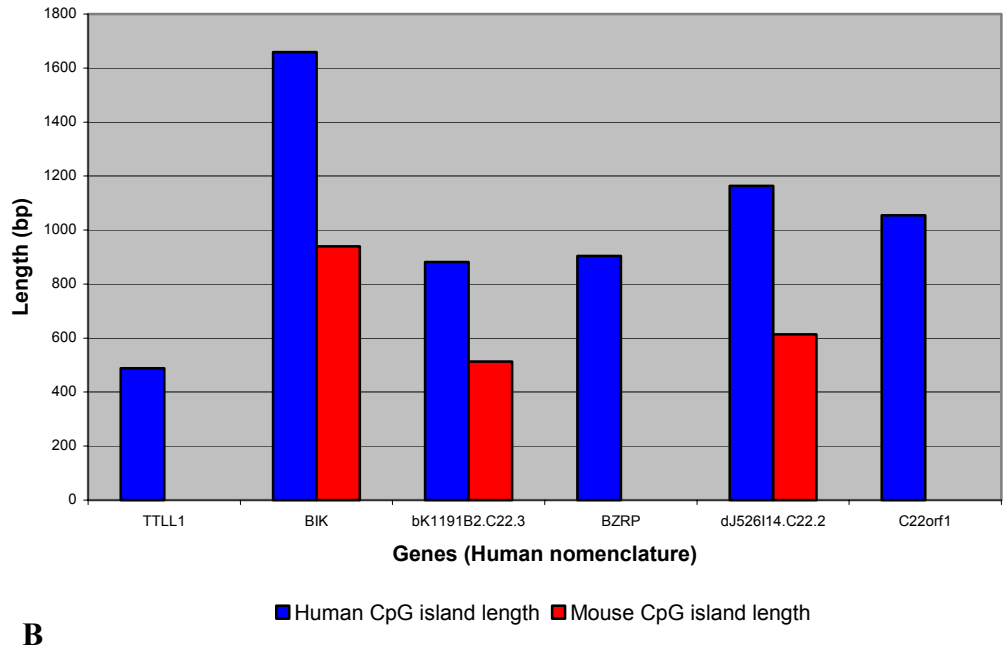
The 5' UTRs of six of the eight genes shown above are contained in the available finished mouse sequence. In human, all six genes contain a CpG island, but four of the mouse genes lack a CpG islands, using the criteria of the CpG island prediction package CPGFIND (Micklem, unpublished) (chapter III). An additional predicted CpG island does correspond to exon 2 of bM121M7.3 however. Antequera and Bird (1993) suggested that approximately 20% of mouse genes lack a CpG island. In this region, 66% of genes lack a CpG island at the 5'UTR, although the sample size is very small and figure 4.13 indicates that there are still peaks in the GC content associated with the starts of all genes. Details of the CpG islands are summarised in figure 4.14.

4.6.4 Repeat content

The repeat content of the human and mouse regions was analysed using RepeatMasker (Smit and Green, unpublished), with human- and rodent-specific repeats as appropriate. Figure 4.15 shows that the human and mouse SINE density are similar. The coverage of the SINEs in human, however, is four times that of mouse. This greater genomic coverage contributes to the difference in size noted between the equivalent regions of the human and mouse genomes: the human region is 1.6X larger than the mouse region. One third of this difference is caused by the greater coverage of the human SINE repeats. Simple sequence repeats and MaLRs are far more abundant in the mouse sequence. The MaLRs in mouse are still actively expanding, which is the most likely reason for the higher density of these repeats in mouse (Smit & Riggs, 1995).



A



B

Figure 4.14: Comparison of human and mouse CpG island GC content (A) and length (B). CpG islands were predicted using CPGFIND (Micklem, unpublished).

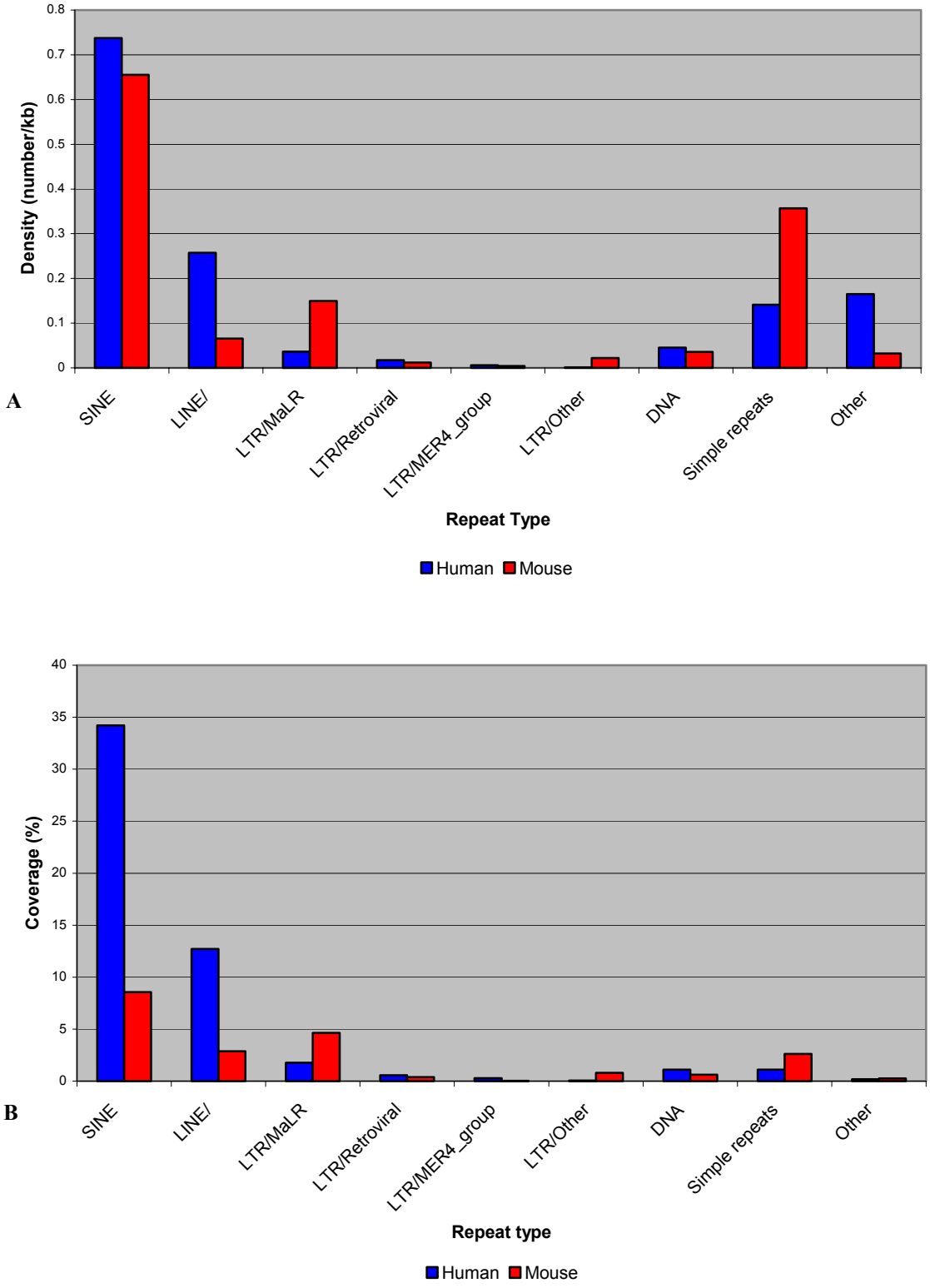
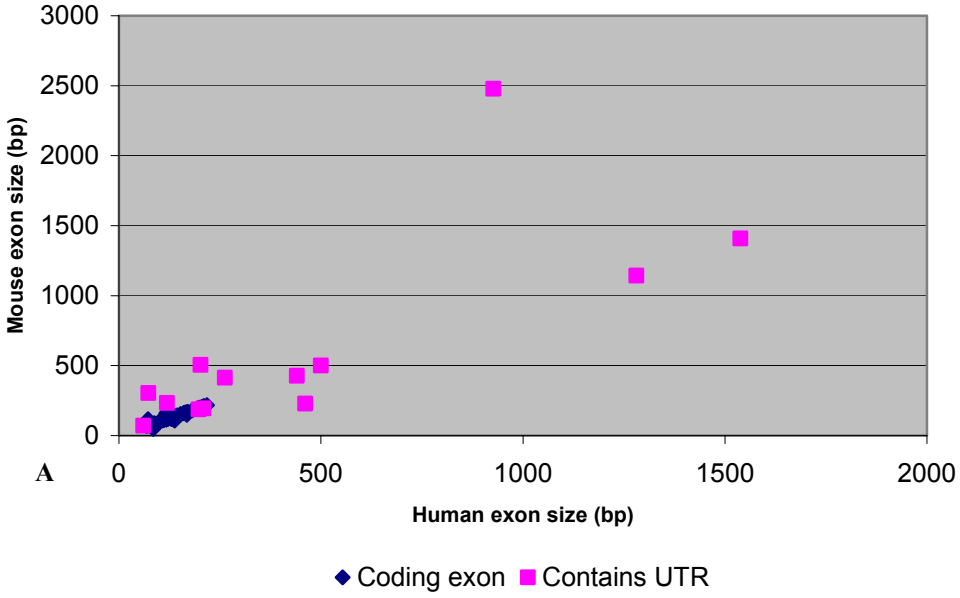


Figure 4.15: Repeat density (A) and genomic coverage by repeats (B) for human and mouse.

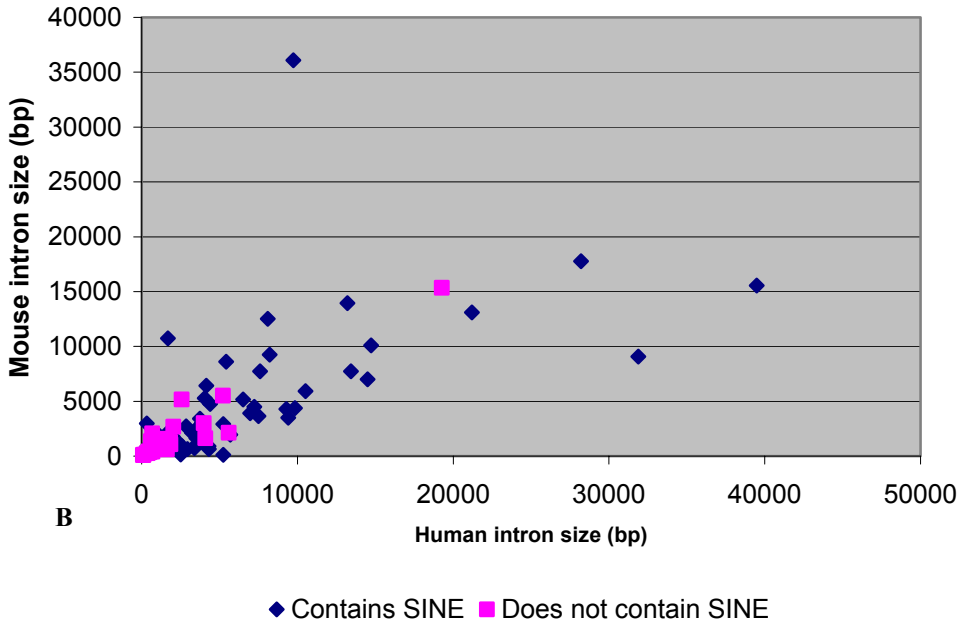
4.6.5 Comparison of coding regions

Exon number is conserved for all of the complete genes shown in table 4.9. The conservation of exon and intron sizes between mouse and human was examined by plotting the mouse exon sizes against the human (figure 4.16a); the equivalent comparison was carried out for intron size (figure 4.16n), and included analysis of the SINE content of the intron. A more detailed depiction of the 500 bp window of the human-mouse exon sizes is shown in figure 4.16c.

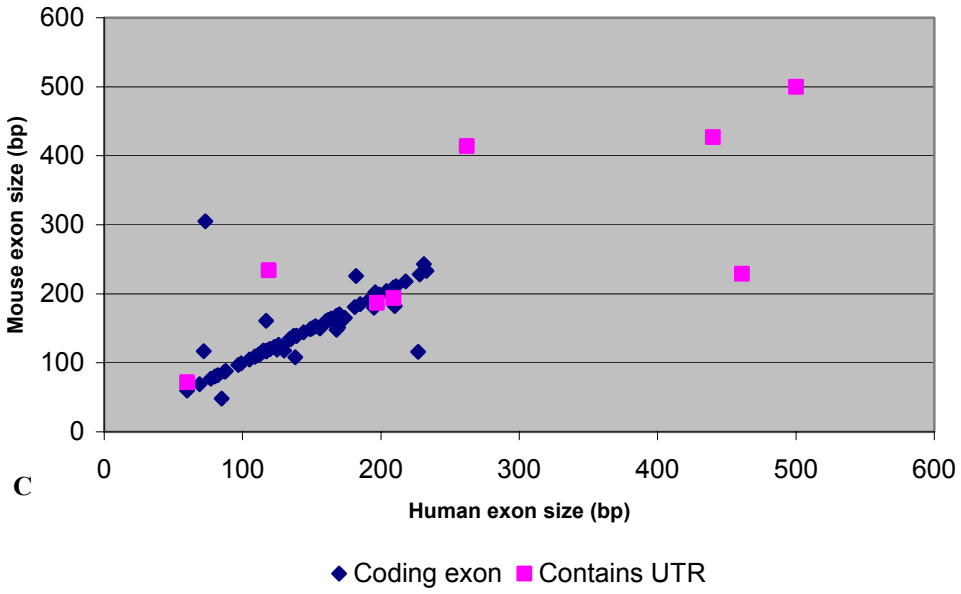
Generally, most of the internal coding exons are exactly the same length. The lengths of the 5' and 3' UTR exons, however, do show differences, as illustrated in table 4.9. The intron sizes are less well correlated (figure 4.16b). Introns containing SINEs generally tend to be larger in human genes, which contributes to the difference in sizes of the two equivalent regions (section 4.6.4). This is also reflected in figure 4.10 where the intron-exon structures are shown for all the genes. Together, this evidence reflects a high degree of conservation of the coding exons, with a lesser degree of conservation of gene structure.


3000
2500
2000
1500
1000
500
0
Mouse exon size (bp)
0
500
1000
1500
2000
Human exon size (bp)
Coding exon
Contains UTR

A


40000
35000
30000
25000
20000
15000
10000
5000
0
Mouse intron size (bp)
0
10000
20000
30000
40000
50000
Human intron size (bp)
Contains SINE
Does not contain SINE

B



C

Figure 4.16: Scatter plots depicting (A) exon sizes and (B) intron sizes between human and mouse gene structures. (C) A more detailed view of the 500 bp exon interval is also shown.

Nucleotide and amino acid sequence conservation was examined using clustalw (Thompson *et al.*, 1994) and sequence identities calculated (belvu; Sonnhammer, unpublished). These results are shown below.

Table 4.10: Percentage identities of mouse and human gene sequences

Orthologous gene pair	mRNA nt. sequence identity (%)	ORF nt. sequence identity (%)	Amino acid sequence identity (%)
bM121M7.1 & TTLL1	79.4	86.7	96.9
Biklk & BIK	57.6	64.0	41.3
bM121M7.3 & bK1191B2.C22.3	69.7	78.1	75.9
Bzrp & BZRP	75.5	81.8	81.1
bM121M7.5 & dJ526I14.C22.2	76.4	85.7	86.2
Scube1 and dJ526I14.C22.3*	81.7	87.8	87.1
bM150J22.1 & C22ORF1	70.8	90.2	98.2
bM150J22.2* & dJ345P10.C22.4*	72.4	72.6	78.0

*Gene currently incomplete; only partial sequences aligned

The percentage identity of all nucleotide sequences was increased by the exclusion of the 5' and 3' UTR sequences, which contain more divergent sequences. In four cases the level of conservation of the predicted amino acid sequence was lower than the equivalent nucleotide value. This was most marked between the human BIK gene and mouse Biklk (figure 4.17). This is due to a reading frame shift, caused by the insertion or deletion of a 7 bp sequence (highlighted in red). The conserved reading frame is restored by a 2 bp insertion/deletion downstream of the 7 bp difference. Five other in-frame insertions/deletions are also present. Altogether, these changes have the effect of lengthening the human protein, or shortening the mouse protein, by 10 amino acids. Additionally, there are 142 nucleotide changes (excluding deletions/insertions), of which only 28 are synonymous changes (do not alter the amino acid sequence). However, the number of amino acid changes that result from non-synonymous nucleotide changes is less than 114, as some changes occur in two different positions within the same codon. The existence of insertions/deletions in the sequence means that other, although perhaps less parsimonious, codon alignments exist in addition to the one shown below.

Figure 4.17: A) Alignment of Biklk and BIK including 5' and 3' UTRs. B) Greater conservation is shown in the alignment of the cDNA sequences without the UTRs. An insertion/deletion of 7bp causes a frameshift, which is corrected downstream by a further 2bp insertion/deletion (red box). C) Alignment of Biklk and BIK peptide sequences. Alignments were created with clustalw (Thompson *et al.*, 1994). The alignments were formatted for printing using belvu (Sonnhammer, unpublished).

4.6.6 Splice site comparison

The splice sites of both the human and mouse genes were compared using the sequence logo technique described in chapter III. Eighty splice acceptor and donor sequences from equivalent

introns were extracted from gff files and used to generate sequence logos (D. Beare). The cumulative height of each position reflects the importance of this position in the splice consensus sequence. The height of each nucleotide reflects the frequency of that nucleotide at that particular position. Figure 4.18 shows the human splice donor and acceptor (A),and mouse splice donor and acceptor (B). This shows that, overall, the splice consensus is well conserved between human and mouse. The important GT nucleotides (positions 7 and 8) in the splice donor and AT (24 and 25) in the acceptor are well conserved between human and mouse. Differences are limited to the C/T tail where a C is more commonly found at position 14 in mouse whereas T is commonly found in human. These results support those of a previous study of 84 human and mouse introns (Smink, 2001).

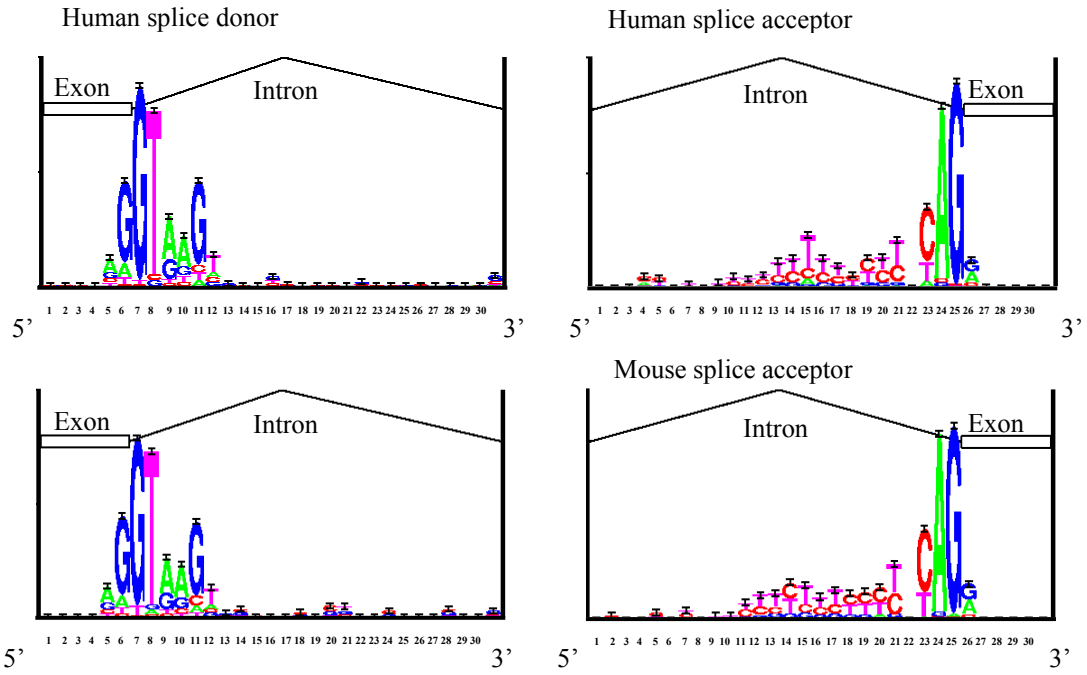


Figure 4.18: The splice acceptor and donor sites for human (A) and mouse (B). The splice site sequences were extracted by D. Beare (Sanger Institute) and visualised using Sequence Logo (Steven Brenner) (http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi).

4.6.7 Regulatory regions

Sequence conservation between human and mouse DNA in inter- and intragenic regions may indicate the existence of functional features, such as exons or regulatory regions, or may be non-functional sequence inherited from a common ancestor. 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). General conservation of the GC profile and peaks seems to suggest conservation of possible CpG islands (see section 4.6.3). The PIP (figure4.12), however, demonstrated conservation upstream of only one gene, TTLL1.

DBA (DNA Block Aligner) (Jareborg *et al.*, 1999) is an alignment algorithm designed to identify conserved collinear blocks in two DNA sequences. The main difference between DBA and PIP alignments is that DBA identifies gapped blocks. Also, blocks identified by DBA can be shorter than 50 bp, although the nucleotide identity must be greater than 60%, whereas PIPs will highlight only ungapped alignments longer than 50 bp with an identity >50% (section 4.1.3.1). Jarebourg *et al*. propose that these features of DBA make the program particularly suitable to identify small conserved functional motifs whose relative positioning may not be conserved and which may be separated by large pieces of non-functional DNA sequence due to random insertions in one species compared with another.

To investigate whether any further sequence conservation could be observed in these putative regulatory regions, three kilobases of sequence was extracted upstream of the transcription start site for both human and mouse, containing the entire length of any CpG islands predicted at this position. The human and mouse sequences were aligned with DBA. DBA identified significant

alignments 5' of the transcription start sites of the genes TTLL1, BIK, BZRP and C22orf1 (see appendix 6). An example of a region aligned by DBA is shown in figure 4.19.

The consensus sequences were used to scan the TRANSFAC 4.0 transcription factor database (Wingender *et al.*, 2000), using MatInspector V2.2 (Quandt *et al.*, 1995). Thresholds were set so that only exact matches to the core sequence of the matrix (capitalised) and overall matrix similarity >0.9 were listed, in order to enhance accuracy of the search results. The sites found are shown in table 4.11

```
bM121M7.1   -582      CCGCCTGCTTCTGCCTCCCAAAGTGCTGGGATTAAGGCATGCGCCACC
Consensus     D      CC CC GC TCTGCCTCCC AAGTGCTGGGATTA AGGC TG GCCACC
TTLL1      -1559     CCACCCGCCTCTGCCTCCC-AAGTGCTGGGATTACAGGCGTGAGCCACC
```

Figure 4.19: Sequence alignment (DBA, Jareborg *et al.*, 1999) of mouse and human sequence upstream of TTLL1 (human gene) and bM121M7.1 (mouse orthologue). A potential binding site for the zinc finger protein Ik-2 is highlighted in red (Molnar & Georgopoulos, 1994)(see table 4.11).

The expression patterns of the human genes (chapter III) were examined in order to determine if there was a relationship between tissue distribution of the human transcript and what is currently known about the putative functional regions listed in table 4.11. TTLL1, BIK and BZRP are expressed in a wide variety of tissues. Examination of the TRANSFAC sites preceding these genes did not preclude this expression pattern. C22orf1 demonstrated a more limited expression pattern in RT-PCR screens of RNA from human tissues and previous research has shown that C22orf1 is predominantly expressed in adult brain (Schwartz & Ota, 1997). However, examination of the 24 sites found did not suggest specific involvement with adult brain transcription.

Table 4.11: Resulting sites from TRANSFAC screen with consensus sequences from DBA alignment of putative promoter regions.

Gene (human nomenclature)	Matrix	Orientation	Matrix similarity	Sequence
TTLL1	GFI1_01	-	0.905	angcctntAATCccagcacttngg
TTLL1	IK2_01	-	0.911	cttnGGGAggca
TTLL1	IK2_01	+	0.946	tgctGGGAttan
TTLL1	LYF1_01	-	0.911	ttnGGGAgg
TTLL1	RFX1_01	-	0.922	nggngncctnGCAAccn
BIK	IK2_01	+	0.928	cttnGGGAtntt
BZRP	DELTAEF1_01	-	0.954	ncacACCTnta
BZRP	GFI1_01	-	0.911	acacctntAATCccagcacttngn
BZRP	HFH2_01	+	0.911	nttTGTTtnntt
BZRP	HNF3B_01	+	0.908	ttnttTGTTtnnttn
BZRP	IK2_01	+	0.946	tgctGGGAttan
BZRP	SRY_02	-	0.931	nnaaACAAanaa
C22orf1	AP4_Q5	-	0.94	ctCAGCagtt
C22orf1	BRN2_01	+	0.923	aagatttgTAATgagt
C22orf1	BRN2_01	-	0.93	ctcattacAAATcttt
C22orf1	CREL_01	-	0.98	gggnntTTCC
C22orf1	DELTAEF1_01	+	0.953	cnccACCTgcn
C22orf1	E47_01	-	0.933	nnnGCAGgtggngac
C22orf1	FREAC2_01	-	0.912	attttgTAAAcaggnn
C22orf1	GFI1_01	-	0.902	tcattacaAATCtttccanctcag
C22orf1	GKLF_01	-	0.93	aaagagggagAGGG
C22orf1	GKLF_01	-	0.927	aanggagggaGGGG
C22orf1	IK2_01	-	0.917	nntgGGGAacag
C22orf1	LMO2COM_01	-	0.969	nngCAGGtggng
C22orf1	MYOD_01	-	0.926	nngCAGGtggng
C22orf1	MYOD_Q6	+	0.947	ncCACCtgcn
C22orf1	MZF1_01	-	0.975	nntGGGGa
C22orf1	MZF1_01	-	0.982	ggaGGGGa
C22orf1	NFAT_Q6	+	0.944	agntgGAAAgat
C22orf1	NFKAPPAB65_01	-	0.958	gggnntTTCC
C22orf1	NKX25_02	+	0.951	caTAATta
C22orf1	S8_01	+	0.968	ngcacataATTAaaat
C22orf1	S8_01	-	0.968	acattttaATTAtgtg
C22orf1	S8_01	-	0.934	ngacaaaaATTAgaga
C22orf1	S8_01	-	0.948	naaaacaaATTAgatt
C22orf1	SRY_02	-	0.925	naaaACAAatta

Core sequences are capitalised

4.7 Chromosome 22 sequence gap

Figure 4.11 shows that the mouse BAC bM150J22 spans one of the few remaining 'unclonable' gaps in the human genomic sequence of chromosome 22. This gap has been estimated to be approximately 50 kb long by fibre-FISH (Dunham *et al.*, 1999) and is known to contain the 3' end of the C22orf1 gene at the centromeric end. The telomeric end of the gap is adjacent to the gene dJ345P10.C22.4. The mouse sequence spanning the gap is approximately 34 kb long. The sequence was analysed in more detail in order to identify any possible reasons why the region may be unclonable in human. To obtain equal start- and end-points for this comparison, sequence from bM150J22.1 to the 3' exons of bM150J22.2 was analysed. These features are equivalent to the closest gene features annotated in the human genome sequence flanking the gap. The mouse 'gap' region, shown in figure 4.20, contains the 3' end of the murine C22orf1 gene and provides evidence that the full human gene may be arranged in six exons. No further mouse EST or cDNA evidence was found to map to this region.

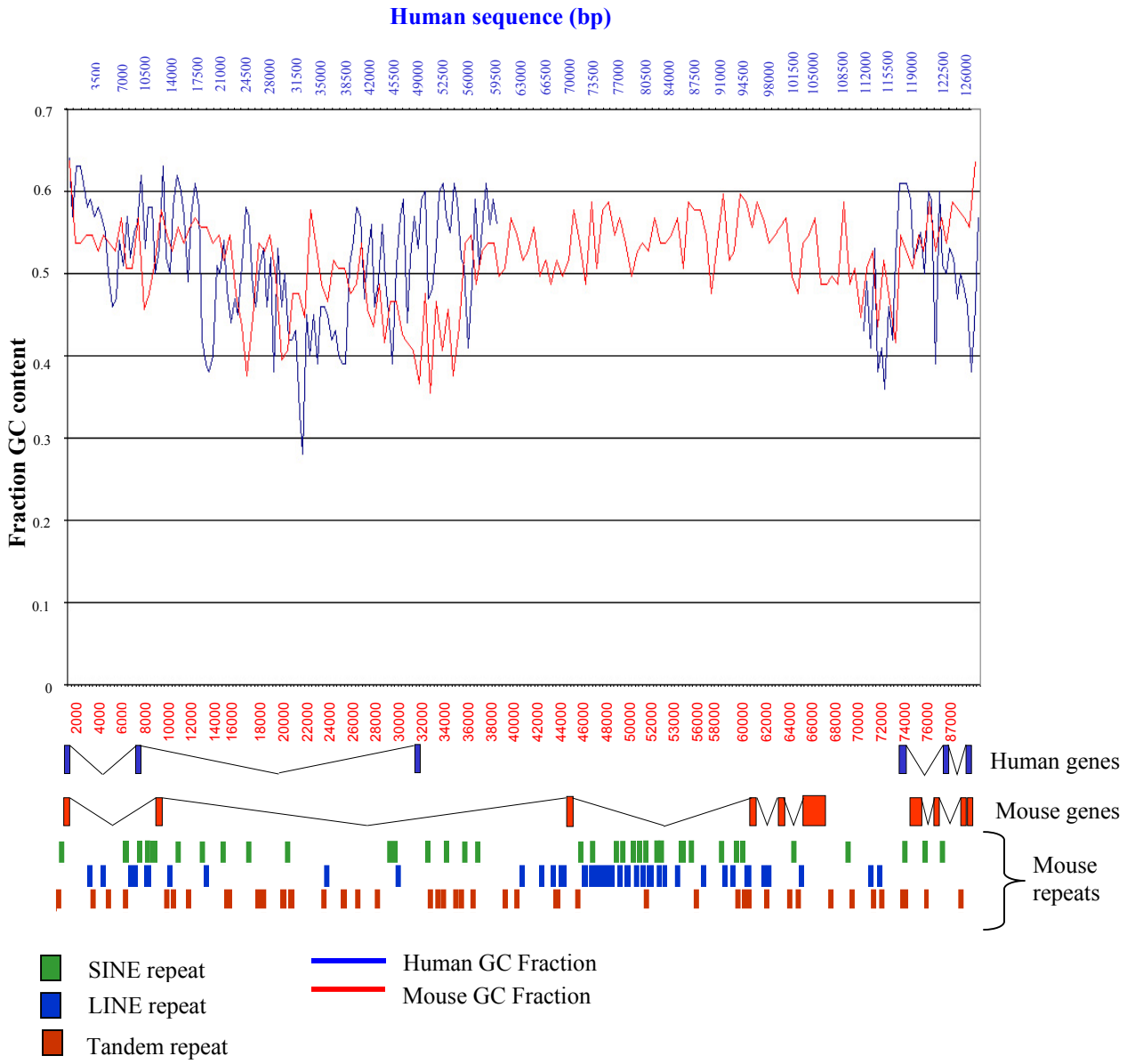


Figure 4.20: Diagram showing GC content, gene content and repeat content (mouse sequence only) of sequence spanning an 'unclonable' sequence gap in human chromosome 22. Human GC content and genes are shown in blue and mouse GC content and genes in red. GC fraction was calculated for 1kb windows using gc profile (Gillian Durham, unpublished). The distribution of mouse SINE, LINE and tandem repeats are also shown.

The graph of mouse GC content shows that a high proportion of GC dinucleotides are found throughout the region spanning the human sequence gap. The overall human GC content of the

region of interest is higher than that of mouse (section 4.6.3.1). Extrapolation of the graph indicates that human GC content is maintained above a level of 50% throughout the gap region. This high GC distribution may have an adverse affect on the 'clonability' of this DNA segment (section 4.9).

The repeat content of the 30216bp of mouse sequence that spans the human sequence gap was analysed in more detail using RepeatMasker. Results are shown below.

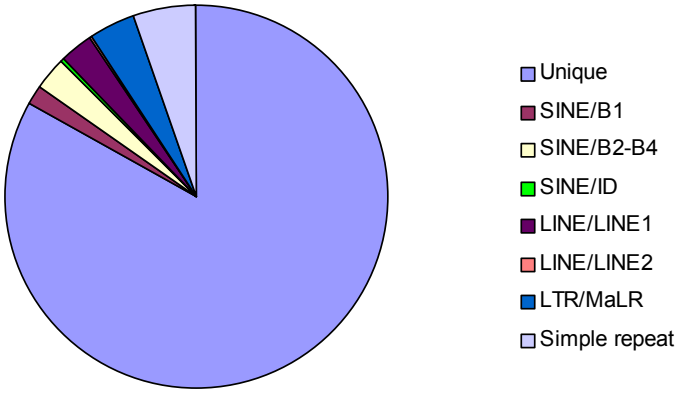


Figure 4.21: Repetitive and non-repetitive DNA distribution of 30216bp of mouse sequence, spanning an equivalent 'unclonable' sequence gap in human chromosome 22.

This region of mouse sequence contains no LTR elements or DNA transposon repeats. Although figure 4.21 shows that this region contains a greater coverage of SINE and LINE repeats than the immediately flanking sequences, the coverage and density of these repeats is comparable to the analysis of 50.2 kb of finished mouse sequence shown in section 4.6.4. No specific repetitive features were identified that could result in instability of this chromosomal region, leading to the difficulties in cloning the equivalent human DNA.

4.8 Localisation of synteny breakpoint

4.8.1 Definition of the junction region

A synteny breakpoint between HSA 22q13.1 and mouse chromosomes 15 and 8 was previously identified by Dunham *et al*. (1999), by combining data from the genomic sequence of HSA22 with information from the Mouse Genome Database (MGD) (http://www.informatics.jax.org/). The genes, HMOX and MB, situated 160 kb apart on HSA22, and their murine orthologues Hmox1 on MMU8 and Mb on MMU15, were identified as flanking the syntenic breakpoint.

In order to further narrow the breakpoint region boundaries, two mouse BAC contigs were constructed across the syntenic regions of mouse chromosomes 8 and 15 (section 4.2). Figure 4.4 shows that marker data from the two contigs localised the synteny breakpoint to a 130 kb region in the human sequence between genes MCM5 and MB. The available sequence from the contig tiling paths was compared with corresponding finished sequence from HSA22 using dot and PIP plots. Mouse BACs were identified that contained both conserved regions and sequence that extended beyond the syntenic breakpoint.

Currently, only unfinished sequence is available from the majority of adjacent mouse clones (see table 4.12) but detailed analysis is still possible.

Table 4.12: Mouse BAC genomic sequence clones adjacent to and spanning the syntenic breakpoint with human chromosome 22q13.1

Clone name	Author	Sequencing Centre	Genomic location	Accession number
bM290L7	Grills *et al.*	AECOM*	MMU8	AC084823.10 (finished)
bM254F2	Sims	Sanger Institute	MMU8	AL603837.2 (unfinished)
bM267J18	Deschamps et al.	UOKNOR#	MMU8	AC076974.23 (unfinished)
bM422F22	Sims	Sanger Institute	MMU15	AL591892.2 (unfinished)
bM412D17	Sims	Sanger Institute	MMU15	AL603843.2 (unfinished)

* AECOM – Albert Einstein College of Medicine. # UOKNOW – University of Oklahoma

A dot plot comparison of these mouse sequences with the finished sequence of the orthologous region of human chromosome 22 is shown below (figure 4.22). The syntenic breakpoint junction is clearly delineated between genes dJ569D19.C22.1 and MB. Gene order and orientation also appear to be conserved. Intergenic sequences are generally divergent, although strong conservation is noted in the genomic sequence 5' to the RBM9 gene, which may denote conserved regulatory regions or a novel gene structure.

The genes APOL5 and APOL6, however, do not appear to be conserved in this dot plot alignment. The nucleotide and protein sequences of these human genes were therefore compared against the available mapped mouse sequence (http://mouse.ensembl.org) using BLAST. The best matches for the protein sequences were found to be within Em:AL603843 (23% and 27% sequence identity respectively), but no matches were found at the nucleotide level, which may explain their absence in the dot plot. Analysis of the finished sequence, when available, may allow annotation of these genes within the mouse sequence. Alternatively, these genes may not exist in mouse, perhaps having arisen from duplication events in the human genome after divergence from the mouse lineage.

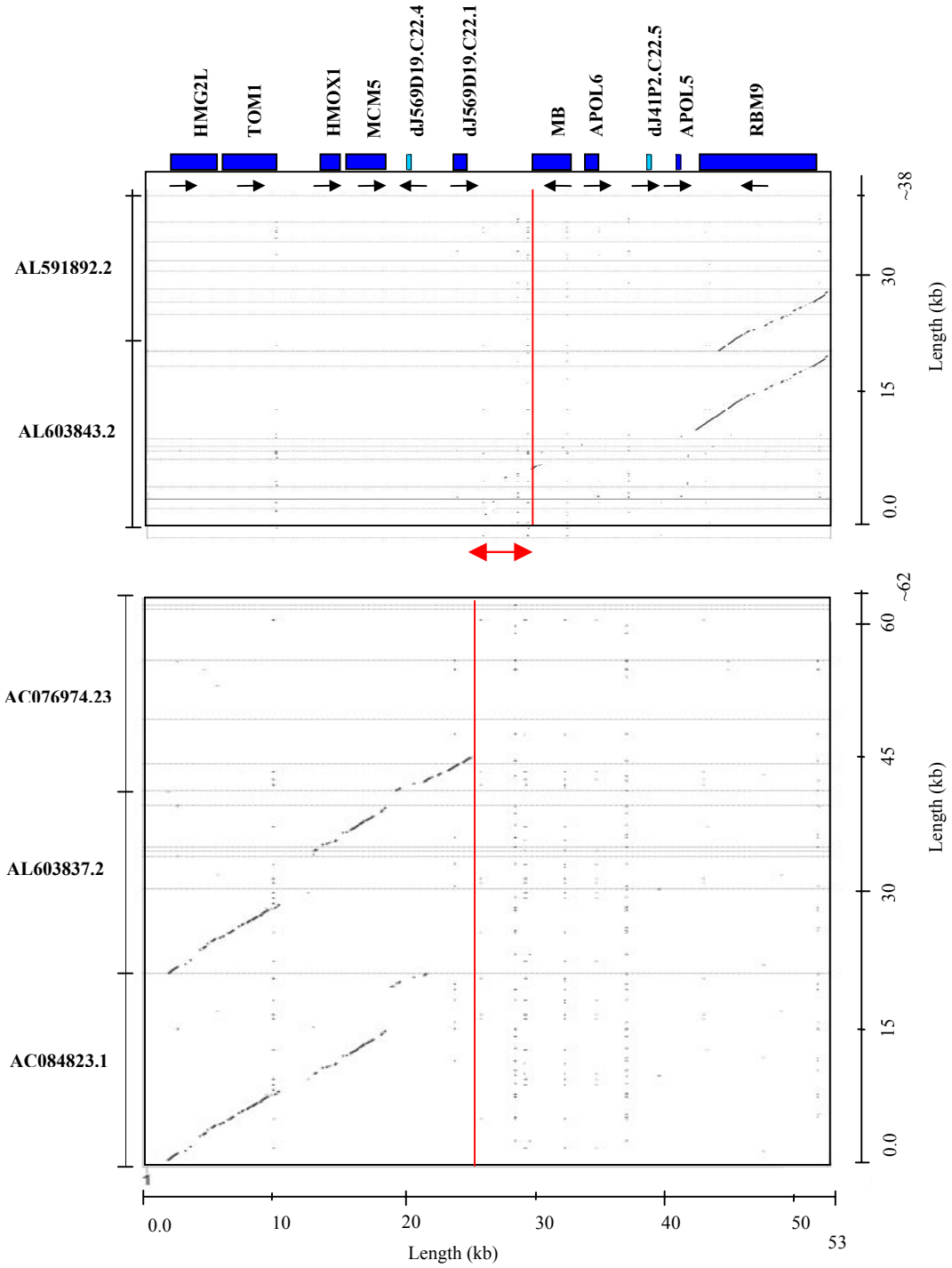


Figure 4.22 : Annotated dot plot of regions of mouse chromosome 8 and 15 available sequences (Y-axis) against the syntenic region of human chromosome 22 sequence (X-axis). The boxes along the X-axis indicate the human genes (dark blue). Human pseudogenes are indicated in light blue. The MMU8:15 syntenic breakpoint on HSA22 lies between dJ569D19.C22.2 and MB (indicated in red). The dot plot was generated using the PipMaker suite of analysis tools (http://bio.cse.psu.edu/pipmaker)

The schematic in Figure 4.23 shows the genes found adjacent to the junction region in the human and mouse chromosomes.

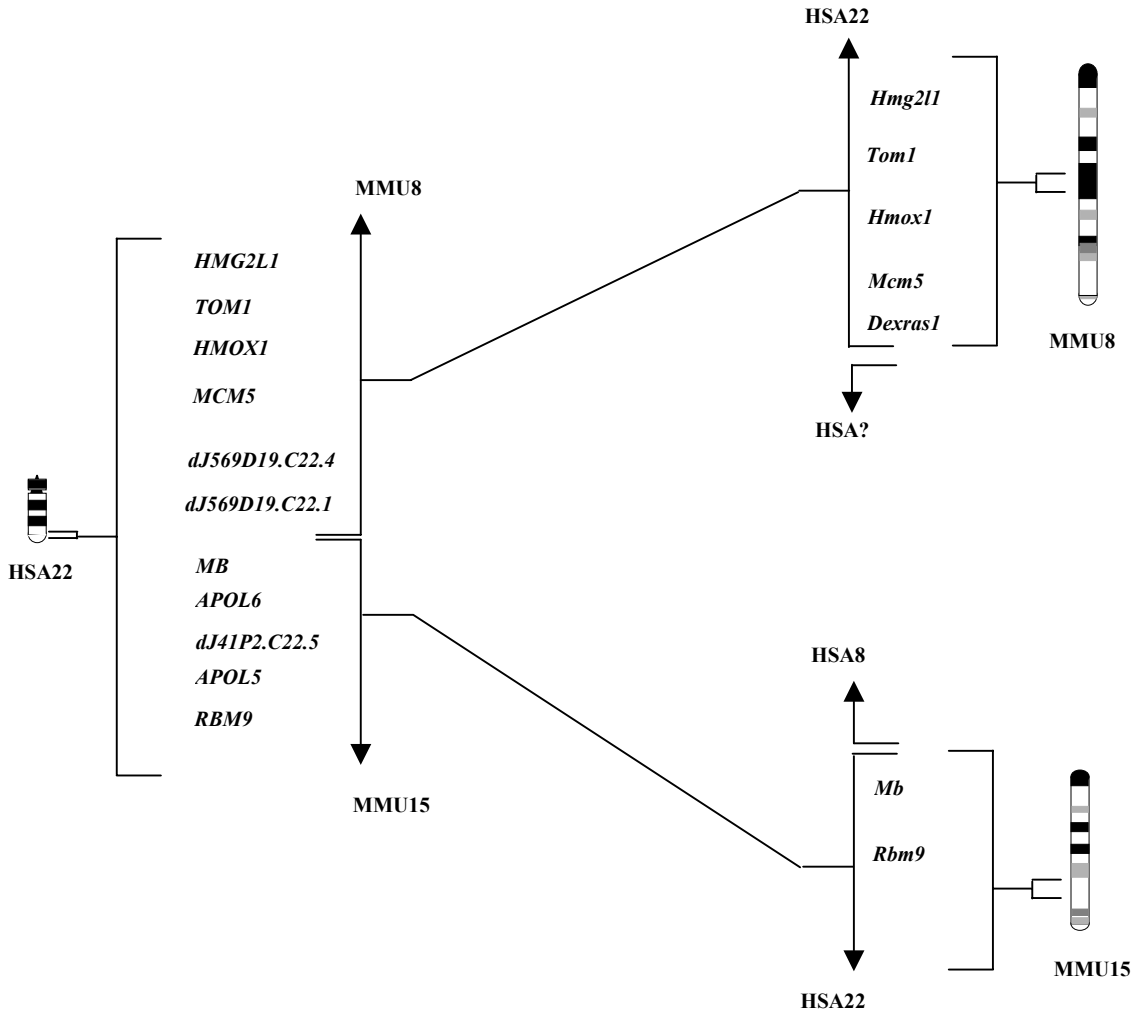


Figure 4.23: Comparative maps define the MMU8:15 chromosome junction region on human chromosome 22. HSA22 gene order is used as the reference. Apart from the apparent absence genes APOL5 and APOL6 and pseudogenes dJ569D19.C22.4 and dJ41P2.C22.5 in the mouse sequence, linkage is conserved within the two mouse chromosomal regions.

Sequence similarity between HSA22 and MMU15 decreases after the gene MB. BLAST experiments using the mouse sequence against the NCBI human genome database show that mouse sequence after this point may be syntenic with HSA8. Additionally, gene predictions in

the unfinished sequence provided by the mouse Ensembl website (http://mouse.ensembl.org) also matched HSA8 sequences in similar BLASTP experiments. This finding correlates with data from the NCBI human-mouse homology map (http://www.ncbi.nlm.nih.gov/Homology).

Similarly, sequence similarity between HSA22 and MMU8 decreases after dJ569D19.C22.1. BLASTP experiments of the mouse sequence against the NCBI human genome database showed low-level similarity to HSA13 and HSA20. However, no genes have been predicted to lie within bM267J18 by Ensembl prediction methods (http://mouse.ensembl.org) and no further information is available on the NCBI human mouse homology map for this region.

4.8.2 The junction region

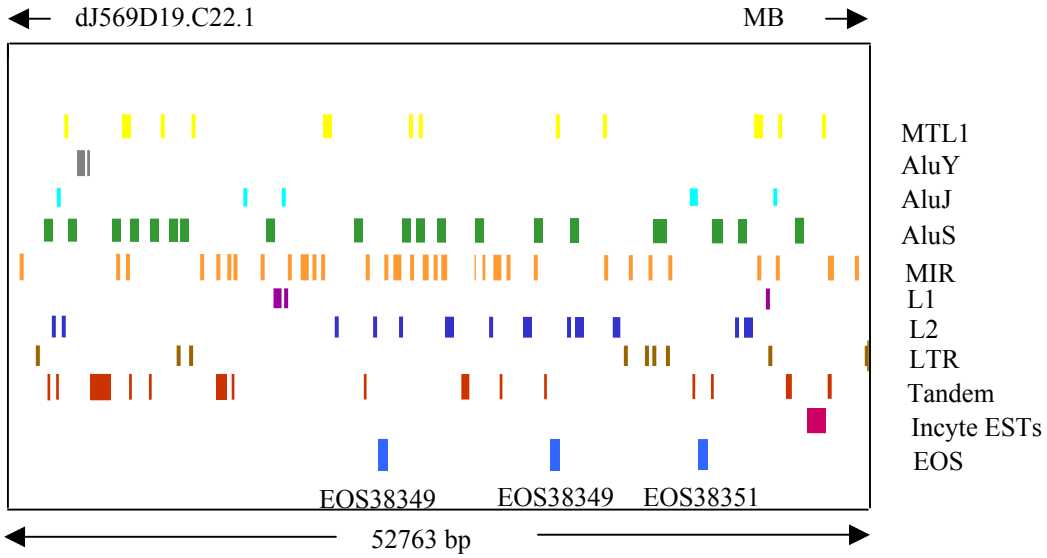


Figure 4.24: Comparative sequence analysis defines the MMU8:15 junction region on human chromosome 22. The junction region is composed of a variety of human repetitive DNA sequences. A cluster of Incyte EST sequences and 3 EOS sequences (see chapter III and appendix 2) are also included within the region.

Repeat sequences make up 40.65% of the 52763 bp MMU8:15 junction region on HSA22 (figure 4.24) and consist of several classes of repetitive DNA elements. Thirty-three

mammalian-wide interspersed repeats (MIRs) were found, distributed throughout the region. The current unfinished nature of much of the mouse sequence in this region, however, makes it difficult to ascertain if these MIR repeats are conserved in the mouse genome. MIRs are believed to have amplified before the radiation of mammals, and their transposition has been implicated in gene control and evolution (Hughes, 2000). A single MIR repeat has also been observed in a HSA21:22 junction region on MMU10 (Pletcher *et al.*, 2000), although no similarity is noted in the distribution of repeat sequences between these two examples.

Three 'EOS' sequences, that have been predicted to be coding by Genscan and which have tested positive for expression by microarray hybridisation (R. Glynne, personal communication) (chapter III and appendix 2), were also contained within the region. Two showed a high level of conservation with sequences on mouse chromosomes 5 (EOS38349), 15, 11, 3, 18 and 6 (EOS38350). EOS38351, along with seven overlapping ESTs from the Incyte database (J. Seilhamer, personal communication) (chapter III and appendix 2) identified in this region, but did not show significant similarity to any other human or mouse DNA or protein sequence by using BLASTN and BLASTX. The remaining 27980 kb of unique sequence was not similar to any known human or mouse sequences.

The sequence analysis of this region and of evolutionary chromosomal breakpoints previously described at the sequence level by both Lund *et al*. (2000) and Pletcher *et al*. (2000), has so far revealed no unusual sequences or repeat structure that might suggest chromosomal instability underlying the rearrangements. As increasing amount of mouse genomic sequence become available, perhaps further examination of similar regions will identify common features of evolutionary chromosomal breakpoint regions.

4.9 Discussion

This chapter has described the construction, sequencing and comparative sequence analysis of approximately 3.5 Mb of the mouse genome, spanning regions of conserved synteny with human chromosome 22q13.31 and with a syntenic breakpoint between mouse chromosomes 8 and 15, within a region of human chromosome 22q13.1.

The use of both fingerprinting and landmark content mapping initially contributed to the construction of three contigs across regions of interest on mouse chromosomes 15 and 8. Restriction enzyme fingerprinting allows analysis over the length of the clone and the construction of contigs relies on the number of bands shared between overlapping clones. The disadvantage of fingerprinting is that it does not allow the orientation of the contigs relative to each other, nor does it allow integration with the framework map. Initial landmark STSs were designed from known orthologous mouse mRNA sequences. Increased marker density was achieved by including STSs to mouse ESTs that demonstrated high similarity to the remaining human genes. The increasing availability of marker and fingerprint data from the mouse physical mouse mapping effort (MGSC, unpublished) anchored the initial contigs to existing mouse framework maps. This combined approach offered the best strategy for contig construction, determining accurately the overlap between clones and integration of the constructed contigs with the framework maps. The resulting BAC maps from this effort provide a resource for the genomic sequencing of these regions of mouse chromosomes 15 and 8 and have been incorporated into the mouse physical map produced by the MGSC (http://mouse.ensembl.org).

PIP analysis of regions of available sequence, show that approximately 90% of annotated gene features within 22q13.31 and 22q13.1 are conserved. 76% of the annotated exons within these regions of HSA22 demonstrate >50% sequence identity with mouse genomic sequence. Interestingly, no mouse sequence homology was noted, outside of repeat regions, of the 18 human pseudogenes annotated in these regions. It may be that these non-functional sequences have diverged more quickly in the mouse genome, possibly because of the much shorter generation time of mouse. Alternatively, some, or all, of the pseudogenes may have arisen in the human lineage after divergence from the common mouse-human ancestor. Otherwise, gene order is generally conserved in these regions. Exceptions were seen with the genes APOL5 and APOL6, which were not found in the available mouse sequence and the APOL2 gene, which may be inverted in mouse. However, a large part of this analysis is based on unfinished sequence and is therefore unconfirmed.

A percentage identity level of 80% was selected for alignment of the mouse genomic sequence generated from this project against the sequence of 22q13.31 and incorporation into 22ace for further analysis. The basis for this choice was the result of preliminary alignment experiments on a subset of the region at a range of identity levels, which suggested that beyond a level of 80% identity, specificity and sensitivity were compromised. This observation is supported by Makalowski and Boguski (1998), who reported that protein-coding exons show an average percent identity of ~85% for many comparisons between human and mouse genes.

The alignment of the 39 annotated gene structures within 22q13.31 (chapter III), with both the mouse genomic sequence generated from this project and other examples of sequence evidence from model organisms, was analysed using MethComp (D. Beare) (chapter II). Higher levels of

specificity and sensitivity were noted for genomic sequence resulting from BLASTN comparison at a level of 80% nucleotide identity of sequence generated by a clone-by-clone shotgun approach than from the WGS mouse project (MSC, unpublished). This may be because the clone-by-clone approach has generated more complete data over the region than the current stage of the WGS project. Interestingly, BLAT alignments (Kent, unpublished) of the output from the WGS project showed greater sensitivity and specificity than alignments from the Exonerate program (Slater, unpublished). The completion of the mouse genome project and reanalysis of these alignments should provide a definitive measure of the correlation of human and mouse sequence in this region.

Overall, these results and those from the equivalent calculations described in chapter II, indicated that the most efficient approach to annotation is through comparison to known gene or protein sequences, both from human and from model organisms. However, this study showed that mouse genomic sequence has the potential to provide an important tool in annotation of the human genome sequence, although comparative sequence analysis utilising mouse genomic sequence supported, but did not add to, the annotation of this already well-studied region (see below). The utility of mouse genomic sequence in this field may therefore lie in the annotation of human genes in previously unstudied regions.

The two regions of human chromosome 22, unlike other examples (Epp *et al.*, 1995; Koop & Hood, 1994; Oeltjen *et al.*, 1997) do not show extensive conservation of intronic and intergenic sequences with mouse, although several isolated examples were noted. Only six conserved regions were also predicted to contain exons by the gene prediction program Genscan (Burge & Karlin, 1997). Three of these predicted exons had already tested negative for expression by

PCR screening of cDNA libraries (chapter III). The remaining three predictions, together with a further 25 candidate exons identified from the human-mouse alignment were tested for expression in seven cDNA libraries. No new exons were confirmed. It is possible, however, that these conserved regions could be transcribed in different tissues or under different conditions than the seven cDNA populations tested. A benefit of mouse sequence comparison is that, unlike EST and cDNA evidence, identification of putative coding regions is not limited by spatial or temporal restrictions on transcription. However, this also means that expression of these regions is difficult to confirm. Analysis of the finished mouse sequence, using techniques similar to those described in chapter III, including detailed comparison to the related human sequence, additional homology searches and use of gene prediction algorithms, may provide additional evidence that these conserved regions encode genes.

The conserved non-coding sequences may also indicate the presence of regulatory elements. The putative promoter regions of six genes, present in both human and mouse finished sequences, were examined for the presence of potential transcription factor binding sites. Thirty-six putative sites were identified in conserved sequences upstream of the annotated transcription start sites of four genes. This investigation represents only a preliminary *in silico* analysis and identification of these regions represents a starting point for further analysis (see chapter I). Many of the consensus sequences listed for possible transcription factor binding sites are very short – only a few nucleotides long in some cases. These could be expected to occur frequently in both functional and non-functional genomic sequence. Recent studies by Göttgens *et al*. (2000) and Frazer *et al*. (2001) have demonstrated the utility of including a third vertebrate species in comparisons of non-coding sequences. Potentially, inclusion of, for example, genomic sequence from chicken or dog, will increase the specificity of this analysis of

potential regulatory regions. Non-coding sequences conserved in all three species will provide strong candidates for future investigation.

Investigation of a 0.5 Mb region of finished mouse sequence showed that the gene structure overall is well conserved in this region between the two species. Comparison of exon and intron size in mouse and human shows that coding regions are more stringently conserved. Increased variation is noted in the sizes of UTR exons. Within coding regions, most insertions/deletions of nucleotides occur in multiples of three, so the reading frame is maintained. Exceptions, such as the shift in reading frame shown between the human and murine versions of BIK, result in a decrease in identity between the predicted protein sequences. It would be interesting to determine if this change has an affect on the functions of the orthologous BIK genes.

The comparison of splice donor and acceptor sites has shown that human and mouse splice sites in this region are highly conserved. The consensus donor and acceptor sites reported in this study are very similar to those reported by Stephens and Schneider(1992) from a study of 1800 human introns, and by Smink (2001) from a study of 84 human and mouse introns. It is therefore clear from the studies that the core splice donor and acceptor sites are strongly conserved in mouse and human.

The repeat density of the 0.5 Mb finished sequence region in mouse (1.33 repeats/kb) is higher than in human (1.26 repeats/kb). This may be explained by the faster murine generation time. Most of the higher repeat density is attributable to the increase in numbers of simple and MaLR repeats. MaLRs retrotransposons are known to be still active within the mouse genome (Smit, 1996). The overall repeat coverage is greater in the human (41.28%) than in mouse (31.90%). This is mainly attributable to the larger size of the human *Alu* repeat, in comparison to the

mouse B1 and B2 repeats (Ansari-Lari *et al.*, 1998). The increased coverage of human repeats contributes to the 1.6X expansion of the sequence length in human compared to mouse. The overall coverage of the repeats in this region are slightly higher than those found in other comparative studies. Ansari-Lari *et al*., (1998) have shown an overall repeat coverage of 33.36% (human) and 26.39% (mouse) whilst Oeltjen *et al*., (1997)(1997) have shown values in the BTK region to be 31.22% (human) and 16.49% (mouse). An additional study by Smink (2000), found repeat coverage to be 39.2% (human) and 11% (mouse) over a 150kb region of human 22q13.3/mouse 15.

The GC content of both human and mouse genomes in this region follow a similar pattern, although the difference in length of the equivalent sequences prohibits direct comparison. This is also reflected by the distribution of predicted CpG islands in the region. All of the six human genes fully annotated in the mouse sequence are associated with a CpG island at the 5' end, whereas only two of the mouse genes start in a predicted CpG island. Peaks in GC content can still be observed for the genes lacking a CpG island, indicating that these regions are relatively GC rich, but not sufficiently so to be predicted as a CpG island. Erosion of mouse CpG islands is generally observed due to deamination of methylated cytosine to thymidine (Cooper & Krawczak, 1989; Coulondre *et al.*, 1978). This also occurs in humans, but the shorter generation time of mouse may account for the faster rate of cytosine deamination and CpG island erosion observed in this and other studies (Aissani & Bernardi, 1991; Antequera & Bird, 1993; Matsuo *et al.*, 1993).

This region of finished mouse sequence was also interesting as it was found to span an 'unclonable' gap in the sequence of human chromosome 22. Analysis of the repeat content of

the mouse 'gap' subregion showed no obvious deviation from that of the total analysed sequence. GC content, however, was maintained at a high level throughout this section. The human GC levels are estimated to be maintained above 50% throughout the gap region. This observation could be a reason why efforts to identify a clone containing the equivalent region in human have so far been unsuccessful. In *Escherichia coli*, (CpG)n repetitive sequences have been shown to be deletion prone (Bichara *et al.*, 1995, 2000). Two pathways have been suggested by which this could occur;

(1) (CpG)n tracts are potential Z-forming DNA sequences and this DNA structure could be processed by an unknown cellular mechanism to give rise to the observed deletions

(2) (CpG)n monotonous runs can be considered as a succession of direct or palindromic repeats, allowing formation of DNA structures that are known to participate in frameshift mutagenesis.

The sequence of the mouse clone and putative structure of the human C22orf1 gene identified by this study could be used in the design of new hybridisation experiments in attempts to identify a human genomic clone spanning this gap from the available libraries.

Examination of unfinished sequence from mouse chromosomes 8 and 15 enabled a more precise definition of the MMU8:15 synteny junction on human chromosome 22q13.1. Investigation of the finished mouse sequence, when available, may further reduce this region. Analysis of the finished human sequence of this junction region identified a range of different repetitive features, including MIR repeats. MIRs are thought to have arisen before the radiation of mammals, and their transposition has been implicated in gene control and evolution (Hughes, 2000). Comparison of this region with the synteny breakpoints analysed by Pletcher *et al*. (2000) and Lund *et al*. (2000), identified no similarity in the distribution of repetitive

sequences. As additional comparative sequence information becomes available, analyses of a range of such synteny breakpoint junction sequences may enable identification of common elements.

In summary, this chapter has shown that comparative sequencing is a powerful tool for the annotation of genomic sequence. Although all the genes annotated during this project were identified without the aid of mouse genomic sequence, the high levels of correlation of the mouse-human sequence alignments with the human transcript map indicate that a completed mouse genome sequence resource will provide a useful gene-finding resource. Comparison of human and mouse genomic sequence will therefore speed the annotation of both genomes. Comparative sequence analysis also enhances *in silico* prediction of conserved regulatory sequences. As the genomic sequence from other vertebrate model organisms becomes available, this process may become more efficient. Comparative analysis also enables detailed, sequence-level analysis of chromosome evolution. This study showed that the availability of genomic sequence permits a level of definition of evolutionary breakpoints that was previously unavailable. An understanding of the mechanism behind these evolutionary changes may develop as more of these detailed comparisons are perfomed.

Chapter V Functional characterisation of protein coding genes from 22q13.31

5.1 Introduction

The ultimate goal of the post-genomic era is to determine the function and biological role of each newly determined sequence (Orengo *et al.*, 1999). Traditionally, small-scale functional characterisation has been successfully carried out on single genes and proteins. Functional genomics is an emerging field, which seeks to establish functional information for all genes or proteins at once in a systematic fashion. Large-scale, high throughput experimental and bioinformatic methods are being developed to further this aim (chapter I).

The starting point for such analyses is ideally a high quality transcript map, providing experimentally verified gene sequences. The previous two chapters have described the production and analysis of such a transcript map of human chromosome 22q13.31. The aim of this chapter is therefore to systematically explore the potential functions of the genes identified in this region, starting with an investigation of the range of data that can be derived *in silico* from the genomic, cDNA and predicted protein sequences, before moving on to preliminary experimental studies of protein function.

Current strategies to functionally characterise proteins generally fall into one of two classes: bioinformatic (*in silico*) analysis and experimental investigation. These approaches are outlined below.

5.1.1 *In silico* methods

5.1.1.1 Database searching

Bioinformatic techniques normally assign functional data by searching for well-characterised relatives in sequence databases. This approach has proven successful although, from a formal point of view, the hypotheses generated must be experimentally verified (Eisenhaber *et al.*, 1995). Information transfer from well-studied proteins to uncharacterised gene products has to

be done carefully, since (i) a similar sequence does not always imply similar protein structure (Sander & Schneider, 1991) or function and (ii) the annotation of the database protein may be incomplete or even wrong. Standard database searches may also fail to pick up distant structural relationships. These may only be recognised from comparison of the 3D structure if available, which is highly conserved during evolution. For these reasons, many resources that aid computational functional characterisation of a protein at different levels have been developed, but there is still a need for more programs to be designed. Output from such programs provides a large amount of information, which needs to be experimentally verified to obtain preliminary data.

5.1.1.2 Domain analysis

Many proteins are modular and have a multidomain architecture. Protein domains are multiply adapted by evolutionary processes and often re-used in a different context. Several databases exist that comprise of patterns or profiles of classified domains, including Pfam (Bateman *et al.*, 1999), PRINTS (Attwood *et al.*, 2002), PROSITE (Hofmann *et al.*, 1999), ProDom (Corpet *et al.*, 2000), SMART (Schultz *et al.*, 2000) and SWISS-PROT and TrEMBL data (Bairoch & Apweiler, 2000). Although somewhat redundant, they each have different strengths (reviewed by Bork & Koonin, 1998). Several resources exist which allow the user to search several of these databases at once and integrate the output. The current release of the InterPro database (3.2) (Apweiler *et al.*, 2000) is built from Pfam 6.2, PRINTS 30.0, PROSITE 16.37, ProDom 2001.1, SMART 3.1 and the current SWISS-PROT + TrEMBL data. This release of InterPro contains 3939 entries, representing 1009 domains, 2850 families, 65 repeats and 15 post-translational modification sites.

5.1.1.3 Intrinsic feature analysis

Protein sequences can contain low complexity regions with a reduced residue alphabet. These common regions can generate spurious matches between otherwise non-related proteins and therefore must be filtered out from database searches. However, these residues may contain useful functional and structural information and several programs exist that are designed to predict their presence (section 5.3.1). The results must be treated with caution though as different prediction algorithms can produce different results. Several major classes of intrinsic features are described here.

Transmembrane regions contain helical structures with a hydrophobic exterior, adapted for a lipid-bilayer environment. Membrane proteins often mediate communication across cell membranes. Despite their biological and medical importance, there is very little experimental information about their 3D structures: <1% of the proteins of known structure are membrane proteins (Liu & Rost, 2001).

Coiled-coil proteins, containing heptarepeats with patterns of hydrophobic and polar residues, are typically formed as bundles of several right-handed alpha helices twisted around each other, forming a left-handed super helix (Lupas, 1996). Coiled-coils often mediate protein-protein interaction, or form filaments and other microscopic structures.

Proteins may also contain small repeats that lead to a bias in amino acid composition and other regions with biases towards one or several amino acids, such as proline-rich regions. Signal peptides are an additional feature of interest and are predicted fairly accurately (Emanuelsson *et al.*, 2000; Nielsen *et al.*, 1997), although signal peptides from different proteins may have diverse sequences. Signal peptides at the N-terminal end target many prokaryotic and eukaryotic proteins to the secretory pathway or membrane organelles (Cleves, 1997; Nakai & Ishikawa, 2000; Nielsen *et al.*, 1997; Thanassi & Hultgren, 2000).

5.1.1.4 Similarity analysis

A database search using BLAST often reveals significant similarities. A recent BLASTP search by Lander *et al*.(2001) revealed that 74% of known human proteins had significant matches to other known proteins. Only in a minority of cases, however, can functional and structural features of a homologue be transferred to the query sequence because often only some of the features are shared.

Functional equivalence is only likely for orthologues: genes whose independent evolution reflects a speciation event rather than a gene duplication event (Fitch, 1970). They are likely to perform the same function in various species and hence represent a refinement over homologues in sequence analysis and annotation. Orthologues are expected to have the highest level of pairwise similarity between all the genes in two genomes (Huynen & Bork, 1998; Tatusov *et al.*, 1997; Tatusov *et al.*, 1996). However, unambiguous assignment of human gene orthologues on this basis alone is difficult. Current database search techniques are not able to discriminate whether the best hit is an orthologue (and therefore potentially functionally equivalent) or only a paralogue, i.e. a homologous member of a multigene family that shares, at best, only some functional features with the query sequence. A large-scale 'all-against-all' sequence comparison of human*, C. elegans* and *D. melanogaster* proteins has shown that most human proteins do not exhibit simple 1:1:1 orthologous relationships and only a minor fraction of homologous relations could be classed as orthologues (Lander *et al.*, 2001).

Subsequent phylogenetic analysis to derive the evolutionary relationships of the identified similar proteins can identify potential orthologous genes, but phylogenetic approaches have inherent limitations. Different methods can produce conflicting results because of ambiguities in identifying homologous characters of alignments, sensitivity of tree-making methods to unequal evolutionary rates, biases in species sampling, unrecognised paralogy, functional

differentiation, loss of phylogenetic informational content due to fast evolution and difficulties with the assumptions and approximations used to infer phylogenetic relationships (reviewed by Brocchieri, 2001). Additionally, phylogenetic analyses are computationally expensive and so difficult to perform on large data sets.

5.1.2 Experimental approaches to determining protein function

Generally, only the molecular function of a protein can be transferred by analogy: it is rare that a particular sequence motif strongly correlates with cellular function. Sometimes, only the expression pattern and the tissue context determine the final functionality (for example, high sequence identity and even sequence equivalence between metabolic medium-chain dehydrogenases and eye lens crystallins (Persson *et al.*, 1994; Piatigorsky & Wistow, 1991)). EST databases can provide information on the tissue distribution of genes, but transcripts that have low levels of expression, or limited spatial or temporal distribution, may escape detection (chapter III). Large-scale expression analysis techniques have been developed, (chapter I). However, the power of such analyses is limited by the current lack of a full catalogue of human genes, once again highlighting the need for full and accurate annotation of the human transcriptome. In addition, accessibility to, and analysis of, the mass of new data is limited, as there is a lack of sufficiently powerful mathematical and visualisation tools for whole-genome expression studies and most is not available on the web, or may not be publicly available.

Knowledge of the mRNA expression pattern alone, however, does not necessarily indicate protein function. Several methods, that have been adapted for large-scale analysis of expression and function at the protein level, have also been described, for example, mass spectrometry of protein complexes, structural analysis and two-hybrid protein-protein interactions. However, improved techniques are still needed for the global analysis of protein expression, post-translational modification, protein subcellular localisation, protein-protein

interactions and chemical inhibition of pathways. New computational technologies will be needed to use such information to model cellular circuitry (chapter I).

5.1.2.1 Subcellular protein localisation

This chapter concentrates on techniques for analysis of subcellular localisation. The eukaryotic cell achieves spatial and temporal regulation of biochemical reactions by a high degree of compartmentalisation. Localisation of proteins involved in a specific network to a particular organelle or compartment both facilitates interactions and allows the segregation of different networks. Information is exchanged between the compartments by active transport of material to ensure that the cell functions properly.

Bioinformatic tools have been developed with the aim of predicting protein localisation based on features within the amino acid sequence. For instance, PSORT (Nakai & Horton, 1999) detects in sequences the signals required for sorting proteins to particular subcellular compartments and generates a prediction of protein localisation. However, as with all the bioinformatic approaches described above, these predictions require experimental confirmation.

Several papers have been published that describe efforts to generate large-scale subcellular protein localisation techniques and are reviewed by Pepperkok *et al*. (2001). The techniques described by Ding *et al.* (2000), Merkulov & Boeke (1998), Pichon *et al.* (2000), Rolls *et al.* (1999), Sawin & Nurse (1996), involve the fusion of the coding sequence of green fluorescent protein (GFP) to either fragments from genomic libraries or individual clones from cDNA libraries. The fusions are then expressed in cells or tissues and their subcellular localisation determined by microscope inspection. Subsequently the respective cDNA was rescued, cloned and sequenced. Although this research has resulted in the localisation of many previously uncharacterised proteins, at least 50% of the cDNAs were already known and had been

characterised (Merkulov & Boeke, 1998; Pichon *et al.*, 2000; Rolls *et al.*, 1999). The genome projects have resulted in the identification of many previously unknown proteins. Individual tagging of the full-length cDNAs encoding only these genes enhances the efficiency of these approaches (Hoja *et al.*, 2000; Simpson *et al.*, 2000).

A major drawback of GFP-fusion techniques is that the reporter protein could mask targeting signals contained within the expressed protein. For example, amino-terminal fusions of GFP to target proteins have been shown in some cases to block signal sequences associated with import into mitochondria and endoplasmic reticulum (Simpson *et al.*, 2000). Different versions of full-length GFP-fusions, tagged at either the amino or carboxyl terminus, can be generated and compared to try to circumvent this risk (Simpson *et al.*, 2000) but it is unclear what affect the position of the GFP fusion has on less well-characterised signal sequences.

5.1.3 Summary

This chapter describes the use of a variety of approaches to functionally characterise 27 complete protein-coding genes, including the initial characterisation of 15 previously unstudied novel genes. Bioinformatic approaches, including domain and secondary structure predictions and phylogenetic analyses, were combined with expression and subcellular localisation studies, to increase understanding of the function of the proteins encoded within 22q13.31.

5.2 Previously published functional data

This thesis has described the production of a high quality transcript map of human chromosome 22q13.31. Thirty-nine genes have been found within this genomic region. One of these, dJ222E13.C22.7, encodes a snoRNA involved in splicing of U12-dependent introns (Montzka & Steitz, 1988). The remaining 38 gene structures potentially encode peptide sequences. Eleven of these structures, however, remain only partially complete. The remaining

27 'full' genes, which have an experimentally verified, unambiguous ORF with a defined start and stop codon, are included in the preliminary study of functional characterisation described in this chapter. Additionally, 15 different gene isoforms have been identified from expressed sequence evidence and are included for functional characterisation. In all, analysis of 42 potential protein sequences is described in this chapter.

Database searches with the nucleotide and predicted amino acid sequence of the 27 full genes showed that 12 of them had previously been cloned and the mRNAs and/or encoded proteins have undergone a range of functional classification analyses. A brief description of what is currently known about each of the mRNAs and/or proteins is contained in table 5.1. Where possible, the SwissProt protein accession number has been listed. SwissProt entries are not yet available for cB33B7.C22.1, and PACSIN2, but some functional characterisation of these proteins has previously been described. ARFGAP1 and TTLL1 also do not have a SwissProt entry, but have been analysed at the mRNA level. EMBL accession numbers for these genes are listed overleaf.

Table 5.1: The available functional information for 12 mRNAs and/or proteins encoded within human chromosome 22q13.31. Functional descriptions are summarised from the referenced papers.

Gene	Accession	A brief description of functional characterisation	References
DIA1	Sw:P00387	Desaturation and elongation of fatty acids, cholesterol biosynthesis, drug metabolism. Methemoglobin reduction in erythrocytes (functional assay).	Yubisui *et al.*, 1984; Shirabe *et al.*, 1991
cB33B7.C22.1	Em:AB037883	Globotriaosylceramide (Gb3)/CD77 synthase (α1,4-galactosyltransferase). Transfection in L cells produces neosynthesis of Gb3/CD77 and sensitivity to Shiga-like toxins. Cell extracts show α1,4-galactosyltransferase activity (functional assay). The genetic basis of the p histo-blood group phenotype.	Kojima *et al.*, 2000; Steffensen *et al.*, 2000
ARFGAP1	Em:AF111847	Possible role in the function of sperm (by similarity).	Zhang *et al.*, 2000
PACSIN2	Em:AAD41781	Binds to endocytic proteins, inhibits endocytosis (functional assay).	Ritter *et al.*, 1999
TTLL1	Em:AL58967; Em:AL096883; Em:AL096886; Em:AF104927	Possible role in the post-translational modification of α-tubulins (by similarity).	Trichet *et al.*, 2000
BIK	Sw:Q13323; Sw:Q16582	Accelerates programmed cell death. Binding to BCL-X, BHRF1 or BCL-2 represses this death-promoting activity (functional assay).	Boyd *et al.*, 1995; Castells *et al.*, 1999; Chittenden *et al.*, 1995; Han *et al.*, 1996
BZRP	Sw:P30536	Manifestation of peripheral-type benzodiazepine recognition sites. Contains binding domains for benzodiazepines and isoquinoline carboxamides. Role in the transport of porphyrins and heme (functional assay).	Riond *et al.*, 1991
C22orf1	Sw:O15442	Possible role in CNS development (by similarity).	Schwartz & Ota, 1997
NUP50	Sw:Q9UKX7	Associated with the nuclear pore (by similarity).	Trichet *et al.*, 2000
UPK3	Sw:O75631	Part of the asymmetric unit membrane (AUM). Possible role in AUM-cytoskeleton interaction in terminally differentiated urothelial cells. Role in the formation of urothelial glycocalyx, which may be involved in preventing bacterial adherence (by similarity).	Yuasa *et al.*, 1998
FBLN1	Sw:P23142; Sw:P23143; Sw:P23144; Sw:P37888; Sw:Q9UGR4	Secreted into the extracellular matrix (functional assay).	Argraves *et al.*, 1990
E46L	Sw:Q9UBB4; Sw:O14998; Sw:O15009;	Defects in SCA10 (E46L) result in spinocerebellar ataxia type 10, an autosomal dominant disorder characterised by cerebellar ataxia seizures. The molecular basis of the disease is due to an ATTCT nucleotide repeat expansion in intron IX.	Matsuura *et al.*, 2000

5.3 *In silico* analysis

The remaining fifteen full genes were not identified in database searches of previously characterised nucleotide or protein sequences. A range of *in silico* analyses was therefore performed on the predicted protein sequences to investigate the presence of any domains or intrinsic sequence features that may give an indication of potential function. The twelve

previously characterised sequences were included in these analyses to provide a useful control and to possibly uncover additional information about them.

5.3.1.1 Intrinsic feature analysis

A large number of programs are available, which recognise features of a protein sequence that may be consistent with a range of secondary structural characteristics. The PIX suite of protein analysis programs provides predictions of secondary structures (DSC, Simpa96), low complexity regions and long/short globular domains (Seg), coiled coil predictions (Coils), transmembrane regions (Tmpred, Tmap and DAS), helix-turn-helix predictions (HTH), signal peptide predictions (Signal, Sigcleave), antigenic regions (Antigenic) and enzyme digest predictions (Digest) (http://www.hgmp.mrc.ac.uk/Registered/Webapp/pix/). An advantage of using PIX is that results from these programs are displayed together, so similarities and differences from different algorithms can be noted. Individual amino acid properties including acidity, polarity, hydrophobicity, aromaticity, charge and size are also included in the PIX display output.

An example of the PIX output is displayed in figure 5.1. The results from the other 41 complete protein sequences are catalogued in appendix 7 and an overview is provided in figure 5.2. PIX can also provide output from limited domain and motif database searches of the SPTREMBL, ProDom, Pfam, Blocks and Prosite databases. The results of a more comprehensive search are addressed in section 5.3.2 and so are not included here.
This analysis of dJ222E13.C22.1 (isoform a) shows that the two predictions of secondary structure prediction provided by DSC and Simpa96 generally agree, although several discrepancies are noted in the sizes of the predicted α-helix and β-strand regions. Both programs also predict beta strand regions that are not supported by the other. Two possible transmembrane regions are supported by more than one prediction program (Tmpred, TMAP

and DAS, and Tmpred and DAS respectively). These, as expected, correspond with hydrophobic regions of the peptide sequence. The existence of supporting predictions provides additional confidence in predictions of secondary structure. Use of two different matrices of the Coils prediction program supports the existence of a coiled coil region between the two transmembrane sections of the peptide. Again, as expected, this corresponds to a low complexity segment of the sequence. Of further note in this analysis is the consensus reached by Sig and Sigcleave of a potential signal sequence at the N terminal of the peptide. This occurs between amino acids 47 and 48 and may indicate the existence of a signal peptide.

5.3.1.2 Overall results

An overview of this analysis is shown as part of figure 5.2. Thirty-seven of the 42 protein sequences (86%) contained at least one consensus prediction of a transmembrane region. BZRP contains the most (five) and has previously been described as an integral membrane protein (Riond *et al.*, 1991). Similarly, UPK3 is predicted to contain three transmembrane regions and has previously been shown to be a type I membrane protein (Yuasa *et al.*, 1998) found in the asymmetric unit membrane (table 5.1). The remaining 35 proteins contain between one and four consensus predictions of transmembrane regions and might play a wide variety of roles in transmembrane communication, cell signalling etc.

Twelve protein sequences (29%) contained coiled-coil regions that were predicted by more than one program. Ten of these also contained transmembrane regions. Involvement of coiled-coil proteins with protein-protein interactions and formation of structural microfilaments has previously been noted (Creighton, 1993). The proteins in which coiled-coil regions form more than 50% of the predicted structure, ARFGAP1, PACSIN2, bK414D7.C22.1, dJ671O14.C22.2 and dJ102D24.C22.2, may be particularly likely to be involved in these processes.

Interestingly, no helix-turn-helix regions were predicted in any of the protein sequences queried. The helix-turn-helix motif is often observed in proteins that have no other structural similarities. Often found in transcription factor proteins, it protrudes from the protein structure in order to penetrate the DNA major groove (Creighton, 1993).

N-terminal signal peptides were predicted to be present in dJ222E13.C22.1a, ARFGAP1, bK268H5.C22.4, UPK3, and all four isoforms of FBLN1. The export of FBLN1 to the extracellular matrix and UPK3 to the asymmetric unit membrane has previously been experimentally confirmed (table 5.1). The subcellular location of dJ222E13.C22.1a, ARFGAP1 and bK268H5.C22.4 may also be directed by possible signal peptide motifs. The subcellular location of all the proteins described here is investigated more fully in section 5.4.

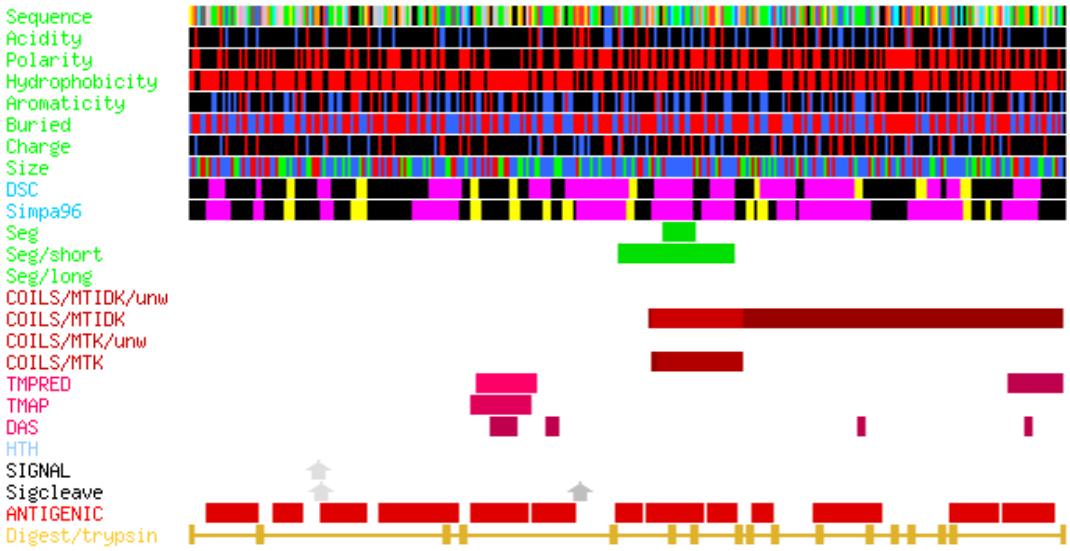


Figure 5.1: PIX display out put showing analysis of the translated coding sequence of dJ222E13.C22.1 (isoform a).
The sequence is displayed in several colour schemes in order to highlight various aspects of the sequence. The key is shown below: letters refer to amino acid symbols.
Sequence: Normal (rasmol) colouring. DE bright red; KR blue; G light grey; A dark grey; H pale blue; CM yellow; ST orange; NQ cyan; LVI green; W pink; P flesh.
Acidity: Acidic/Basic (Red=acidic, Blue=basic). DE red; RKH blue.
Polarity: Polar (Red=Polar). RNDQEHKSTWY red.
Hydrophobicity: Hydrophobic (Red=Hydrophobic). ACGILMFPSTWYV red.
Aromaticity: Aromatic/Aliphatic (Red=Aromatic, Blue=Aliphatic). HFWY red; ILV blue.
Buried: Surface/Buried (Red=Surface, Blue=Buried). RNDEQGHKPSTY red; ACILMFWV blue.
Charge: Positive/Negative charge (Red=Positive, Blue=Negative). RHK red; DE blue.
Size: Tiny/Small/Large (Red=Tiny, Green=Small, Blue=Large). AGS red; NDCPTV green; REQHILKMFWY blue.
DSC & Simpa96: Prediction of protein secondary structure. Coil region white; alpha helix magenta; beta strand yellow.
Seg: segment sequence by local complexity. Low complexity region green.
Seg short/long: prediction of short/long non-globular regions. Non-globular region green.
Coils MTK/MTIDK, wt/uwt: prediction of solvent-exposed left-handed coiled coils. 'Excellent' prediction light brown; 'good' prediction mid-brown; 'marginal' prediction dark brown.
Tmpred, TMAP, DAS: prediction of transmembrane segments. 'Excellent' prediction light purple; 'good' prediction mid-purple; 'marginal' prediction dark purple.
HTH: Helix-turn-helix prediction.
Signal/Sigcleave: Signal sequence prediction.
Antigenic: prediction of antigenic regions of protein sequence. antigenic red.
Digest/trypsin: prediction of peptide fragments produced by digestion with trypsin. (Key adapted from Williams and Faller M. (1999) (http://www.hgmp.mrc.ac.uk/Registered/Webapp/pix/).

5.3.2 Domain Analysis

InterPro 3.2 was searched to identify possible domains, families, repeats or post-translational modification sites contained within the translated full coding sequences annotated within

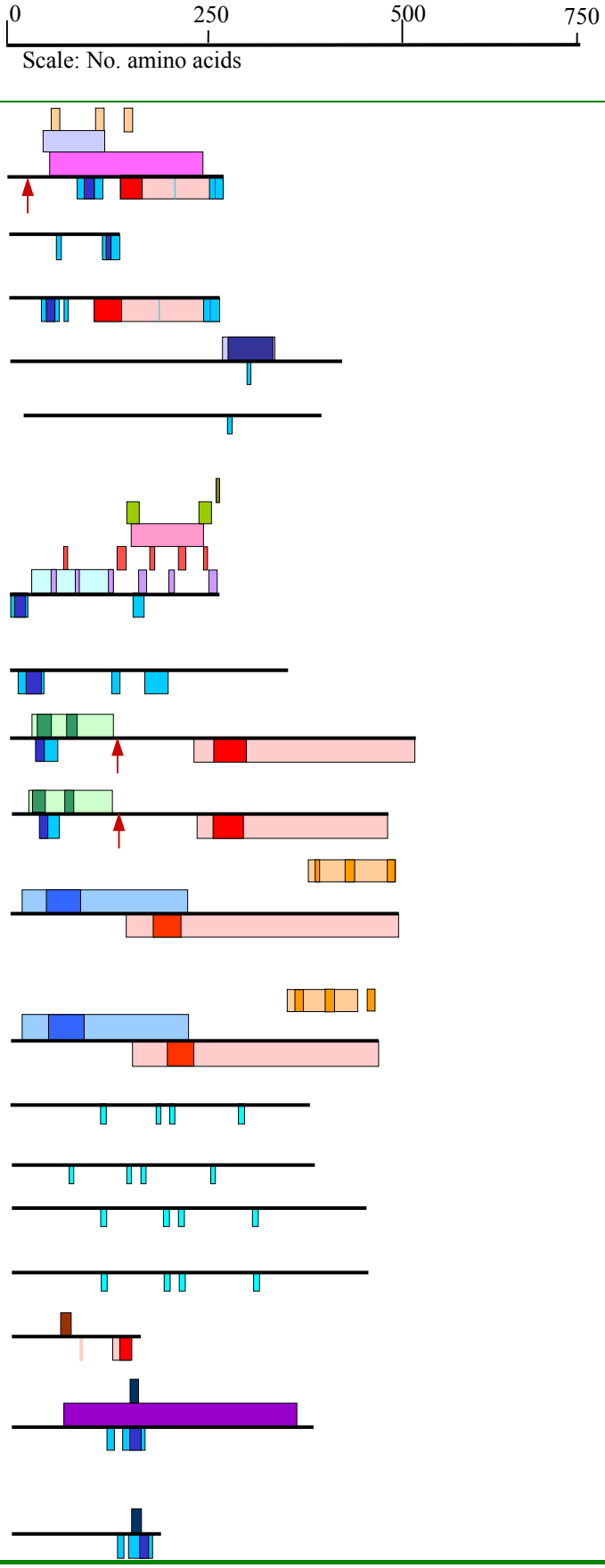
22q13.31. Where possible, InterPro attaches potential functions to the domains. The coding sequences of known alternatively spliced gene structures were included in the search to identify whether the alternative splice altered the domain content of the protein sequence.

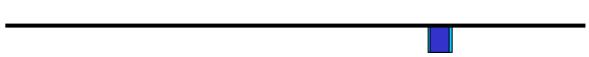
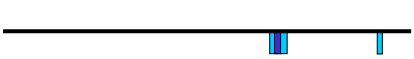
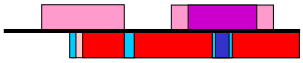
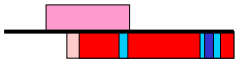
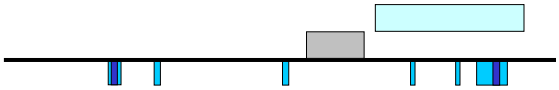
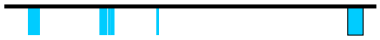
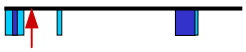
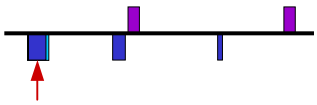
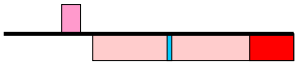
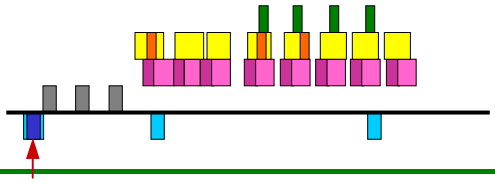
A diagram of each peptide showing InterPro features (depicted above the line in each diagram) and transmembrane, coiled coil regions and potential N-terminal signal peptides (depicted below the line in each diagram), is shown below in figure 5.2. Minimum (dark shades) and maximum (light shades) lengths of a particular predicted protein feature are shown where two or more prediction programs gave conflicting results. The domain descriptions are listed in table 5.2 below.

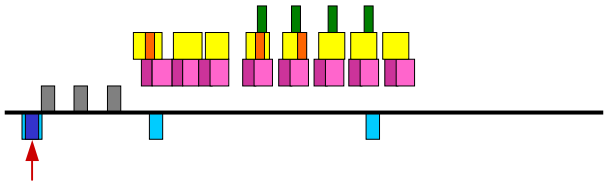
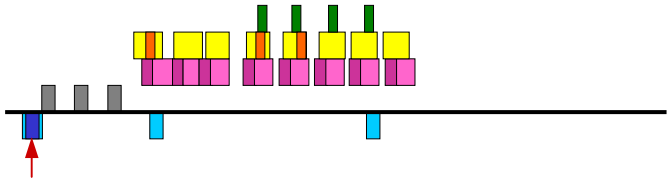
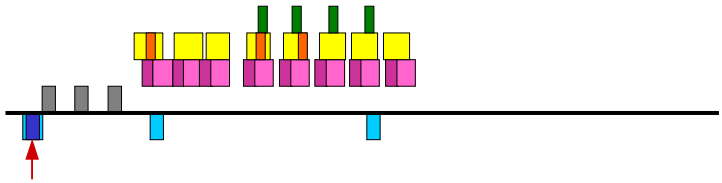
Table 5.2 shows that overall 16 of the 27 protein coding (ignoring alternative splice forms) contained a domain or other InterPro feature. Six of these were identified as multidomain genes. Interestingly, the alternative splice forms of dJ222E13.C22.1, dJ222E13.C22.3 and bK1191B2.C22.3 contained different numbers of domains. This could mean that the different alternative splice forms encode proteins with different, or modified, functions. Domains noted in the genes DIA1, BIK, C22orf1, and FBLN1 support previously published functional studies (table 5.1). No known domains were found in BZRP, NUP50, UPK3 or E46L, which is also consistent with published reports.

Figure 5.2 incorporates results from both the secondary structure and domain analysis to allow an overview of the amino acid sequence contents.

Protein	Length (No. amino acids)	Interpro feature
dJ222E13.C22.1a	314	IPR000073 IPR000379 IPR003089
dJ222E13.C22.1b	134	
dJ222E13.C22.1c	250	
dJ222E13.C22.3a	451	IPR0000504
dJ222E13.C22.3.b	391	
DIA1	301	IPR001230 IPR001221 IPR001433 IPR001709 IPR001834
cB33B7.C22.1	353	
ARFGAP1a	516	IPR001164
ARFGAP1b	472	IPR001164
PACSIN2a	486	IPR001452 IPR001060
PACSIN2b	445	IPR001452 IPR001060
TTLL1a	394	
TTLL1b	385	
TTLL1c	423	
TTLL1d	422	
BIK	160	IPR000712
bK1191B2.C22.3a	357	IPR001617 IPR001227
bK1191B2.C22.3b	180	IPR001617



BZRP	169		
dJ526I14.C22.2a	644		
dJ526I14.C22.2b	273		
C22orf1	301	IPR000934	
SULTX3a	284	IPR000863	
SULTX3b	260	IPR000863	
dJ549K18.C22.1	481		
CGI-51	469		
bK414D7.C22.1	364	IPR001715	
dJ671O14.C22.2a	331	IPR001715	
dJ671O14.C22.2b	273	IPR001715	
dJ1033E15.C22.2	113		
ARHGAP8	643	IPR000198 IPR001251	
dJ127B20.C22.3	432	IPR001965	
NUP50	468		
bK268H5.C22.1	404		
UPK3	287		
bK268H5.C22.4	357	IPR001472	
dJ102D24.C22.2	309		
FBLN1a	566	IPR000152 IPR000561 IPR000020 IPR001881	

FBLN1b	601	IPR000152 IPR000561 IPR000020 IPR001881	
FBLN1c	683	IPR000152 IPR000561 IPR000020 IPR001881	
FBLN1d	703	IPR000152 IPR000561 IPR000020 IPR001881	
E46L	475		

Different isoforms are denoted by a, b, c etc. InterPro features are shown above the line.

Transmembrane regions, coiled coil regions and predicted signal peptides are denoted below the line.

- **Maximal predicted transmembrane region, predicted by >=1 program**
- **Minimal predicted transmembrane region, predicted by >1 program**
- **Maximal predicted coiled coil region, predicted by >=1 program**
- **Minimal predicted coiled coil region, predicted by >1 program**
- **N terminal signal peptide, predicted by >1 program**

Table 5.2: Domain-containing proteins. The domain, InterPro accession number and potential function are listed.

Protein	InterPro accession	Title	InterPro function
dJ222E13.C22.1a	IPR000073	Alpha/beta hydrolase fold	
	IPR000379	Esterase/lipase/thioesterase family active site	enzyme
	IPR003089	Hydrolases	hydrolase
dJ222E13.C22.1b	IPR000379	Esterase/lipase/thioesterase family active site	enzyme
	IPR003089	Hydrolases	Hydrolase
dJ222E13.C22.3.a	IPR0000504	RNA-binding region RNP-1(RNA recognition motif)	nucleic acid binding
DIA1	IPR001230	Prenyl group binding site (CAAX box)	
	IPR001221	Phenol hydroxylase reductase family	
	IPR001433	Oxidoreductase FAD/NAD-binding domain	electron transfer flavoprotein
	IPR001709	Flavoprotein pyridine nucleotide cytochrome reductase	electron transfer flavoprotein
	IPR001834	FAD/NAD-binding cytochrome reductase/cytochrome B5 reductase	
	IPR001834	FAD/NAD-binding Cytochrome reductase/cytochrome B5 reductase	electron transfer flavoprotein
ARFGAP1a & b	IPR001164	Zinc-finger GCS-type	DNA binding
PACSIN2a & b	IPR001060	Cell division control protein 15 (CDC15)	
	IPR001452	Src homology 3 (SH3) domain	
BIK	IPR000712	Apoptosis regulator protein, Bcl-2 family BH domain	apoptosis regulator
bK1191B2.C22.3a	IPR001227	Acyl transferase domain	transferase
	IPR001617	ABC transporters family	
bK1191B2.C22.3b	IPR001617	ABC transporters family	
C22orf1	IPR000934	Serine/threonine specific protein phosphatase	phosphatase
SULTX3a & b	IPR000863	Sulfotransferase	sulfotransferase
bK414D7.C22.1	IPR001715	Calponin homology (CH) domain	actin binding
dJ671O14.C22.2a & b	IPR001715	Calponin homology (CH) domain	actin binding
ARHGAP8	IPR000198	RhoGAP domain	
	IPR0001251	Cellular retinaldehyde binding protein (CRAL)/Triple function domain (TRIO)	
dJ127B20.C22.3	IPR001965	PHD-finger	DNA binding
bK268H5.C22.4	IPR001472	Bipartite nuclear localisation signal	
FBLN1a, b, c & d	IPR000020	Anaphylotoxin domain	plasma glycoprotein
	IPR000152	Aspartic acid and asparagine hydroxylation site	
	IPR000561	EGF_like domain	
	IPR001881	Calcium-binding EGF_like domain	calcium binding

5.3.3 Orthologues

Additional functional information about a protein can be derived from a previously characterised orthologous gene. Potential orthologues of the 27 full protein sequences from 22q13.31 were identified as described below. The full criteria for database searches and tree construction are listed in chapter II.

Refined data sets of homologous sequences from BLASTP searches of the NCBI nonredundant protein sequence database showed that 25 of the proteins had significant matches to known proteins. These sequences were aligned using clustalw (Thompson *et al.*, 1994) and results visualised using belvu (Sonnhammer, unpublished)**.** Neighbour Joining (NJ)-tree analyses of the datasets were then produced using the Phylowin package (Galtier *et al.*, 1996), in order to distinguish between potential orthologues and paralogues amongst the similar sequences. Additionally, the chromosomal position of potential mouse orthologues was verified as far as possible by searching with the nucleotide sequence against the available mapped mouse genomic sequence (http://mouse.ensembl.org) using BLAST. In all cases, the potential mouse orthologues were positioned on mouse chromosome 15, within a region that demonstrates conserved synteny to human chromosome 22 (chapter IV). Literature searches were then undertaken to ascertain if any of the candidate orthologues had previously been functionally characterised. An example of this analysis is provided by dJ222E13.C22.1, shown below.

Figure 5.3 shows an alignment of five similar protein sequences identified from BLASTP searches of the NCBI protein sequence database with the predicted protein sequence of the human gene dJ222E13.C22.1. The phylogenetic comparison of dJ222E13.C22.1 and the similar proteins, shown in figure 5.4, segregates the human and mouse proteins into a potentially orthologous group. Comparison of the *Mus musculus* protein sequence NP_075964.1, against the NCBI nonredundant protein sequence database using BLAST, confirmed that the potentially orthologous pair were the two most similar known proteins found between the two organisms. Additionally, the nucleotide sequence of NP_075964.1 was compared against the available mouse genomic sequence using BLAST, in order to verify, as far as possible, chromosomal position using the Ensembl mouse database

(http://mouse.ensembl.org). NP_075964.1 was localised to the region of mouse chromosome 15 with conserved synteny to human chromosome 22q13.31.

Literature searches were then carried out to determine if NP_075964.1 had previously been characterised. Sadusky *et al*. (2001) describe that this murine protein encodes Serhl, which immunolocalises to perinuclear vesicles when transiently expressed in muscle cells *in vitro.* The mRNA is expressed in murine skeletal muscle and undergoes increased expression in response to passive stretch. In comparison, expression of dJ222E13.C22.1 is noted in skeletal muscle and a range of other tissue (chapter III), although analysis of subcellular localisation (see section 5.4) does not indicate localisation to perinuclear vesicles, but instead suggests localisation in the cytoplasm. Both the mouse and human proteins contain putative α/β hydrolase folds and a serine hydrolase active centre (figure 5.2). Sadusky *et al*. (2001) conclude that Serhl's expression pattern and response to passive stretch indicate that it may play a role in normal peroxisome function and skeletal muscle growth, in response to mechanical stimuli.

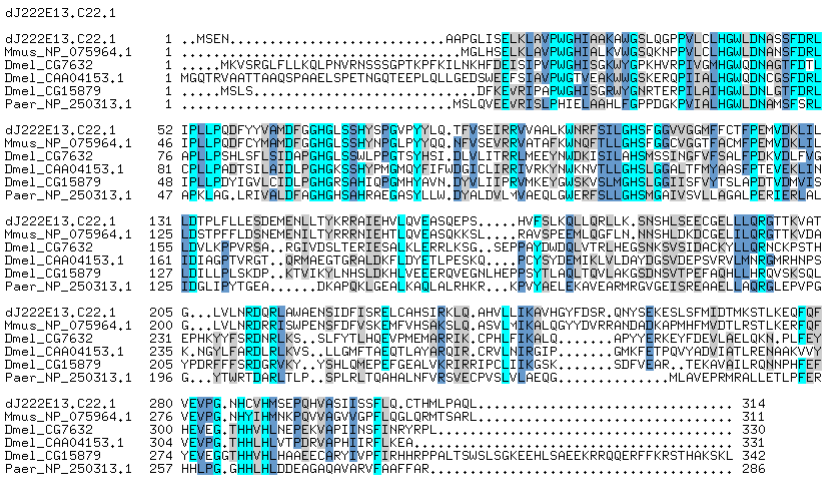


Figure 5.3: Clustalw alignment of the amino acid sequence of dJ222E13.C22.1 against five homologous protein sequences identified from a BLASTP search of the NCBI nonredundant protein sequence database.

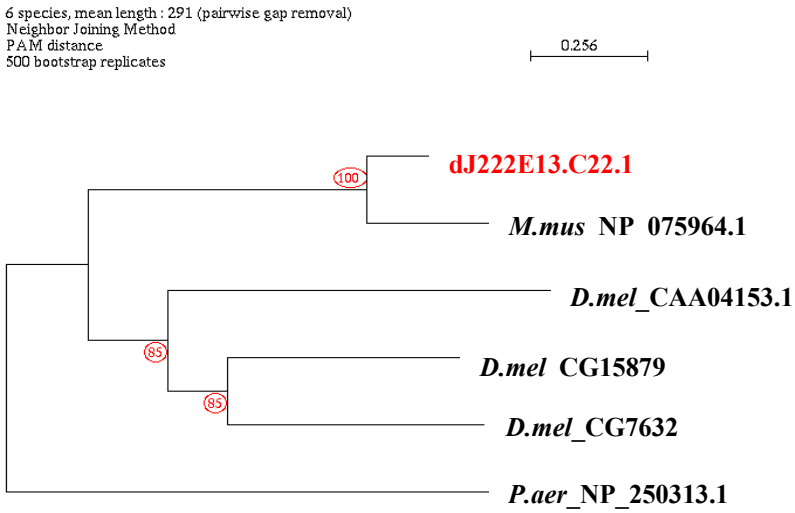


Figure 5.4: Phylogenetic tree derived from the above alignment using the Phylowin package (Galtier *et al.*, 1996). The human protein dJ222E13.C22.1 from chromosome 22q13.31 is highlighted in red. The distance-based tree making method used was the Neighbour-joining method (Saitou & Nei, 1987). The numbers circled in red show the percentage number of times each branch was reproduced from 500 bootstrap replications (chapter II).

Table 5.3: Key to figures 5.3 and 5.4, showing title, organism and accession number of protein sequences.

Gene	Title	Organism	NCBI Accession
dJ222E13.C22.1		*H. sapiens*	
*M.mus*_NP_075964.1	serine hydrolase protein	*M. musculus*	gi|13443008|ref|NP_075964.1|
*D.mel*_CAA04153.1	kraken	*D. melanogaster*	gi|2274926|emb|CAA04153.1|
*P.aer*_NP_250313.1	probable hydrolase	*P. aeruginosa*	gi|15596819|ref|NP_250313.1|
*D.mel*_CG15879	CG15879	*D. melanogaster*	gi|7292201|gb|AAF47611.1|
*D.mel*_CG7632	CG7632	*D. melanogaster*	gi|7296419|gb|AAF51706.1|

Similar analyses were carried out on all proteins (appendix 8). Interestingly, in two cases, pairs of proteins from 22q13.31 were shown to be similar to each other and were therefore included in the same phylogenetic trees (figure 5.5). dJ671O14.C22.2 and bK414D7.C22.1 share 42.8% identity at the protein level. Olski *et al.* (2001) has named these genes β- and γ-parvin, part of the parvin subfamily, but these members have not previously been functionally characterised. Similarly, dJ549K18.C22.1 and the partial gene dJ388M5.C22.4 were shown to share 38.3% identity at the protein level. Phylogenetic analysis of these four proteins showed that each 22q13.31 protein clustered with its potential mouse orthologue. More distantly related proteins from *C.elegans* and *D.melanogaster* were not shown to cluster in this way.

Several trees also identified segregated orthologous groups, clearly distinguishing the 22q13.31 protein from similar paralogous human genes: DIA1, cB33B7.C22.1, ARFGAP1, PACSIN2, TTLL1, C22orf1, SULTX3, ARHGAP8 and bK268H5.C22.4. The separate groups may serve distinct cellular functions in the human body. Other trees highlighted groups of similar proteins whose sequences were highly conserved across different species: BIK, bK1191B2.C22.3, BZRP, dJ526I14.C22.2, CGI-51, NUP50, bK268H5.C22.1, UPK3 and E46L. Interestingly, bK1191B2.C22.3 demonstrated extensive potential orthology with both eukaryotes and prokaryotes, suggesting that the gene encodes an essential protein conserved throughout evolution.

Overall, these results showed that paralogous sequences identified within each organism are generally more different from each other than they are from their orthologues in other species. This suggests that the paralogues have differing functions within a species, which may be conserved in orthologous proteins in other species.

Over 20 potential orthologues had undergone some functional characterisation, which could be potentially transferred to 14 genes from 22q13.31. These results are shown in table 5.4. In two cases (dJ222E13.C22.1 and bK1191B2.C22.3), identification of these orthologues provides the first preliminary functional characterisation of these novel genes from 22q13.31. In other cases, these results confirm, update and extend previous phylogenetic analyses of these protein groups. The domain and secondary structures of the potential orthologous proteins were reanalysed using the InterPro database and PIX analysis programs, to allow comparison between the human protein and its putative functionally characterised orthologue. An overview of these results is shown in figure 5.6.

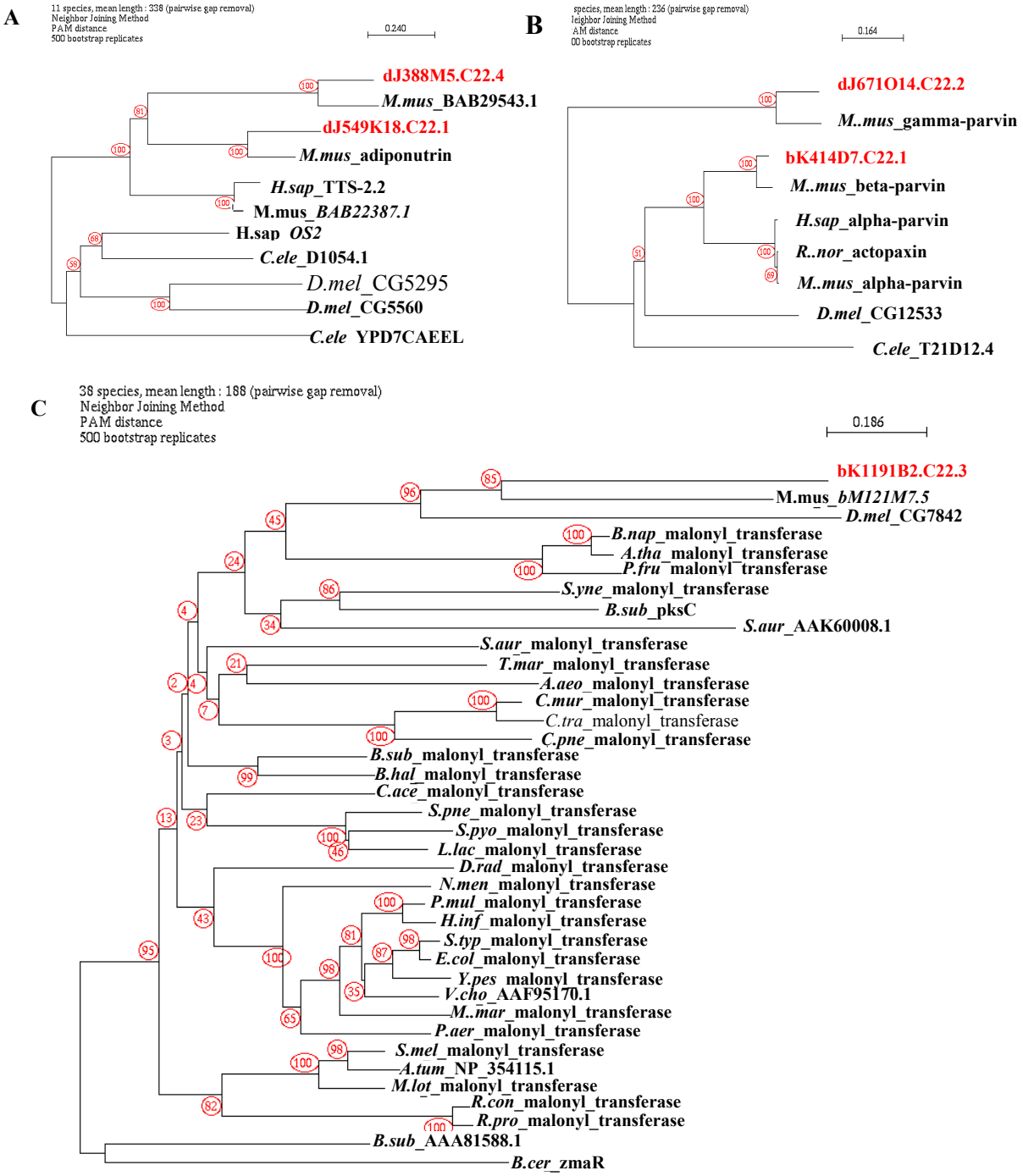


Figure 5.5: Phylogenetic trees derived using NJ methodology from clustalw protein alignments (Phylowin, Galtier *et al.* 1996). Proteins from 22q13.31 are highlighted in red. Other protein identifiers are listed in appendix 8.

A: Potential phylogenetic relationship between dJ388M5.C22.4 and dJ549K198.C22.1

B: Potential phylogenetic relationship between dJ671O14.C22.2 and bK414D7.C22.1

C: Phylogenetic tree showing relationship of bK1191B2.C22.3 to >30 potential orthologues

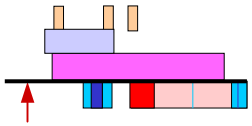
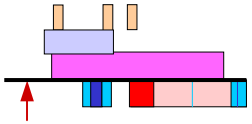
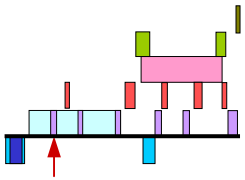
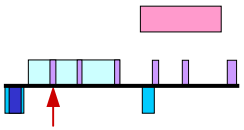
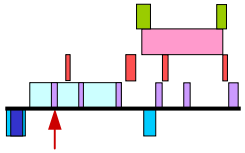
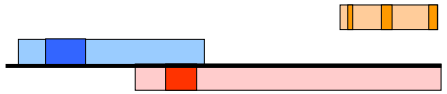
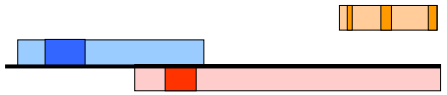
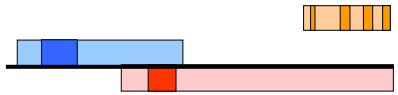
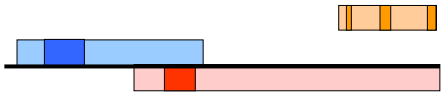
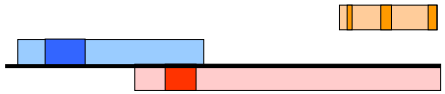
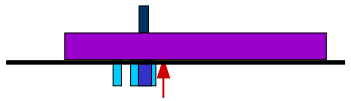
Table 5.4: Potential orthologues of proteins from 22q13.31 identified by phylogenetic analysis. Sequence identifiers are provided in appendix 8.

Gene	Functionally characterised putative orthologue		Function	Author
	Accession	Organism		
dJ222E13.C22.1	Em:CAC20674	*M. mus*	Immunolocalises to perinuclear vesicles; induced by passive stretch of skeletal muscle *in vivo*	Sadusky *et al.*, 2001
DIA1	SW:P07514	*B. tau*	Sequence analysis and functional assays to test catalytic activity are consistent with the function of human protein.	Ozols *et al.*, 1984; Strittmatter *et al.*, 1992; Tamura *et al.*, 1987
	SW:P20070	*R. nor*	Sequence analysis is consistent with the function of human protein.	Murakami *et al.*, 1989; Pietrini *et al.*, 1988; Zenno *et al.*, 1990
cB33B7.C22.1	Gb:AAF82758	*R. nor*	Gb3 synthase activity consistent with the human protein.	Keusch *et al.*, 2000
bK1191B2.C22.3	>30	>30	Essential enzyme in the biosynthesis of fatty acids. Catalyses the transacylation of malonate from malonyl-CoA to activated holo-ACP, to generate malonyl-ACP, an elongation substrate in fatty acid biosynthesis.	
PACSIN2	Gb:AAD41780	*M. mus*	Protein localised to cytoplasm.	Ritter *et al.*, 1999
	Gb:AAF22213	*R. nor*	Colocalises with proteins involved in endocytosis and actin dynamics.	Qualmann & Kelly, 2000
	Em:CAA90678	*G. gal*	Localises to focal adhesion sites.	Merilainen *et al.*, 1997
	Em:CAC17814	*X. lae*	Localised to cytoplasm and membrane ruffles. Colocalises with ADAM13 in migrating neural crest cells during embryonic development. Binds ADAM13 *in vitro* and rescues developmental alterations induced by over expression of ADAM13.	Cousin *et al.*, 2000
BZRP	SW:P50637	*M. mus*	Manifestation of peripheral-type benzodiazepine binding sites. Possible role in porphyrin transport.	Taketani *et al.*, 1994
	SW:P16257	*R. nor*	Manifestation of peripheral-type benzodiazepine binding sites. Contains benzodiazepine and isoquinoline carboxamide binding domains.	Casalotti *et al.*, 1992; Sprengel *et al.*, 1989
	SW:P30535	*B. tau*	Manifestation of peripheral-type benzodiazepine binding sites.	Parola *et al*., 1991
dJ549K18.C22.1	Gb:AAK68636	*M. mus*	mRNA restricted to adipose tissues. mRNA levels fall under fasting conditions, but increase under high carbohydrate diet. Protein localises to membranes, absent from the cytosol.	Baulande *et al.*, 2001
bK414D7.C22.1	Gb:AAG27172	*M. mus*	Member of parvin family. Other members may be involved in cell matrix adhesion	Olski *et al.*, 2001
dJ671O14.C22.2	Gb:AAG29542	*M. mus*	See bK414D7.C22.1	
NUP50	Gb:AAF70057	*M. mus*	Cyclin E-mediated elimination of p27.	Muller *et al.*, 2000
UPK3	SW:P38574	*B. tau*	Sequence analysis consistent with the function of the human protein.	Wu & Sun, 1993
FBLN1	SW:Q08879	*M. mus*	Sequence analysis consistent with the function of the human protein. Calcium-dependent binding to basement ligands (functional assay.	Pan *et al*., 1993

Figure 5.6 shows that putative functional domains are generally conserved between the 22q13.31 proteins and their functionally characterised putative orthologues. Exceptions were observed in the bovine version of DIA1, which appears to lack a prenyl group-binding site (IPR001230) and prokaryotic versions of bK1191B2.C22.3, which lack a sequence feature conserved between ABC transporter proteins (IPR001617). Additionally, small differences are seen in the number of SH3 domains present in orthologues of PACSIN2.

Discrepancies are also seen in the results of similar functional assays previously carried out on the orthologous proteins (table 5.4). For example, the subcellular localisation of the PACSIN2 chicken orthologue FAP52 to focal adhesion sites (Merilainen *et al.*, 1997), has not been reported in similar experiments involving the mouse, rat and *Xenopus* orthologues (Cousin *et al.*, 2000; Qualmann & Kelly, 2000; Ritter *et al.*, 1999).

Figure 5.6: The predicted domain and secondary structures of both proteins from 22q13.31 and functionally characterised potential orthologues.

InterPro accession	H. Sapiens *protein*	Orthologues	
IPR000073 IPR000379 IPR003089	dJ222E13.C22.1 	*M. mus*. Em:CAC20674 	
IPR001230 IPR001221 IPR001433 IPR001709 IPR001834	DIA1 	*B. tau*. SW:P07514 	*R. nor*. SW:P20070 
	cB33B7.C22.1 	*R nor*. Gb:AAF82758 	
IPR001452 IPR001060	PACSIN2 	*M. mus*. Gb:AAD41780  *G. gal*. Em:CAA90678 	*R. nor*. Gb:AAF22213  *X. lae*. Em:CAC17814 
IPR001617 IPR001227	bK1191B2.C22.3 	Over 30 orthologues in different species. Each contains IPR001227. 	

	BZRP	*M. mus*. SW:P50637	*R. nor*. SW:P16257	*B. tau*. SW:P30535
	dJ549K18.C22.1	*M. mus* Gb:AAK68636		
IPR001715	bK414D7.C22.1	*M. mus*. Gb:AAG27172		
IPR001715	dJ671O14.C22.2	*M. mus*. Gb:AAG29542		
	NUP50	*M. mus*. Gb:AAF70057		
	UPK3	*B. tau*. SW:P38574		
IPR000152 IPR000561 IPR000020 IPR001881	FBLN1	*M. mus*. SW:Q08879		

Potentially, the functional evidence derived from orthologous proteins could be transferred to the human versions. However, this approach must be tentative for several reasons. The techniques described here identify only putative orthologues – confirmation requires completion and accurate annotation of the model organism genomes. Even then, complications arising from gene duplication and other evolutionary mechanisms mean that, for many genes, simple orthologous relationships cannot be discerned (Lander *et al.*, 2001). In addition, as shown above, differences can exist between a protein and its putative orthologue, which may or may not affect function. Potential functional characteristics transferred between orthologous proteins must therefore be experimentally verified. Nevertheless, this study of putative orthology provides a starting point for future investigation of the functional characteristics of the proteins encoded within 22q13.31.

5.3.4 *In silico* prediction of subcellular localisation

The subcellular localisation of a protein can have a large affect on function (section 5.1). It was therefore decided to experimentally determine the subcellular localisation of a subset of the proteins encoded within 22q13.31 (section 5.4). An additional *in silico* investigation was undertaken (see below), in order to compare the results to those generated from the experimental system.

5.3.4.1 PSORT prediction of protein subcellular localisation

The program PSORT (Nakai & Horton, 1999) was used to detect sorting signals in the 42 peptide sequences (including known alternative splice forms) and predict their subcellular localisation. These results are shown in figure 5.7. The horizontal bars depict the probability of protein localisation at a particular location: the longest bar shows the most likely subcellular localisation according to the PSORT algorithm. Protein localisations that generated a probability value of less than 0.12 were classed together as ‘Other’.

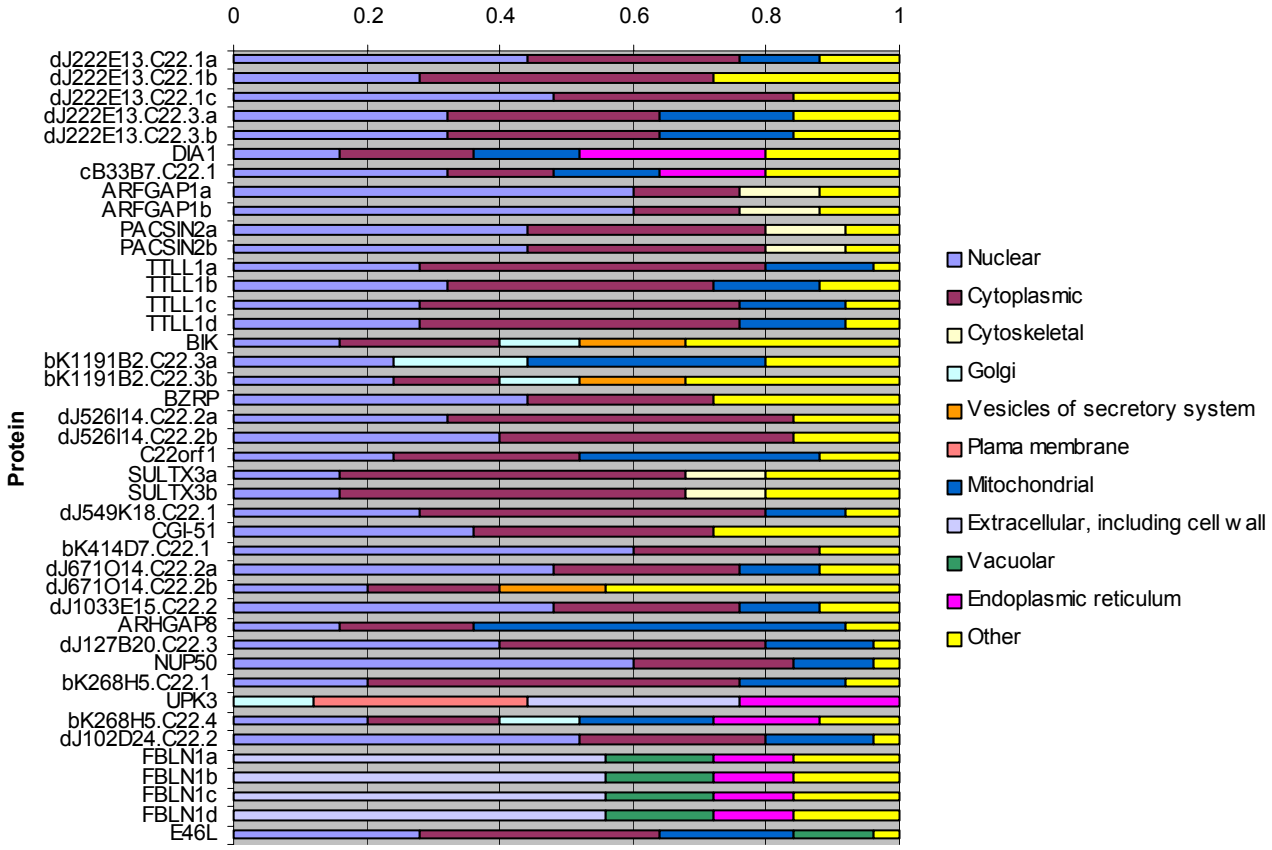


Figure 5.7: Predicted subcellular localisation (PSORT, Nakai & Horton, 1999). The length of each coloured horizontal bar depicts the probability of localisation at a particular cellular location.

The most common predicted subcellular locations for this group of 42 proteins is the nucleus (33%), the cytoplasm (30%) or both (9%). One protein (bK268H5.C22.4) was predicted as equally likely to be found in the nucleus, cytoplasm or mitochondria. A further three proteins, bK1191B2.C22.3a, C22orf1 and ARHGAP8, were predicted to contain mitochondrial localisation signals, whilst DIA1 was predicted to be localised to the endoplasmic reticulum (ER). FBLN1 and UPK3 localisation was predicted to be the extracellular matrix and, in the case of UPK3, also in the plasma membrane.

The subcellular localisations of seven of these proteins are already known from experimental data in the literature. The PSORT predictions were compared with these experimental derived localisations. The four isoforms of FBLN1 were correctly predicted as secreted into the

extracellular matrix (Argraves *et al.*, 1990). DIA1 was correctly predicted to be localised to the endoplasmic reticulum but is also found in mitochondrial and other membranes, as well as existing in a soluble form in erythrocytes (reviewed in OMIM Accession: 250800). PSORT predicted BZRP to be localised in the nucleus, whereas it has previously been shown to be an integral membrane protein in the mitochondria (Hirsch *et al.*, 1998; Mukherjee & Das, 1989), although other localisation results have been noted for this protein (Olson *et al.*, 1988). BIK was predicted as being localised in the cytoplasm, but has previously been placed around the nuclear envelope and cytoplasmic membranes (Han *et al.*, 1996).

PSORT therefore demonstrated an accuracy of 71% in these seven cases. However, four of these peptides are isoforms of the same gene, FBLN1. If these are excluded, the success rate falls to 50%, highlighting the necessity for experimental verification of predicted protein characteristics.

5.4 Experimental analysis of subcellular localisation

5.4.1 Overall strategy

The approach used included the cloning of full-length cDNAs, generated by RT-PCR, derived from the genes encoded within 22q13.31. The generated clone inserts were sequenced in order to identify possible PCR errors or SNPs. The cloned ORFs of these genes provide a valuable resource for all future work on the proteins of this region. Initial experiments of protein subcellular localisation are described here, but these clones are available for research on all aspects of protein function.

For the experimental investigation of subcellular localisation, it was intended to individually tag the N- and C- termini of the encoded protein with a T7 amino acid tag (T7.Tag), to which monoclonal antibodies are commercially available. The T7.Tag encodes the peptide sequence

Met-Ala-Ser-Met-Thr-Gly-Gly-Gln-Gln-Met-Gly and is the natural amino terminal end of the T7 major capsid protein. Since the T7.Tag mouse monoclonal antibody used reacts specifically with this peptide sequence, it can be used as an epitope tag to follow target proteins by sensitive immunological procedures (Lutz-Freyermuth *et al.*, 1990; Tsai *et al.*, 1992).

Dr. B. Aguado (HGMP Resource Centre, Cambridge) kindly provided vectors suitable for the C-terminus tagging process. A further novel vector was created containing the T7.Tag sequence in a context suitable for N-terminal tagging (chapter II and figures 5.11 and 5.12). Tagged protein expression constructs were individually transfected into COS-7 cells (SV40 transformed African Green monkey kidney cell line). The cells were used in immunofluorescence experiments to determine subcellular localisation and the cell protein extracts were used in Western blot experiments to confirm the size of the expressed protein product.

5.4.2 Selection and generation of full-length cDNA clones

At the time of investigation, 23 of the 27 full protein-coding genes analysed in this thesis had annotated 5' and 3' UTRs enclosing an ORF. Nested PCR (see chapter II) was used to amplify the ORF from the start to the stop codon. Seventeen different PCR products, representing 13 genes and splice variants, were successfully generated from 13 of the 23 nested primer pairs. These were cloned into a 'holding' vector, pGEMEasyT (Promega), to provide a resource both for this project and for future research. The clone inserts were then sequenced (E. Huckle) and compared to the genomic human DNA. These results are summarised in table 5.5.

Attempts to generate full-length cDNA sequences from dJ222E13.C22.3, DIA1, ARFGAP1, ARHGAP8, NUP50, bK268H5.C22.1, UPK3, bK268H5.C22.4, FBLN1 and E46L by nested PCR failed. This was probably due to the large size of the ORFs involved (up to 2.1 kb in the

case of FBLN1) and the difficulty of designing primers in the GC-rich DNA frequently found at the 5' end of the gene.

Table 5.5: cDNAs from 22q13.31 were generated by nested PCR, cloned and sequenced.

Locus	Isoform amplified	RNA source	Accession no.	ORF	Remark
dJ222E13.C22.1	dJ222E13.C22.1c	Testis	AL590120	203	Novel isoform
	dJ222E13.C22.1e	Kidney	**AL590118**	250	Novel isoform
cB33B7.C22.1	cB33B7.C22.1	F. liver	AB037883	353	Possible SNPs
PACSIN2	PACSIN2a	F. brain	AF128536	486	
TTLL1	TTLL1a	Lung	AL096886	423	Possible SNP
	TTLL1c	Kidney	**AL0589867**	394	Novel isoform
BIK	BIK	F. liver	X89986, U34584	160	
bK1191B2.C22.3	bK1191B2.C22.3a	Kidney	AL359401	390	
	bK1191B2.C22.3b	Kidney	AL359403	180	
BZRP	BZRP	Kidney	M36035	169	
dJ526I14.C22.2	dJ526I14.C22.2b	Kidney	AL590888	210	Novel isoform
SULTX3	SULTX3a	F. brain	**AL590119**	284	
	SULTX3b	Testis	**AL590119**	260	Novel isoform
dJ549K18.C22.1	dJ549K18.C22.1	F. brain	**AK025665**	481	Possible SNPs
CGI-51	dJ796I17.C22.2	Kidney	AF151809	469	Possible SNPs
dJ671O14.C22.2	dJ671O14.C22.2b	Testis	AL590887	273	Novel isoform
dJ102D24.C22.2	dJ102D24.C22.2	Testis	AL442116	309	

Novel isoforms were submitted to EMBL. These, and other cDNAs previously identified in this study are shown in bold.

5.4.2.1 SNP analysis

Sequence reads from the cDNA clone inserts were imported into ACeDB and the aligned sequences were examined using blixem (Sonnhammer & Durbin, 1994). The quality of the reads was examined using trev (Staden, unpublished) in order to identify discrepancies between the cDNA sequence and genomic sequences. Differences were found to be restricted to five clones. Available cDNA and EST sequences were also examined at these positions using blixem (Sonnhammer and Durbin, 1994) to determine if the discrepancies also existed in other expressed sequence evidence.

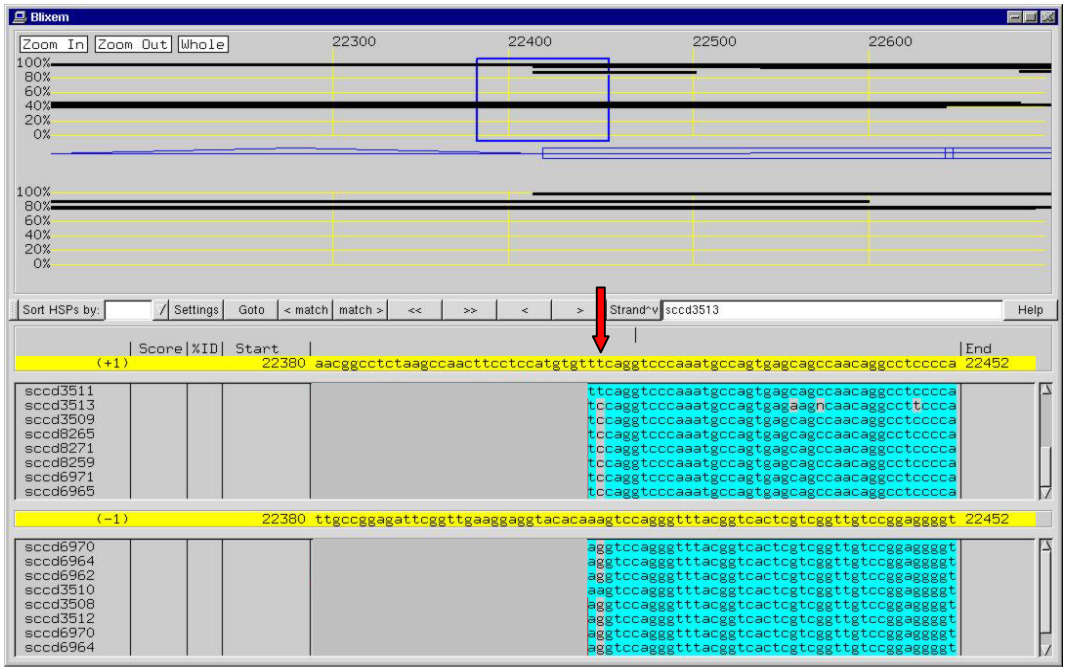


Figure 5.8: Blixem alignment of dJ549K18.C22.1 cDNA clone sequencing reads. The arrow indicates a discrepancy between the cDNA and genomic sequence.

Both transitions (pyrimidine to pyrimidine or purine to purine substitutions) and transversions (purine to pyrimidine or pyrimidine to purine substitutions) were noted. Twelve variations were identified in total, several of which altered the amino acid code. These variations are listed in table 5.6.

Table 5.6: Discrepancies discovered between cDNA clone and genomic sequences.

Cloned gene	DNA change	Type	Amino acid change	cDNA/EST evidence
dJ549K18.C22.1	**A**AG-**G**AG	Transition substitution	Lys-Glu	AAG & GAG
TTLL1a	AT**C**-AT**T**	Transition substitution	-	ATC only
dJ671O14.C22.2a	C**T**C-C**C**C	Transition substitution	Leu – Pro	CTC & CCC
	CC**C**-CC**G**	Transversion substitution	-	CCC only
cB33B7.C22.1	CCC G-ACC TCC CCA	Multiple alterations	Disrupts original ORF	CCC G only
CGI-51	GC**G**-GC**T**	Transversion substitution	-	GCG only
	GG**A**-GG**G**	Transition substitution	-	GGA & GGG
	A**G**T-A**A**T	Transition substitution	Ser-Asn	AGT only
	TTC-**A**TC	Transversion substitution	Phe-Ile	TTC only
	C**G**G-C**A**G	Transition substitution	Arg-Gln	GGG only
	C**C**C-C**T**C	Transition substitution	Pro-Leu	CCC only
	TT**A**-TT**G**	Transition substitution	-	TTA only

To determine whether these changes were the results of genomic polymorphisms, or instead the results of PCR errors, PCR primers were designed and used to amplify fragments containing the candidate variations from the DNA of 24 different individuals (set M24PDR of 24 human DNAs, Coriell cell repository). Samples from each product were electrophoresed and visualised to confirm amplification. The remainder were purified (chapter II) and sequenced (E. Huckle).

The trace files were imported into a Gap4 sequence-editing database (generated by K. Rice) (Bonfield *et al.*, 1995). Sequences flanking the cDNA discrepancies were highlighted for ease of analysis. All differences between the clone and genomic sequences were examined. Discrepancies are shown as dashes in the consensus sequence at the bottom of the Gap4 graphical user interface (figure 5.9). The traces were inspected at the positions of the discrepancies (figure 5.10).

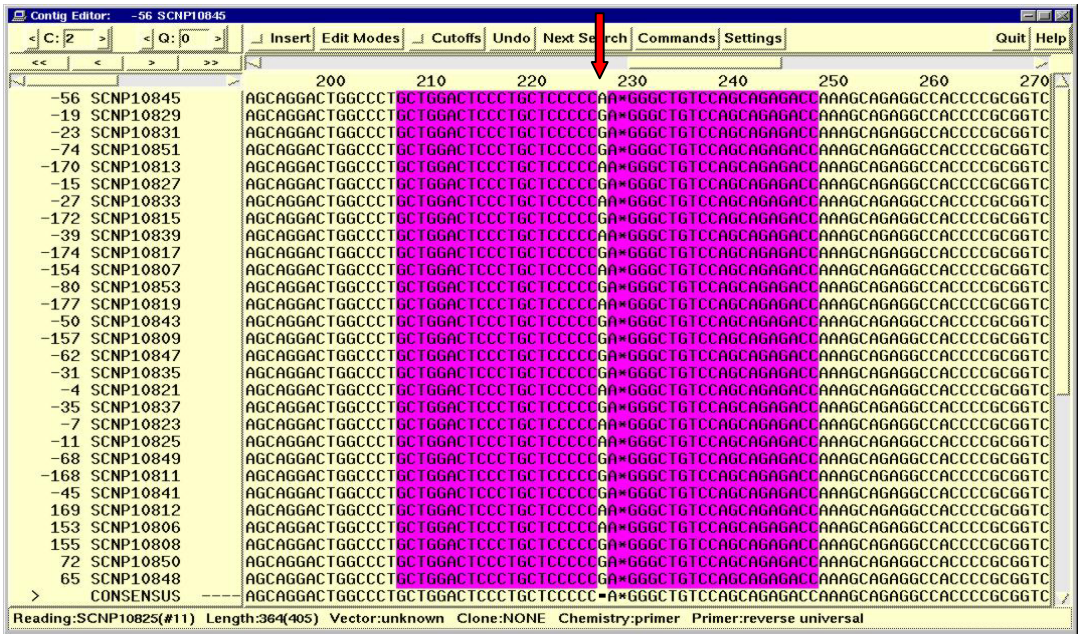


Figure 5.9: Visual display from Gap4 database. The red arrow indicates a potential SNP within the cDNA sequence of dJ549K18.C22.1.

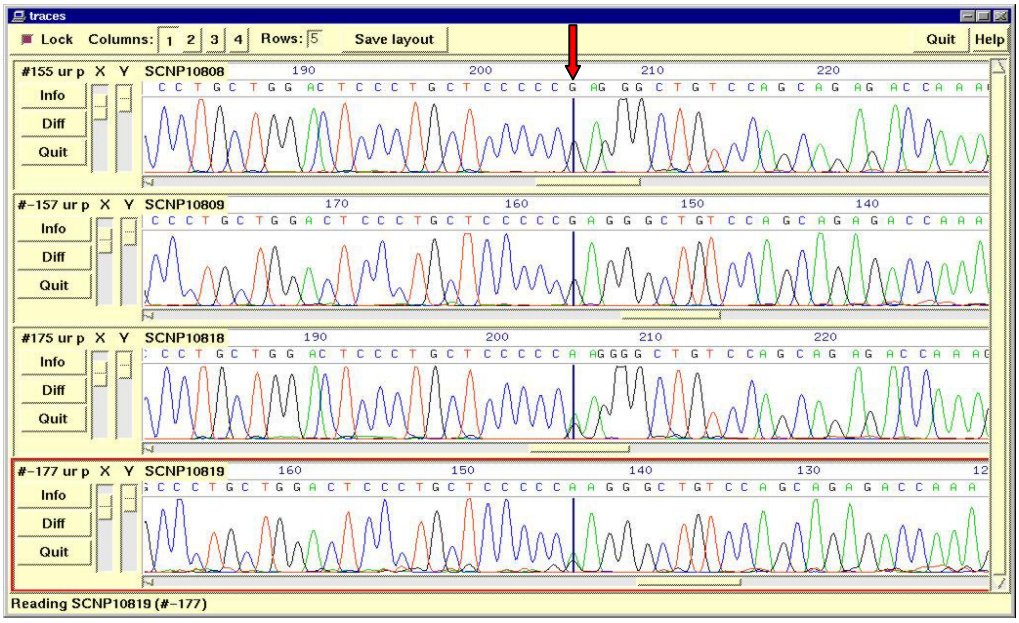


Figure 5.10: Inspection of the forward and reverse traces from two (of 24) individuals. The red arrow indicates a candidate variation in the cDNA sequence from dJ549K18.C22.1.

5.4.2.2 Genomic variation

Of the 12 candidate variations, only one (dJ549K18.C22.1) was supported by genomic evidence from the twenty individuals and confirmed as a SNP. Additionally, one non-coding variation (C-T) was identified within an intron of CGI-51.

This clone was therefore included in further studies, together with those containing discrepancies that not alter the amino acid sequence (TTLL1 and dJ671O14.C22.2a), or that were evident in independent cDNA or EST evidence (dJ671O14.C22.2a).

New clones were generated by nested PCR (see chapter II) to represent the only two genes that would otherwise have been excluded from these investigations (cB33B7.C22.1 and CGI-51), due to discrepancies not supported by other evidence that resulted in a changed amino acid sequence. The annealing temperature used was increased by 2^{o}C in order to enhance specificity of primer-template binding. The new clones did not contain any discrepancies between the cDNA insert and genomic sequence. The original discrepancies may therefore

have been generated by PCR errors, or amplification from paralogous sequences, which was not repeated in the second attempt to amplify these cDNAs.

5.4.3 Addition of T7.Tag

A schematic showing the strategy used to incorporate the T7.Tag at the N- and C- termini of each ORF is shown in figure 5.12.

5.4.3.1 C-terminal T7.Tag

PCR primers were designed to amplify the cDNA from the start ATG to the stop codon from the holding vector. The amplified cDNA was subcloned into pBlue-CT7 (a kind gift from B. Aguado), removing the stop codon and incorporating the T7.Tag, in-frame, at the C-terminus. The construct was then subcloned into the mammalian expression vector pCDNA3 (Invitrogen) (chapter II) and sequenced to ensure that PCR errors had not been introduced and the T7.Tag was correctly positioned in-frame. Ninety-four percent of the experiments to tag the ORFs at the C-terminal end and insert into an expression vector were successful (table 5.7). The experiment to clone the tagged dJ671O14.C22.2b construct failed despite repeated attempts to transfer the insert into the expression vector. Later sequencing of the tagged construct showed that the 5' restriction site was corrupted. This may have been caused by an error in primer design or generation.

5.4.3.2 Modification of pcDNA3 expression vector to include N-terminal T7.Tag

In order to eliminate the possibility of deriving spurious results from steric interference of the C-terminal T7.tag or masking of internal protein localisation signals, it was desirable to position a T7.Tag at the N-terminal of the proteins. To avoid repeated digestion and ligation steps (see above) it was decided to modify the pCDNA3 expression vector to include the T7.Tag in an appropriate context so that the cDNA of interest could be inserted by just one round of digestion and ligation.

The plasmid pcDNA3-NT7 was designed to include an additional unique restriction enzyme site (*Not*I*)* at the end of the T7.Tag to allow in-frame insertion of the cDNA of choice. The creation of this restriction site was necessary to produce a wide enough choice of restriction enzyme sites for later cDNA insertion; none of the genes of interest contained an internal *Not*I site. The resulting vector pCDNA3-NT7 also contained the T7.Tag in a modified context, including the incorporation of a strong Kozak consensus sequence. A vector diagram is shown in figure 5.11.

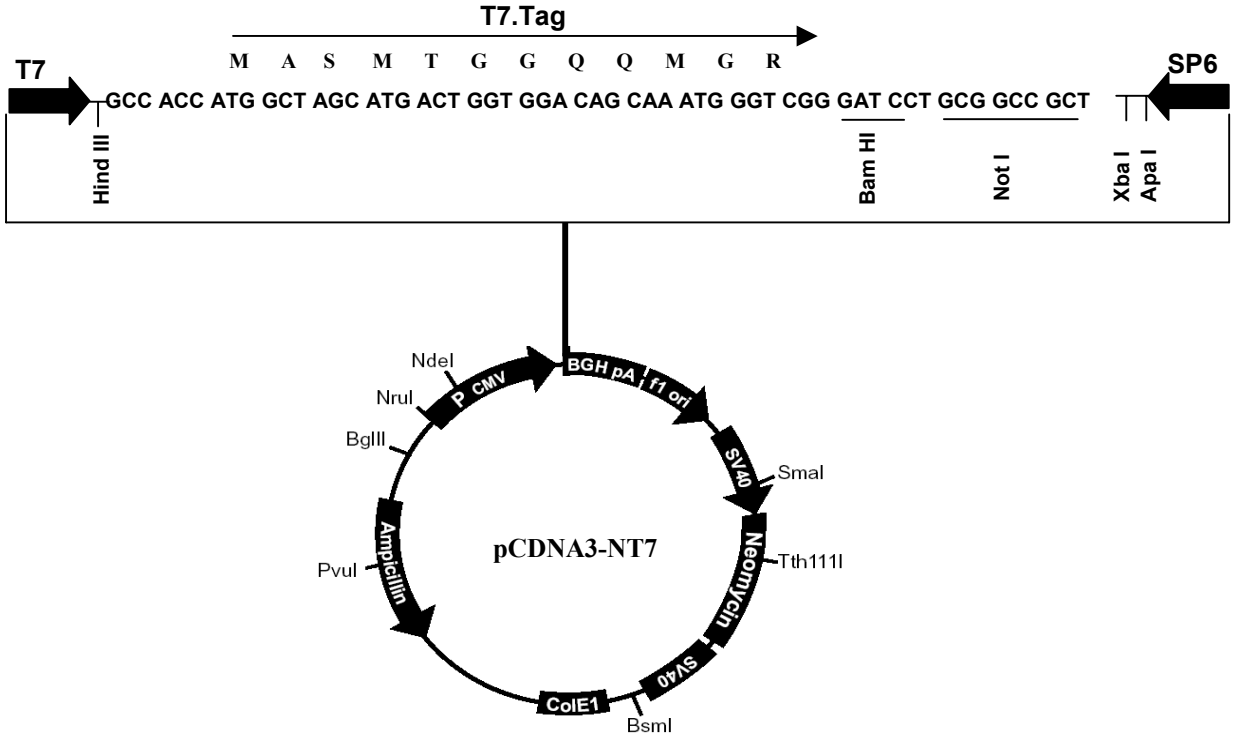


Figure 5.11: Schematic of the mammalian cell expression vector pCDNA3-T7-C. The PCDNA3 vector (Invitrogen) polylinker site was modified as described in the text.

5.4.3.3 N-terminal T7.Tag

Appropriate primers were designed to amplify the ORF of each cDNA from the start to the stop codon, incorporating suitable restriction sites. The amplified cDNAs were incorporated into the pcDNA3-NT7 vector via one round of restriction enzyme digestion and ligation. The constructs were then sequenced to confirm that the expression vector inserts were correct.

These results are summarised in table 5.7. Again, dJ671O14.C22.2b failed attempts to clone it into the expression vector, due to corruption of the 5' restriction site.

Table 5.7: Outcome of restriction, ligation and transformation reactions to generate N- and C-terminally T7 tagged cDNA inserts.

Gene	N-terminal T7.Tag construct Successfully generated?	C-terminal T7.Tag construct Successfully generated?
dJ222E13.C22.1a	Y	Y
dJ222E13.C22.1 b	Y	Y
cB33B7.C22.1	Y	Y
PACSIN2a	Y	Y
TTLL1a	Y	Y
TTLL1c	Y	Y
BIK	Y	Y
bK1191B2.C22.3a	Y	Y
bK1191B2.C22.3b	Y	Y
BZRP	Y	Y
dJ526I14.C22.2b	Y	Y
SULTX3a	Y	Y
SULTX3b	Y	Y
dJ549K18.C22.1	Y	Y
dJ796I17.C22.2	Y	Y
dJ671O14.C22.2b	Failed ligation reaction	Failed ligation reaction
dJ102D24.C22.2	Y	Y

5.4.4 Expression in COS-7 cells

To confirm expression and elucidate the sizes of the protein products, COS-7 cells were transiently transfected and the proteins analysed by SDS-PAGE, three days post-transfection. Figure 5.13 shows the results of western blot analysis of the protein constructs and table 5.8 summarises the expected and obtained protein sizes.

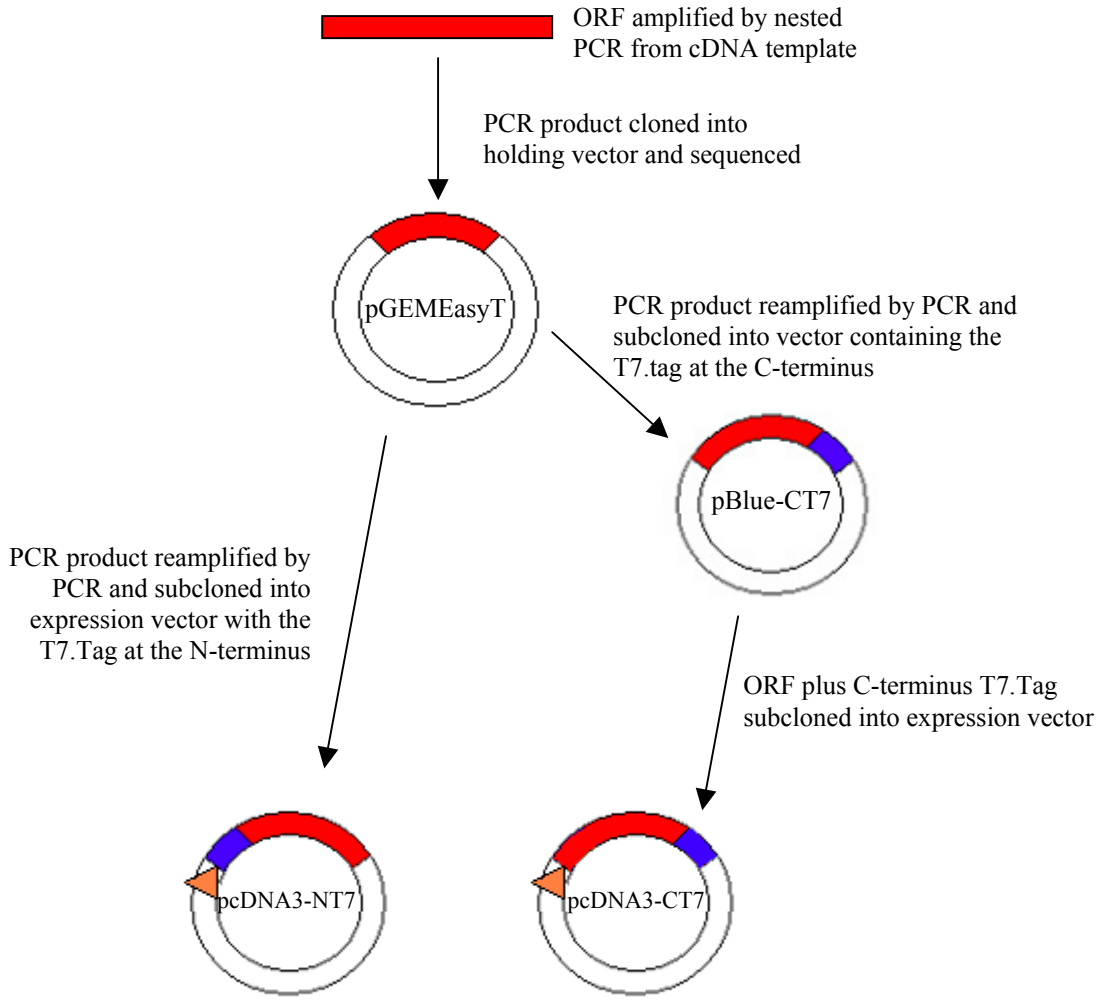


Figure 5.12: Schematic showing strategy used to generate N- and C- terminally T7-tagged clones. The ORF of the gene under investigation (shown in red) was amplified from a cDNA template by nested PCR. The PCR product was then cloned into a holding vector, pGEMEasyT (Promega) and the insert sequenced (E. Huckle).

For C-terminal tagging, the clone insert was reamplified by PCR using primers that removed the stop codon and incorporated specific restriction enzyme sites flanking the ORF. The PCR product was then digested and subcloned into the vector pBlue-CT7 (a kind gift from B. Aguado), thus incorporating the C-terminal T7.Tag (shown in blue), in-frame with the gene ORF. The ORF plus T7.Tag were then subcloned into the expression vector pCDNA3, containing a promoter sequence (yellow) (Invitrogen) and sequenced.

For N-terminal tagging, the holding clone insert was reamplified by PCR using primers that incorporated specific restriction enzyme sites flanking the ORF. The PCR product was then digested with appropriate enzymes and subcloned into the vector pcDNA3-NT7 (figure 5.11). The clone insert was then sequenced (E. Huckle).

See chapter II for more details.

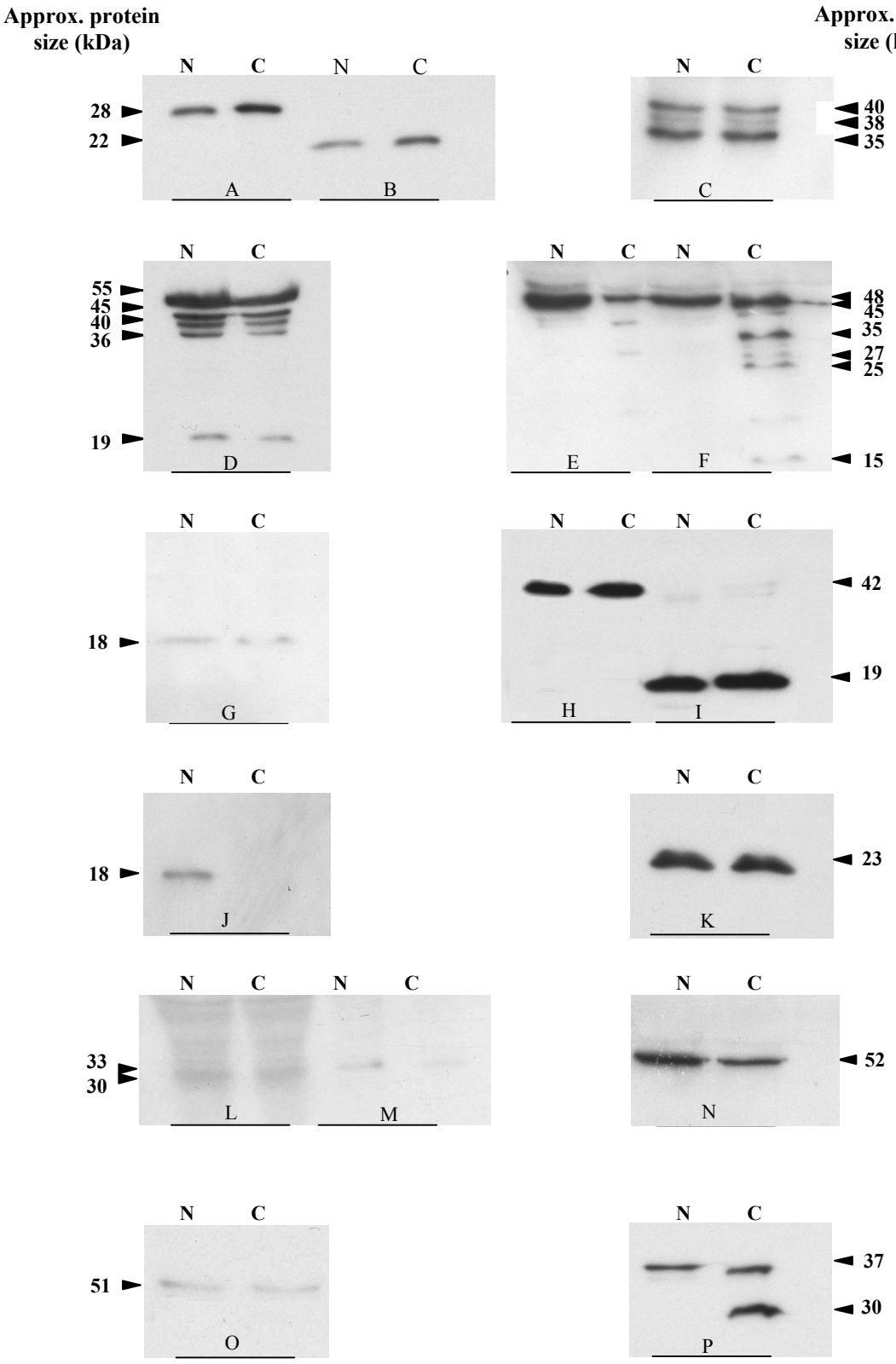


Figure 5.13: Western blot analysis of transiently transfected COS-7 cells. N- and C- terminally tagged constructs are shown. A) dJ222E13.C22.1a; B) dJ222E13.C22.1b; C) cB33B7.C22.1; D) PACSIN2a; E) TTLL1a; F)TTLL1c; G) BIK; H) bK1191B2.C22.3a; I) bK1191B2.C22.3b; J) BZRP; K dJ526I14.C22.2b; L) SULTX3a; M) SULTX3b ; N) dJ549K18.C22.1; O) CGI-51; P) dJ102D24.C22.2

Table 5.8: Expected and obtained protein sizes, estimated from SDS-PAGE. Obtained sizes that are equivalent to expected, according to the limit of gel resolution, are highlighted in blue.

	Protein	Expected size (kDa)	Obtained size (kDa)	
			N-terminal T7.Tag	C-terminal T7.Tag
A	dJ222E13.C22.1.a	28.3	28	28
B	dJ222E13.C22.1.b	22.5	22	22
C	cB33B7.C22.1	40.5	40	40, 38, 35
D	PACSIN2a	55.6	55	55, 45, 40, 36, 19
E	TTLL1a	48.9	48, 45, 27	48, 45, 27
F	TTLL1c	45.4	45, 36, 27, 25, 15	45, 36, 27, 25, 15
G	BIK	18.0	18	18
H	bK1191B2.C22.3a	42.9	42	42
I	bK1191B2.C22.3b	19.1	42, 19	42, 19
J	BZRP	18.8	18	-
K	dJ526I14.C22.2b	23.6	23	23
L	SULTX3a	33.0	33	33
M	SULTX3b	30.2	30	30
N	dJ549K18.C22.1	52.8	55	55
O	CGI-51	51.9	51	51
P	dJ102D24.C22.2	37.0	37	37, 30

Proteins of expected size were expressed from both the C- and N-terminally tagged constructs. No overall difference in expression levels was noted between the two construct types. However, no bands were observed from the western blot experiment using the C-terminal construct of BZRP (Figure 5.13.J). Repeated transfections using fresh DNA preparations also failed. An attempt to resequence the insert sequence also failed, so it may be that the insert was lost after plasmid construction.

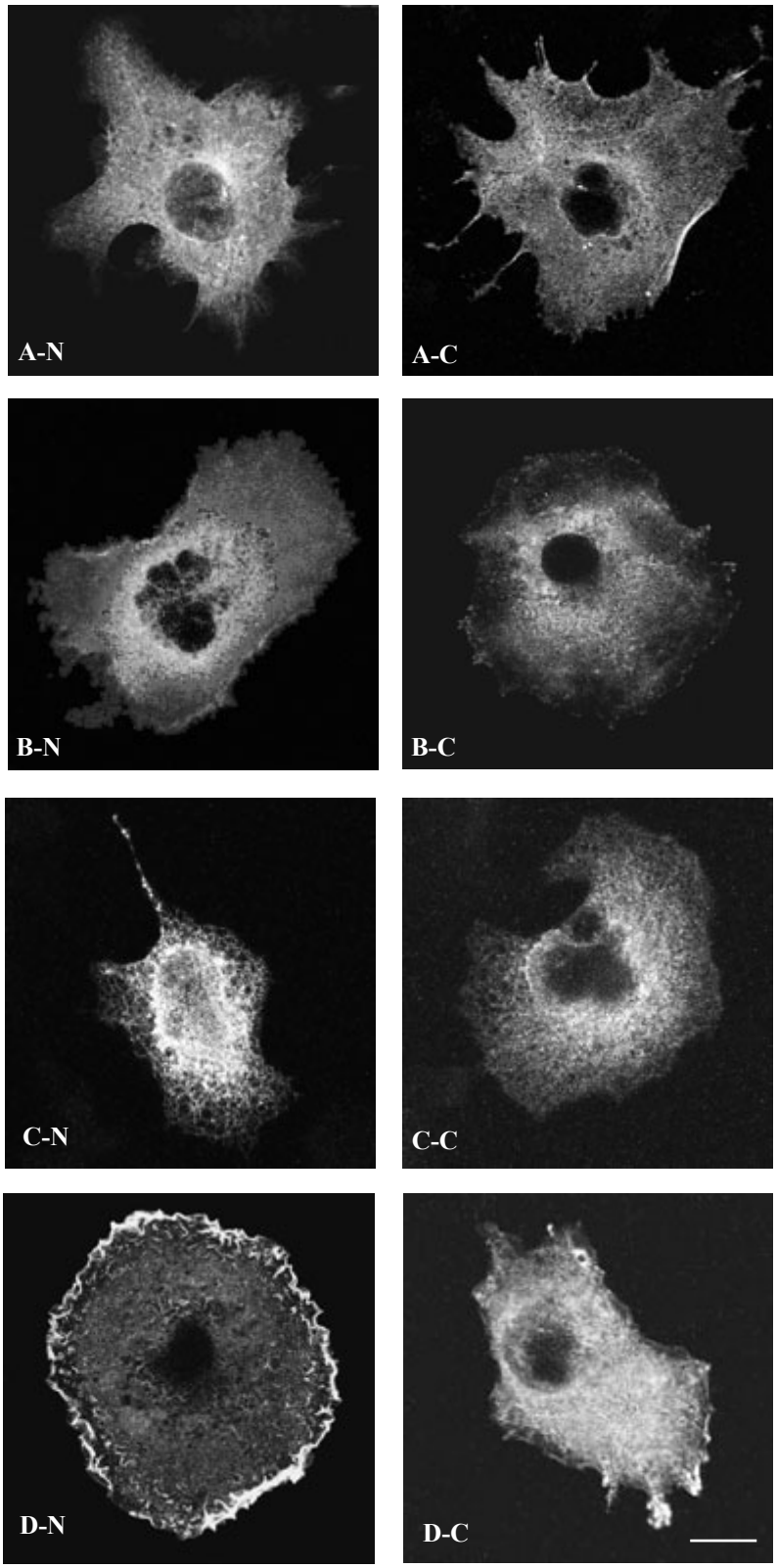
The presence of extra bands, smaller than the expected size of the protein construct, was noted in several cases (figure 5.13.C, D, E, F and P). These bands may be caused by partially degraded copies of the protein construct, or could be the result of post-translational modifications. Interestingly in bK1191B2.C22.3b (figure 5.13.I), faint bands of approximately twice the expected size of both N- and C-terminal constructs were observed. These bands were also noted in two repeat transfections (data not shown). These may indicate dimerisation,

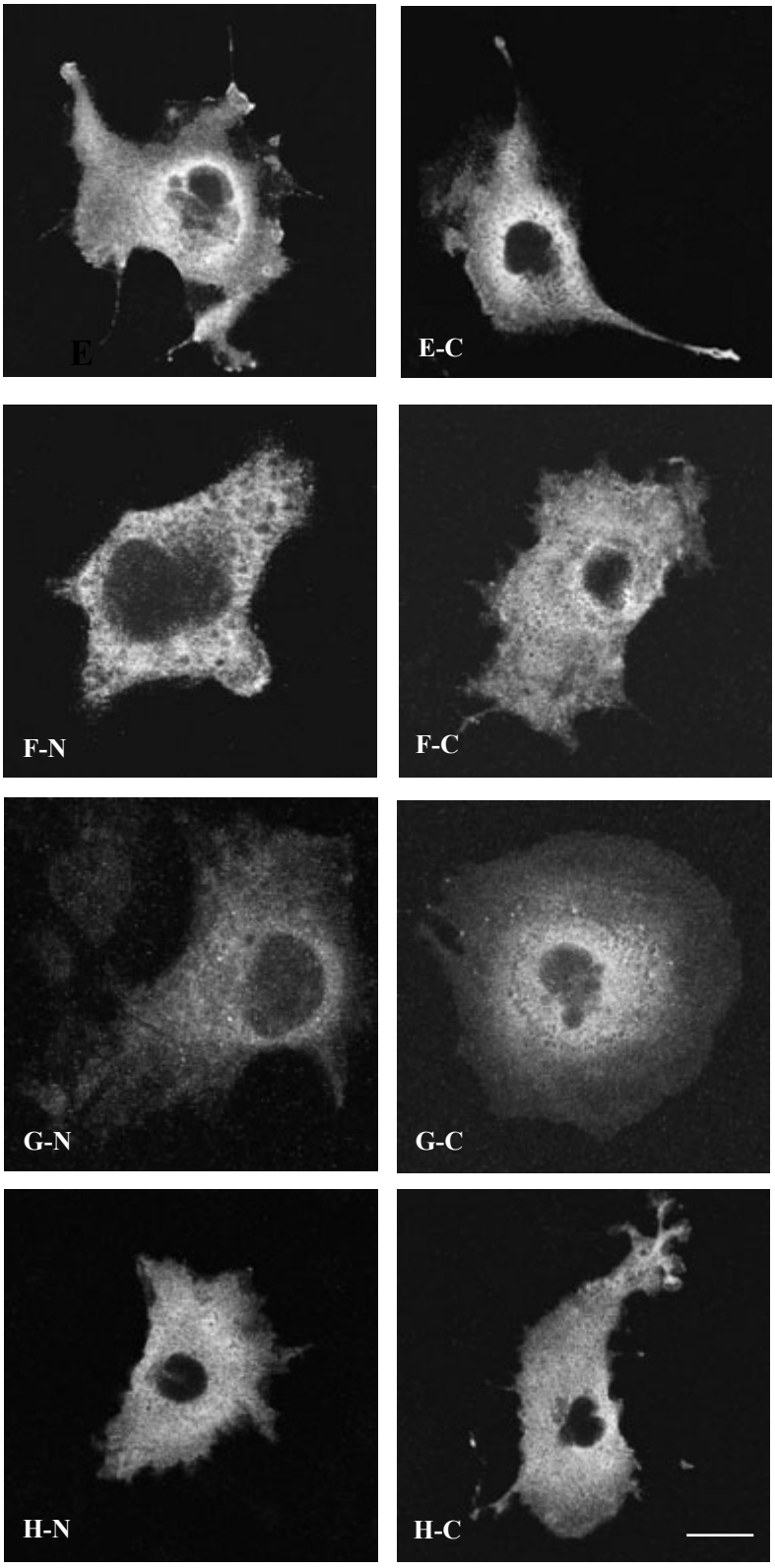
although the use of β-mercaptoethanol in the sample preparation should preclude this. Alternatively, these larger bands could result from glycosylation, or similar post-translational modification, of the expressed protein construct.

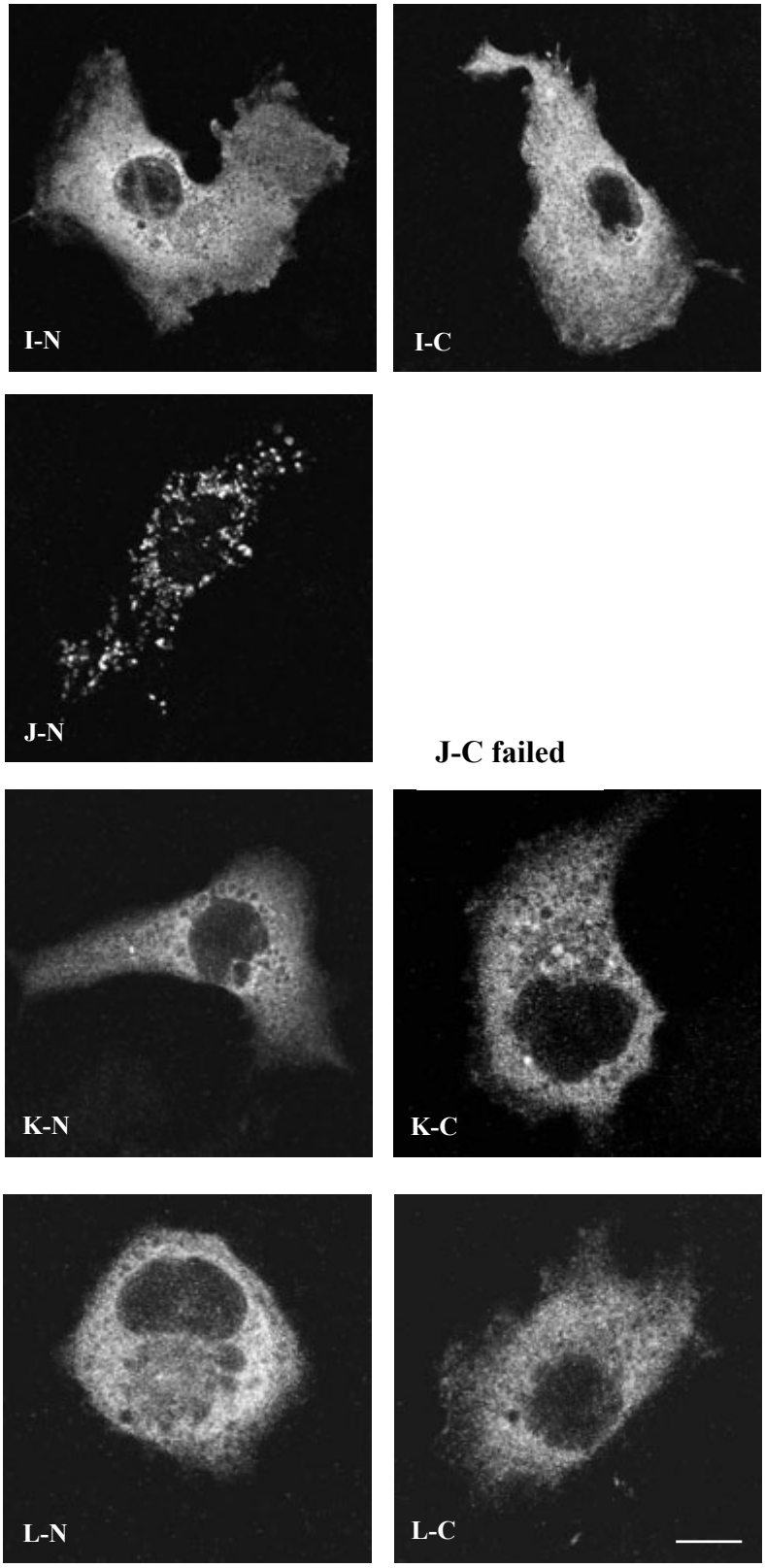
5.4.5 Analysis of T7.Tag protein subcellular location

To investigate the subcellular localisation of the fusion-protein T7.Tag constructs, immunofluorescence experiments in transiently transfected COS-7 cells were performed, under permeabilising conditions (chapter II). Permeabilising the cell allows entry of antibody and thus permits detection of intracellular proteins. A selection of the images obtained from these experiments using is shown in figure 5.14. Each image was examined to determine subcellular localisation. An electronic library of images from previous subcellular localisation experiments (Simpson *et al.*, 2000) (http://www.dkfz-heidelberg.de/abt0840/GFP/) was used to aid categorisation of the observed localisation patterns.

Subcellular localisation could be determined for 14 pairs of N- and C- terminal tagged cDNAs. BZRP (fig. 5.14.J) was successfully transfected only in the C-terminally tagged form, but an unidentified, but distinct, localisation pattern is observed from the N-tagged construct. The majority of the images demonstrate fluorescence of the expressed protein construct in the cytoplasm (fig. 5.14.A, B, D, E, F, G, H, I, K, L, M, N, O and P). Nuclear and vesicular exclusion is also noted in these images. In fig. 5.14.D and N, high levels of fluorescence are also seen in the ruffles of the cell membrane. Fig. 5.14.C demonstrates subcellular protein localisation at the endoplasmic reticulum.


A-N
A-C
B-N
B-C
C-N
C-C
D-N
D-C


E
E-C
F-N
F-C
G-N
G-C
H-N
H-C


I-N
I-C
J-N
J-C failed
K-N
K-C
L-N
L-C

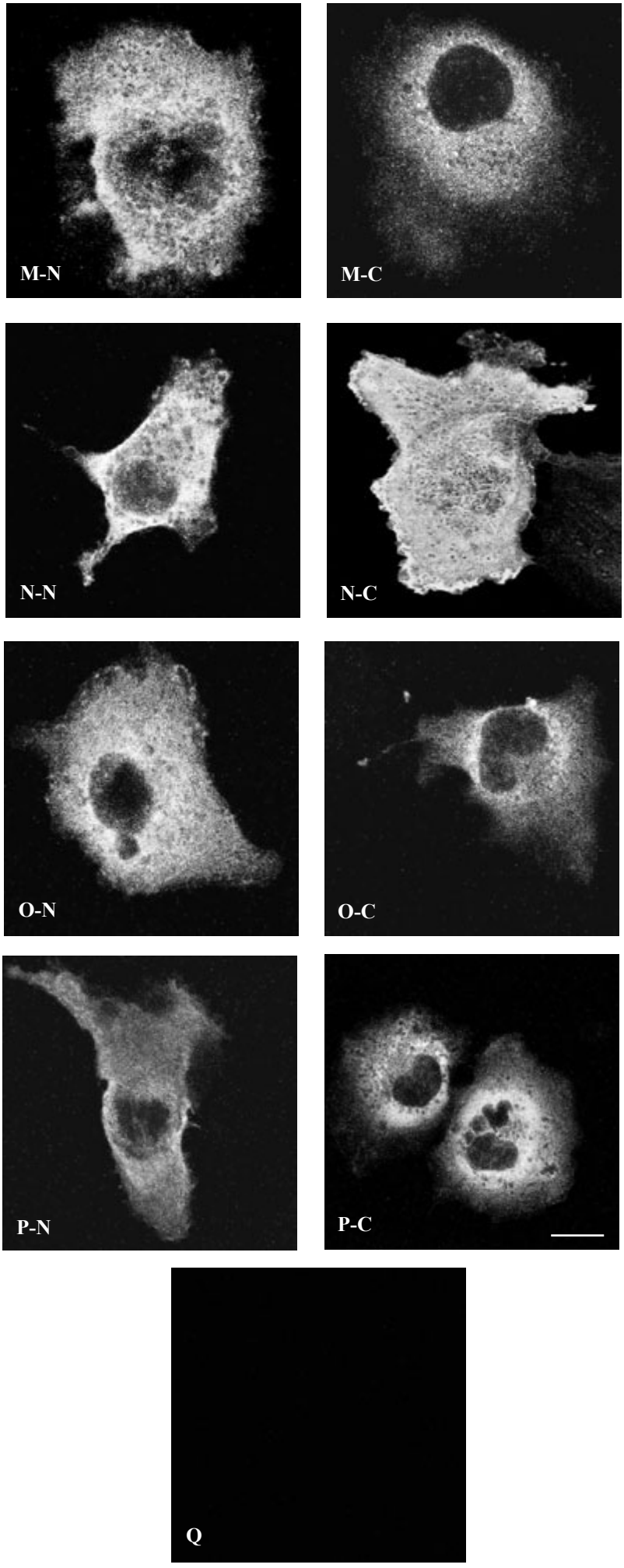

M-N
M-C
N-N
N-C
O-N
O-C
P-N
P-C
Q

Figure 5.14 (previous page): Examples of immunofluorescence experiments of COS-7 cells, transiently transfected with N- and C- terminally T7 tagged constructs of :
A) dJ222E13.C22.1a; B) dJ222E13.C22.1b; C) cB33B7.C22.1; D) PACSIN2a; E) TTLL1a; F)TTLL1c; G) BIK; H) bK1191B2.C22.3a; I) bK1191B2.C22.3b; J) BZRP(N-terminal T7.Tag construct only); K dJ526I14.C22.2b; L) SULTX3a; M) SULTX3b ; N) dJ549K18.C22.1; O) CGI-51; P) dJ102D24.C22.2; Q) Negative control (pcDNA3 empty vector). The bar indicates 10 µm.

Table 5.9: Subcellular localisation of 16 proteins encoded within 22q13.31

Image	Protein	Predicted Localisation	Localisation in COS7 cells	Remark
A	dJ222E13.C22.1a	Nu	Cy	
B	dJ222E13.C22.1b	Cy	Cy	
C	cB33B7.C22.1	Nu	ER	
D	PACSIN2	Nu	CM	~70% of transfected cells showed localisation at the cell membrane
E	TTLL1a	Cy	Cy	
F	TTLL1c	Cy	Cy	
G	BIK	Cy	Cy	
H	bK1191B2.C22.3a	Mi	Cy	
I	bK1191B2.C22.3b	Nu	Cy	
J	BZRP	Nu	Unknown	
K	dJ526I14.C22.2	Cy	Cy	
L	SULTX3a	Cy	Cy	
M	SULTX3b	Cy	Cy	
N	dJ549K18.C22.1	Cy	CM	~80% of transfected cells showed localisation at the cell membrane
O	CGI-51	Nu/Cy	Cy	
P	dJ102D24.C22.2	Nu	Cy	

Nu=nucleus; Cy=cytoplasm; Mi=mitochondria; ER=endoplasmic reticulum; CM=cytoplasm and cell membrane

Table 5.9 shows the PSORT correctly predicted 56% of the experimentally determined subcellular localisations. Interestingly however, these experimental results did not agree with small amount of available published data. BIK has previously been localised to the nuclear and cell membranes (Han *et al.*, 1996), but, in this experiment, a cytoplasmic localisation pattern was observed. Similarly, BZRP has previously been localised to mitochondrial tissues,

but may reside in other organelles (Hirsch *et al.*, 1998; Mukherjee & Das, 1989; Olson *et al.*, 1988). The localisation pattern shown by the BZRP construct in these experiments could not be classified, but was distinctly not nuclear in origin. These differences in localisation patterns may result from the different expression vectors and cell lines used in these experiments. Further experiments, such as co-expression of the tagged proteins with proteins of known localisation, or subcellular fractionation of transformed cells could be performed to verify and investigate these differences.

Some of the transfected cell samples showed examples of possible aggresome formation. Aggresomes are structures that have been observed to form peripherally and travel on microtubules in a minus-end direction to the microtubule organising centre (MTOC) regions, where they remain as distinct but closely apposed particulate structures. They are formed when production of misfolded proteins exceeds the cellular capacity to degrade them (Garcia-Mata *et al.*, 1999). Possible aggresome structures were noted in several cells, which, by their bright fluorescence, appeared to be expressing the tagged protein at high levels. (figure 5.15).

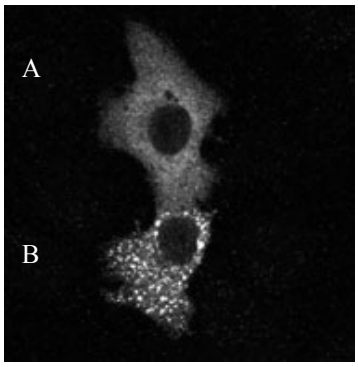


Figure 5.15: An example of possible aggresome formation. COS-7 cells were transfected with the N-terminal T7 tag construct of dJ222E13.C22.1a. The construct displays a cytosolic localisation pattern in cell A, which was also observed in the majority of other transfectants. Possible over-expression of the construct in cell B leads to aggresome formation.

5.5 Data integration

Table 5.10 below contains an overview of the accumulated data about each protein.

Table 5.10: Overall functional characteristics of 27 protein coding genes encoded within human chromosome 22q13.31.

Gene	Previously known functional information? (see table 5.1)	Characterised orthologue? (see table 5.4)	Subcellular localisation (see table 5.9)	Domains (largest isoform) (see figure 5.2)	Predicted N terminal signal peptide (figure 5.2)	Predicted transmembrane regions (figure 5.2)	Predicted coiled coil regions (figure 5.2)	Possible posttranslational modification (figure 5.13)	Expression pattern (section 3.5)	Alternative isoforms (section 3.8.6)
dJ222E13.C22.1		•	**Cy**	3	•	2	1			4
dJ222E13.C22.3			Cy	1						2
DIA1	•	•	ER	5		1				
cB33B7.C22.1	•	•	**ER**			1		•		
ARFGAP1	•		Nu	1	•	1	1			2
PACSIN2	•	•	**CyM**	1			1	•		2
TTLL1	•		**Cy**					•		4
BIK	•		**Cy**	2		1				
bK1191B2.C22.3		•	**Cy**	2		1		•		2
BZRP	•	•	*			5				
dJ526I14.C22.2			**Cy**			1				2
C22orf1	•		Mi	1						
SULTX3			**Cy**	1						2
dJ549K18.C22.1		•	**Cy**	1		4				
CGI-51			**Cy**			1				
bK414D7.C22.1		•	Nu	1		2	1			
dJ671O14.C22.2		•	Nu	1		1	1			2
dJ1033E15.C22.2			Nu			2				
ARHGAP8			Mi			2	1			
dJ127B20.C22.3			Cy/Nu	1		1				
NUP50	•		Nu			1				
bK268H5.C22.1			Nu							
UPK3	•	•	PM/Ex		•	2				
bK268H5.C22.4			Cy/Nu/ Mi	1	•	1				
dJ102D24.C22.2			**Cy**				1	•		
FBLN1	•	•	Ex	3	•		1			4
E46L	•		Cy				1			

Further details can be found in the tables indicated. Subcellular localisations shown in bold have been experimentally verified as part of this project, otherwise PSORT (Nakai & Horton, 1999) predictions are given.
* = unknown subcellular localisation pattern
Cy = cytoplasmic; CyM= cytoplasm and cell membrane; Nu = nuclear; Mi = mitochondrial; PM = plasma membrane; Ex = extracellular matrix; ER = endoplasmic reticulum

5.6 Discussion

This chapter has described a preliminary functional characterisation of the 27 full proteins annotated within 22q13.31. A selection of *in silico* analyses was performed to illustrate intrinsic sequence features and putative domains within the protein sequences. Database searches and phylogenetic tree analysis were used to identify putative orthologues. Extensive literature searches were performed to discover what was previously known about the proteins encoded within 22q13.31 and their putative orthologues. A subset of the genes was cloned, providing a valuable resource for future experimental analyses of protein function. Using these clones, an experimental analysis of subcellular localisation was performed. This study provides the first preliminary functional characterisation of 15 protein-coding genes encoded within 22q13.31 and reviews and extends the analysis of 12 previously studied genes in this region.

Identification of a previously characterised orthologue proved to be an efficient way to identify possible protein functions. For example, domain analysis of bK1191B2.C22.3a identified the presence of an acyl transferase domain in the larger isoform (bK1191B2.C22.3a). This functional evidence was corroborated and extended through phylogenetic analysis, which identified bK1191B2.C22.3 as an orthologue of an enzyme extensively conserved in evolution, malonyl CoA-acyl carrier protein transacylase. bK1191B2.C22.3a may therefore encode an essential enzyme in the biosynthesis of fatty acids, which catalyses the transacylation of malonate from malonyl-CoA to activated holo-ACP to generate malonyl-ACP, an elongation substrate in fatty acid biosynthesis. The localisation of bK1191B2.C22.3a to the cytoplasm also supports this hypothesis, as this is where fatty acid synthesis occurs in eukaryotes (reviewed in Stryer, 1988). Should this functional characterisation be correct, it seems likely that the prediction of a transmembrane

region within bK1191B2.C22.3a is incorrect. It remains unclear what role the isoform bK1191B2.C22.3b may play in this process. Western blot evidence, from both the N- and C-terminally tagged expressed protein bK1191B2.C22.3b, consistently produced bands both at the expected size of 19 kDa, and a weaker band, approximately 42 kDa in size. Potentially this could indicate that this isoform forms a dimer, although the use of β-mercaptoethanol in the sample preparation should preclude this. The larger band size could also derive from glycosylation of the expressed protein. Further experiments, such as two-hybrid analysis in yeast to identify protein-protein interactions, or investigation of post-translational modifications using mass spectroscopy or chemical assays, could be performed to investigate these hypotheses.

Phylogenetic analysis also indicated that the mouse gene adiponutrin (Baulande *et al.*, 2001) could be the potential orthologue of the previously uncharacterised gene dJ549K18.C22.1. BLAST searches of the protein and nucleotide sequence of this gene against the mouse Ensembl database (http://mouse.ensembl.org) indicate that the best match against the available mapped mouse sequence is the region of mouse chromosome 15 identified as syntenic with human chromosome 22q13.31 (chapter II). The nucleotide (coding exons only) and protein identities are 75% and 68% respectively. Additionally, analysis of the intrinsic sequence features of both proteins illustrated that both contained four putative transmembrane domains. Immunolocalisation assays of the transiently expressed adiponutrin and dJ549K18.C22.1 proteins in COS cells or COS-7 cells respectively (Baulande *et al.*, 2001, section 5.4) showed similar staining of the overexpressed proteins in the cytosol and appeared brighter close to the cell membrane. By fractionation of cell homogenates and immunoblotting of the membrane and cytosolic fractions, Baulande *et al*. were able to demonstrate localisation of the adiponutrin protein to the cell membrane. It would be interesting to perform this assay for comparison in the case of the human protein. Baulande *et al*. demonstrated by Northern

blotting that the mRNA expression pattern of adiponutrin is limited to adipose tissues. Furthermore, they note that a 3.2 kb transcript is undetectable in adipose tissue from fasting mice, but the level is dramatically increased when fasted mice are returned to a high carbohydrate diet. These analyses lead Baulande *et al*. to postulate that adiponutrin may be involved in adipocyte function. However, the expression studies of dJ549K18.C22.1 described in this thesis (chapter III), which include a range of tissues tested by Baulande *et al*. do not show a restricted expression pattern, although adipose tissue was not specifically tested.

Putative orthologues were identified for a further three novel genes. dJ222E13.C22.1 , bK414D7.C22.1 and dJ671O14.C22.2 were discussed briefly in the text (section 5.3.3), but in these cases, functional characterisation was less well advanced. The expression, domain and secondary structure analysis of each of the novel genes listed in table 5.10, as well as subcellular localisation where experimentally verified, should contribute to future analysis of both these proteins and their orthologues.

The results of these analyses were also compared with studies carried out on previously described genes. Subcellular localisation experiments added to functional knowledge in the case of TTLL1 and localisation of cB33B7.C22.1 to the ER indicates that this may be the site where this α1,4-galactosyltransferase acts in synthesis pathway of globo-series glycosphingolipids (Keusch *et al.*, 2000).

This approach also highlighted several examples of conflicting evidence from human proteins and their orthologues. Putative orthologues of the PACSIN2 protein have previously been localised to the cytoplasm (*M. musculus* protein PACSIN2, Ritter *et al.*, 1999), focal adhesion regions (*G. gallus* protein FAP52, Merilainen *et al.*, 1997) and to membrane ruffles and cytoplasmic vesicles (*X. laevis* protein X-PACSIN2, Cousin *et al.*, 2000). In this study, the human orthologue of PACSIN2 was localised to the cytoplasm and in ~70% of cases, to the

cell membrane. Like its putative orthologues, PACSIN2 contains up to three SH3 domains, which are often found in intracellular or membrane-associated proteins and may mediate assembly of specific protein complexes. An extensive coiled-coil secondary structure was also predicted. Interestingly, no signal peptide was recognised in the PACSIN2 protein sequence to enable localisation to the plasma membrane. However, this finding may be explained by the hypothesis put forward by Cousin *et al*.(2000) (figure 5.16). In a study of the *X. laevis* protein, this group proposed that X-PACSIN2 binds to the membrane-bound protein ADAM13, a metalloprotease, via SH3 domain regions in both proteins. X-PACSIN2 was also thought to interact with another 'repressor' protein via coiled coil regions, which affected ADAM13 activity when brought into close proximity via X-PACSIN2. *H. sapiens* PACSIN2 may also interact with a membrane bound protein in this way, leading to the localisation observed at the cell membrane described in this thesis.

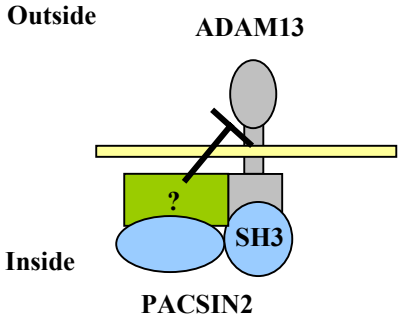


Figure 5.16: Adapted from Cousin *et al.* (2000). Schematic representation of the regulation of ADAM13 by X-PACSIN2. ADAM13 is in grey, the plasma membrane in yellow and X-PACSIN2 in blue. In this model, X-PACSIN2 binds to the ADAM13 cytoplasmic domain through its SH3 domain and to a putative repressor (green) with its coiled coil domain.

Several limitations of the subcellular localisation experiments described here are demonstrated by the results obtained for BZRP. Expression of the N-terminally T7 tagged BZRP protein in COS-7 cells resulted in an unidentified localisation pattern. Future work could include co-expression of the tagged BZRP protein together with proteins of known localisation, or subcellular fractionation experiments, in order to determine the origin of the unknown localisation pattern.

Chang *et al*. (1992), showed that the BZRP ligand [3H] PK 11195 had high affinity for an expressed BZRP construct in COS-1 cells, but the affinities of a pair of isoquinoline propanamide enantiomers differed remarkably in expressed and endogenous human BZRP. They suggested that the host cell and/or post-translational modification might have an important influence on BZRP function. In this case, the cell line used in the expression system, COS-7, may influence the localisation of the fusion-protein BZRP and therefore may not be truly representative of the human BZRP protein. Future work could therefore include transfer of the procedure to human cell lines.

Knowledge of the subcellular localisation of a protein provides an important clue to potential function. Many known biochemical reactions, signalling pathways and structural features are localised to different regions of the cell structure. Information derived from subcellular localisation experiments provides a starting point for further work to determine the role of a particular protein in that location.

This thesis illustrates a subcellular localisation protocol that could be streamlined for high throughput studies. The construction of the pCDNA3-NT7 vector removed one restriction digestion and ligation step from the protocol. Construction of an equivalent pCDNA3-CT7 plasmid and cloning of the PCR products directly into these expression vectors would reduce the number of restriction digests and ligation steps to a minimum and thus increase the overall efficiency of the protocol. Alternatively, a recombination-based cloning system could be introduced, such as the Gateway™ cloning system (Invitrogen). Recent studies (Simpson *et al.*, 2000; Wiemann *et al.*, 2001) have described the systematic tagging with GFP of full-length cDNAs that have been identified and sequenced by large-scale genome projects. The procedures described are amenable to automation and other characterisation studies (for example, mutagenesis, protein dynamics and identification of interacting partners) could

follow the localisation screen immediately without further generation of new reagents. Similar expression vectors could be designed that incorporate the T7 .Tag into a recombination-based system, thus avoiding the problems of steric hindrance associated with GFP-fusion proteins discussed in the introduction to this chapter.

Difficulties were encountered in the design of nested PCR primers from the frequently GC-rich 5' UTR sequences. Additionally, several genes from 22q13.31 have ORFs that are several kilobases long and are thus more difficult to amplify by PCR. In some cases, these problems could be overcome by utilising existing cloned cDNA resources (for example, IMAGE clones Lennon *et al.*, 1996) or by PCR amplification and subsequent ligation of sections of the ORF.

Several occurrences of possible post-translational modification were identified from Western blot experiments. Further investigation is needed to explain these observations. For example, two-dimensional gel electrophoresis coupled to mass spectroscopy and appropriate software allows not only peptide mass fingerprinting for low quantities (Kuster & Mann, 1998) but also specific detection of amino acid modifications on a large scale, including phosphorylation, acetylation and non-standard amino acid residues such as hydroxyproline and hydroxylysine (Dongre *et al.*, 1997).

Efficient identification of orthologues is currently hindered by redundancy and poor annotation in protein sequence databases. Several identical or near-identical 'versions' of nearly all the proteins in this study currently exist in the NCBI non-redundant protein sequence database and accompanying annotation does not often reveal the origin of the sequence. Manual removal of redundant sequences for phylogenetic analysis is therefore fairly arduous. Additionally, some proteins have acquired several different names (see appendix 4), or are named after similar proteins that are not true orthologues. Worryingly, some names seem to be wrongly transferred by similarity. For example, the *C. pneumoniae* malonyl acyl

carrier transacyclase, which has extensive homology to other malonyl acyl carrier transacyclase proteins, appears to have been misspelt as a transcyclase and thus is not found in database searches for transacylases. Some of these problems could be avoided by the use of an extensively curated protein database such as SwissProt (Bairoch & Apweiler, 2000). However, it was found that many of the orthologues described in this chapter did not yet have SwissProt entries and so would have been missed.

In summary, this chapter has described preliminary functional characterisation of 27 protein-coding genes within 22q13.31. Successful identification of a characterised putative orthologue proved to be the most efficient analysis used, as this established a basis upon which the results of other analyses could be compared and evaluated. It has also described a pilot project for further possible subcellular localisation studies, which, with further streamlining as described above, could be scaled-up for higher-throughput investigations.

Chapter VI Discussion

6.1 Summary

This thesis describes a structural, comparative and functional study of a 3.4 Mb region of human chromosome 22 (22q13.31). Production of a high quality transcript map of the region enabled extensive analysis of the genes encoded within the DNA sequence. Mapping and sequencing of syntenic regions of the mouse genome allowed detailed comparative sequence analysis both of this region and of an upstream syntenic breakpoint junction. The study was extended to include an *in silico* functional characterisation of the proteins encoded within 22q13.31. Additional experimental investigation of the subcellular localisation of a subset of these proteins was also performed.

6.2 Genomic sequence

The work presented in this thesis emphasises the importance of the availability of genomic sequence, as it enables detailed analysis of genes in relation to their genomic environment. The assembly of an experimentally verified transcript map was described in chapter III. In total, 39 genes and 17 pseudogenes were annotated across this region. The high quality transcript data has been integrated with analysis of the surrounding genome to produce a base pair-resolution map that will provide a durable reference for all kinds of future studies.

The genomic sequence provided the basis for a systematic approach to the experimental verification of putative coding sequences. Both positive and negative expression results were referenced against the region of DNA sequence tested, in order to provide a clear picture of sequence transcription. Extensive sequence analysis of the transcribed regions supported earlier studies of the conservation of splice site sequences, but highlighted discrepancies between the sequence context of the putative start codons and the scanning model of translation initiation.

Analysis of these and other features examined in this thesis will soon become possible on a genomic scale with the advent of the finished human sequence. Previously, this type of analysis was only possible on fragmented sequences, often biased towards particular genes or gene families. The availability of the genomic sequence enables a more structured, organised and, once sequencing and annotation are complete, unbiased approach to investigation of both the genomic sequence and its encoded protein products. It will be interesting to determine if the theories previously based on study of fragmented sequences are supported by these studies.

Detailed sequence analysis is reliant on high quality finished sequence. Although the publication and analysis of the draft genome sequence highlighted the utility of unfinished sequence for large-scale analysis of broad features of the human genome, such as GC and repeat content (Lander *et al.*, 2001), resolution of errors and gaps in the draft sequence will enable an unambiguous analysis of these features to be performed. The current imperfect state of the draft genome sequence causes more serious problems in the annotation of human genes. High accuracy is essential in delineation of the protein-coding regions, as ambiguities and errors in unfinished sequence can result in annotation of partial or fragmented genes: predicted genes may also be incorrectly fused or even spurious. Errors leading to alteration of the protein code may affect predictions of function and design of future experiments. The provision of finished sequence will therefore provide a valuable and essential resource for the annotation of human genes.

6.3 Gene annotation

The sequence data generated by the human genome project is paving the way for the identification of the entire complement of human genes. The vast amount of data produced has prompted development of fully automated annotation systems. The Ensembl approach

(Hubbard & Birney, 2000) is based on confirmation of *ab initio* predictions by homology and provides functional annotation via Pfam (Bateman *et al.*, 1999). However, such systems have several limitations. Depending on annotation criteria, if no overlapping similarity information is found, multiple genes may be annotated for what may in fact be a single gene. Artefacts in EST data, arising from unspliced mRNAs, genomic DNA contamination and nongenic transcription (for example from the promoter of a transposable element) detailed both in this project and in the study by Wolfberg and Landman (1997), only confounds this problem as spurious EST data may be used to support incorrect predictions.

A semi-automatic approach, based on sequence similarity information, is utilised in the annotation of the clone-by-clone output of the genome centres. However, this means that genes spanning multiple clones are partially annotated multiple times, a practice that can lead to confusion and redundancy in sequence database entries.

This thesis describes the assembly of a high quality transcript map of human chromosome 22q13.31. Central to the approach used was the availability of high quality, linked, finished genomic sequence spanning nearly the entire region under investigation. Availability of this resource prevented misannotation of genes spanning multiple clones and avoided ambiguities arising from unfinished sequence. All available data, including both expressed sequence evidence and *ab initio* predictions, were manually inspected and gene structures were annotated only when supported by experimental evidence. Where this was absent, additional cDNA sequencing was undertaken to confirm the intron-exon structure. This approach, although arduous, is necessary to ensure high levels of accuracy. Ambiguities generated by inclusion of unsupported gene predictions are thus avoided, although retention of these predictions within the transcript map database (22ace) ensures that this data is easily accessible, if required, together with a record of all cDNA library screens performed. The

transcript map of 22q13.31 therefore provides a strong foundation for future research into this region.

6.4 Mouse genomics

Mapping and sequencing of three regions of the mouse genome with conserved synteny to both 22q13.31 and to a synteny breakpoint junction on 22q13.1, demonstrated the potential utility of mouse genomic sequence for gene annotation and analysis of chromosome evolution. The study also illustrated the justification of the early data release policy implemented by the Sanger Institute and other public domain sequencing centres, as a large amount of information was derived from unfinished mouse genomic sequence.

Although comparative analysis of the mouse sequence with 22q13.31 did not result in the annotation of any further genes, conserved regions were found to correlate closely with the gene annotation. This implies that mouse genomic sequence could provide a powerful tool for gene annotation in less well-studied areas of the human genome. Identification of functionally conserved coding regions could also be useful in the identification of genes that are not represented in the available RNA or cDNA resources, perhaps because of a spatially or temporally limited expression pattern. As this study was largely based on unfinished mouse sequence, identification of potentially conserved coding regions outside of the existing annotation was not considered strong enough grounds for inclusion. However, the increasing availability of finished mouse sequence should permit a more detailed examination of these regions, including the use of sequence similarity searches and gene prediction programs.

Over 30 putative regulatory sequences were identified within the conserved sequences upstream of four annotated transcription start sites. Increased specificity in this study could perhaps be achieved by the inclusion of a third genomic sequence from a vertebrate organism

in the comparison. Recent studies (Frazer *et al.*, 2001; Göttgens *et al.*, 2000) have shown that identification of a conserved non-coding sequence in three vertebrate genomes increases confidence that the putative regulatory sequence is not a false positive. Experimental assays, such as gel retardation, DNase footprinting or methylation interference, could then be carried out to identify protein-binding sites within the region of interest.

This analysis also redefined the boundaries of the syntenic junction of mouse chromosomes 8 and 15 on human chromosome 22q13.1, to a 50 kb region between two adjacent human genes. No potentially causal similarity could be discerned from comparison with breakpoints previously described at the sequence level (Lund *et al.*, 2000; Pletcher *et al.*, 2000). However, as finished mouse sequence for this, and other, syntenic junctions becomes available, a clearer picture of mammalian chromosomal evolution may develop.

An area of finished mouse sequence spanned a ~50 kb gap in the sequence of human chromosome 22, providing a picture of what the equivalent 'unclonable' human sequence may contain. The region includes the 3' exons of the murine orthologue of C22orf1. The putative human exon sequences could be used to design experiments to screen genomic libraries in an effort to close the gap in the human sequence. The comparison of mouse and human GC profiles showed that the mouse GC profile is very similar to that of the human region, but has a lower GC content overall. Interestingly, GC content is raised throughout the murine 'gap' region, and thus possibly exaggerated in the equivalent human region. The high GC content could cause the region to be deletion-prone though frameshift mutagenesis or other unknown cellular mechanism (Bichara *et al.*, 1995, 2000) and thus difficult to clone.

Possibly the most valuable contribution that the mouse genome sequence will make will be to the functional characterisation of orthologous human genes. This thesis illustrates several examples where identification of a functionally characterised murine orthologue permitted

more efficient characterisation of the human protein. As more mapped mouse genomic sequence becomes available, the identification of murine orthologues will become easier. The high quality transcript map of human chromosome 22q13.31, and the comparative sequencing and annotation of the equivalent murine region of conserved synteny described in this thesis, provides an excellent resource for the study of known and potential genetic diseases in this region. This may include further study of spinocerebellar ataxia type 10 (Matsuura *et al.*, 2000), which is caused by the expansion of a pentanucleotide repeat in an intron of the gene E46L. The accurate annotation of this gene onto the genomic sequence and the near availability of finished mouse sequence surrounding the orthologous gene will provide a vital resource for future study of this disorder, both in human populations and in model populations of the laboratory mouse.

6.5 Functional studies

Annotated sequence is now available for much of the human genome, but in the vast majority of cases, the question of gene and protein function remains unsolved. The determination of function is being addressed in a growing number of ways by the emerging field of functional genomics.

This thesis illustrates a selection of these techniques in a preliminary functional characterisation of 27 protein coding genes encoded within 22q13.31. This study represents the first functional analysis of 15 novel genes identified in this region. The importance of a high quality transcript map was demonstrated by this work. Confidence in the gene annotation enabled the generation of cDNA clones containing the full, unambiguous ORF. Discrepancies in the clone insert sequences were easily identified and were systematically assessed to determine of they were due to PCR error, or were an accurate representation of genomic polymorphisms. Inaccurate gene annotation could have led to these discrepancies being

missed, which could potentially have altered the results of functional analyses. Accuracy in gene annotation was also vital for *in silico* investigations of protein function. Much of this work was based upon sequence similarity searches: errors in unverified transcripts could again have potentially altered functional predictions.

A range of software was used to undertake an *in silico* analysis of secondary structure, domain content and subcellular localisation. This type of investigation is amenable to high throughput analysis, but comparison of the results of different algorithms and experimental verification of subcellular localisation highlighted shortcomings in individual programs. However, as the amount of gene data in the public domain increases as a result of the human genome project, such analysis software is likely to improve.

Identification of a previously characterised orthologue was found to be an effective method of attaching a putative function to a protein. This type of analysis is currently less amenable to large-scale study, as current database search techniques cannot distinguish between orthologous genes or merely paralogous matches. The problem is exacerbated by redundancy and examples of poor description of submitted sequences. Some of these problems will be relieved by the increasing amounts of mapped genomic data emanating from the mouse and other model organism sequencing projects, enabling chromosomal location to be taken into account during orthologue identification. Even so, the example of PACSIN2 discussed in this study, where determination of the subcellular localisation of the protein contradicted previous findings from the mouse and chicken orthologues, emphasises the importance of experimental as well as computational investigation in this field.

This thesis illustrates one experimental approach that could be adapted for the high throughput analysis of protein subcellular localisation. Additional high throughput studies are being developed, or are already underway, in order to accumulate information about DNA sequence,

regulatory regions, mRNA profiles, protein expression and interaction and metabolite concentration. Model organisms are also used in functional studies to manipulate the orthologous gene and observe the functional affect. The ultimate aim of functional genomics is to integrate information from all these 'levels' in order to generate effective models of biological systems. The mass of data generated from these projects will necessitate a combination of bioinformatic and experimental approaches. The future challenge to the bioinformatics community will be the integration and finding of patterns in the combined datasets, such as the linking of expression data to genotype and the deduction of genetic pathways from available functional information and expression data.

The generation of a capable and reliable bioinformatic infrastructure is essential to ensure success in this future work to define the functions of the human genome. The beginnings of this infrastructure are already in place through the development of sequence databases and, more recently, whole genome browsers such as Ensembl (Hubbard & Birney, 2000). However, in order to provide a firm foundation for higher-level, interconnecting databases containing functional information, it is necessary to ensure that complete, non-redundant gene and protein information is accurately catalogued. Extensive curation of the existing sequence databases is required to 'clean-up' the thousands of redundant entries that have been generated by the continuous release of genomic sequence over the past few years. A large amount of functional information is already available, both from small-scale investigations of individual genes described in the literature and from high throughput studies. Integration of the existing data into a readily accessible bioinformatic infrastructure will greatly enhance the utility of the previous research and enable an accurate assessment of current functional understanding. Further data can easily, and usefully, be derived using existing *in silico* approaches. For example, a large-scale bioinformatic analysis to establish a database of orthologous protein

relationships across whole genomes, will provide important preliminary information for future research.

It is unlikely that high throughput projects alone will provide all the answers. The functional characterisation of just 38 protein-coding genes described in this thesis illustrates that group analysis generates many different avenues for further study of individual proteins. In this case, further investigation could include individual biochemical assays of the functional domains identified in the protein sequences, experimental confirmation that the functional characteristics of particular orthologous proteins are retained in the human version, or further analysis of the possible post-transcriptional modifications noted from the experimental expression of the protein in COS-7 cells. The accurately annotated human genome sequence and preliminary functional studies described in this thesis, provide an excellent resource for future functional characterisation of these genes.

6.6 Conclusion

This thesis demonstrates the utility of the human genomic sequence in the generation of a high quality transcription map. The availability of the genomic sequence enabled extensive sequence analysis of the annotated genes and their environment. The value of comparative sequence analysis was illustrated through investigation of regions of the mouse genome syntenic to human chromosome 22.The study also illustrates the utility of these genomic resources for functional analysis in the post-genomic era. Both the transcript map and comparative mouse data will provide a valuable tool for future research to further characterise the proteins encoded within 22q13.31.

Chapter VII References

Aach J., Bulyk M. L., Church G. M., Comander J., Derti A., and Shendure J. (2001). Computational comparison of two draft sequences of the human genome. *Nature* **409:** 856-9.

Achaz G., Netter P., and Coissac E. (2001). Study of intrachromosomal duplications among the eukaryote genomes. *Mol Biol Evol* **18:** 2280-8.

Adams M. D., Celniker S. E., Holt R. A., Evans C. A., Gocayne J. D., Amanatides P. G., Scherer S. E., Li P. W., Hoskins R. A., Galle R. F., George R. A., Lewis S. E., Richards S., Ashburner M., Henderson S. N., Sutton G. G., Wortman J. R., Yandell M. D., Zhang Q., Chen L. X., Brandon R. C., Rogers Y. H., Blazej R. G., Champe M., Pfeiffer B. D., Wan K. H., Doyle C., Baxter E. G., Helt G., Nelson C. R., Gabor G. L., Abril J. F., Agbayani A., An H. J., Andrews-Pfannkoch C., Baldwin D., Ballew R. M., Basu A., Baxendale J., Bayraktaroglu L., Beasley E. M., Beeson K. Y., Benos P. V., Berman B. P., Bhandari D., Bolshakov S., Borkova D., Botchan M. R., Bouck J., Brokstein P., Brottier P., Burtis K. C., Busam D. A., Butler H., Cadieu E., Center A., Chandra I., Cherry J. M., Cawley S., Dahlke C., Davenport L. B., Davies P., de Pablos B., Delcher A., Deng Z., Mays A. D., Dew I., Dietz S. M., Dodson K., Doup L. E., Downes M., Dugan-Rocha S., Dunkov B. C., Dunn P., Durbin K. J., Evangelista C. C., Ferraz C., Ferriera S., Fleischmann W., Fosler C., Gabrielian A. E., Garg N. S., Gelbart W. M., Glasser K., Glodek A., Gong F., Gorrell J. H., Gu Z., Guan P., Harris M., Harris N. L., Harvey D., Heiman T. J., Hernandez J. R., Houck J., Hostin D., Houston K. A., Howland T. J., Wei M. H., Ibegwam C., et al. (2000). The genome sequence of Drosophila melanogaster. *Science* **287:** 2185-95.

Adams M. D., Kelley J. M., Gocayne J. D., Dubnick M., Polymeropoulos M. H., Xiao H., Merril C. R., Wu A., Olde B., Moreno R. F., and et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252:** 1651-6.

Adams M. D., Soares M. B., Kerlavage A. R., Fields C., and Venter J. C. (1993). Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat Genet* **4:** 373-80.

Aissani B., and Bernardi G. (1991). CpG islands, genes and isochores in the genomes of vertebrates. *Gene* **106:** 185-95.

Altschul S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W., and Lipman D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-402.

Amadou C., Ribouchon M. T., Mattei M. G., Jenkins N. A., Gilbert D. J., Copeland N. G., Avoustin P., and Pontarotti P. (1995). Localization of new genes and markers to the distal part of the human major histocompatibility complex (MHC) region and comparison with the mouse: new insights into the evolution of mammalian genomes. *Genomics* **26:** 9-20.

Anderson L., and Seilhamer J. (1997). A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* **18:** 533-7.

Anderson S., Bankier A. T., Barrell B. G., de Bruijn M. H., Coulson A. R., Drouin J., Eperon I. C., Nierlich D. P., Roe B. A., Sanger F., Schreier P. H., Smith A. J., Staden R., and Young I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature* **290:** 457-65.

Ansari-Lari M. A., Oeltjen J. C., Schwartz S., Zhang Z., Muzny D. M., Lu J., Gorrell J. H., Chinault A. C., Belmont J. W., Miller W., and Gibbs R. A. (1998). Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res* **8:** 29-40.

Antequera F., and Bird A. (1993). Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A* **90:** 11995-9.

Apweiler R., Attwood T. K., Bairoch A., Bateman A., Birney E., Biswas M., Bucher P., Cerutti L., Corpet F., Croning M. D., Durbin R., Falquet L., Fleischmann W., Gouzy J., Hermjakob H., Hulo N., Jonassen I., Kahn D., Kanapin A., Karavidopoulou Y., Lopez R., Marx B., Mulder N. J., Oinn T. M., Pagni M., Servant F., Sigrist C. J., and Zdobnov E. M. (2000). InterPro--an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16:** 1145-50.

Argraves W. S., Tran H., Burgess W. H., and Dickerson K. (1990). Fibulin is an extracellular matrix and plasma glycoprotein with repeated domain structure. *J Cell Biol* **111:** 3155-64.

Ashburner M., Misra S., Roote J., Lewis S. E., Blazej R., Davis T., Doyle C., Galle R., George R., Harris N., Hartzell G., Harvey D., Hong L., Houston K., Hoskins R., Johnson G., Martin C., Moshrefi A., Palazzolo M., Reese M. G., Spradling A., Tsang G., Wan K., Whitelaw K., Celniker S., and et al. (1999). An exploration of the sequence of a 2.9-Mb region of the genome of Drosophila melanogaster: the Adh region. *Genetics* **153:** 179-219.

Attwood T. K., Blythe M. J., Flower D. R., Gaulton A., Mabey J. E., Maudling N., McGregor L., Mitchell A. L., Moulton G., Paine K., and Scordis P. (2002). PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Res* **30:** 239-41.

Audic S., and Claverie J. M. (1998). Visualizing the competitive recognition of TATA-boxes in vertebrate promoters. *Trends Genet* **14:** 10-1.

Bairoch A., and Apweiler R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28:** 45-8.

Baker W., van den Broek A., Camon E., Hingamp P., Sterk P., Stoesser G., and Tuli M. A. (2000). The EMBL nucleotide sequence database. *Nucleic Acids Res* **28:** 19-23.

Bankier A. T., Weston K. M., and Barrell B. G. (1987). Random cloning and sequencing by the M13/dideoxynucleotide chain termination method. *Methods Enzymol* **155:** 51-93.

Bargmann C. I. (2001). High-throughput reverse genetics: RNAi screens in Caenorhabditis elegans. *Genome Biol* **2**.

Bateman A., Birney E., Durbin R., Eddy S. R., Finn R. D., and Sonnhammer E. L. (1999). Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res* **27:** 260-2.

Baudin A., Ozier-Kalogeropoulos O., Denouel A., Lacroute F., and Cullin C. (1993). A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae. *Nucleic Acids Res* **21:** 3329-30.

Baulande S., Lasnier F., Lucas M., and Pairault J. (2001). Adiponutrin, a transmembrane protein corresponding to a novel dietary- and obesity-linked mRNA specifically expressed in the adipose lineage. *J Biol Chem* **276:** 33336-44.

Beaudoing E., Freier S., Wyatt J. R., Claverie J. M., and Gautheret D. (2000). Patterns of variant polyadenylation signal usage in human genes. *Genome Res* **10:** 1001-10.

Bell C. J., Budarf M. L., Nieuwenhuijsen B. W., Barnoski B. L., Buetow K. H., Campbell K., Colbert A. M., Collins J., Daly M., Desjardins P. R., and et al. (1995). Integration of physical, breakpoint and genetic maps of chromosome 22. Localization of 587 yeast artificial chromosomes with 238 mapped markers. *Hum Mol Genet* **4:** 59-69.

Bentley D. R., Deloukas P., Dunham A., French L., Gregory S. G., Humphray S. J., Mungall A. J., Ross M. T., Carter N. P., Dunham I., Scott C. E., Ashcroft K. J., Atkinson A. L., Aubin K., Beare D. M., Bethel G., Brady N., Brook J. C., Burford D. C., Burrill W. D., Burrows C., Butler A. P., Carder C., Catanese J. J., Clee C. M., Clegg S. M., Cobley V., Coffey A. J., Cole C. G., Collins J. E., Conquer J. S., Cooper R. A., Culley K. M., Dawson E., Dearden F. L., Durbin R. M., de Jong P. J., Dhami P. D., Earthrowl M. E., Edwards C. A., Evans R. S., Gillson C. J., Ghori J., Green L., Gwilliam R., Halls K. S.,

Hammond S., Harper G. L., Heathcott R. W., Holden J. L., Holloway E., Hopkins B. L., Howard P. J., Howell G. R., Huckle E. J., Hughes J., Hunt P. J., Hunt S. E., Izmajlowicz M., Jones C. A., Joseph S. S., Laird G., Langford C. F., Lehvaslaiho M. H., Leversha M. A., McCann O. T., McDonald L. M., McDowall J., Maslen G. L., Mistry D., Moschonas N. K., Neocleous V., Pearson D. M., Phillips K. J., Porter K. M., Prathalingam S. R., Ramsey Y. H., Ranby S. A., Rice C. M., Rogers J., Rogers L. J., Sarafidou T., Scott D. J., Sharp G. J., Shaw-Smith C. J., Smink L. J., Soderlund C., Sotheran E. C., Steingruber H. E., Sulston J. E., Taylor A., Taylor R. G., Thorpe A. A., Tinsley E., Warry G. L., Whittaker A., Whittaker P., Williams S. H., Wilmer T. E., Wooster R., et al. (2001). The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* **409:** 942-3.

Bernardi G. (1993). The isochore organization of the human genome and its evolutionary history--a review. *Gene* **135:** 57-66.

Bernardi G., Olofsson B., Filipski J., Zerial M., Salinas J., Cuny G., Meunier-Rotival M., and Rodier F. (1985). The mosaic genome of warm-blooded vertebrates. *Science* **228:** 953-8.

Bernot A., Heilig R., Clepet C., Smaoui N., Da Silva C., Petit J. L., Devaud C., Chiannilkulchai N., Fizames C., Samson D., Cruaud C., Caloustian C., Gyapay G., Delpech M., and Weissenbach J. (1998). A transcriptional Map of the FMF region. *Genomics* **50:** 147-60.

Bichara M., Pinet I., Schumacher S., and Fuchs R. P. (2000). Mechanisms of dinucleotide repeat instability in Escherichia coli. *Genetics* **154:** 533-42.

Bichara M., Schumacher S., and Fuchs R. P. (1995). Genetic instability within monotonous runs of CpG sequences in Escherichia coli. *Genetics* **140:** 897-907.

Bird A., Taggart M., Frommer M., Miller O. J., and Macleod D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40:** 91-9.

Bird A., Tate P., Nan X., Campoy J., Meehan R., Cross S., Tweedie S., Charlton J., and Macleod D. (1995). Studies of DNA methylation in animals. *J Cell Sci Suppl* **19:** 37-9.

Bird A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature* **321:** 209-13.

Blackstock W. P., and Weir M. P. (1999). Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol* **17:** 121-7.

Blattner F. R., Plunkett G., 3rd, Bloch C. A., Perna N. T., Burland V., Riley M., Collado-Vides J., Glasner J. D., Rode C. K., Mayhew G. F., Gregor J., Davis N. W., Kirkpatrick H. A., Goeden M. A., Rose D. J., Mau B., and Shao Y. (1997). The complete genome sequence of Escherichia coli K-12. *Science* **277:** 1453-74.

Blechschmidt K., Schweiger M., Wertz K., Poulson R., Christensen H. M., Rosenthal A., Lehrach H., and Yaspo M. L. (1999). The mouse Aire gene: comparative genomic sequencing, gene organization, and expression. *Genome Res* **9:** 158-66.

Boguski M. S., Lowe T. M., and Tolstoshev C. M. (1993). dbEST--database for "expressed sequence tags". *Nat Genet* **4:** 332-3.

Boguski M. S., and Schuler G. D. (1995). ESTablishing a human transcript map. *Nat Genet* **10:** 369-71.

Bonfield J. K., Smith K., and Staden R. (1995). A new DNA sequence assembly program. *Nucleic Acids Res* **23:** 4992-9.

Bork P., and Koonin E. V. (1998). Predicting functions from protein sequences--where are the bottlenecks? *Nat Genet* **18:** 313-8.

Botstein D., White R. L., Skolnick M., and Davis R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32:** 314-31.

Bouffard G. G., Idol J. R., Braden V. V., Iyer L. M., Cunningham A. F., Weintraub L. A., Touchman J. W., Mohr-Tidwell R. M., Peluso D. C., Fulton R. S., Ueltzen M. S., Weissenbach J., Magness C. L., and Green E. D. (1997). A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79 kb. *Genome Res* **7:** 673-92.

Boyd J. M., Gallo G. J., Elangovan B., Houghton A. B., Malstrom S., Avery B. J., Ebb R. G., Subramanian T., Chittenden T., Lutz R. J., and et al. (1995). Bik, a novel death-inducing protein shares a distinct sequence motif with Bcl-2 family proteins and interacts with viral and cellular survival-promoting proteins. *Oncogene* **11:** 1921-8.

Brett D., Hanke J., Lehmann G., Haase S., Delbruck S., Krueger S., Reich J., and Bork P. (2000). EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett* **474:** 83-6.

Brocchieri L. (2001). Phylogenetic inferences from molecular sequences: review and critique. *Theor Popul Biol* **59:** 27-40.

Brosius J. (1999). Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* **107:** 209-38.

Buckler A. J., Chang D. D., Graw S. L., Brook J. D., Haber D. A., Sharp P. A., and Housman D. E. (1991). Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. *Proc Natl Acad Sci U S A* **88:** 4005-9.

Burge C., and Karlin S. (1997). Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268:** 78-94.

Burge C. B., and Karlin S. (1998). Finding the genes in genomic DNA. *Curr Opin Struct Biol* **8:** 346-54.

Burke D. T., Carle G. F., and Olson M. V. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236:** 806-12.

Burns N., Grimwade B., Ross-Macdonald P. B., Choi E. Y., Finberg K., Roeder G. S., and Snyder M. (1994). Large-scale analysis of gene expression, protein localization, and gene disruption in Saccharomyces cerevisiae. *Genes Dev* **8:** 1087-105.

Burset M., and Guigo R. (1996). Evaluation of gene structure prediction programs. *Genomics* **34:** 353-67.

Burt D. W., Bruley C., Dunn I. C., Jones C. T., Ramage A., Law A. S., Morrice D. R., Paton I. R., Smith J., Windsor D., Sazanov A., Fries R., and Waddington D. (1999). The dynamics of chromosome evolution in birds and mammals. *Nature* **402:** 411-3.

Carver E. A., and Stubbs L. (1997). Zooming in on the human-mouse comparative map: genome conservation re- examined on a high-resolution scale. *Genome Res* **7:** 1123-37.

Casalotti S. O., Pelaia G., Yakovlev A. G., Csikos T., Grayson D. R., and Krueger K. E. (1992). Structure of the rat gene encoding the mitochondrial benzodiazepine receptor. *Gene* **121:** 377-82.

Castells A., Ino Y., Louis D. N., Ramesh V., Gusella J. F., and Rustgi A. K. (1999). Mapping of a target region of allelic loss to a 0.5-cM interval on chromosome 22q13 in human colorectal cancer. *Gastroenterology* **117:** 831-7.

Centola M., Chen X., Sood R., Deng Z., Aksentijevich I., Blake T., Ricke D. O., Wood G., Zaks N., Richards N., Krizman D., Mansfield E., Apostolou S., Liu J., Shafran N., Vedula A., Hamon M., Cercek A., Kahan T., Gumucio D., Callen D. F., Richards R. I., Moyzis R. K., Kastner D. L., and et al. (1998). Construction of an approximately 700-

kb transcript map around the familial Mediterranean fever locus on human chromosome 16p13.3. *Genome Res* **8:** 1172-91.

Chambers D. M., Rouleau G. A., and Abbott C. M. (2001). Comparative genomic analysis of genes encoding translation elongation factor 1B(alpha) in human and mouse shows EEF1B1 to be a recent retrotransposition event. *Genomics* **77:** 145-8.

Chan M. F., Liang G., and Jones P. A. (2000). Relationship between transcription and DNA methylation. *Curr Top Microbiol Immunol* **249:** 75-86.

Chang Y. J., McCabe R. T., Rennert H., Budarf M. L., Sayegh R., Emanuel B. S., Skolnick P., and Strauss J. F., 3rd (1992). The human "peripheral-type" benzodiazepine receptor: regional mapping of the gene and characterization of the receptor expressed from cDNA. *DNA Cell Biol* **11:** 471-80.

Cheung V. G., Nowak N., Jang W., Kirsch I. R., Zhao S., Chen X. N., Furey T. S., Kim U. J., Kuo W. L., Olivier M., Conroy J., Kasprzyk A., Massa H., Yonescu R., Sait S., Thoreen C., Snijders A., Lemyre E., Bailey J. A., Bruzel A., Burrill W. D., Clegg S. M., Collins S., Dhami P., Friedman C., Han C. S., Herrick S., Lee J., Ligon A. H., Lowry S., Morley M., Narasimhan S., Osoegawa K., Peng Z., Plajzer-Frick I., Quade B. J., Scott D., Sirotkin K., Thorpe A. A., Gray J. W., Hudson J., Pinkel D., Ried T., Rowen L., Shen-Ong G. L., Strausberg R. L., Birney E., Callen D. F., Cheng J. F., Cox D. R., Doggett N. A., Carter N. P., Eichler E. E., Haussler D., Korenberg J. R., Morton C. C., Albertson D., Schuler G., de Jong P. J., and Trask B. J. (2001). Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409:** 953-8.

Chittenden T., Flemington C., Houghton A. B., Ebb R. G., Gallo G. J., Elangovan B., Chinnadurai G., and Lutz R. J. (1995). A conserved domain in Bak, distinct from BH1 and BH2, mediates cell death and protein binding functions. *Embo J* **14:** 5589-96.

Chomczynski P., and Sacchi N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate- phenol-chloroform extraction. *Anal Biochem* **162:** 156-9.

Chumakov I., Rigault P., Guillou S., Ougen P., Billaut A., Guasconi G., Gervy P., LeGall I., Soularue P., Grinas L., and et al. (1992). Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* **359:** 380-7.

Chumakov I. M., Rigault P., Le Gall I., Bellanne-Chantelot C., Billault A., Guillou S., Soularue P., Guasconi G., Poullier E., Gros I., and et al. (1995). A YAC contig map of the human genome. *Nature* **377:** 175-297.

Church D. M., Banks L. T., Rogers A. C., Graw S. L., Housman D. E., Gusella J. F., and Buckler A. J. (1993). Identification of human chromosome 9 specific genes using exon amplification. *Hum Mol Genet* **2:** 1915-20.

Claverie J. M. (1997). Computational methods for the identification of genes in vertebrate genomic sequences. *Hum Mol Genet* **6:** 1735-44.

Cleves A. E. (1997). Protein transports: the nonclassical ins and outs. *Curr Biol* **7:** R318-20.

Coldwell M. J., Mitchell S. A., Stoneley M., MacFarlane M., and Willis A. E. (2000). Initiation of Apaf-1 translation by internal ribosome entry. *Oncogene* **19:** 899-905.

Colgan D. F., and Manley J. L. (1997). Mechanism and regulation of mRNA polyadenylation. *Genes Dev* **11:** 2755-66.

Collins F. S., and Weissman S. M. (1984). The molecular genetics of human hemoglobin. *Prog Nucleic Acid Res Mol Biol* **31:** 315-462.

Collins J., and Hohn B. (1978). Cosmids: a type of plasmid gene-cloning vector that is packageable in vitro in bacteriophage lambda heads. *Proc Natl Acad Sci U S A* **75:** 4242-6.

Collins J., Saari B., and Anderson P. (1987). Activation of a transposable element in the germ line but not the soma of Caenorhabditis elegans. *Nature* **328:** 726-8.

Collins J. E., Cole C. G., Smink L. J., Garrett C. L., Leversha M. A., Soderlund C. A., Maslen G. L., Everett L. A., Rice K. M., Coffey A. J., and et al. (1995). A high-density YAC contig map of human chromosome 22. *Nature* **377:** 367-79.

Cooper D. N., and Krawczak M. (1989). Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Hum Genet* **83:** 181-8.

Cooper P. R., Smilinich N. J., Day C. D., Nowak N. J., Reid L. H., Pearsall R. S., Reece M., Prawitt D., Landers J., Housman D. E., Winterpacht A., Zabel B. U., Pelletier J., Weissman B. E., Shows T. B., and Higgins M. J. (1998). Divergently transcribed overlapping genes expressed in liver and kidney and located in the 11p15.5 imprinted domain. *Genomics* **49:** 38-51.

Corpet F., Servant F., Gouzy J., and Kahn D. (2000). ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res* **28:** 267-9.

Coulondre C., Miller J. H., Farabaugh P. J., and Gilbert W. (1978). Molecular basis of base substitution hotspots in Escherichia coli. *Nature* **274:** 775-80.

Coulson A. (1996). The Caenorhabditis elegans genome project. C. elegans Genome Consortium. *Biochem Soc Trans* **24:** 289-91.

Cousin H., Gaultier A., Bleux C., Darribere T., and Alfandari D. (2000). PACSIN2 is a regulator of the metalloprotease/disintegrin ADAM13. *Dev Biol* **227:** 197-210.

Cox D. R. (1992). Radiation hybrid mapping. *Cytogenet Cell Genet* **59:** 80-1.

Cox D. R., Burmeister M., Price E. R., Kim S., and Myers R. M. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* **250:** 245-50.

Craig J. M., and Bickmore W. A. (1993). Chromosome bands--flavours to savour. *Bioessays* **15:** 349-54.

Creighton T. E. (1993). Proteins. Second edition. W. H. Freeman and Company.

Crick F. (1957). On protein synthesis. *Symp. Soc. Exp. Biol*. **12**: 138-163.

Crollius H. R., Jaillon O., Dasilva C., Ozouf-Costaz C., Fizames C., Fischer C., Bouneau L., Billault A., Quetier F., Saurin W., Bernot A., and Weissenbach J. (2000). Characterization and repeat analysis of the compact genome of the freshwater pufferfish Tetraodon nigroviridis. *Genome Res* **10:** 939-49.

Cross S. H., and Bird A. P. (1995). CpG islands and genes. *Curr Opin Genet Dev* **5:** 309-14.

Cross S. H., Clark V. H., Simmen M. W., Bickmore W. A., Maroon H., Langford C. F., Carter N. P., and Bird A. P. (2000). CpG island libraries from human chromosomes 18 and 22: landmarks for novel genes. *Mamm Genome* **11:** 373-83.

Das M., Burge C. B., Park E., Colinas J., and Pelletier J. (2001). Assessment of the total number of human transcription units. *Genomics* **77:** 71-8.

de Souza S. J., Camargo A. A., Briones M. R., Costa F. F., Nagai M. A., Verjovski-Almeida S., Zago M. A., Andrade L. E., Carrer H., El-Dorry H. F., Espreafico E. M., Habr-Gama A., Giannella-Neto D., Goldman G. H., Gruber A., Hackel C., Kimura E. T., Maciel R. M., Marie S. K., Martins E. A., Nobrega M. P., Paco-Larson M. L., Pardini M. I., Pereira G. G., Pesquero J. B., Rodrigues V., Rogatto S. R., da Silva I. D., Sogayar M. C., de Fatima Sonati M., Tajara E. H., Valentini S. R., Acencio M., Alberto F. L., Amaral M. E., Aneas I., Bengtson M. H., Carraro D. M., Carvalho A. F., Carvalho L. H., Cerutti J. M., Correa M. L., Costa M. C., Curcio C., Gushiken T., Ho P. L., Kimura E., Leite L. C., Maia G., Majumder P., Marins M., Matsukuma A., Melo A. S., Mestriner C. A., Miracca E. C., Miranda D. C., Nascimento A. N., Nobrega F. G., Ojopi E. P., Pandolfi J. R., Pessoa L. G., Rahal P., Rainho C. A., da Ros N., de Sa R. G., Sales M. M., da Silva N. P., Silva T. C., da Silva W., Jr., Simao D. F., Sousa J. F., Stecconi D., Tsukumo F., Valente V., Zalcbeg H., Brentani R. R., Reis F. L., Dias-Neto E., and Simpson A. J. (2000). Identification of human chromosome 22

transcribed sequences with ORF expressed sequence tags. *Proc Natl Acad Sci U S A* **97:** 12690-3.

Deak P., Omar M. M., Saunders R. D., Pal M., Komonyi O., Szidonya J., Maroy P., Zhang Y., Ashburner M., Benos P., Savakis C., Siden-Kiamos I., Louis C., Bolshakov V. N., Kafatos F. C., Madueno E., Modolell J., and Glover D. M. (1997). P-element insertion alleles of essential genes on the third chromosome of Drosophila melanogaster: correlation of physical and cytogenetic maps in chromosomal region 86E-87F. *Genetics* **147:** 1697-722.

Deloukas P., Schuler G. D., Gyapay G., Beasley E. M., Soderlund C., Rodriguez-Tome P., Hui L., Matise T. C., McKusick K. B., Beckmann J. S., Bentolila S., Bihoreau M., Birren B. B., Browne J., Butler A., Castle A. B., Chiannilkulchai N., Clee C., Day P. J., Dehejia A., Dibling T., Drouot N., Duprat S., Fizames C., Bentley D. R., and et al. (1998). A physical map of 30,000 human genes. *Science* **282:** 744-6.

den Dunnen J. T., van Neck J. W., Cremers F. P., Lubsen N. H., and Schoenmakers J. G. (1989). Nucleotide sequence of the rat gamma-crystallin gene region and comparison with an orthologous human region. *Gene* **78:** 201-13.

Deng A. Y., and Rapp J. P. (1994). Evaluation of the angiotensin II receptor AT1B gene as a candidate gene for blood pressure. *J Hypertens* **12:** 1001-6.

DeRisi J. L., Iyer V. R., and Brown P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278:** 680-6.

Dib C., Faure S., Fizames C., Samson D., Drouot N., Vignal A., Millasseau P., Marc S., Hazan J., Seboun E., Lathrop M., Gyapay G., Morissette J., and Weissenbach J. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380:** 152-4.

Dietrich W. F., Miller J., Steen R., Merchant M. A., Damron-Boles D., Husain Z., Dredge R., Daly M. J., Ingalls K. A., O'Connor T. J., and et al. (1996). A comprehensive genetic map of the mouse genome. *Nature* **380:** 149-52.

Ding D. Q., Tomita Y., Yamamoto A., Chikashige Y., Haraguchi T., and Hiraoka Y. (2000). Large-scale screening of intracellular protein localization in living fission yeast cells by the use of a GFP-fusion genomic DNA library. *Genes Cells* **5:** 169-90.

Doggett N. A., Goodwin L. A., Tesmer J. G., Meincke L. J., Bruce D. C., Clark L. M., Altherr M. R., Ford A. A., Chi H. C., Marrone B. L., and et al. (1995). An integrated physical map of human chromosome 16. *Nature* **377:** 335-65.

Dongre A. R., Eng J. K., and Yates J. R., 3rd (1997). Emerging tandem-mass-spectrometry techniques for the rapid identification of proteins. *Trends Biotechnol* **15:** 418-25.

Donis-Keller H., Green P., Helms C., Cartinhour S., Weiffenbach B., Stephens K., Keith T. P., Bowden D. W., Smith D. R., Lander E. S., and et al. (1987). A genetic linkage map of the human genome. *Cell* **51:** 319-37.

Dulbecco R. (1986). A turning point in cancer research: sequencing the human genome. *Science* **231:** 1055-6.

Dunham I., Shimizu N., Roe B. A., Chissoe S., Hunt A. R., Collins J. E., Bruskiewich R., Beare D. M., Clamp M., Smink L. J., Ainscough R., Almeida J. P., Babbage A., Bagguley C., Bailey J., Barlow K., Bates K. N., Beasley O., Bird C. P., Blakey S., Bridgeman A. M., Buck D., Burgess J., Burrill W. D., O'Brien K. P., and et al. (1999). The DNA sequence of human chromosome 22. *Nature* **402:** 489-95.

Duret L., Mouchiroud D., and Gautier C. (1995). Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J Mol Evol* **40:** 308-17.

Duyk G. M., Kim S. W., Myers R. M., and Cox D. R. (1990). Exon trapping: a genetic screen to identify candidate transcribed sequences in cloned mammalian genomic DNA. *Proc Natl Acad Sci U S A* **87:** 8995-9.

Eisenhaber F., Persson B., and Argos P. (1995). Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence. *Crit Rev Biochem Mol Biol* **30:** 1-94.

Emanuelsson O., Nielsen H., Brunak S., and von Heijne G. (2000). Predicting subcellular localization of proteins based on their N- terminal amino acid sequence. *J Mol Biol* **300:** 1005-16.

Epp T. A., Wang R., Sole M. J., and Liew C. C. (1995). Concerted evolution of mammalian cardiac myosin heavy chain genes. *J Mol Evol* **41:** 284-92.

Eppig J. T., and Nadeau J. H. (1995). Comparative maps: the mammalian jigsaw puzzle. *Curr Opin Genet Dev* **5:** 709-16.

Etzold T., Ulyanov A., and Argos P. (1996). SRS: information retrieval system for molecular biology data banks. *Methods Enzymol* **266:** 114-28.

Falany C. N., Xie X., Wang J., Ferrer J., and Falany J. L. (2000). Molecular cloning and expression of novel sulphotransferase-like cDNAs from human and rat brain. *Biochem J* **346 Pt 3:** 857-64.

Favello A., Hillier L., and Wilson R. K. (1995). Genomic DNA sequencing methods. *Methods Cell Biol* **48:** 551-69.

Feinberg A. P., and Vogelstein B. (1983). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal Biochem* **132:** 6-13.

Fichant G. A., and Burks C. (1991). Identifying potential tRNA genes in genomic DNA sequences. *J Mol Biol* **220:** 659-71.

Fitch W. M. (1970). Distinguishing homologous from analogous proteins. *Syst Zool* **19:** 99-113.

Fodor S. P., Read J. L., Pirrung M. C., Stryer L., Lu A. T., and Solas D. (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science* **251:** 767-73.

Foote S., Vollrath D., Hilton A., and Page D. C. (1992). The human Y chromosome: overlapping DNA clones spanning the euchromatic region. *Science* **258:** 60-6.

Footz T. K., Brinkman-Mills P., Banting G. S., Maier S. A., Riazi M. A., Bridgland L., Hu S., Birren B., Minoshima S., Shimizu N., Pan H., Nguyen T., Fang F., Fu Y., Ray L., Wu H., Shaull S., Phan S., Yao Z., Chen F., Huan A., Hu P., Wang Q., Loh P., Qi S., Roe B. A., and McDermid H. E. (2001). Analysis of the cat eye syndrome critical region in humans and the region of conserved synteny in mice: a search for candidate genes at or near the human chromosome 22 pericentromere. *Genome Res* **11:** 1053-70.

Foy C., Newton V., Wellesley D., Harris R., and Read A. P. (1990). Assignment of the locus for Waardenburg syndrome type I to human chromosome 2q37 and possible homology to the Splotch mouse. *Am J Hum Genet* **46:** 1017-23.

Frazer K. A., Sheehan J. B., Stokowski R. P., Chen X., Hosseini R., Cheng J. F., Fodor S. P., Cox D. R., and Patil N. (2001). Evolutionarily conserved sequences on human chromosome 21. *Genome Res* **11:** 1651-9.

Galtier N., Gouy M., and Gautier C. (1996). SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* **12:** 543-8.

Garcia-Mata R., Bebok Z., Sorscher E. J., and Sztul E. S. (1999). Characterization and dynamics of aggresome formation by a cytosolic GFP- chimera. *J Cell Biol* **146:** 1239-54.

Gardiner K. (1995). Human genome organization. *Curr Opin Genet Dev* **5:** 315-22.

Gardiner K. (1996). Base composition and gene distribution: critical patterns in mammalian genome organization. *Trends Genet* **12:** 519-24.

Gardiner K., and Mural R. J. (1995). Getting the message: identifying transcribed sequences. *Trends Genet* **11:** 77-9.

Gardiner-Garden M., and Frommer M. (1987). CpG islands in vertebrate genomes. *J Mol Biol* **196:** 261-82.

Gautheret D., Poirot O., Lopez F., Audic S., and Claverie J. M. (1998). Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res* **8:** 524-30.

Gelfand M. S. (1995). Prediction of function in DNA sequence analysis. *J Comput Biol* **2:** 87-115.

Gemmill R. M., Chumakov I., Scott P., Waggoner B., Rigault P., Cypser J., Chen Q., Weissenbach J., Gardiner K., Wang H., and et al. (1995). A second-generation YAC contig map of human chromosome 3. *Nature* **377:** 299-319.

Gianfrancesco F., Esposito T., Ruini L., Houlgatte R., Nagaraja R., D'Esposito M., Rocchi M., Auffray C., Schlessinger D., D'Urso M., and Forabosco A. (1997). Mapping of 59 EST gene markers in 31 intervals spanning the human X chromosome. *Gene* **187:** 179-84.

Goffeau A., Barrell B. G., Bussey H., Davis R. W., Dujon B., Feldmann H., Galibert F., Hoheisel J. D., Jacq C., Johnston M., Louis E. J., Mewes H. W., Murakami Y., Philippsen P., Tettelin H., and Oliver S. G. (1996). Life with 6000 genes. *Science* **274:** 546, 563-7.

Goss S. J., and Harris H. (1975). New method for mapping genes in human chromosomes. *Nature* **255:** 680-4.

Gottgens B., Barton L. M., Gilbert J. G., Bench A. J., Sanchez M. J., Bahn S., Mistry S., Grafham D., McMurray A., Vaudin M., Amaya E., Bentley D. R., Green A. R., and Sinclair A. M. (2000). Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat Biotechnol* **18:** 181-6.

Goulding M. D., Chalepakis G., Deutsch U., Erselius J. R., and Gruss P. (1991). Pax-3, a novel murine DNA binding protein expressed during early neurogenesis. *Embo J* **10:** 1135-47.

Graves J. A. (1996). Mammals that break the rules: genetics of marsupials and monotremes. *Annu Rev Genet* **30:** 233-60.

Gray N. K., and Wickens M. (1998). Control of translation initiation in animals. *Annu Rev Cell Dev Biol* **14:** 399-458.

Green E. D., and Olson M. V. (1990). Chromosomal region of the cystic fibrosis gene in yeast artificial chromosomes: a model for human genome mapping. *Science* **250:** 94-8.

Green E. D., Riethman H. C., Dutchik J. E., and Olson M. V. (1991). Detection and characterization of chimeric yeast artificial-chromosome clones. *Genomics* **11:** 658-69.

Gregory S. G., Howell G. R., and Bentley D. R. (1997). Genome mapping by fluorescent fingerprinting. *Genome Res* **7:** 1162-8.

Guigo R., Agarwal P., Abril J. F., Burset M., and Fickett J. W. (2000). An assessment of gene prediction accuracy in large DNA sequences. *Genome Res* **10:** 1631-42.

Gumucio D. L., Wiebauer K., Caldwell R. M., Samuelson L. C., and Meisler M. H. (1988). Concerted evolution of human amylase genes. *Mol Cell Biol* **8:** 1197-205.

Gyapay G., Morissette J., Vignal A., Dib C., Fizames C., Millasseau P., Marc S., Bernardi G., Lathrop M., and Weissenbach J. (1994). The 1993-94 Genethon human genetic linkage map. *Nat Genet* **7:** 246-339.

Gyapay G., Schmitt K., Fizames C., Jones H., Vega-Czarny N., Spillett D., Muselet D., Prud'Homme J. F., Dib C., Auffray C., Morissette J., Weissenbach J., and Goodfellow P. N. (1996). A radiation hybrid map of the human genome. *Hum Mol Genet* **5:** 339-46.

Gygi S. P., Rochon Y., Franza B. R., and Aebersold R. (1999). Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* **19:** 1720-30.

Haig D. (1999). A brief history of human autosomes. *Philos Trans R Soc Lond B Biol Sci* **354:** 1447-70.

Han J., Sabbatini P., and White E. (1996). Induction of apoptosis by human Nbk/Bik, a BH3-containing protein that interacts with E1B 19K. *Mol Cell Biol* **16:** 5857-64.

Hardison R. C., Oeltjen J., and Miller W. (1997). Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res* **7:** 959-66.

Heller R. A., Schena M., Chai A., Shalon D., Bedilion T., Gilmore J., Woolley D. E., and Davis R. W. (1997). Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci U S A* **94:** 2150-5.

Hess G. F., Drong R. F., Weiland K. L., Slightom J. L., Sclafani R. A., and Hollingsworth R. E. (1998). A human homolog of the yeast CDC7 gene is overexpressed in some tumors and transformed cell lines. *Gene* **211:** 133-40.

Hieter P., and Boguski M. (1997). Functional genomics: it's all how you read it. *Science* **278:** 601-2.

Hillier L. D., Lennon G., Becker M., Bonaldo M. F., Chiapelli B., Chissoe S., Dietrich N., DuBuque T., Favello A., Gish W., Hawkins M., Hultman M., Kucaba T., Lacy M., Le M., Le N., Mardis E., Moore B., Morris M., Parsons J., Prange C., Rifkin L., Rohlfing T., Schellenberg K., Marra M., and et al. (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Res* **6:** 807-28.

Hirsch T., Decaudin D., Susin S. A., Marchetti P., Larochette N., Resche-Rigon M., and Kroemer G. (1998). PK11195, a ligand of the mitochondrial benzodiazepine receptor, facilitates the induction of apoptosis and reverses Bcl-2-mediated cytoprotection. *Exp Cell Res* **241:** 426-34.

Hodgson C. P., and Fisk R. Z. (1987). Hybridization probe size control: optimized 'oligolabelling'. *Nucleic Acids Res* **15:** 6295.

Hofmann K., Bucher P., Falquet L., and Bairoch A. (1999). The PROSITE database, its status in 1999. *Nucleic Acids Res* **27:** 215-9.

Hogenesch J. B., Ching K. A., Batalov S., Su A. I., Walker J. R., Zhou Y., Kay S. A., Schultz P. G., and Cooke M. P. (2001). A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106:** 413-5.

Hoja M. R., Wahlestedt C., and Hoog C. (2000). A visual intracellular classification strategy for uncharacterized human proteins. *Exp Cell Res* **259:** 239-46.

Hood L., Rowen L., and Koop B. F. (1995). Human and mouse T-cell receptor loci: genomics, evolution, diversity, and serendipity. *Ann N Y Acad Sci* **758:** 390-412.

Houlgatte R., Mariage-Samson R., Duprat S., Tessier A., Bentolila S., Lamy B., and Auffray C. (1995). The Genexpress Index: a resource for gene discovery and the genic map of the human genome. *Genome Res* **5:** 272-304.

Huang X. Q., Hardison R. C., and Miller W. (1990). A space-efficient algorithm for local similarities. *Comput Appl Biosci* **6:** 373-81.

Huang S. H., Yang A. Y. and Holcenberg J. (1993). Amplification of gene ends from gene libraries by polymerase chain reaction with single-sided specificity. *Methods in Molecular Biology, PCR Protocols: Current Methods and Applications* (ed. White B. A.) 357-363 (Humana Press, Totowa, New Jersey).

Hubbard T., and Birney E. (2000). Open annotation offers a democratic solution to genome sequencing. *Nature* **403:** 825.

Hudson T. J., Engelstein M., Lee M. K., Ho E. C., Rubenfield M. J., Adams C. P., Housman D. E., and Dracopoli N. C. (1992). Isolation and chromosomal assignment of 100 highly informative human simple sequence repeat polymorphisms. *Genomics* **13:** 622-9.

Hudson T. J., Stein L. D., Gerety S. S., Ma J., Castle A. B., Silva J., Slonim D. K., Baptista R., Kruglyak L., Xu S. H., and et al. (1995). An STS-based map of the human genome. *Science* **270:** 1945-54.

Hughes D. C. (2000). MIRs as agents of mammalian gene evolution. *Trends Genet* **16:** 60-2.

Hurst L. D., and Eyre-Walker A. (2000). Evolutionary genomics: reading the bands. *Bioessays* **22:** 105-7.

Huynen M. A., and Bork P. (1998). Measuring genome evolution. *Proc Natl Acad Sci U S A* **95:** 5849-56.

Ioannou P. A., Amemiya C. T., Garnes J., Kroisel P. M., Shizuya H., Chen C., Batzer M. A., and de Jong P. J. (1994). A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nat Genet* **6:** 84-9.

Jackson R. J., and Kaminski A. (1995). Internal initiation of translation in eukaryotes: the picornavirus paradigm and beyond. *Rna* **1:** 985-1000.

Jang W., Hua A., Spilson S. V., Miller W., Roe B. A., and Meisler M. H. (1999). Comparative sequence of human and mouse BAC clones from the mnd2 region of chromosome 2p13. *Genome Res* **9:** 53-61.

Jareborg N., Birney E., and Durbin R. (1999). Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res* **9:** 815-24.

Jones C. T., Morrice D. R., Paton I. R., and Burt D. W. (1997). Gene homologs on human chromosome 15q21-q26 and a chicken microchromosome identify a new conserved segment. *Mamm Genome* **8:** 436-40.

Jurka J., and Kapitonov V. V. (1999). Sectorial mutagenesis by transposable elements. *Genetica* **107:** 239-48.

Kamb A., Wang C., Thomas A., DeHoff B. S., Norris F. H., Richardson K., Rine J., Skolnick M. H., and Rosteck P. R., Jr. (1995). Software trapping: a strategy for finding genes in large genomic regions. *Comput Biomed Res* **28:** 140-53.

Kawai J., Shinagawa A., Shibata K., Yoshino M., Itoh M., Ishii Y., Arakawa T., Hara A., Fukunishi Y., Konno H., Adachi J., Fukuda S., Aizawa K., Izawa M., Nishi K., Kiyosawa H., Kondo S., Yamanaka I., Saito T., Okazaki Y., Gojobori T., Bono H., Kasukawa T., Saito R., Kadota K., Matsuda H. A., Ashburner M., Batalov S., Casavant T., Fleischmann W., Gaasterland T., Gissi C., King B., Kochiwa H., Kuehl P., Lewis S., Matsuo Y., Nikaido I., Pesole G., Quackenbush J., Schriml L. M., Staubli F., Suzuki R., Tomita M., Wagner L., Washio T., Sakai K., Okido T., Furuno M., Aono H., Baldarelli R., Barsh G., Blake J., Boffelli D., Bojunga N., Carninci P., de Bonaldo M. F., Brownstein M. J., Bult C., Fletcher C., Fujita M., Gariboldi M., Gustincich S., Hill D., Hofmann M., Hume D. A., Kamiya M., Lee N. H., Lyons P., Marchionni L., Mashima J., Mazzarelli J., Mombaerts P., Nordone P., Ring B., Ringwald M., Rodriguez I., Sakamoto N., Sasaki H., Sato K., Schonbach C., Seya T., Shibata Y., Storch K. F., Suzuki H., Toyo-oka K., Wang K. H., Weitz C., Whittaker C., Wilming L., Wynshaw-Boris A., Yoshida K., Hasegawa Y., Kawaji H., Kohtsuki S., and Hayashizaki Y. (2001). Functional annotation of a full-length mouse cDNA collection. *Nature* **409:** 685-90.

Keusch J. J., Manzella S. M., Nyame K. A., Cummings R. D., and Baenziger J. U. (2000). Cloning of Gb3 synthase, the key enzyme in globo-series glycosphingolipid synthesis, predicts a family of alpha 1, 4- glycosyltransferases conserved in plants, insects, and mammals. *J Biol Chem* **275:** 25315-21.

Khan A. S., Wilcox A. S., Polymeropoulos M. H., Hopkins J. A., Stevens T. J., Robinson M., Orpana A. K., and Sikela J. M. (1992). Single pass sequencing and physical and genetic mapping of human brain cDNAs. *Nat Genet* **2:** 180-5.

Khorana H. G., Buchi H., Ghosh H., Gupta N., Jacob T. M., Kossel H., Morgan R., Narang S. A., Ohtsuka E., and Wells R. D. (1966). Polynucleotide synthesis and the genetic code. *Cold Spring Harb Symp Quant Biol* **31:** 39-49.

Kim U. J., Shizuya H., de Jong P. J., Birren B., and Simon M. I. (1992). Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res* **20:** 1083-5.

Kojima Y., Fukumoto S., Furukawa K., Okajima T., Wiels J., Yokoyama K., Suzuki Y., Urano T., and Ohta M. (2000). Molecular cloning of globotriaosylceramide/CD77 synthase, a glycosyltransferase that initiates the synthesis of globo series glycosphingolipids. *J Biol Chem* **275:** 15152-6.

Koop B. F. (1995). Human and rodent DNA sequence comparisons: a mosaic model of genomic evolution. *Trends Genet* **11:** 367-71.

Koop B. F., and Hood L. (1994). Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nat Genet* **7:** 48-53.

Korn B., Sedlacek Z., Manca A., Kioschis P., Konecki D., Lehrach H., and Poustka A. (1992). A strategy for the selection of transcribed sequences in the Xq28 region. *Hum Mol Genet* **1:** 235-42.

Kozak M. (1980). Evaluation of the "scanning model" for initiation of protein synthesis in eucaryotes. *Cell* **22:** 7-8.

Kozak M. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res* **15:** 8125-48.

Kozak M. (1999). Initiation of translation in prokaryotes and eukaryotes. *Gene* **234:** 187-208.

Kozak M. (2000). Do the 5'untranslated domains of human cDNAs challenge the rules for initiation of translation (or is it vice versa)? *Genomics* **70:** 396-406.

Krauter K., Montgomery K., Yoon S. J., LeBlanc-Straceski J., Renault B., Marondel I., Herdman V., Cupelli L., Banks A., Lieman J., and et al. (1995). A second-generation YAC contig map of human chromosome 12. *Nature* **377:** 321-33.

Kremer E. J., et al. (1991). Mapping of DNA instability ath the fragil X to a trinucleotide repeat sequence p(CCG)n. *Science* **252**, 1711-4.

Krichevsky A. M., Metzer E., and Rosen H. (1999). Translational control of specific genes during differentiation of HL-60 cells. *J Biol Chem* **274:** 14295-305.

Kruglyak S., Durrett R. T., Schug M. D., and Aquadro C. F. (1998). Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci U S A* **95:** 10774-8.

Kulp D., Haussler D., Reese M. G., and Eeckman F. H. (1996). A generalized hidden Markov model for the recognition of human genes in DNA. *Proc Int Conf Intell Syst Mol Biol* **4:** 134-42.

Kuster B., and Mann M. (1998). Identifying proteins and post-translational modifications by mass spectrometry. *Curr Opin Struct Biol* **8:** 393-400.

Lamerdin J. E., Montgomery M. A., Stilwagen S. A., Scheidecker L. K., Tebbs R. S., Brookman K. W., Thompson L. H., and Carrano A. V. (1995). Genomic sequence comparison of the human and mouse XRCC1 DNA repair gene regions. *Genomics* **25:** 547-54.

Lamerdin J. E., Stilwagen S. A., Ramirez M. H., Stubbs L., and Carrano A. V. (1996). Sequence analysis of the ERCC2 gene regions in human, mouse, and hamster reveals three linked genes. *Genomics* **34:** 399-409.

Lander E. S. (1996). The new genomics: global views of biology. *Science* **274:** 536-9.

Lander E. S., Linton L. M., Birren B., Nusbaum C., Zody M. C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J.

P., Miranda C., Morris W., Naylor J., Raymond C., Rosetti M., Santos R., Sheridan A., Sougnez C., Stange-Thomann N., Stojanovic N., Subramanian A., Wyman D., Rogers J., Sulston J., Ainscough R., Beck S., Bentley D., Burton J., Clee C., Carter N., Coulson A., Deadman R., Deloukas P., Dunham A., Dunham I., Durbin R., French L., Grafham D., Gregory S., Hubbard T., Humphray S., Hunt A., Jones M., Lloyd C., McMurray A., Matthews L., Mercer S., Milne S., Mullikin J. C., Mungall A., Plumb R., Ross M., Shownkeen R., Sims S., Waterston R. H., Wilson R. K., Hillier L. W., McPherson J. D., Marra M. A., Mardis E. R., Fulton L. A., Chinwalla A. T., Pepin K. H., Gish W. R., Chissoe S. L., Wendl M. C., Delehaunty K. D., Miner T. L., Delehaunty A., Kramer J. B., Cook L. L., Fulton R. S., Johnson D. L., Minx P. J., Clifton S. W., Hawkins T., Branscomb E., Predki P., Richardson P., Wenning S., Slezak T., Doggett N., Cheng J. F., Olsen A., Lucas S., Elkin C., Uberbacher E., Frazier M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Larsen F., Gundersen G., Lopez R., and Prydz H. (1992). CpG islands as gene markers in the human genome. *Genomics* **13:** 1095-107.

Lee L. G., Connell C. R., Woo S. L., Cheng R. D., McArdle B. F., Fuller C. W., Halloran N. D., and Wilson R. K. (1992). DNA sequencing with dye-labeled terminators and T7 DNA polymerase: effect of dyes and dNTPs on incorporation of dye-terminators and probability analysis of termination fragments. *Nucleic Acids Res* **20:** 2471-83.

Lennon G., Auffray C., Polymeropoulos M., and Soares M. B. (1996). The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* **33:** 151-2.

Lewin B. (1997). Genes VI. Oxford University Press and Cell Press.

Li W. (2001). Delineating relative homogeneous G+C domains in DNA sequences. *Gene* **276:** 57-72.

Liang P., and Pardee A. B. (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257:** 967-71.

Litt M., and Luty J. A. (1989). A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am J Hum Genet* **44:** 397-401.

Liu C. C., Simonsen C. C., and Levinson A. D. (1984). Initiation of translation at internal AUG codons in mammalian cells. *Nature* **309:** 82-5.

Liu J., and Rost B. (2001). Comparing function and structure between entire proteomes. *Protein Sci* **10:** 1970-9.

Loots G. G., Locksley R. M., Blankespoor C. M., Wang Z. E., Miller W., Rubin E. M., and Frazer K. A. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136-40.

Lovett M., Kere J., and Hinton L. M. (1991). Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proc Natl Acad Sci U S A* **88:** 9628-32.

Lund J., Chen F., Hua A., Roe B., Budarf M., Emanuel B. S., and Reeves R. H. (2000). Comparative sequence analysis of 634 kb of the mouse chromosome 16 region of conserved synteny with the human velocardiofacial syndrome region on chromosome 22q11.2. *Genomics* **63:** 374-83.

Lundin L. G. (1993). Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* **16:** 1-19.

Lupas A. (1996). Coiled coils: new structures and new functions. *Trends Biochem Sci* **21:** 375-82.

Lutz-Freyermuth C., Query C. C., and Keene J. D. (1990). Quantitative determination that one of two potential RNA-binding domains of the A protein component of the U1 small

nuclear ribonucleoprotein complex binds with high affinity to stem-loop II of U1 RNA. *Proc Natl Acad Sci U S A* **87:** 6393-7.

Maglott D. R., Katz K. S., Sicotte H., and Pruitt K. D. (2000). NCBI's LocusLink and RefSeq. *Nucleic Acids Res* **28:** 126-8.

Makalowski W., and Boguski M. S. (1998). Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc Natl Acad Sci U S A* **95:** 9407-12.

Marra M. A., Kucaba T. A., Dietrich N. L., Green E. D., Brownstein B., Wilson R. K., McDonald K. M., Hillier L. W., McPherson J. D., and Waterston R. H. (1997). High throughput fingerprint analysis of large-insert clones. *Genome Res* **7:** 1072-84.

Martindale D. W., Wilson M. D., Wang D., Burke R. D., Chen X., Duronio V., and Koop B. F. (2000). Comparative genomic sequence analysis of the Williams syndrome region (LIMK1-RFC2) of human chromosome 7q11.23. *Mamm Genome* **11:** 890-8.

Matsuo K., Clay O., Takahashi T., Silke J., and Schaffner W. (1993). Evidence for erosion of mouse CpG islands during mammalian evolution. *Somat Cell Mol Genet* **19:** 543-55.

Matsuura T., Yamagata T., Burgess D. L., Rasmussen A., Grewal R. P., Watase K., Khajavi M., McCall A. E., Davis C. F., Zu L., Achari M., Pulst S. M., Alonso E., Noebels J. L., Nelson D. L., Zoghbi H. Y., and Ashizawa T. (2000). Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nat Genet* **26:** 191-4.

Maxam A. M., and Gilbert W. (1977). A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74:** 560-4.

McPherson J. D., Marra M., Hillier L., Waterston R. H., Chinwalla A., Wallis J., Sekhon M., Wylie K., Mardis E. R., Wilson R. K., Fulton R., Kucaba T. A., Wagner-McPherson C., Barbazuk W. B., Gregory S. G., Humphray S. J., French L., Evans R. S., Bethel G., Whittaker A., Holden J. L., McCann O. T., Dunham A., Soderlund C., Scott C. E., Bentley D. R., Schuler G., Chen H. C., Jang W., Green E. D., Idol J. R., Maduro V. V., Montgomery K. T., Lee E., Miller A., Emerling S., Kucherlapati, Gibbs R., Scherer S., Gorrell J. H., Sodergren E., Clerc-Blankenburg K., Tabor P., Naylor S., Garcia D., de Jong P. J., Catanese J. J., Nowak N., Osoegawa K., Qin S., Rowen L., Madan A., Dors M., Hood L., Trask B., Friedman C., Massa H., Cheung V. G., Kirsch I. R., Reid T., Yonescu R., Weissenbach J., Bruls T., Heilig R., Branscomb E., Olsen A., Doggett N., Cheng J. F., Hawkins T., Myers R. M., Shang J., Ramirez L., Schmutz J., Velasquez O., Dixon K., Stone N. E., Cox D. R., Haussler D., Kent W. J., Furey T., Rogic S., Kennedy S., Jones S., Rosenthal A., Wen G., Schilhabel M., Gloeckner G., Nyakatura G., Siebert R., Schlegelberger B., Korenberg J., Chen X. N., Fujiyama A., Hattori M., Toyoda A., Yada T., Park H. S., Sakaki Y., Shimizu N., Asakawa S., et al. (2001). A physical map of the human genome. *Nature* **409:** 934-41.

Meisler M. H. (2001). Evolutionarily conserved noncoding DNA in the human genome: how much and what for? *Genome Res* **11:** 1617-8.

Merilainen J., Lehto V. P., and Wasenius V. M. (1997). FAP52, a novel, SH3 domain-containing focal adhesion protein. *J Biol Chem* **272:** 23278-84.

Merkulov G. V., and Boeke J. D. (1998). Libraries of green fluorescent protein fusions generated by transposition in vitro. *Gene* **222:** 213-22.

Mironov A. A., Fickett J. W., and Gelfand M. S. (1999). Frequent alternative splicing of human genes. *Genome Res* **9:** 1288-93.

Molnar A., and Georgopoulos K. (1994). The Ikaros gene encodes a family of functionally diverse zinc finger DNA-binding proteins. *Mol Cell Biol* **14:** 8292-303.

Monaco A. P., Neve R. L., Colletti-Feener C., Bertelson C. J., Kurnit D. M., and Kunkel L. M. (1986). Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. *Nature* **323:** 646-50.

Montgomery K. T., Lee E., Miller A., Lau S., Shim C., Decker J., Chiu D., Emerling S., Sekhon M., Kim R., Lenz J., Han J., Ioshikhes I., Renault B., Marondel I., Yoon S. J., Song K., Murty V. V., Scherer S., Yonescu R., Kirsch I. R., Ried T., McPherson J., Gibbs R., and Kucherlapati R. (2001). A high-resolution map of human chromosome 12. *Nature* **409:** 945-6.

Montzka K. A., and Steitz J. A. (1988). Additional low-abundance human small nuclear ribonucleoproteins: U11, U12, etc. *Proc Natl Acad Sci U S A* **85:** 8885-9.

Moore M. J., and Sharp P. A. (1993). Evidence for two active sites in the spliceosome provided by stereochemistry of pre-mRNA splicing. *Nature* **365:** 364-8.

Morgan J. G., Dolganov G. M., Robbins S. E., Hinton L. M., and Lovett M. (1992). The selective isolation of novel cDNAs encoded by the regions surrounding the human interleukin 4 and 5 genes. *Nucleic Acids Res* **20:** 5173-9.

Mori I., Benian G. M., Moerman D. G., and Waterston R. H. (1988). Transposable element Tc1 of Caenorhabditis elegans recognizes specific target sequences for integration. *Proc Natl Acad Sci U S A* **85:** 861-4.

Morton N. E. (1991). Parameters of the human genome. *Proc Natl Acad Sci U S A* **88:** 7474-6.

Mott R. (1997). EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput Appl Biosci* **13:** 477-8.

Mukherjee S., and Das S. K. (1989). Subcellular distribution of "peripheral type" binding sites for [3H]Ro5- 4864 in guinea pig lung. Localization to the mitochondrial inner membrane. *J Biol Chem* **264:** 16713-8.

Muller D., Thieke K., Burgin A., Dickmanns A., and Eilers M. (2000). Cyclin E-mediated elimination of p27 requires its interaction with the nuclear pore-associated protein mNPAP60. *Embo J* **19:** 2168-80.

Muller S., Stanyon R., O'Brien P. C., Ferguson-Smith M. A., Plesker R., and Wienberg J. (1999). Defining the ancestral karyotype of all primates by multidirectional chromosome painting between tree shrews, lemurs and humans. *Chromosoma* **108:** 393-400.

Mullikin J. C., Hunt S. E., Cole C. G., Mortimore B. J., Rice C. M., Burton J., Matthews L. H., Pavitt R., Plumb R. W., Sims S. K., Ainscough R. M., Attwood J., Bailey J. M., Barlow K., Bruskiewich R. M., Butcher P. N., Carter N. P., Chen Y., Clee C. M., Coggill P. C., Davies J., Davies R. M., Dawson E., Francis M. D., Joy A. A., Lamble R. G., Langford C. F., Macarthy J., Mall V., Moreland A., Overton-Larty E. K., Ross M. T., Smith L. C., Steward C. A., Sulston J. E., Tinsley E. J., Turney K. J., Willey D. L., Wilson G. D., McMurray A. A., Dunham I., Rogers J., and Bentley D. R. (2000). An SNP map of human chromosome 22. *Nature* **407:** 516-20.

Mungall A. J., Edwards C. A., Ranby S. A., Humphray S. J., Heathcott R. W., Clee C. M., East C. L., Holloway E., Butler A. P., Langford C. F., Gwilliam R., Rice K. M., Maslen G. L., Carter N. P., Ross M. T., Deloukas P., Bentley D. R., and Dunham I. (1996). Physical mapping of chromosome 6: a strategy for the rapid generation of sequence-ready contigs. *DNA Seq* **7:** 47-9.

Mungall A. J., Humphray S. J., Ranby S. A., Edwards C. A., Heathcott R. W., Clee C. M., Holloway E., Peck A. I., Harrison P., Green L. D., Butler A. P., Langford C. F., William R. G., Huckle E. J., Baron L., Smith A., Leversha M. A., Ramsey Y. H., Clegg S. M., Rice C. M., Maslen G. L., Hunt S. E., Scott C. E., Soderlund C. A., Dunham I., and et al. (1997). From long range mapping to sequence-ready contigs on human chromosome 6. *DNA Seq* **8:** 151-4.

Murakami K., Yubisui T., Takeshita M., and Miyata T. (1989). The NH2-terminal structures of human and rat liver microsomal NADH- cytochrome b5 reductases. *J Biochem (Tokyo)* **105:** 312-7.

Murray J. C., Buetow K. H., Weber J. L., Ludwigsen S., Scherpbier-Heddema T., Manion F., Quillen J., Sheffield V. C., Sunden S., Duyk G. M., and et al. (1994). A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC). *Science* **265:** 2049-54.

Nadeau J. H., and Sankoff D. (1998). Counting on comparative maps. *Trends Genet* **14:** 495-501.

Nagaraja R., Kere J., MacMillan S., Masisi M. J., Johnson D., Molini B. J., Halley G. R., Wein K., Trusgnich M., Eble B., and et al. (1994). Characterization of four human YAC libraries for clone size, chimerism and X chromosome sequence representation. *Nucleic Acids Res* **22:** 3406-11.

Nagaraja R., MacMillan S., Kere J., Jones C., Griffin S., Schmatz M., Terrell J., Shomaker M., Jermak C., Hott C., Masisi M., Mumm S., Srivastava A., Pilia G., Featherstone T., Mazzarella R., Kesterson S., McCauley B., Railey B., Burough F., Nowotny V., D'Urso M., States D., Brownstein B., and Schlessinger D. (1997). X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content. *Genome Res* **7:** 210-22.

Nagase T., Seki N., Tanaka A., Ishikawa K., and Nomura N. (1995). Prediction of the coding sequences of unidentified human genes. IV. The coding sequences of 40 new genes (KIAA0121-KIAA0160) deduced by analysis of cDNA clones from human cell line KG-1. *DNA Res* **2:** 167-74, 199-210.

Nakai A., and Ishikawa T. (2000). A nuclear localization signal is essential for stress-induced dimer-to- trimer transition of heat shock transcription factor 3. *J Biol Chem* **275:** 34665-71.

Nakai K., and Horton P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* **24:** 34-6.

Nielsen H., Engelbrecht J., Brunak S., and von Heijne G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* **10:** 1-6.

Nirenberg M., Caskey T., Marshall R., Brimacombe R., Kellogg D., Doctor B., Hatfield D., Levin J., Rottman F., Pestka S., Wilcox M., and Anderson F. (1966). The RNA code and protein synthesis. *Cold Spring Harb Symp Quant Biol* **31:** 11-24.

Nusbaum C., Slonim D. K., Harris K. L., Birren B. W., Steen R. G., Stein L. D., Miller J., Dietrich W. F., Nahf R., Wang V., Merport O., Castle A. B., Husain Z., Farino G., Gray D., Anderson M. O., Devine R., Horton L. T., Jr., Ye W., Wu X., Kouyoumjian V., Zemsteva I. S., Wu Y., Collymore A. J., Courtney D. F., and et al. (1999). A YAC-based physical map of the mouse genome. *Nat Genet* **22:** 388-93.

Oakey R. J., Watson M. L., and Seldin M. F. (1992). Construction of a physical map on mouse and human chromosome 1: comparison of 13 Mb of mouse and 11 Mb of human DNA. *Hum Mol Genet* **1:** 613-20.

Oeltjen J. C., Malley T. M., Muzny D. M., Miller W., Gibbs R. A., and Belmont J. W. (1997). Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res* **7:** 315-29.

O'Farrell P. H. (1975). High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **250:** 4007-21.

Okada N., and Hamada M. (1997). The 3' ends of tRNA-derived SINEs originated from the 3' ends of LINEs: a new example from the bovine genome. *J Mol Evol* **44:** S52-6.

Okubo K., Hori N., Matoba R., Niiyama T., Fukushima A., Kojima Y., and Matsubara K. (1992). Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat Genet* **2:** 173-9.

Olski T. M., Noegel A. A., and Korenbaum E. (2001). Parvin, a 42 kDa focal adhesion protein, related to the alpha-actinin superfamily. *J Cell Sci* **114:** 525-38.

Olson J. M., Ciliax B. J., Mancini W. R., and Young A. B. (1988). Presence of peripheral-type benzodiazepine binding sites on human erythrocyte membranes. *Eur J Pharmacol* **152:** 47-53.

Olson M., Hood L., Cantor C., and Botstein D. (1989). A common language for physical mapping of the human genome. *Science* **245:** 1434-5.

Olson M. V., Dutchik J. E., Graham M. Y., Brodeur G. M., Helms C., Frank M., MacCollin M., Scheinman R., and Frank T. (1986). Random-clone strategy for genomic restriction mapping in yeast. *Proc Natl Acad Sci U S A* **83:** 7826-30.

Orengo C. A., Todd A. E., and Thornton J. M. (1999). From protein structure to function. *Curr Opin Struct Biol* **9:** 374-82.

Osoegawa K., Tateno M., Woon P. Y., Frengen E., Mammoser A. G., Catanese J. J., Hayashizaki Y., and de Jong P. J. (2000). Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. *Genome Res* **10:** 116-28.

Ozols J., Carr S. A., and Strittmatter P. (1984). Identification of the NH2-terminal blocking group of NADH-cytochrome b5 reductase as myristic acid and the complete amino acid sequence of the membrane-binding domain. *J Biol Chem* **259:** 13349-54.

Pan T. C., Kluge M., Zhang R. Z., Mayer U., Timpl R., and Chu M. L. (1993). Sequence of extracellular mouse protein BM-90/fibulin and its calcium- dependent binding to other basement-membrane ligands. *Eur J Biochem* **215:** 733-40.

Parimoo S., Patanjali S. R., Shukla H., Chaplin D. D., and Weissman S. M. (1991). cDNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc Natl Acad Sci U S A* **88:** 9623-7.

Parola A. L., Stump D. G., Pepperl D. J., Krueger K. E., Regan J. W., and Laird H. E., 2nd (1991). Cloning and expression of a pharmacologically unique bovine peripheral- type benzodiazepine receptor isoquinoline binding protein. *J Biol Chem* **266:** 14082-7.

Pepperkok R., Simpson J. C., and Wiemann S. (2001). Being in the right location at the right time. *Genome Biol* **2**.

Peri S., and Pandey A. (2001). A reassessment of the translation initiation codon in vertebrates. *Trends Genet* **17:** 685-7.

Persson B., Zigler J. S., Jr., and Jornvall H. (1994). A super-family of medium-chain dehydrogenases/reductases (MDR). Sub- lines including zeta-crystallin, alcohol and polyol dehydrogenases, quinone oxidoreductase enoyl reductases, VAT-1 and other proteins. *Eur J Biochem* **226:** 15-22.

Piatigorsky J., and Wistow G. (1991). The recruitment of crystallins: new functions precede gene duplication. *Science* **252:** 1078-9.

Pichon B., Mercan D., Pouillon V., Christophe-Hobertus C., and Christophe D. (2000). A method for the large-scale cloning of nuclear proteins and nuclear targeting sequences on a functional basis. *Anal Biochem* **284:** 231-9.

Pietrini G., Carrera P., and Borgese N. (1988). Two transcripts encode rat cytochrome b5 reductase. *Proc Natl Acad Sci U S A* **85:** 7246-50.

Pletcher M. T., Roe B. A., Chen F., Do T., Do A., Malaj E., and Reeves R. H. (2000). Chromosome evolution: the junction of mammalian chromosomes in the formation of mouse chromosome 10. *Genome Res* **10:** 1463-7.

Prober J. M., Trainor G. L., Dam R. J., Hobbs F. W., Robertson C. W., Zagursky R. J., Cocuzza A. J., Jensen M. A., and Baumeister K. (1987). A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238:** 336-41.

Qiu Y., Cavelier L., Chiu S., Yang X., Rubin E., and Cheng J. F. (2001). Human and mouse ABCA1 comparative sequencing and transgenesis studies revealing novel regulatory sequences. *Genomics* **73:** 66-76.

Qualmann B., and Kelly R. B. (2000). Syndapin isoforms participate in receptor-mediated endocytosis and actin organization. *J Cell Biol* **148:** 1047-62.

Quandt K., Frech K., Karas H., Wingender E., and Werner T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* **23:** 4878-84.

Reese M. G., Hartzell G., Harris N. L., Ohler U., Abril J. F., and Lewis S. E. (2000). Genome annotation assessment in Drosophila melanogaster. *Genome Res* **10:** 483-501.

Reichhardt T. (1999). It's sink or swim as a tidal wave of data approaches. *Nature* **399:** 517-20.

Rettenberger G., Klett C., Zechner U., Bruch J., Just W., Vogel W., and Hameister H. (1995). ZOO-FISH analysis: cat and human karyotypes closely resemble the putative ancestral mammalian karyotype. *Chromosome Res* **3:** 479-86.

Rhodes M., Straw R., Fernando S., Evans A., Lacey T., Dearlove A., Greystrong J., Walker J., Watson P., Weston P., Kelly M., Taylor D., Gibson K., Mundy C., Bourgade F., Poirier C., Simon D., Brunialti A. L., Montagutelli X., Gu'enet J. L., Haynes A., and Brown S. D. (1998). A high-resolution microsatellite map of the mouse genome. *Genome Res* **8:** 531-42.

Riley J., Butler R., Ogilvie D., Finniear R., Jenner D., Powell S., Anand R., Smith J. C., and Markham A. F. (1990). A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. *Nucleic Acids Res* **18:** 2887-90.

Riond J., Mattei M. G., Kaghad M., Dumont X., Guillemot J. C., Le Fur G., Caput D., and Ferrara P. (1991). Molecular cloning and chromosomal localization of a human peripheral- type benzodiazepine receptor. *Eur J Biochem* **195:** 305-11.

Ritter B., Modregger J., Paulsson M., and Plomann M. (1999). PACSIN 2, a novel member of the PACSIN family of cytoplasmic adapter proteins. *FEBS Lett* **454:** 356-62.

Rivas E., Klein R. J., Jones T. A., and Eddy S. R. (2001). Computational identification of noncoding RNAs in E. coli by comparative genomics. *Curr Biol* **11:** 1369-73.

Roest Crollius H., Jaillon O., Bernot A., Dasilva C., Bouneau L., Fischer C., Fizames C., Wincker P., Brottier P., Quetier F., Saurin W., and Weissenbach J. (2000). Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nat Genet* **25:** 235-8.

Rolls M. M., Stein P. A., Taylor S. S., Ha E., McKeon F., and Rapoport T. A. (1999). A visual screen of a GFP-fusion library identifies a new type of nuclear envelope membrane protein. *J Cell Biol* **146:** 29-44.

Rommens J. M., Iannuzzi M. C., Kerem B., Drumm M. L., Melmer G., Dean M., Rozmahel R., Cole J. L., Kennedy D., Hidaka N., and et al. (1989). Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245:** 1059-65.

Ross-Macdonald P., Coelho P. S., Roemer T., Agarwal S., Kumar A., Jansen R., Cheung K. H., Sheehan A., Symoniatis D., Umansky L., Heidtman M., Nelson F. K., Iwasaki H., Hager K., Gerstein M., Miller P., Roeder G. S., and Snyder M. (1999). Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402:** 413-8.

Rushforth A. M., Saari B., and Anderson P. (1993). Site-selected insertion of the transposon Tc1 into a Caenorhabditis elegans myosin light chain gene. *Mol Cell Biol* **13:** 902-10.

Saccone S., Caccio S., Kusuda J., Andreozzi L., and Bernardi G. (1996). Identification of the gene-richest bands in human chromosomes. *Gene* **174:** 85-94.

Saccone S., De Sario A., Della Valle G., and Bernardi G. (1992). The highest gene concentrations in the human genome are in telomeric bands of metaphase chromosomes. *Proc Natl Acad Sci U S A* **89:** 4913-7.

Saccone S., De Sario A., Wiegant J., Raap A. K., Della Valle G., and Bernardi G. (1993). Correlations between isochores and chromosomal bands in the human genome. *Proc Natl Acad Sci U S A* **90:** 11929-33.

Sadusky T. J., Kemp T. J., Simon M., Carey N., and Coulton G. R. (2001). Identification of Serhl, a new member of the serine hydrolase family induced by passive stretch of skeletal muscle in vivo. *Genomics* **73:** 38-49.

Saiki R. K., Gelfand D. H., Stoffel S., Scharf S. J., Higuchi R., Horn G. T., Mullis K. B., and Erlich H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239:** 487-91.

Saiki R. K., Scharf S., Faloona F., Mullis K. B., Horn G. T., Erlich H. A., and Arnheim N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230:** 1350-4.

Saitou N., and Nei M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406-25.

Salamov A. A., and Solovyev V. V. (2000). Ab initio gene finding in Drosophila genomic DNA. *Genome Res* **10:** 516-22.

Sander C., and Schneider R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9:** 56-68.

Sanger F., Coulson A. R., Hong G. F., Hill D. F., and Petersen G. B. (1982). Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* **162:** 729-73.

Sanger F., Nicklen S., and Coulson A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74:** 5463-7.

Sargent M. G., and Bennett M. F. (1986). Identification of a specific membrane-particle-associated DNA sequence in Bacillus subtilis. *J Bacteriol* **166:** 38-43.

Sawin K. E., and Nurse P. (1996). Identification of fission yeast nuclear markers using random polypeptide fusions with green fluorescent protein. *Proc Natl Acad Sci U S A* **93:** 15146-51.

Scherf M., Klingenhoff A., Frech K., Quandt K., Schneider R., Grote K., Frisch M., Gailus-Durner V., Seidel A., Brack-Werner R., and Werner T. (2001). First pass annotation of promoters on human chromosome 22. *Genome Res* **11:** 333-40.

Scherf M., Klingenhoff A., and Werner T. (2000). Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J Mol Biol* **297:** 599-606.

Scherthan H., Cremer T., Arnason U., Weier H. U., Lima-de-Faria A., and Fronicke L. (1994). Comparative chromosome painting discloses homologous segments in distantly related mammals. *Nat Genet* **6:** 342-7.

Schuler G. D. (1997). Sequence mapping by electronic PCR. *Genome Res* **7:** 541-50.

Schuler G. D. (1998). Sequence alignment and database searching. *Methods Biochem Anal* **39:** 145-71.

Schuler G. D., Boguski M. S., Stewart E. A., Stein L. D., Gyapay G., Rice K., White R. E., Rodriguez-Tome P., Aggarwal A., Bajorek E., Bentolila S., Birren B. B., Butler A., Castle A. B., Chiannilkulchai N., Chu A., Clee C., Cowles S., Day P. J., Dibling T., Drouot N., Dunham I., Duprat S., East C., Hudson T. J., and et al. (1996a). A gene map of the human genome. *Science* **274:** 540-6.

Schuler G. D., Epstein J. A., Ohkawa H., and Kans J. A. (1996b). Entrez: molecular biology database and retrieval system. *Methods Enzymol* **266:** 141-62.

Schultz J., Copley R. R., Doerks T., Ponting C. P., and Bork P. (2000). SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res* **28:** 231-4.

Schulz R. A., and Butler B. A. (1989). Overlapping genes of Drosophila melanogaster: organization of the z600- gonadal-Eip28/29 gene cluster. *Genes Dev* **3:** 232-42.

Schwartz F., and Ota T. (1997). The 239AB gene on chromosome 22: a novel member of an ancient gene family. *Gene* **194:** 57-62.

Schwartz S., Zhang Z., Frazer K. A., Smit A., Riemer C., Bouck J., Gibbs R., Hardison R., and Miller W. (2000). PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res* **10:** 577-86.

Sealey P. G., Whittaker P. A., and Southern E. M. (1985). Removal of repeated sequences from hybridisation probes. *Nucleic Acids Res* **13:** 1905-22.

Seed B., and Aruffo A. (1987). Molecular cloning of the CD2 antigen, the T-cell erythrocyte receptor, by a rapid immunoselection procedure. *Proc Natl Acad Sci U S A* **84:** 3365-9.

Sehgal A., Briggs J., Rinehart-Kim J., Basso J., and Bos T. J. (2000). The chicken c-Jun 5' untranslated region directs translation by internal initiation. *Oncogene* **19:** 2836-45.

Senior K. (1999). Fingerprinting disease with protein chip arrays. *Mol Med Today* **5:** 326-7.

Shaikh T. H., Kurahashi H., Saitta S. C., O'Hare A. M., Hu P., Roe B. A., Driscoll D. A., McDonald-McGinn D. M., Zackai E. H., Budarf M. L., and Emanuel B. S. (2000). Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum Mol Genet* **9:** 489-501.

Sharp P. A., and Burge C. B. (1997). Classification of introns: U2-type or U12-type. *Cell* **91:** 875-9.

Shehee W. R., Loeb D. D., Adey N. B., Burton F. H., Casavant N. C., Cole P., Davies C. J., McGraw R. A., Schichman S. A., Severynse D. M., and et al. (1989). Nucleotide sequence of the BALB/c mouse beta-globin complex. *J Mol Biol* **205:** 41-62.

Shirabe K., Yubisui T., Borgese N., Tang C. Y., Hultquist D. E., and Takeshita M. (1992). Enzymatic instability of NADH-cytochrome b5 reductase as a cause of hereditary methemoglobinemia type I (red cell type). *J Biol Chem* **267:** 20416-21.

Shirabe K., Yubisui T., Nishino T., and Takeshita M. (1991). Role of cysteine residues in human NADH-cytochrome b5 reductase studied by site-directed mutagenesis. Cys-273 and Cys-283 are located close to the NADH-binding site but are not catalytically essential. *J Biol Chem* **266:** 7531-6.

Shizuya H., Birren B., Kim U. J., Mancino V., Slepak T., Tachiiri Y., and Simon M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc Natl Acad Sci U S A* **89:** 8794-7.

Shoemaker D. D., Lashkari D. A., Morris D., Mittmann M., and Davis R. W. (1996). Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat Genet* **14:** 450-6.

Shoemaker D. D., Schadt E. E., Armour C. D., He Y. D., Garrett-Engele P., McDonagh P. D., Loerch P. M., Leonardson A., Lum P. Y., Cavet G., Wu L. F., Altschuler S. J., Edwards S., King J., Tsang J. S., Schimmack G., Schelter J. M., Koch J., Ziman M., Marton M. J., Li B., Cundiff P., Ward T., Castle J., Krolewski M., Meyer M. R., Mao M., Burchard J., Kidd M. J., Dai H., Phillips J. W., Linsley P. S., Stoughton R., Scherer S., and Boguski M. S. (2001). Experimental annotation of the human genome using microarray technology. *Nature* **409:** 922-7.

Simpson J. C., Wellenreuther R., Poustka A., Pepperkok R., and Wiemann S. (2000). Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep* **1:** 287-92.

Sinden R. R. (1999). Biological implications of the DNA structures associated with disease-causing triplet repeats. *Am J Hum Genet* **64:** 346-53.

Slusher L. B., Gillman E. C., Martin N. C., and Hopper A. K. (1991). mRNA leader length and initiation codon context determine alternative AUG selection for the yeast gene MOD5. *Proc Natl Acad Sci U S A* **88:** 9789-93.

Smink L. J. (2000). Genome studies of human chromosome 22q13.31. PhD Thesis, Open University.

Smit A. F. (1996). The origin of interspersed repeats in the human genome. *Curr Opin Genet Dev* **6:** 743-8.

Smit A. F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9:** 657-63.

Smit A. F., and Riggs A. D. (1995). MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res* **23:** 98-102.

Smith L. M., Kaiser R. J., Sanders J. Z., and Hood L. E. (1987). The synthesis and use of fluorescent oligonucleotides in DNA sequence analysis. *Methods Enzymol* **155:** 260-301.

Smith V., Chou K. N., Lashkari D., Botstein D., and Brown P. O. (1996). Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* **274:** 2069-74.

Soderlund C., Longden I., and Mott R. (1997). FPC: a system for building contigs from restriction fingerprinted clones. *Comput Appl Biosci* **13:** 523-35.

Solovyev V., and Salamov A. (1997). The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *Proc Int Conf Intell Syst Mol Biol* **5:** 294-302.

Solovyev V. V., Salamov A. A., and Lawrence C. B. (1994). The prediction of human exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Proc Int Conf Intell Syst Mol Biol* **2:** 354-62.

Somma M. P., Gambino I., and Lavia P. (1991). Transcription factors binding to the mouse HTF9 housekeeping promoter differ between cell types. *Nucleic Acids Res* **19:** 4451-8.

Sonnhammer E. L., and Durbin R. (1994). A workbench for large-scale sequence homology analysis. *Comput Appl Biosci* **10:** 301-7.

Sonnhammer E. L., and Durbin R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167:** GC1-10.

Spradling A. C., Stern D. M., Kiss I., Roote J., Laverty T., and Rubin G. M. (1995). Gene disruptions using P transposable elements: an integral component of the Drosophila genome project. *Proc Natl Acad Sci U S A* **92:** 10824-30.

Sprengel R., Werner P., Seeburg P. H., Mukhin A. G., Santi M. R., Grayson D. R., Guidotti A., and Krueger K. E. (1989). Molecular cloning and expression of cDNA encoding a peripheral-type benzodiazepine receptor. *J Biol Chem* **264:** 20415-21.

Steffensen R., Carlier K., Wiels J., Levery S. B., Stroud M., Cedergren B., Nilsson Sojka B., Bennett E. P., Jersild C., and Clausen H. (2000). Cloning and expression of the histo-blood group Pk UDP-galactose: Ga1beta-4G1cbeta1-cer alpha1, 4-galactosyltransferase. Molecular genetic basis of the p phenotype. *J Biol Chem* **275:** 16723-9.

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

Stein L. D., and Thierry-Mieg J. (1998). Scriptable access to the Caenorhabditis elegans genome sequence and other ACEDB databases. *Genome Res* **8:** 1308-15.

Stephens R. M., and Schneider T. D. (1992). Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J Mol Biol* **228:** 1124-36.

Strachan T., and Read A. P. (1999). Human molecular genetics 2. Second edition. BIOS Scientific Publishers Ltd.

Strausberg R. L., Feingold E. A., Klausner R. D., and Collins F. S. (1999). The mammalian gene collection. *Science* **286:** 455-7.

Strittmatter P., Kittler J. M., and Coghill J. E. (1992). Characterization of the role of lysine 110 of NADH-cytochrome b5 reductase in the binding and oxidation of NADH by site-directed mutagenesis. *J Biol Chem* **267:** 20164-7.

Stryer L. (1988).Biochemistry. Third edition. W.H. Freeman and Company.

Sved J., and Bird A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc Natl Acad Sci U S A* **87:** 4692-6.

Tagle D. A., Koop B. F., Goodman M., Slightom J. L., Hess D. L., and Jones R. T. (1988). Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol* **203:** 439-55.

Taketani S., Kohno H., Okuda M., Furukawa T., and Tokunaga R. (1994). Induction of peripheral-type benzodiazepine receptors during differentiation of mouse erythroleukemia cells. A possible involvement of these receptors in heme biosynthesis. *J Biol Chem* **269:** 7527-31.

Tamura M., Yubisui T., Takeshita M., Kawabata S., Miyata T., and Iwanaga S. (1987). Structural comparison of bovine erythrocyte, brain, and liver NADH- cytochrome b5 reductase by HPLC mapping. *J Biochem (Tokyo)* **101:** 1147-59.

Tassabehji M., Read A. P., Newton V. E., Harris R., Balling R., Gruss P., and Strachan T. (1992). Waardenburg's syndrome patients have mutations in the human homologue of the Pax-3 paired box gene. *Nature* **355:** 635-6.

Tatusov R. L., Koonin E. V., and Lipman D. J. (1997). A genomic perspective on protein families. *Science* **278:** 631-7.

Tatusov R. L., Mushegian A. R., Bork P., Brown N. P., Hayes W. S., Borodovsky M., Rudd K. E., and Koonin E. V. (1996). Metabolism and evolution of Haemophilus influenzae deduced from a whole- genome comparison with Escherichia coli. *Curr Biol* **6:** 279-91.

Tautz D. (1989). Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res* **17:** 6463-71.

Thanassi D. G., and Hultgren S. J. (2000). Multiple pathways allow protein secretion across the bacterial outer membrane. *Curr Opin Cell Biol* **12:** 420-30.

Thompson J. D., Higgins D. G., and Gibson T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673-80.

Tilford C. A., Kuroda-Kawaguchi T., Skaletsky H., Rozen S., Brown L. G., Rosenberg M., McPherson J. D., Wylie K., Sekhon M., Kucaba T. A., Waterston R. H., and Page D. C. (2001). A physical map of the human Y chromosome. *Nature* **409:** 943-5.

Tomita M. (2001). Whole-cell simulation: a grand challenge of the 21st century. *Trends Biotechnol* **19:** 205-10.

Toth G., Gaspari Z., and Jurka J. (2000). Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* **10:** 967-81.

Touchman J. W., Bouffard G. G., Weintraub L. A., Idol J. R., Wang L., Robbins C. M., Nussbaum J. C., Lovett M., and Green E. D. (1997). 2006 expressed-sequence tags derived from human chromosome 7-enriched cDNA libraries. *Genome Res* **7:** 281-92.

Tran H., Mattei M., Godyna S., and Argraves W. S. (1997). Human fibulin-1D: molecular cloning, expression and similarity with S1- 5 protein, a new member of the fibulin gene family. *Matrix Biol* **15:** 479-93.

Trichet V., Ruault M., Roizes G., and De Sario A. (2000). Characterization of the human tubulin tyrosine ligase-like 1 gene (TTLL1) mapping to 22q13.1. *Gene* **257:** 109-17.

Trichet V., Shkolny D., Dunham I., Beare D., and McDermid H. E. (1999). Mapping and complex expression pattern of the human NPAP60L nucleoporin gene. *Cytogenet Cell Genet* **85:** 221-6.

Trofatter J. A., Long K. R., Murrell J. R., Stotler C. J., Gusella J. F., and Buckler A. J. (1995). An expression-independent catalog of genes from human chromosome 22. *Genome Res* **5:** 214-24.

Trofatter J. A., MacCollin M. M., Rutter J. L., Murrell J. R., Duyao M. P., Parry D. M., Eldridge R., Kley N., Menon A. G., Pulaski K., and et al. (1993). A novel moesin-, ezrin-, radixin-like gene is a candidate for the neurofibromatosis 2 tumor suppressor. *Cell* **75:** 826.

Tsai D. E., Kenan D. J., and Keene J. D. (1992). In vitro selection of an RNA epitope immunologically cross-reactive with a peptide. *Proc Natl Acad Sci U S A* **89:** 8864-8.

Uberbacher E. C., and Mural R. J. (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc Natl Acad Sci U S A* **88:** 11261-5.

Vagner S., Gensac M. C., Maret A., Bayard F., Amalric F., Prats H., and Prats A. C. (1995). Alternative translation of human fibroblast growth factor 2 mRNA occurs by internal entry of ribosomes. *Mol Cell Biol* **15:** 35-44.

Valdes J. M., Tagle D. A., and Collins F. S. (1994). Island rescue PCR: a rapid and efficient method for isolating transcribed sequences from yeast artificial chromosomes and cosmids. *Proc Natl Acad Sci U S A* **91:** 5377-81.

Van Etten W. J., Steen R. G., Nguyen H., Castle A. B., Slonim D. K., Ge B., Nusbaum C., Schuler G. D., Lander E. S., and Hudson T. J. (1999). Radiation hybrid map of the mouse genome. *Nat Genet* **22:** 384-7.

Venter J. C., Adams M. D., Myers E. W., Li P. W., Mural R. J., Sutton G. G., Smith H. O., Yandell M., Evans C. A., Holt R. A., Gocayne J. D., Amanatides P., Ballew R. M., Huson D. H., Wortman J. R., Zhang Q., Kodira C. D., Zheng X. H., Chen L., Skupski M., Subramanian G., Thomas P. D., Zhang J., Gabor Miklos G. L., Nelson C., Broder S., Clark A. G., Nadeau J., McKusick V. A., Zinder N., Levine A. J., Roberts R. J., Simon M., Slayman C., Hunkapiller M., Bolanos R., Delcher A., Dew I., Fasulo D., Flanigan M., Florea L., Halpern A., Hannenhalli S., Kravitz S., Levy S., Mobarry C., Reinert K., Remington K., Abu-Threideh J., Beasley E., Biddick K., Bonazzi V., Brandon R., Cargill M., Chandramouliswaran I., Charlab R., Chaturvedi K., Deng Z., Di Francesco V., Dunn P., Eilbeck K., Evangelista C., Gabrielian A. E., Gan W., Ge W., Gong F., Gu Z., Guan P., Heiman T. J., Higgins M. E., Ji R. R., Ke Z., Ketchum K. A., Lai Z., Lei Y., Li Z., Li J., Liang Y., Lin X., Lu F., Merkulov G. V., Milshina N., Moore H. M., Naik A. K., Narayan V. A., Neelam B., Nusskern D., Rusch D. B., Salzberg S., Shao W., Shue B., Sun J., Wang Z., Wang A., Wang X., Wang J., Wei M., Wides R., Xiao C., Yan C., et al. (2001). The sequence of the human genome. *Science* **291:** 1304-51.

Verkerk A. J., et al. (1991). Identification of a gene (FMR-1) containing a CGG repeat conicident with a breakpoint cluster region exhibiting langth variation in fragil X syndrome. *Cell* **65**: 904-14.

Verma S., Budarf M. L., Emanuel B. S., and Chinnadurai G. (2000). Structural analysis of the human pro-apoptotic gene Bik: chromosomal localization, genomic organization and localization of promoter sequences. *Gene* **254:** 157-62.

Vulpe C., Levinson B., Whitney S., Packman S., and Gitschier J. (1993). Isolation of a candidate gene for Menkes disease and evidence that it encodes a copper-transporting ATPase. *Nat Genet* **3:** 7-13.

Walter G., Bussow K., Cahill D., Lueking A., and Lehrach H. (2000). Protein arrays for gene expression and molecular interaction screening. *Curr Opin Microbiol* **3:** 298-302.

Walter M. A., Spillett D. J., Thomas P., Weissenbach J., and Goodfellow P. N. (1994). A method for constructing radiation hybrid maps of whole genomes. *Nat Genet* **7:** 22-8.

Watson J. D. (1990). The human genome project: past, present, and future. *Science* **248:** 44-9.

Weber J. L., Polymeropoulos M. H., May P. E., Kwitek A. E., Xiao H., McPherson J. D., and Wasmuth J. J. (1991). Mapping of human chromosome 5 microsatellite DNA polymorphisms. *Genomics* **11:** 695-700.

Weissenbach J., Gyapay G., Dib C., Vignal A., Morissette J., Millasseau P., Vaysseix G., and Lathrop M. (1992). A second-generation linkage map of the human genome. *Nature* **359:** 794-801.

Weterman M. A., Wilbrink M., Dijkhuizen T., van den Berg E., and Geurts van Kessel A. (1996). Fine mapping of the 1q21 breakpoint of the papillary renal cell carcinoma-associated (X;1) translocation. *Hum Genet* **98:** 16-21.

Wiemann S., Weil B., Wellenreuther R., Gassenhuber J., Glassl S., Ansorge W., Bocher M., Blocker H., Bauersachs S., Blum H., Lauber J., Dusterhoft A., Beyer A., Kohrer K., Strack N., Mewes H. W., Ottenwalder B., Obermaier B., Tampe J., Heubner D., Wambutt R., Korn B., Klein M., and Poustka A. (2001). Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res* **11:** 422-35.

Wilcox A. S., Khan A. S., Hopkins J. A., and Sikela J. M. (1991). Use of 3' untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications for an expression map of the genome. *Nucleic Acids Res* **19:** 1837-43.

Wingender E., Chen X., Hehl R., Karas H., Liebich I., Matys V., Meinhardt T., Pruss M., Reuter I., and Schacherer F. (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* **28:** 316-9.

Wolfe K. H., Sharp P. M., and Li W. H. (1989). Mutation rates differ among regions of the mammalian genome. *Nature* **337:** 283-5.

Wolfsberg T. G., and Landsman D. (1997). A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res* **25:** 1626-32.

Wu X. R., and Sun T. T. (1993). Molecular cloning of a 47 kDa tissue-specific and differentiation- dependent urothelial cell surface glycoprotein. *J Cell Sci* **106:** 31-43.

Xiao Z., Jiang X., Beckett M. L., and Wright G. L., Jr. (2000). Generation of a baculovirus recombinant prostate-specific membrane antigen and its use in the development of a novel protein biochip quantitative immunoassay. *Protein Expr Purif* **19:** 12-21.

Xu Y., Mural R. J., and Uberbacher E. C. (1995). Correcting sequencing errors in DNA coding regions using a dynamic programming approach. *Comput Appl Biosci* **11:** 117-24.

Yang S., Tutton S., Pierce E., and Yoon K. (2001). Specific double-stranded RNA interference in undifferentiated mouse embryonic stem cells. *Mol Cell Biol* **21:** 7807-16.

Yeh R. F., Lim L. P., and Burge C. B. (2001). Computational inference of homologous gene structures in the human genome. *Genome Res* **11:** 803-16.

Yu S. et al. (1991). Fragile X genotype characterised by an unstable region of DNA. *Science* **252**: 1179-81 (1991).

Yuasa T., Yoshiki T., Tanaka T., Kim C. J., Isono T., and Okada Y. (1998). Expression of uroplakin Ib and uroplakin III genes in tissues and peripheral blood of patients with transitional cell carcinoma. *Jpn J Cancer Res* **89:** 879-82.

Yubisui T., Miyata T., Iwanaga S., Tamura M., Yoshida S., Takeshita M., and Nakajima H. (1984). Amino acid sequence of NADH-cytochrome b5 reductase of human erythrocytes. *J Biochem (Tokyo)* **96:** 579-82.

Yubisui T., Naitoh Y., Zenno S., Tamura M., Takeshita M., and Sakaki Y. (1987). Molecular cloning of cDNAs of human liver and placenta NADH-cytochrome b5 reductase. *Proc Natl Acad Sci U S A* **84:** 3609-13.

Zenno S., Hattori M., Misumi Y., Yubisui T., and Sakaki Y. (1990). Molecular cloning of a cDNA encoding rat NADH-cytochrome b5 reductase and the corresponding gene. *J Biochem (Tokyo)* **107:** 810-6.

Zhang C., Yu Y., Zhang S., Liu M., Xing G., Wei H., Bi J., Liu X., Zhou G., Dong C., Hu Z., Zhang Y., Luo L., Wu C., Zhao S., and He F. (2000). Characterization, chromosomal assignment, and tissue expression of a novel human gene belonging to the ARF GAP family. *Genomics* **63:** 400-8.

Zhuo D., Zhao W. D., Wright F. A., Yang H. Y., Wang J. P., Sears R., Baer T., Kwon D. H., Gordon D., Gibbs S., Dai D., Yang Q., Spitzner J., Krahe R., Stredney D., Stutz A., and Yuan B. (2001). Assembly, annotation, and integration of UNIGENE clusters into the human genome draft. *Genome Res* **11:** 904-18.

Ziauddin J., and Sabatini D. M. (2001). Microarrays of cells expressing defined cDNAs. *Nature* **411:** 107-10.

Zoubak S., Clay O., and Bernardi G. (1996). The gene distribution of the human genome. *Gene* **174:** 95-102.

Zwaal R. R., Broeks A., van Meurs J., Groenen J. T., and Plasterk R. H. (1993). Target-selected gene inactivation in Caenorhabditis elegans by using a frozen transposon insertion mutant bank. *Proc Natl Acad Sci U S A* **90:** 7431-5.

Appendix 1: Sequences of primers used in this project.

STS Name	Primer 1	Primer 2	Product size (genomic)	Product size (cDNA)
st105046C1A10snp	ATGCCGATGTGAGTCGTGTA	TTTCATTGTGTTCGTCAGGC	380	
st105047B5snp	AGCTGTGTCTCTGGCTGTGA	TCTGCTAGACTCGCCTCCTC	396	
st118609D5snp	CCCATTTCGTAGGTGAGCAG	GGATCTCTGCTGGCAGGTTA	400	
st118622E3snp	CTCCTGGGGAAGAAAGACAA	GACCTGGTTGGAGGCTAGAA	393	
st118626E9B5asnp	AAAATAGAATGAACCTTCAGATTGCT	ACCCAAAACATATTATCACCAATTC	390	
st118626E9B5bsnp	TCCAATCCAAAGTTCTCATTCC	AACAGTGATTAGAACTTGCACACA	395	
st118626E9B5csnp	CTTCTTCCCCTTCTCCCTCT	CACTGTGGACACCCCATCT	399	
st118626E10B6snp	TCCTTCTGTTTTTATGAAAGTCCC	CAGACTTAACACTACGCCCG	404	
st118634F4snp	TCACAGAGGGCAGAGACCAC	CCCAGGACTCTATGCCAGAA	398	
stAA000919	ATGACAAGGGCTTTGTGGAC	CCTGGTAGACCAGTAGCCCA	156	
stAA000919.2	TATCTGCTCTTTGCCAACCA	TGTACCAGAGCTTGAAGCGA	100	
stAA017949.2	GCTATTGCCGACGAGGTTAC	AACAGTTATGGATGCCGGTC	222	
stAA020167	AGAACCAGCTGCTCACGAAG	GGCAGCACATCACTGAAGAA	121	
stAA020167	AGAACCAGCTGCTCACGAAG	AGACCGCGACTACTCCGTTA	121	
stAA020167	CTCAGTGACCTGGGCAAGAT	AGACCGCGACTACTCCGTTA	121	
stAA020167	CTCAGTGACCTGGGCAAGAT	GGCAGCACATCACTGAAGAA	121	
stAA020167.1	ACGCCACCTGGAACAGTATC	TCAGTGAAGAAGGACCCCAG	136	
stAA020167.2	AGAACCAGCTGCTCACGAAG	GGCAGCACATCACTGAAGAA	121	
stAA020167.2	GAAGCTGCTGGACTCATTGG	GAAGATGGCCTGCAGGGTT	121	
stAA020167.2	AGAACCAGCTGCTCACGAAG	GAAGATGGCCTGCAGGGTT	121	
stAA020167.2	GAAGCTGCTGGACTCATTGG	GGCAGCACATCACTGAAGAA	121	
stAA034567.2	ACATCCTGGAAGTACGGCAG	TTCAGTTTGGGCTGGTTTTC	134	
stAA036311.2	CTCTCACACAGCGAGGGACT	CACCAGCAGGCATTTTCTCT	102	
stAA038856	AGAAAAAGGATGCCTGAGCA	AGGAAGGGAGAAGAGCTTCG	215	
stAA073404.2	CTGGATCATGAAGCCATGTG	CCGATGAGCAACGGATTATT	171	
stAA103214.2	AAAGGAACTTCATGGCCCTC	AACCCTGGAGCACTCAGCTA	155	
stAA104556.2	CACAGAGACCGCTGATGCTA	TCTGCTTTACCGTCGACCTC	174	
stAA209454	AATAGCCTAGCCCCCTGCTA	AAGCACTTTCTGCTTTTGGC	123	
stAA284306	TCTCACAAGCTGGAATGTGC	CAGGATAAAGTCCAGGCAGC	122	
stAA308620.2	GCTTCTGCCTTTCACCAAGA	CTACTGGCAGCCCAACAAGT	106	
stAA316228	CATGGATTTTTCATCTTCTACCG	ATGCACGTTCTATCCAGGG	202	
stAA316829.1	GGCAGGGAAAGTAAAATGGG	TCACTTGGACCCATCCTCTC	83	
stAA316829.2	GCTCTGAGCCCACAGGGA	CATAGGCACTGACAGAGCACC	100	
stAA316883.2	AGGAGATGGAAGACCTGCG	TCCTCCTCCTCCTCCAGTT	108	
stAA316883.3	TGCAGAAGGACTGTGACCTG	TTATTTTTGCAGAGGAGCCC	103	
stAA390181.2	TGGAGCAGCACATTCAGAAC	CTCTTGCCCACTGAAGAACC	193	
stAA390181.2	GTGTTTGACAGCTTGGAGCA	CTCTTGCCCACTGAAGAACC	193	
stAA451754.1	CATCTCATTCGTCTGATTGGAA	TGATGTTGCAGTTGTCCAGG	106	
stAA451754.2	CGTGAGAGCAGAAGGTGACA	CCATCCAGCTCATTTACCTTG	123	
stAA474253.2	CCCTGTGGAAGACCAGTCAC	GAGTGTCCACTTTCCTTCCG	106	
stAA474253.2	CCTGTGGAAGACCAGTCACA	ACTTTGAACAAGGCCCCTCT	106	
stAA474253.2	CCCTGTGGAAGACCAGTCAC	ACTTTGAACAAGGCCCCTCT	106	
stAA474253.2	CCTGTGGAAGACCAGTCACA	GAGTGTCCACTTTCCTTCCG	106	

stAA482326	TGGTGTCTCTCAGGCTTTCC	TGCCCACGATATGCTAAGTG	210	
stAA535365	AGCTGCCTCTATCCCTGACA	TGGGTCTGTTAAGTCCTGGG	132	
stAA726210.2	AGAGGTGACGCAGTCGAGAT	ATGCTCAGGAAGGTGAATGG	198	
stAA889354	GGGTTTGGGTCCATTTGTTA	GCTCATGTTCATGGAGTCCC	122	
stAA911743	CATCAACATGGGTGTTCCAC	CTTCCGTGTTGCTTGAGAGG	108	
stAA913461	GGTTTCAGTTGGTCTCTTTTGC	CTGCATATCCCTCATCCACC	83	
stAF009246	TGGTCATTTGCGGTAACAAA	GGTCCAAGCTGCTGTTCTTC	138	
stAF048838	TTACTGGCTAAGTGCCCCTG	ACTCTCCACCCCAAAGGACT	174	
stAF059257	CAGATTCCTGGGTGTGTCCT	AACGTGAGGTCACACTTCCC	207	
stAF059257.2	GAAGAGTTGGGCTGTGGAAG	CTAGTAGAGCGTTGGGACCG	161	
stAF070556	TACTCTTCGGGGGTTTCCTT	CCATTTCAATTTGTGACCCC	143	
stAF086048	ACCACGTCTTTCAGTGGTCC	TGGTGGTATTGCTTGGTGTG	151	
stAI168007	CAGGACTGCTGAAGCACAAG	GCAAATTCCGACCCTACTGA	239	
stAI208666	CCAGATGAGTAGAGAGGCGG	AGCGGCTTTGGGAGAGTAAT	238	
stAI225490.2	CATCAAGGTTGTGTCTCCCA	CAATCTTCACCTCCTCAGGC	116	
stAI333015	CAAGCCACTGGGTAGCATTT	GCGACGTGTGTCTTTTCAGA	151	
stAI538474	AGCTCTCCTCTTGATGCCAA	GCTCACTGTTGCAAGGGTTT	101	
stAI623320	GAAGGGATGGATCCTGAGGT	GGTGTCCAGGATGCTCTCAT	159	
stAI647249.2	GGTATTGCTCCTGCGACATT	CTACCAGAGAACTGCCCAGC	296	
stAI904287	TGCAAACCTACTCCCAAAGG	CCTGAGACATGGCTTCTGTG	454	
stAJ006972	AGACCCAGGAGAAGGACGAT	TAATACCAAGGGAGGGAGGG	117	
stAJ006972.2	CCCTTACGGTCTTGGAAACA	ATGGTCCTCACCAGCACATT	100	
stAL050282.1	GCACCCAGGACTATCTTCCA	TCAACAGCAACACCAATCGT	215	
stAL050282.2	TGTGAAATTGCAAGTGTTTGAA	AATGCACGTTCTATCCAGGG	177	
stAL050282.3	TCCTACTCCTCCCTCCTCGT	GGTGCTGTTTCTCGGTTCC	216	
stARFGAP1.3tag	GGCCGATATCCTGACGATGGGGGACCCCAGC	GGCCGCTAGCAGAACCGTATCGATC CTG		1584
stARFGAP1.5tag	GGCCGCGGCCGCGATGGGGGACCCCAGC	GGCCTCTAGATTAAGAACCGTATCG ATCCTG		1584
stARFGAP1na	CCGCTTTTCGTCGACTCTTA	CCGCCTGAGATGTGGTTACT		1686
stARFGAP1nb	CACAGCTGACGATGGG	CTCCAGGAAATACACATC		1592
stARFGAP1r	TGCTGGAAAACTCTCCGTCT	CCGCCTGAGATGTGGTTACT		138
stARHGAP8	CTACCCTACCTCCGAGTCCC	AAGTGCACAGACAGGTGTGG	101	
stARHGAP8.2	ATGAGGACTCTCCGCAGGTT	TCTGAATTCATCTAATGGCTGG		2042
stARHGAP8.nest1	GCTCGGCTTCCTGCTCTC	TCTGAATTCATCTAATGGCTGGT		2171
stARHGAP8.nest2	GTCACCATGAGGACTCTCCG	GCTCGAAATATACAGAGTGTTCG		1966
stARHGAP8b	AAGTTCATGAGTTCGCCCAG	CTGTTCCAGGTGGCGTTG	113	
stARHGAP8c	CAGCAGCTCACAGGACTCAG	GGGAACAAGGCTAGGACACA	114	
stARHGAP8n	CTCCGTGCAATGATTAACCC	AAGTGCACAGACAGGTGTGG		2255
stARHGAP8r	AGCCTCCCAGAGCACAACTA	ATGGCCAGATCAAATTCAGC		136
stAW006377	AGAAATTCTCAGCCTCGGG	GGCAAGGGGATGGTTCTC	872	
stAZ039089	AACACCCAAGCAGTAATATCCC	AAGGAGTGGGGAGAAAGAATG	122	
stAZ053629	GTGTCAGCAAACCAAGCAGA	CGATCCTAAATGTGAGATGCTG	135	
stAZ053645	TTCCTGGAGCTTGAGAAATAGC	AGGGCTTCAAGTCCCCTG	132	
stAZ069460	CATGAGCACCTGAAATGGG	TGCACTTTACAGCTGAGAGCA	139	
stAZ122996	ACCAGGGAAAGGAGGAACAT	GCTGCTTAGCAAAATCAGTCG	128	
stAZ122997	AATGTTACCCCCTTTCCTGG	GTGGGTGAGTGGGTGGAG	97	

stAZ223795	CAGACAGTGCTAGAAGGGGC	GAGACTCCATCACGAGCTCC	162	
stAZ223797	GGATCTGAGTCCGCCTCC	ATCGCTAAATGGCTCTTACTCG	90	
stAZ224850	CAGACCAGAGAAACCCTACCC	AACTGGTAAAGCTAGTGCTGCC	133	
stAZ224853	GATGGTGGCAATATTTTGGG	GAGGCAGAGAGAGCCAGCTA	137	
stAZ235760	CTGACCACAGCCATGCTG	TTGCCTCTCCTGTCAGGG	138	
stAZ259729	GCCAGTCAATCACCACACC	ATCTCGTGTTTATGGGGCTG	167	
stAZ259730	ACAGAATACCAGTGGCCTGG	TACCCAGGCTGTGTATCTTGG	165	
stAZ273863	CTGAGCGCTTTATGGTCACA	TATCTATGAGGGTGCAGGGG	175	
stAZ280505	ATGAAGAGCCCAGGGTCC	AAGACCCTCAGAGTCTCCTGTG	172	
stAZ280507	TGGTTTCCTGACTTCCGATC	CTAGGCACCCCTAATACCTGC	140	
stAZ287193	CACCACCACACCTGGTGATA	GAGATGAGGCATGTGTCCAA	166	
stAZ287194	TGGTAGATGGGTTTAGCATGC	GTTATGGGCTTTGCCTTAAGG	156	
stAZ297741	AGAGGCAGGCAGATTTCTGA	GGGTTTCTCTGTGTAGTCCTGG	83	
stAZ297742	TTCTTCTCTTTGAAGTCCAGCC	GATTCCTGACCCACAGAAACA	161	
stAZ696681	AGTCTTAGGGGTTGGAGACTCC	GAGGGCTGCAATTCTGAAAG	180	
stAZ696682	CCTGGACGTCCAGTCCCT	AGCCCTGGGTCTCCAGAC	122	
stBIK.3tag	GGCCGATATCGGAGAAATGTCTGAAGTAAGACCC	GGCCGCTAGCCTTGAGCAGCAGGTGCAGGCC		560
stBIK.5tag	GGCCGCGGCCGCAATGTCTGAAGTAAGACCC	GGCCTCTAGATCACTTGAGCAGCAGGTGCAGGCC		562
stBIKna	GTGGCTTACAGACGCTGCC	AACCTCAGCAGTGTTCCAGC		674
stBIKnb	GAGAAATGTCTGAAGTAAGAC	GATAACAGCAGCAGG		565
stBIKr	GAAGTTTCATGGACGGTTTCA	AGCAGCACCTGTTCGCAG		100
stbK116F5mus1	AGTGACCTGGGCAAGAGAGA	AAGAAGGACCCCAGCTCTGT		124
stbK217C2mus1	AAATGCCGAGAAGGAACTGA	GCGCTTTGCTTTCTTTATGG		434
stbK268H5.C22.1	GCGCACATAGAAAAGCATGT	GCTTCAAGGATGACCGCAT	101	
stbK268H5.C22.1cds	CACATGCTGCGTGCCATAG	GGTGGACAGGTAGGCACTTG		
stbK268H5.C22.1n	GAGGCCGAGGAGCGCTC	CGGCATCTCTAGTCCTCTTTG		1529
stbK268H5.C22.1r	GAAGTTTCATGGACGGTTTCA	AGCAGCACCTGTTCGCAG		132
stbK268H5.C22.4	CAGATATGCTTCTGCACGGA	TGTTTGCAGATCTGAGTGGC	100	
stbK268H5.C22.4r	CGTGGACAGCACCACATTAT	CAGTATCTGATTGTGTGCGTTTT		107
stbK268H5.C22.7	GCAAAAAGAACCTATGGCAA	GCCTCTCTGGTCCTTAGCCT	104	
stbK268H5.C22.9a	GGCCCACAACTAGCTCTGAC	GTTGTATTTATGCCCCGACG	107	
stbK268H5.C22.9b	AACCCACTCGTCCTCACATC	CTGCCACAGCTACCATCTCA	129	
stbK268H5.C22.10	GCAGTAGCAGACTCGGAAGG	ACGGAAGATGAGGTGTTTGC	134	
stbK268H5.C22.11	CCTGCTGAAGCCCAATATGT	CAAGCTCTATGCCACGATGA	151	
stbK268H5mus1	TTCAGAGTAGCATAAAATTTGGC	CTGTGTGGACTGGGGTCTTT		107
stbK268H5mus2	GGATGGGGAGAAGAAGGC	CCTGTGGGACTGAAGAGGAC		116
stbK414D7.C22.1	TGCAGACGTGGTTAACTTGG	CAACACAGCTTGGACCAGAA	196	
stbK414D7.C22.1a	CCACATAAAGAAAGTGGCCC	GGGACCAAAAACCCCAGTAA	84	
stbK414D7.C22.1B	GCCTGGAGGAAGTGAGCA	ACTGAGTGCAAAGGGTGGTC	80	
stbK414D7.C22.1c	GGTCGCCCTTTCAGTAAAGC	TCCTAGAAGGGAGGTTGCAG	81	
stbK414D7.C22.1d	AGCACCAAAAGTGGTCTGAA	CATGGGCCCTCATCCTTC	80	
stbK414D7.C22.1e	TGAATTATTTAAGCTGGCTCCG	CAGATTCTGCACAGACAGGC	239	106
stbK414D7.C22.1f	CTCTCCAGATGAGAGACGGTG	GTAGGGGCACTCAGAAGGCT	80	
stbK414D7.C22.1na	CGCGGCCCATGTCC	GATGCCGCCTTTCTTGCT		1154
stbK414D7.C22.1nb	CGCGGCCCATGTCC	GCCACCATCCACAGC		1125

stbK414D7.C22.1r	ATGTGTCCTTCGCCTTTGAG	CTCAGGGTGGATTTGAGGTC		100
stbK941F9.C22.2	CAAGATGGCGGCGCCCAGGC	ACACTTGCAATTTCACATACGG		1516
stbK941F9.C22.3	CATTTCCATTCACGTTTCCC	CCTTCAAGAGTGAGTTCCCG	108	
stbK941F9.C22.4	TTGTCCGTGCTGATTGATGT	CTTGTCTCCACTGGAATCGG	123	
stbK941F9.C22.5	ACAGGTTCCCTAAGGCAAGG	CATTCCTGGGAAAAGGAGTGT	101	
stbK941F9.C22.6	TTCCTTGTGACTTGGCTTCA	TTCCTTGTGACTTGGCTTCA	101	
stbK941F9mus1	CATCATCTTCGACATCACGG	ACAACATTTCGGTGGGAGAC		183
stbK941F9mus2	CCAATTTTTCAACAATGTCTCCA	AAGCTCAGGTTACTCCCTTGAA		68
stbK1191B2.C22.3	CTACGACCTGCTGGAACTGA	GGGCTGCAGGTGATGTAGTT	86	
stbK1191B2.C22.3A	TGACCATGCTCCGTGCTGCTCTTCCCG	GATCTTCGTGGCATCGCT	200	
stbK1191B2.C22.3a.3tag	GGCCGATATCCCGACCATGAGCGTCC	GGCCGCTAGCTCTCGGGGGCTCC		1173
stbK1191B2.C22.3a.5tag	GGCCGCGGCCGCCATGAGCGTCC	GGCCTCTAGATCATCTCGGGGGCTCC		1173
stbK1191B2.C22.3B	TGACCATGGCTCACGTTCTGCATCTGTC	CTTCAGCAAATTCCATGGCT	126	
stbK1191B2.C22.3b.5tag	GGCCGCGGCCGCCATGAGCGTCCGGGTCGC	GGCCTCTAGATTAGAGGAATTCTTCTGGAG		543
stbK1191B2.C22.3C	TGACCATGTATGCAGTGAAAATCCGAGC	GAAATCACCCTGCAATCTGG	200	
stbK1191B2.C22.3D	TGACCATGCGCAAGCTTTAAAGGCAGTC	TACTTCGAAAGTTTGGGGGA	200	
stbK1191B2.C22.3r	TTGCAGCCCTAGTGTTTGC	TCCTCAGGAGGACAGA		749
stbK1191B2.C22.5	TCCGAAACGTACCCTCTCTG	CAGAGGGACCACCCCATT	112	
stbK1191B2.C22.6	TCAGCCCCACTCACACTG	GGAGCATCAGACTGGAGCG	102	
stbK1191B2.mus1	CAGCAGTGGCCATCACTGTA	CAGGAATCCAGGGTCAGAAA	100	
stbK1191B2_4600	GCTGCACTCTACTTCCAGGG	AGGTATGGAGCCCAAAGGAG	109	
stbK1191B2_48486	CTGCATGTTACAGCTCTTCAGG	CGGTTTCTCCAGAAGAATTCC	357	
stbK1191B2_58212	CTACGACCTGCTGGAACTGA	GGGCTGCAGGTGATGTAGTT	86	
stBZRP	GCGGCTGCCAGAGTGAGTGC	TCCATGTTCCAAGAACATGC	231	
stBZRP.3tag	GGCCGATATCGCAGCCATGGCCCCGCCCTGG	GGCCGCTAGCCTCTGGCAGCCGCCGTCCCCC		515
stBZRP.5tag	GGCCGCGGCCGCCATGGCCCCGCCCTGG	GGCCTCTAGATCACTCTGGCAGCCGCCGTCCCCC		510
stBZRPna	CCTGGCTAACTCCTGCCA	AAGGCCCTGACAGACTAGCA		702
stBZRPnb	GCAGCAGCCATGGC	CCAGTGGTCATGAAAGC		606
stBZRPr	TACGGCTCCTACCTGGTCTG	CGCCATACGCAGTAGTTGAG		284
stC22orf1r	ACTTCACTGAGCTGGGGCT	GTCAAAGGTCAGCTCGTGGT		113
stcB13C9mus1	GAAGTTCAACGAGTGGCTGG	TTCAGCTTCGACACAGATGG		144
stcB20F6_4357	CAAAGAGGCTTTCCCTTGAA	CTCTGCAGACCAGCACAGC	79	
stcB33B7.C22.1	ATCCACATAGGCTGAGGGTG	TACCAGCCAGTGACATCAGC	106	
stcB33B7.C22.1.3tag	GGCCGATATCGATACCATGTCCAAGCCC	GGCCGCTAGCCAAGTACATTTTCATGGCC		1062
stcB33B7.C22.1.5tag	GGCCGCGGCCGCCATGTCCAAGCCC	GGCCTCTAGATCACAAGTACATTTTCATGGCC		1062
stcB33B7.C22.1na	CCAGCCGGTTCCTG	CAGCTCCTCAACAGCC		1241
stcB33B7.C22.1nb	GATACCATGTCCAAGCCC	CAGGTTGGGGAGGTG		1097
stcB33B7.C22.1r	GTTCGTGTTGGCAAAGAAGG	TGGACATGGTATCCCCAGAT		156
stcB33B7.C22.3	CCACAACCTCCACGCTCT	TGCAGGGCTCCCCGTAG	82	
stcB33B7_3867	AGCCCTGAGCATCACTGACT	AACTGGCATGGTGCTTATCC	358	
stcB33B7_3867	AGCCCTGAGCATCACTGACT	GCCTCGCAGTAAAAGATTTTAA	358	
stcB33B7_3867	CCTGACGATTTCCTGCAAGT	AACTGGCATGGTGCTTATCC	358	
stcB33B7_3867	CCTGACGATTTCCTGCAAGT	GCCTCGCAGTAAAAGATTTTAA	358	
stcB79B4.C22.1	CTCAAGCTCAGAGAAGGGGC	ACCTGATTCTCTCCGGCTTT	181	

stcN75B3.C22.2	AGGACGATGGATGAAACAGG	AGGCAGACGAACTCTGGAAG	185	
stcN75B3.C22.3	GGTATGTGACCAAGGCGTGT	CAGCTCATCAGCCCACAGTA	122	
stcN75B3.C22.4	TGGGGGAGATGAGATCTGAAG	GGCTGTCCCTGGCATCTCT	75	
stcN75B3_28674	CGTGAAGCAGGAAAGAAAGG	AAGGGCAGCTCTTGGATGT	67	
stcN128A12.C22.2	AGACACAAGGAAGGCCACAC	ATGGCAGAGCTGACC	246	
stcN128A12.C22.3a	GGAGTCCAGGGGAACCTAAA	CAAGGACATTCAGTGGTTGC	123	
stcN128A12.C22.3b	CTACAGCACAGATGAGGCCC	GCTCTGGACAGAAAGATGCC	141	
stcN128A12.C22.4	AAGTGGCTGCTGAGGAGGTA	TCCAGGAATTGCTCTTCCAC	158	
stcN128A12A	TACAGGCTGGAACTGCCAG	TGGGAGTTAGTGAAGGGGTG	260	
stD22S1012E.1	AAGGGAGTGTGATGATTC	AATACGGGGTTGAAAAGG	170	
stD22S1015E.1	CACGCACCCCTCTTAATTGAAC	GAAGCCGCACCAAGGAAC	85	
stD21207	CTTCTGCTTGTCAGTGGGGT	TAGTAGTTGAGCACGGTGGC	125	
stD26090.3	ACAAGATGCGGGAAGATGAC	AGAGCAGCGAGAGTTCAAGG	107	
stDIA1	AAACACTGAGGCAGCCTTGG	AAAGCCAATACGCAAGAGGC	272	
stDIA1na	GACAGAGCGAGCGCGG	GGGCAGGCCAGGCTG		1048
stDIA1nb	GCCACCATGGGGGCC	CGTGTGACCGTGCCC		932
stDIA1r	ACCTGAGCTGGAGGAACTCA	ATCTCCTCATTCACGAAGCC		114
stdJ32I10.C22.1	AGGGTGCTATTGAAGGGATG	CTGCAAGCAATAGGTCCCTC	90	
stdJ32I10.C22.1b	AGACCTGGAGTGGGAGGAAG	TCATCTTCTCCAGGCCCC	111	
stdJ32I10.C22.5a	AAACCTGCTACAGCCCCC	AGGACAGGGTGCAGGTGTCT	122	
stdJ32I10.C22.5b	AGGCCATCTCAGGCTGTTAC	CAGCTAGCACACGTTTCCAC	111	
stdJ32I10.C22.6	CTGTCAGGCCGCTCAGATA	TCTCCATGGAAACCACAAGC	103	
stdJ32I10.C22.7	CACATGGTCTTGATTGCTTGA	TCCACCATCTCCTTGAGACC	109	
stdJ32I10.C22.8a	GCTCTTTAGGCAGAAAGCCA	CCCGATGGACTCAAGGTAGA	103	
stdJ32I10.C22.8b	GTGAAGACGCCAGACTCACA	TTGCTCTGTTAGCTCCAGCA	118	
stdJ37M3.C22.2	CTCTGAGAACACCCAGCTCC	CAGAGGGAGTGGTCCAGG	122	
stdJ37M3.C22.3	TGTATGGGAAGGAATGCTGTC	CTGGGATATGGGAAGAGCC	79	
stdJ37M3.C22.4	TCTGTCTCCTGCCTGAAGGT	CCAGTGGGATCTTTGCTGAT	101	
stdJ37M3mus1	CGTGAGAGCAGAAGGTGACA	GATCAACGGGATACCATCCA		136
stdJ41A17.C22.1A	TGACCATGCCGCTTTTCGTCGACTCTTA	CTTGTTAGTGGGCACCGAG	130	
stdJ41A17.C22.1B	TGACCATGCAGGTGTGTTTTGATTGTGG	CGAATAAAACTCAAGTGAACACCA	85	
stdJ41A17.C22.1C	TGACCATGTTCTTTCCATTTAGATCTACAGAG TTG	CACTAGCGTTTCCTCCGACT	85	
stdJ41A17.C22.1D	TGACCATGTTCATCAACATGGGTGTTCC	ATCAGTGCCATGCTTCCG	122	
stdJ41A17.C22.1E	TGACCATGGCTCAGCCCTGTGTTGTTTT	CTCAGGAGAAACGTGAGAGG	114	
stdJ41A17.C22.1F	TGACCATGTGACTTTTCTCGCCTTGTTTC	TCCAAAGTGGTTTCCACAGG	100	
stdJ41A17.C22.1G	TGACCATGTGCGCTAATATCAATGTTTTGAA	AAGTAGCCTTTGTTGGTACATTAAG AC	87	
stdJ41A17.C22.1I	TGACCATGAAAAAGGAAGTTTGGGAGCTCAG	ATTGATTCTTCTTTAGATACCACCTT G	130	
stdJ41A17.C22.1J	TGACCATGTGTTTCATCATTACGATTAGCCT	CTTCTGCAATTTCCAAATCC	129	
stdJ41A17.C22.1K	TGACCATGTGTTATTTCACATTCAGTGACTTC AG	GAGCTGGAAGTAAAATATGAATCG	120	
stdJ41A17.C22.1L	TGACCATGACTTTGACGAGCCAGTGGAG	CTGTCTGAATAGCCTGTGGTTTT	130	
stdJ41A17.C22.1M	TGACCATGCGCCGCAAGCCAGATTAT	ATCAGCCTGGGATTGTCTTC	114	
stdJ41A17.C22.1N	TGACCATGACATTTCCTTTTCCTTGTCTTCTC	CTGCTTCCTCGGCTCCTC	111	
stdJ41A17.C22.1O	TGACCATGCTACAGCCTGTCCAGTGTGC	CGACTCCATTAGCAAAGACG	104	
stdJ41A17.C22.1P	TGACCATGCAAGTAACCACATCTCAGGCGG	GCAAGGATATACACAGAGACGC	200	
stdJ47A17.C22.2	AGACGCAGGAAGAGGAGAGG	CTCCAGCTCGTGGTTCTAGG	148	

stdJ47A17.C22.2b	GCCCAAAGACCATGCTGC	GAGGTGCCGCCATGCTAC	101	
stdJ47A17.C22.2c	GGGGGAGCAGGTGAAGG	CTCCAGCTCGTGGTTCTAGG	97	
stdJ47A17.C22.3	CAGGGTGTACAACAGAGGCA	CAAGCCCTTCGACCTCTAAA	207	
stdJ47A17.C22.4	TAAGGGCCCATAGCACAGAT	AATGCCATTGAGGAGGAGG	117	
stdJ47A17.C22.4b	CCAGGGTGTCACAGTGCTC	GTGTGCAGAGGCAGAGTGTG	75	
stdJ47A17.C22.6a	CAGGAGAGCCAGAGGCTATT	TCCTGGAGTTCCAGTTCCTG	108	
stdJ47A17.C22.6b	GCAGGCCCTGGCTATAACTC	TGTTGTCATTGTTCAGTATATCTCCC	100	
stdJ47A17.C22.7	GCAGGATGGAGGCCTGAG	GTTCTCACACACACGCCC		
stdJ100N22.1	AGGACAATAATGGTGGCTGC	GCAGGTATGCTGGTTGTCAC	104	
stdJ100N22.C22.1	AGGACAATAATGGTGGCTGC	GCAGGTATGCTGGTTGTCAC	104	
stdJ100N22.C22.3	CGGCAGCTGCTCTAAAGTTC	CTGTGTTGTCACGGGTCTTG	120	
stdJ100N22.C22.4	GAACAAGCTGGAATCTTCGG	TCAACTCCCTTCACCCTGTT	120	
stdJ100N22mus1	TTGTTTATGGAGTGGGAGGC	TTATGGTGCTACCGGAGGAG	139	
stdJ102D24.C22.2	TTCCTTATTTTCGTCGCCTG	CCTCCTCTTTGCCAGGTTG	134	
stdJ102D24.C22.2.5tag	GGCCGCGGCCGCAATGAGGCAAAATGAC	GGCCTCTAGATTATCGACTTCCTG		
stdJ102D24.C22.2b	CCATCGAGTCAGTGGAAAGG	TTCTCCTGGATTTGCTGCTT	223	
stdJ102D24.C22.2r	GAGCTCTGGACAGCAGCAAC	TGAAATAGTCTCCCGTGGGT		102
stdJ102D24.C22.4	CTCGGAACCTGCAGAAAGAG	CACAGACGGTGCTCAGTGAC	122	
stdJ102D24.C22.5	CAATTAGAGGTTTGTGGGGC	AAAAAGGAAACACAGCTCCAA	101	
stdJ102D24mus1	GAGGGCTGCTGCAATACTTC	ATTTCCAGTCATTTCCACCG		129
stdJ127B20.C22.3	GATGAGCACTGTGCCGC	AGGCAGCTGAGGTGGTAGG		
stdJ127B20.C22.3b	GCCGGTCTTCCTTCTTCTTT	TCATTCAGCCTCAAGTGCAG	111	
stdJ127B20.C22.3r	CCTACCACCTCAGCTGCCT	GCACACCCTCGTCTTTCTTT		105
stdJ127B20.C22.4	ATGCCTTATGTTCCTGCACC	AGGTGGAGATGATGGCGTAG	134	
stdJ127B20.C22.4	ATGCCTTATGTTCCTGCACC	ACCTTTTTGCTGGAAACCCT	134	
stdJ127B20.C22.4	GGCCAAAGACTCTGATTCCA	AGGTGGAGATGATGGCGTAG	134	
stdJ127B20.C22.4	GGCCAAAGACTCTGATTCCA	ACCTTTTTGCTGGAAACCCT	134	
stdJ127B20.C22.5	AGGAAGCCCATGAGGTAGC	CTTCCAGGTGTGGAGAGCAG	102	
stdJ127B20.C22.6	GGGAGAAGACCCTGATTTCC	ATGATTTCAAAGCCCTGTGG	111	
stdJ127B20.C22.7	GAGACCTTCCAGGAGCGTG	ATATCAAAGGCCCTGGGAAC	104	
stdJ127B20mus1	TTCCAAATGTTCCGTGGTAA	CTCAGGCCACCAAAAAGAAG		107
stdJ127B20mus2	AACCTCTGCAAGGAGCTGAC	GAGACTCTGGCCACAGGAAG		147
stdJ127B20mus3	GCAGCTTGGAGAGAGAATCC	GCTTCCTTTAAAATGTTTATTGCC		106
stdJ181C9.C22.2.b	AGTGACCTGGGCAAGAGAGA	GAGGCTGAAGAGCTCCTGGT		156
stdJ181C9.C22.2.c	AGCTACAACACGCCTCTGCT	GACGTCATCTCCGACACAGA	104	
stdJ181C9.C22.2.d	GGACCAATACGTTGAGAACGA	ACTCCTTGTATGCGCTCTGG	104	
stdJ181C9.C22.2.e	AGAAGCTCCAGAGCCTGCAC	TATTGCAGACTGACGCCAAA	103	
stdJ181C9.C22.4	ATCTCCGTCTCATCCTCTGG	GGTATGGAGGGCAGAGAGACT	82	
stdJ181C9.C22.5a	CCTTCCCCAATTCAATGCT	ACGCACTGACAACTCGATCA	100	
stdJ181C9.C22.5b	ACCTGGCCAGTTCCTTCTGT	ACCCTATGCAGGGAAGCC	136	
stdJ185D5.C22.1	GTTTTCAACCAAAGGGATG	GAAAATTCATGAAACCCCCA		1196
stdJ185D5mus1	TGGCAAACTCTTCTTGCTCA	CTGGGGAACATTTGCCTG	83	
stdJ185D5T7	TCCATTTTGTTTGTTTGTTTGC	GGACAGAAATCACAGCTGCA	180	
stdJ222E13.3.1	CCCATGCCTCTATGTTACCC	AATAACAAATGTTTATTCAGAAATGGA	115	
stdJ222E13.3.2	CCCTCTCACCAAAGTGGTTC	GGGTTCTTCCTTGTTCACCA	180	
stdJ222E13.3.nest1	GTAGTTCGTCGCTCCCTAGC	CACAGGGCAGAACACAACAC		1523

stdJ222E13.3.nest2	GACTGCTTTCGGCTTGCTC	GAAACAGAGCCACCCTCCTC		1348
stdJ222E13.C22.1	TTCCAGTTTGTGGAAGTCCC	GGAGCATGTGTGTGCACTGT	103	
stdJ222E13.C22.1.c.3tag	GGCCGAATTCAGAGCGATGAGTGAGAACGCC	GGCCGCTAGCACAGGACATGGGCCTGC		610
stdJ222E13.C22.1.c.5tag	GGCCGCGGCCGCGATGAGTGAGAACGCC	GGCCTCTAGATCAACAGGACATGGGCCTGC		610
stdJ222E13.C22.1.e.3tag	GGCCGAATTCAGAGCGATGAGTGAGAACGC	GGCCGCTAGCCAGCTGGGCTGGGAGC		753
stdJ222E13.C22.1.e.5tag	GGCCGCGGCCGCGATGAGTGAGAACGC	GGCCTCTAGACTACAGCTGGGCTGGGAGC		753
stdJ222E13.C22.1b	AGCAGTCCACGGATATTTTGA	GGGTGGATTTCATCGTGTCT	86	
stdJ222E13.C22.1n	GACGAGAGCGATGAGTGAGA	CAGACTCTGTCTCCCCCTTG		1151
stdJ222E13.C22.1r	CTTCATCAGCAGGGAGCTGT	CATGAACGACAGGGACTCCT		130
stdJ222E13.C22.3	TGACAGCCCATCAATGAAAA	GAGTGCCTTCAGGATGGTGT	109	
stdJ222E13.C22.3n	GTAGTTCGTCGCTCCCTAGC	AAACAGAGCCACCCTCCTCT		1413
stdJ222E13.C22.3r	TGGTGAACAAGGAAGAACCC	AGTGACTCGAGGGTGCAGAT		107
stdJ222E13.C22.6	CTTGGTTTAGGAAGTGGGGG	AGTCTTAAAGCCCCTGCACC	126	
stdJ222E13.C22.7	GGAAAATAACGATTCGGGGT	CAAAGTAGGCGGGTCACCT	100	
stdJ222E13mus1	AGGGGGCTTCAGTGTTCAG	CTGCCTGCTGGGATGAATAG	81	
stdJ323M22.C22.2.b	GACCACATTGACCAGTGGAA	CTTCAGGATGGCACCTTCTC	117	
stdJ323M22.C22.2B	TGACCATGATTTCATCTGTGTGTTAACTTTTAGGG	CATAGGCACTGACAGAGCACC	200	
stdJ323M22.C22.2C	TGACCATGTATTTTGATCTCTTCCAGGATTATG	CAGTAAAAATTCCAGTCCTCGT	135	
stdJ323M22.C22.2D	TGACCATGAAATAACCCCTGTGATTTTGTTTT	CAGACTCTGAAGTCAGGAAACTCA	95	
stdJ323M22.C22.2E	TGACCATGGCAAACCATCCGAAATGTGT	CCAGATAGAGGTATTTTCCATTTTCA	200	
stdJ323M22.C22.2F	TGACCATGAGTCACCTATATGCTGCCCG	GAAGATGTCTTGCTGTCCCG	171	
stdJ323M22.C22.2G	TGACCATGTGTGTCTCAATCTAATAAGGAAGCC	GTAACAGCGCAGTGGACG	130	
stdJ323M22.C22.2H	TGACCATGAATCTACATTCCCTCCTAGGTACAAG	GTGTTTCTGGATGGCGACG	125	
stdJ323M22.C22.2I	TGACCATGACATCCATGGGGGCAAGT	GCCACAGCCTTCAGGGACT	130	
stdJ323M22.C22.2J	TGACCATGCATGTCACCCACTGGTCACT	GATCAGCCAGGGCTTCAG	114	
stdJ323M22.C22.2K	TGACCATGGTGAATGCGTCCCCGTCT	TCTCGTAATTGCCGAGGACT	160	
stdJ323M22.C22.2L	TGACCATGATGATGAAGAATTGGCCCAG	GGGAAATTTCAAAATAGGGCTT	200	
stdJ323M22.C22.5	AAGGAAACAAGCTGATGACCA	GCTGCTGATTCTGATAGCCC	96	
stdJ323M22mus1	CTGACCAGTCGGTGGGATAG	ATGACCTGGAGCAGAGCATC		105
stdJ323M22mus2	TTTTTCAAAGGCTGAGGGAG	GCAGCCTGAACGGAGTGT		110
stdJ323M22_103421	ACCCAGACATCACTTCCCTG	ACAGATACTGTGGCTTGGGG	149	
stdJ345P10.C22.3r	TTGGCATGCATAGGAATTGA	TGGTTTTATCCTCCGTGAGC		145
stdJ345P10.C22.4	CAGCGTTGCAAAATTGATTG	AGCCGGCATGATAATGCTAT	120	
stdJ345P10.C22.4r	CTGCGTATTCAGCCCAAAAT	CGGAGGAAGTCGTTGTAGGA		215
stdJ345P10.C22.6	ACTATTCGCCTGCTGTCCAT	CCTTCTCTGTGCTCTGCTGC	120	
stdJ345P10.C22.7	CGCTCATATGACCTTGGAGA	TTATGAAATGTGCATTCGTTAAAAA	145	
stdJ345P10.C22.7b	AAGGGATGGAAGTGTTGCAG	ATGTGCAGATGGCACAGAAC	145	
stdJ345P10.C22.9	CACACTTTCGTGGTACGCAG	CATCCTCCAGCAATGTCCTT	201	
stdJ345P10mus1	CAGCGTTGCAAAATTGATTG	GGCAATGGCATTCTTCGACG		576
stdJ345P10_6233	GCATCATTGCTTACCTGCAA	CGGAATGATGCCTGTGTGTA	271	
stdJ345P10_7536	AATGACTGCAACCCAGCTCT	CCAAACATACGGCACACAAC	205	
stdJ345P10_8395	TGGGCTCAAGTGTCAACGTA	CTGTGGGTTGGAAAAGCTCA	123	
stdJ345P10_124665	TCCAGAGTCCCCCTCAGTTA	GACAGAGGCAAAGGGACTGA	69	
stdJ388M5.C22.2	GATTATCAGCGTGTCAGGCA	ACCCAAGCATGACTTTCAGG	147	

stdJ388M5.C22.3	TCTGACCTCCACAATGGAGAC	CTTGAAAGGTGCCTCGGTAG	1024	
stdJ388M5.C22.4	TGAAGAAAGCATGTACGAGGG	TGTACTCGAAGGGCAGTGTG	111	
stdJ388M5.C22.4r	GAGGTTTTCTCCAGGACCAAG	TGAGGAGCTATCTGGGGTTG		101
stdJ388M5.C22.6	TTCGTTAGGGTGAGCAGTCA	AGGTTACAGCCTGGGCAAAG	131	
stdJ388M5.C22.8	TAAAGTGGGGATAGGGAGGC	CTGGCTGTCTTCCTGGATGT	134	
stdJ388M5.C22.8b	CGGGATAGGAACCAGATTCA	GGACGTGGAAGGTCCAGAT	111	
stdJ388M5_56861	GATTATCAGCGTGTCAGGCA	ACCCAAGCATGACTTTCAGG	147	
stdJ398C22mus1	TCTTCAAAGAGCAGCGGAAC	AGTTCTGGGAAAGCATGCAC		388
stdJ437M21.C22.1	GCCAGATTATGAGCCAGTTGA	CCTGGGATTGTCTTCCAAAA	101	
stdJ437M21.c22.1b	AGTAACCACATCTCAGGCGG	GAGGGTGTGACAGGAAGGAA	171	
stdJ437M21.C22.2	CATCAACATGGGTGTTCCAC	CTTCCGTGTTGCTTGAGAGG	108	
stdJ437M21.C22.3	GCTTCCCACTCAATCCTTCA	TGTGAAACGCTTCTGGTGAG	116	
stdJ437M21mus1	TCCTTAACTCCACTGGCTCG	ACCCATTATGGCAAAACCAA		101
stdJ474I12.C22.2	CCAGGGCTCTACCCTCCTAC	TTGCCACGAAACATCCAATA	102	
stdJ474I12.C22.2r	TATGGACCTCAAAGGGCAAG	TTGACTTCTGGAAACCTGGG		180
stdJ474I12.C22.3	CAACCTCCAAACTGTGGCTG	TGCCTTGAAATTTTTAGCTCTG	74	
stdJ474I12.C22.3b	CTGAGCTGGGTGTGTCCTC	AACAGGACAAAGCGGAAGTG	116	
stdJ474I12.C22.3c	AACCTTCTGCAGAGCCTGA	AAGTGTTGCTTCCAGCCCT	80	
stdJ474I12.C22.4	GCTAACAATAGCCAAGCACGA	TCTTCCCTGATGAGGCTTTG	81	
stdJ474I12.C22.5	TGGTTACACATCCTGCTTGG	TTCTGTGTCATGCTGTACTGACC	100	
stdJ474I12.C22.5r	TCACTGAACCAGAGGGGAAC	GATGAGGCTTTGGAATGAGC		129
stdJ474I12.C22.6a	GTTACCTGGCAACTAAGCCG	TGGGTTTTCTGAGACAAGGG	115	
stdJ474I12.C22.6b	ATGCGATTTGCCTTTTGAAT	CCTGTGCACAAGGTAATGAAGA	100	
stdJ526I14.C22.2	AGCGGTGTCTCCTCTTTGAA	TATTGCAGCACGAACTGAGG	239	
stdJ526I14.C22.2na	CGGGTGCTGGCG	CAGGGCCGGTGT		2051
stdJ526I14.C22.2nb	GCCATGGAGGCCGAG	CCTGCCCCAAGCACAG		1966
stdJ526I14.C22.2r	ACCTCATGCTGAAGTGGGAC	AGACAAGGCAGGTAACGTGG		183
stdJ526I14.C22.3	TGGTGACTACACCGGCTACA	ATCTCAGGGACCACGATGAG	123	
stdJ526I14.C22.3b	AAGCTGATCAAGGCCCTCTT	AGTTTGATGAAGGACCGTGG	104	
stdJ526I14.C22.3na	CGCCCGCACGCCCAG	CTCGGTCTCCATGGGCTG		3016
stdJ526I14.C22.3nb	GAGCATGGGCGCG	CGGCTTCCCTGAAGGGC		2935
stdJ526I14.C22.3r	CTCATCGTGGTCCCTGAGAT	GCTGTTGCCTTCATTGGATT		189
stdJ526I14mus1	AAAGAGCAATATTTAGGTTCATTTCC	TGCATGTAATGAAACATTTGTGA	70	
stdJ526I14mus2	CCCAACTCATGACTTCTGCTG	CCCCATATGCATGTAATGAAA	107	
stdJ549K18.C22.1	GCTTCCTGGGCTTCTACCAC	GGGATACCGGAGAGGACG	100	
stdJ549K18.C22.1.3tag	GGCCGAATTCGCCGCCATGTACGACGCAGAGC	GGCCGCTAGCCAGACTCTTCTCTAGTG		1446
stdJ549K18.C22.1.5tag	GGCCGCGGCCGCCATGTACGACGCAGAGC	GGCCTCTAGATCACAGACTCTTCTCTAGTG		1446
stdJ549K18.C22.1b	TGGAGGAGTGAGTGACAACG	GAGGCGTAGACTGAGCTTGG	145	
stdJ549K18.C22.1cds	CATGTACGACGCAGAGCG	ACACAGCAATGCGGAGGTAG		
stdJ549K18.C22.1n	ATCCCGACCCAGATCCTAAC	CCATTAATAGGGCCACGAAA		1784
stdJ549K18.C22.1r	GATGTCCTGTGGTTGCAGTG	TGAAAAACTGGGAAAGGTGG		285
stdJ549K18_14129	GCTTCCTGGGCTTCTACCAC	GGGATACCGGAGAGGACG	100	
stdJ549K19.C22.1snp	AGACTGGTGACATGGCTTCC	TGGGAAACTTTAGCACCTCTG	390	
stdJ671O14.2.nest1	CCTTCTCGAACCCTGCTATG	AATCCAACGATGGAGACAGG		1415
stdJ671O14.2.nest2	TGCACTCAGTAGGCCTTTGTT	CACGAATGCACAGGAAACAG		1154

stdJ671O14.C22.2	TTGCAGATGGCGTGTACCT	CTTCTGATCGAAGCTTTCCG	102	
stdJ671O14.C22.2.3atag	GGCCCTCGAGGAGGCGATGGAGCCGGAGTTC	GGCCGCTAGCATTCGGGGCTCCATGGGGCG		996
stdJ671O14.C22.2.3btag	GGCCCTCGAGGAGGCGATGGAGCCGGAGTTC	GGCCGCTAGCCGATATCTGCGGGCCAAAGG		820
stdJ671O14.C22.2.5atag	GGCCGCGGCCGCGATGGAGCCGGAGTTC	GGCCTCTAGATCAATTCGGGGCTCCATGGGGCG		996
stdJ671O14.C22.2.5btag	GGCCGCGGCCGCGATGGAGCCGGAGTTC	GGCCTCTAGATCACGATATCTGCGGGCCAAAGG		820
stdJ671O14.C22.2b	GGCGATGGAGCCGGAGTT	GCAGGAGAGTTGGGAGTGAG		810
stdJ671O14.C22.2na	GGACCACCCTTTGCACTCAG	GAGCAAGGCCAGGCGG		1014
stdJ671O14.C22.2nb	GCGTGGAAAGCGGTTGGG	GGGTCACGAATGCACAGG		1000
stdJ671O14.C22.2r	CTGCACAACGTCACCCTG	AGTGTGCTCTTGGCATCCTT		104
stdJ671O14.C22.4	TTCTGGTCCAAGCTGTGTTG	GCATTTTCAAAGGCTCTTGC	136	
stdJ671O14.C22.7	TATCGTGAACAAGGATGCCA	CTGTCCCTGTGTGCCTTCTG	84	
stdJ671O14.C22.8	CATCCTCATCACCTCCACTG	TGGTGATGATGTCATGGTGA	170	
stdJ671O14.C22.9	CGTTCATGTCACGTCTTGCT	AGCGGCACATGCTCTTTTA	145	
stdJ671O14T7	TGGACAGCCAGGCAGAAT	TGATAGTGCCCAGCTCTGG	122	
stdJ753M9.C22.1	GAGCCAAGGCTCAGAGATGT	TCTCAGGCCAGTTGTCAGTG	103	
stdJ753M9.C22.2	CATGCAGGTCTGCTGATTCT	CAGGCTGCTGATTACGATGA	139	
stdJ753M9.C22.4	AGAGAACAAAGCTTGGGACTTATT	AGACTTGCTAATCAAATCAAACCA	100	
stdJ753M9mus1	AATCAAATCAAACCATTCCACC	CTTGGGACTTATTTATTTCCCGT	80	
stdJ753M9mus2	TTAAAAACACAAGTTCCAATGGC	TTTTTCTCCGGCCTTTTAGA	100	
stdJ754E20A.C22.2a	CTCAGCAGGGATGGAAGAAG	CTCCCTGCTCATCATAACCG	106	
stdJ754E20A.C22.2b	TCTTCCTCCTCTTCACCAGG	GAAAGATCCCTGGGCACTG	101	
stdJ754E20A.C22.3	ATGGCAAAACCCCAATATCA	CTAGAGGGCTGTGGTTGCTC	161	
stdJ786D3.C22.1	TGAAAAGGTTGAAAAGGCCC	TTGGTATACAAAAGTGTCGAGGAC	100	
stdJ786D3.C22.1r	CGCAGGAGCTAAGAAGGGTT	CTGCAGTATCCTTGTGGATGTT		102
stdJ786D3mus1	GGTCGAACCAGTCCAGTCTT	GGAGGTGGGAATGAAATGAA		140
stdJ786D3_34803	TGAAAAGGTTGAAAAGGCCC	TTGGTATACAAAAGTGTCGAGGAC	100	
stdJ796I17.C22.1	AGCAGAGACCAAAGCAGAGG	TGAAAAACTGGGAAAGGTGG	118	
stdJ796I17.C22.1b	ACATGGCTTCCAGATATGCC	GGTCACTACACAGCAATGCG		419
stdJ796I17.C22.2.3tag	GGCCCTCGAGGGAACCATGGGGACTGTGC	GGCCGCTAGCCAGGAACCTTATCCCAGC		1410
stdJ796I17.C22.2.5tag	GGCCGCGGCCGCCATGGGGACTGTGC	GGCCTCTAGACTACAGGAACCTTATCCCAGC		1410
stdJ796I17.C22.2na	CTTCTGCCCTCAGCAGCA	ATTTATTGACGGGGTTTCCC		1612
stdJ796I17.C22.2nb	GGAGAGGGAACCATGGGGAC	GACGGGGTTTCCCACAGGG		1546
stdJ796I17.C22.2r	TCTCAACGCAGGAAACCTCT	ATTAAGTTCCAACCGAGCGA		142
stdJ796I17.C22.4	GGCTGCTCCTCCACAGGT	CTGAGGATGAACAGAACCCG	72	
stdJ796I17.C22.5a	GCACGAGTCAACCATAACCA	GAGGGCAGTCCTTTCCATC	89	
stdJ796I17.C22.5b	GCTCAAAAGGGTCCTCACTG	TTTGCCAAAGACATCACAGC	144	
stdJ1033E15.C22.1	TGTTCAGCTTGGTGTCCAGA	AGAACAGGATGAGGAAGGGG	106	
stE46L.nest1	CGGTTAGGGCTGTGTAGGG	CACTTGCAATTTCACATACGG		1642
stE46L.nest2	CTCCTCGCCTTCCTCCTC	TCCATGAAACAGATTCCAAAAA		1525
stE46La	TCCTACTCCTCCCTCCTCGT	GACAAGTTTGTGCGATTCCA		3092
stE46Lb	CAAGATGGCGGCG	TGCACACATCCATTGAGACA		3050
stE46Lr	TAGGCTTCTCGACGTCCTGT	TACATGAATCACCCGCAAAA		115
stFBLN1na	GTTGGCTGCCGAGGCTC	GGCAGCAATGATTTGGCCTG		2220
stFBLN1nb	GCCGCCCATGGAGCG	CCTGCGGCTGTGCGGCA		2150

stFBLN1r	ATCTCCCACACCGTCATCTC	CCTTCCGTGATGTCGAAGAT		131
stH55062	CACTGCCATGTGACTTTCGT	TTCCATCTTTGTCTCGATCTCA	126	
stH55082	CCTCTGCACTTCTGGTCTCC	AGGTGACTCCGGCAGAAAC	96	
stH55089	TCAGATACTGGTAATGCATGCC	CAACGGAGAATCCCATGC	147	
stH55091	AGTGTTTGAGTATCAGTGCCCC	CCCTCCTCGTTGAGATGGTA	121	
stH55092	TTGCAGCCCTAGTGTTTGC	GTTGAACTTGGACTGAGGCTG	146	
stH55095	AAGCAGCAGTGCCGAGAC	GGAGACACCATCAGACAGCA	81	
stH55162	AAGAGGATGTCAATGGGCAG	TCTCTAGGCCATGGAGCAGT	126	
stH55193	GATGGTGCAGGAGCAGTG	ACAAATGTGGCCTCCAGG	132	
stH55194	ATGCACTTAGCTTTGTAATGGG	CGCCACTTTCCTCCACATAT	141	
stH55200	GATGGCGTGTACCTGGTTCT	CGAAGCTTTCTGGACTCAGG	190	
stH55206	CAGAAGTCATGAGTTGGGGG	CAGTCCTCAATAGCGTCATGA	87	
stH55220	CTACAACCAAGGGAAGCCC	CAAGAATCTGCTCGTAGGCC	130	
stH55227	ACCTGGGACTTCTTCTTCAGTG	ACTGAAGGGTGATGGCATTC	127	
stH55235	GCCTTGCTGTCCTCCTGG	GCTTCAGGTCAGAGAACCATG	120	
stH55238	AGGGTATGAACTGCATGAACA	TGCAGTCCTTCTGGTTTTGG	120	
stH55275	AGCCTCCAAAGGACGTCTTT	AGATTCTGCACAGACAGGCC	126	
stH55328	GATGAACTATGTGGTCGGGG	CAGTCACTGGCAGCAATGAT	144	
stH55333	CTCTGGATGGTCCTCAGCTC	TTACTGTCCTGAGTGATGGGC	187	
stH55345	GATGAGGCTGCAATTGAGC	GAGTAGCCGCCGTTTCAC	81	
stH55366	GATGGGGTCATCTTACTCTTGC	GCAGGAGAGTTGGGAGTGAG	89	
stH55370	CCGGACAAAAAAATTGTTGC	TGTCTCCTAGGGGAGGGG	160	
stH55390	TACGATGAGAAGCTCCAGAGC	TATTGCAGACTGACGCCAAA	110	
stH55405	AACCTGGACATAAGCAAAGAGG	CATCAGTGAGCTTTCCTTTGC	134	
stH55431	ATTTGGACGCGGCGATTA	AAAGCAGTGACTTCCCATTACC	120	
stL11804.2	ATGCCACACCCAGTGGTATC	CCCAAATCAAGTGACCACAGT	87	
stL17306	GCTTCCCAGAAGAGCAGTTG	TGAATACAGTGTGCCCCAGA	202	
stL17306.2	TTGCTGTACCCTTACCTGGC	GGCCCAATAGTCATGAAAGC	187	
stM21019.2	CTCACAAGCTGGTGGTCGTA	ATGGCACCAAATTCCTCTTG	193	
stM21019.3	CTCACAAGCTGGTGGTCGTA	CATGGCACCAAATTCCTCTT	194	
stM57470.2	AGGCTGACCTGACCATCAAG	GCAGCTGCCTTTTATTGAGG	166	
stNUP50na	GCGTTTCTTTCCTCCCTTTT	CAGCAGCAACTTGGCAATAA		1800
stNUP50nb	GTTACAAAATGGCCGCCCGGAGC	GCAGCCGACTTTGCGTGTTC		1722
stNUP50r	AGAATGATGAGCCACCCAAA	CGCACCAAAAGCTGTGTCT		169
stPACSIN2.3tag	GGCCGATATCAAAATGTCTGTCACATATG	GGCCGCTAGCCTGGATCGCCTCCAC		
stPACSIN2.5tag	GGCCGCGGCCGCAATGTCTGTCACATATGATG	GGCCTCTAGATCACTGGATCGCCTC CAC		
stPACSIN2na	GCAGCCTGAACGGAGTGT	AGGAACACCATGAAGCCAAG		1699
stPACSIN2nb	GAAAAAATGTCTGTCACATATGATG	GGAACCATCATCTCTTGCAGG		1569
stPACSIN2r	CGTCAGTGAGAAGGACGACA	TGAAGCTCAGCTCATCATGC		224
stpryM746A3a.C22.1	CTGCTGGTAGCCAGGACTTT	ACGTCATGTCACCTTCCCAC	103	
stpryR7CC1.C22.1	TTCCTGCTCAAGAAAGCCAG	CAGAAACGCACACAGCTCAC	104	
stpryR7CC1.C22.2	AAGGACGTTGCGAGTTCAAT	ATGGACGATGACAGGGTTTG	122	
str14750	GCTCTTGAATGGGTTTTGGA	CTGTCTCCTCCCTCACTTGG	148	
stR45869	AGTCCCGCTACGTCCTCAAC	CACCACTTCTTGAAGACCCG	153	
stR75496.2	AGAGAGCCACAGAGGTGTGC	GTTGTTGTTGGCTCCTGGTT	150	
stRPCI232D12SP6	AAGGTGTGAGGACTTGGGG	AGCATTGGAAATGTAAATGAGC	145	

stRPCI232D12T7	CTGGTTCTAGCGTCCCTGAG	CAATTGCTGGGTAGAAGGGA	144
stRPCI237M16SP6	TCTGGGCTGTTTTGAGTTCC	CCGCCTGCCCTTTTAAAG	83
stRPCI237M16T7	CTTTTATTCATCAAGGCAGCG	AGTGGCACCGAATATTTTGC	124
stRPCI2310A16SP6	AATACAACACCAGCTGAGTCCA	CTGGACCAAGGCAGTGTTTT	190
stRPCI2310A16T7	TGTAAGCCAGCATTGGTGAG	CTACCACCTCCAACAGCCAT	181
stRPCI2313J12SP6	CAAGAGAAGACCGAGCGAAC	CTCTCTGCTTCCTGGTTTGG	127
stRPCI2314N23T7	TATGTGTTGGCATGCATGTG	GTTGAAGTGCTTGTGTAGCAGC	131
stRPCI2318M22T7	TGGCTATGACTGGGGCAT	CTGTCAGAACCCAGCAGTCA	157
stRPCI2322C11SP6	TCTGCACCTGAGATTCGATG	ACACAACCAAACAGGTGAAGG	149
stRPCI2322C11T7	TGTCTCACTTTGTCATGTTCCC	TCCTCTGGCCTTTGACTGAG	186
stRPCI2322L12SP6	TCCTCTGAGGCTCAACCTGT	TCGGAAAACATAAAAATGACCC	152
stRPCI2328O9SP6	CTCCCCATCCTTAGCTGTCA	TTAGCCCCCTCCTCTGAAAT	136
stRPCI2328O9T7	GGCTGGGGGAAGTAATTATAGG	TGGCCTTTAACTCACAGGCT	123
stRPCI2329O14SP6	AGGAAGATAAGCCTTTTGAGGG	GGCAGAAAATCACCTTGGAA	187
stRPCI2332M11SP6	CCTGCTCCCTTCCTCTCTG	TTGAACTTGCCTGCAAACAC	156
stRPCI2333H22T7	GTCTGTTTGGAATTAAGCGTCC	TCCACCTTGTCCATTGTTCA	125
stRPCI2339A17SP6	GTTCAAGCAGGAGAATAGGGG	TGTTGGTTGTACATTGCTTTCC	88
stRPCI2339A17T7	CCTGCCGCTGATAGCACT	ACACATGATTTTCACATCGCA	126
stRPCI2341K23SP6	AGCCTCTGATTCAGAAATCTGC	AGCGTCAGCAACCCAGAC	177
stRPCI2342N24SP6	TTGAGCTCTCAAGCCTCCTC	AAAGAGCAGCACGTGTCCTT	132
stRPCI2342N24T7	TGGGCTGCTGGAGACTAAGT	GACATATGGCAGGCAGGG	165
stRPCI2347E17SP6	GTGGTTACAGCTGACTGTGAGC	CTCAGGCACACAAGGAGACA	137
stRPCI2347E17T7	AACAACGAAATAAAAACAGGGG	AGGATAGTATCGAGGGCGGG	120
stRPCI2347F18SP6	ATCCCAGAATTGATATCAAGCA	GCAGGCAAAACACTCACTGA	162
stRPCI2351J16SP6	GCACAGGTATGAAGAACTCCTG	TTTTGTGCTAAATCTGGTGACA	121
stRPCI2351J16T7	AAGCTCTGCGGCTCTAGAGA	TTGAAACCGAGCAGGCTC	120
stRPCI2352G7SP6	AACCCAGCAGGAGGTATGG	CCAACTATGCCCACGGAG	149
stRPCI2352G7T7	ATGACCAACACACAGAATGCA	CTACTCCTGCCTGGTGTTTTG	160
stRPCI2357E6SP6	AGAGGTTCCTTCCTGTTTAGGG	TCCTTGCTTTCTGTGCACAC	162
stRPCI2357E6T7	CACAAGCCCTGAAATGGG	AAAAAAAAGCCAGCTCAGAGG	159
stRPCI2359G3SP6	GAACCTGTTATCTCTGGCATCC	AAACTTTCACGTGGACCCC	133
stRPCI2363B24SP6	CACAAATGAAGCTCCACGAA	GCTTGGGCTACATGTAAGTTCC	167
stRPCI2363B24T7	AAGAGGGCATCAGATCCCTT	AGCAAGGATGTTGCAGGC	162
stRPCI2364L3SP6	CTGAGTTCAATTCCCAGAAACC	ACCACCAAACGAAAGTGGAG	173
stRPCI2364L3T7	ACTGCCCTACTTGGGGATCT	GCTCAGTCCAATGGTTGGAT	169
stRPCI2367K8SP6	GCCAACTCCACGGGTAACTA	GAGAGCCTGAGTTCTCCCCT	128
stRPCI2367K8T7	CATTTTCCCACAGGTGCC	TAAGTGCTGCCCTTTGGC	121
stRPCI2378B22SP6	GCAGAGGAGGAGTGTCAAGG	TTTCCCCTAGCACCAAAGG	133
stRPCI2378B22T7	CTGTGTGAGATGCACGTCG	GGAAACCCCCACTCAATTTT	121
stRPCI2378M10SP6	ACAAACCCCCCAGAGCTC	CAGAACCAAGGGCTTCTCTG	131
stRPCI2380L7SP6	AATATCACCAAACACCAGGTCC	AAGTAGCCAAGGATGGCCTT	153
stRPCI2382C5SP6	AAACAAATGCCAAACTGATGC	TAGCAAGGACACGGAGCAC	176
stRPCI2382C5T7	ACAGACCAAGCACCACAGC	AGGAAGTCCCTGGGCAAC	146
stRPCI2382M1SP6	TACATCTGGAGCCACCAACA	CTCATAGTGGCCTTATTGGAGG	134
stRPCI2382M1T7	AGATGGAGGAACCCAGGC	CACCCATCTCTATAGGTCAGGG	140
stRPCI2383D12SP6	CAATTCACAATTGCACTGGG	TGACAGAGCCCAGATGAGC	145

stRPCI2383D12T7	AGTCCAGCTTCTGGGTCAGA	GGTCTTCATCTTGCACAGCA	150
stRPCI2389D3SP6	CTGGGGTTATCACAGAGAAAGG	CTTGCTCAGCCTGCTCTCTT	164
stRPCI2389D3T7	TCTTTCAGAACAAGCCATTGG	AGTGGATTGAGGAACTCCACA	131
stRPCI2389E20SP6	CCAACTTCAGAGCTCATGACC	CAGCACACCATATATTTGGGC	189
stRPCI2389E20T7	AAGGGATGGCCAAGTCTCTT	CTGGATGTCCTATGGCTCTACC	181
stRPCI2397J3SP6	GGACAGCAAATCTCTAACACCC	TTTCCCAACAAGATCATCAGG	142
stRPCI2397J3T7	TGTCTCACTTTGTCATGTTCCC	TCCTCTGGCCTTTGACTGAG	186
stRPCI23156C24SP6	AGCAGACTCACCTCCTGGAA	GAAATGATCCGGCCTCACTA	149
stRPCI23156C24T7	CTCCCCCATCCCCTAACTAC	TAAAGCCACCAACCCTGC	81
stRPCI23245E19SP6	GCAGATTGTAGACACCATTTGC	TGGAATAAAGAGGAACTGGACA	125
stRPCI23245E19T7	GGCAGTCCTTGCTTTGTCTC	CGGAACCACCTCAAATGC	123
stRPCI23246J11SP6	GTCTACGGCCAGACTTGGAG	GCCCTCGCAGAACTAACTTG	152
stRPCI23246J11T7	CAGGGCGTCACAGAGAAATT	TGACATGTCTGAGGCTGAGG	124
stRPCI23251B23T7	GTGAAGGAATGTGGACCTTCA	AGATGACACAGCACAGTCACG	129
stRPCI23255A15SP6	TTATTCTCCCTGGCAGCACT	ATTGTCCCGTGCTTTGAAAC	125
stRPCI23255P22T7	TTGTTTGGAAAAGCACAAACC	TTGCTGGACAGAACAAATTCC	190
stRPCI23260D9SP6	ATGGTCCTCCTGTCAACAGG	ATACCTCCAAGGGGAGACTAGC	132
stRPCI23260D9T7	ATTCAAAGGGAGGGCCAG	AGGCCTCAGTGTGTGTGATG	149
stRPCI23267E10SP6	ACATCATCAGGTTCGGGAAG	AATGAAAAGGTTGCGGGAG	131
stRPCI23267E10T7	TTCCAGGTGTCACTCAGCAG	CCAACAGTGCAGCAAGAAAA	148
stRPCI23267H21SP6	ATTCCATCAGGGGGAAAAAG	AAAGGTCCACAGCAGTTTGG	156
stRPCI23267H21T7	CTAATAAAATCCCCAGTGCTGG	TTGGGCAAAATCTTAATGAAGA	135
stRPCI23268I1SP6	TGGTGTGTGGTCCTCCAGTA	GCTGCAGAAGATATCTGGGC	170
stRPCI23274I12SP6	GGCAGAACTCAGCCAATAGC	AGGCAGGTGGAGCAACAG	172
stRPCI23274I12T7	TCTTTGGGTAAAGGTGCGTC	CTAACAGCTGCGAATCCACA	135
stRPCI23275O16SP6	CCTTTCCTTCTGGCCTTTG	AAAGACACGCATGCTCAGG	84
stRPCI23275O16T7	CACATCTCTGCTGCGGTCT	TGTGTGGTGCTGGGTTTTTA	153
stRPCI23277E6SP6	GCTTTTGACCAGCCTGAGAC	CATATTCAATTCTGGGACCACA	177
stRPCI23278I10SP6	ATGTCCTTGCAAAGAATGCC	TGGAAGCTCTTTAACCAACTTG	131
stRPCI23278I10T7	ATCAATGTAAGGTGGAGATCCA	AGATACATGTACACCTGCGCA	82
stRPCI23280B14T7	AAACTTGTGCAGGCTCCTGT	GAGATCTGACACCCTTTTCTGG	122
stRPCI23281D22T7	CAGAGGAGGGCATCAGAAAG	TATATGGGGAGCAAAGTGGC	145
stRPCI23283A19SP6	ATAATGCTTGCCACTCCTGG	TCAGGAGCTAACAGGTTCACG	150
stRPCI23283A19T7	CACGTGTGCTCCTTCACG	CCAGGTAGAGGATGCAGAGC	178
stRPCI23284G6T7	ACCTTTGGGCCATAGCAAG	CGTGTACACACAGATGGTTGTG	136
stRPCI23284H5SP6	AGCCTCTGAGATGACAATGACA	ATGTGGGGAACTCAAGATGC	176
stRPCI23284H5T7	ATTCAAAGCACACAAACAATGC	CAGGTTCGACCTTAGCCTTG	136
stRPCI23290L7SP6	AACATGTCACCAAGCACTTCC	AGCTCTGAAAAGTCTCCAGGG	161
stRPCI23290L7T7	GTGGCTATTCATCGTCCTTAGG	GAATCTCTGCAGTGACAAGGC	143
stRPCI23290M7SP6	TCTGACCATAGATGTGCCCA	CCAAACAACCAAGGAGGAAA	139
stRPCI23290M7T7	ATTGTAAAGCTTTCTTGCCTCG	GTACCATCTTTTAGGCACTCGG	176
stRPCI23292J15SP6	AGCAGAAAGTGGTTAATTGTGG	AAGTCTGGAGTCTACTCCTCGC	126
stRPCI23292J15T7	GAACCTGTTATCTCTGGCATCC	GTACTGCTGCAAACTTTCAACG	144
stRPCI23292L2SP6	GAGCATTAAGGAACAGGGCA	GGCAGCAGCACAGATCATAA	188
stRPCI23309F10SP6	CTTAATCACTGAGCCATCTCCC	ACAAAGCAAACAAACCCCAG	153
stRPCI23323K24SP6	TTTCCAAGTGCAGGGTTTCT	CAATTTGGTTAAGCTGAGTGAA	121

stRPCI23327F21SP6	GTGAAGGAATGTGGACCTTCA	AGATGACACAGCACAGTCACG	129
stRPCI23336I22SP6	GGTGGTTGTGCCTTTGTTG	AAGCCAGAATTCAGGCTCAA	139
stRPCI23336I22T7	CTGCTGGTTGTTGCTGTTGT	AGAAGTGTTAGAGGGGAAAGGG	170
stRPCI23341G1SP6	CTTTGCTCAAGGTCACAAAGG	GAACGACAGGAGAGAGACGG	171
stRPCI23341G1T7	GGCACTTCCTTGTGTTGGTT	GCACAGGTAAGTGCTCCACA	172
stRPCI23341G2SP6	CCTGCCCAGCAAGTACTAGG	CAGTAAAACAGCTGGTGTGCA	175
stRPCI23341G2T7	ACCTAGTGTGTTCCATGGTCG	GGGCAGAACTCCATTAGCAG	131
stRPCI23349K11SP6	GGCTTTTCTTCCTTGATCCC	AGCTGCTGATGGTGCTTTTT	163
stRPCI23349K11T7	CTGGGTCTTGGCTTTGTTTC	CTTTTAAAAAAGGCTGGGGC	127
stRPCI23351H18SP6	TTTTTGGTTACATTGGATGAGG	ATCCTACTCTGCAGAAAATGGC	172
stRPCI23351H18T7	GGATGTGAGTTTCAGCAGAGC	CTTTGCAGGTAACACAGTGCA	177
stRPCI23352H8SP6	ACTTCTGAACCCTCCCCG	GGGTTAGGGGTTGGAAAAGA	88
stRPCI23352H8T7	ACTGAACCCAGCCTGCTG	GACATCAAAGAGTCAGAGGCG	178
stRPCI23360I20SP6	GTCTATCCCCACTTCTGCCA	CAGAGAATGGCCTGGAAGG	84
stRPCI23360I20T7	TTCATAAAAGGCTCTCCTCAGC	GACCCCAGAGATGCCATG	128
stRPCI23364K6SP6	CTGGCACATCGCTGAATAAA	ACAGCTTTCCACCCCCTC	124
stRPCI23364K6T7	TTCCTGTCAGAGCTGTGGTG	GGCAGCGTTCCTGTCAGT	146
stRPCI23367C4SP6	TTTTCCGCAGTTACCTGGTC	TGTAAACTGCCCAAAGTGACC	154
stRPCI23367C4T7	GATCTCACTGCCATTGGCTT	TGGCCCAAGTGGAGTTATG	125
stRPCI23367D23SP6	AACACCCAAGCAGTAATATCCC	AAGGAGTGGGGAGAAAGAATG	122
stRPCI23368B15SP6	TCCCTGTCTTGGGATCTGAG	TTCATCTGGGCAGCAACTC	131
stRPCI23368B15T7	CGTGGATGGGAACTGAGAGT	TGGCCAGGAAGAACGAAG	161
stRPCI23373J12SP6	GTTCTGCCCTGCCCTTTC	AGGTTCATAAAGCTCCTTTCCC	185
stRPCI23373J12T7	CAATCTGCCCTCGAAATGAT	GAATGGAAGCTGTAAAAGGGG	126
stRPCI23373P4SP6	TGTGGTGCTAGTGGGCATAA	CAGCACAGACACTTGAGGACA	174
stRPCI23373P4T7	CTTTCTTTGCTTCCTGGCAG	GCTCTGACAAAAGCATCTTGC	131
stRPCI23379E23SP6	TACATCTGGAGCCACCAACA	CTCATAGTGGCCTTATTGGAGG	134
stRPCI23379E23T7	GAACATGCTTGACAGAGCCA	GGTGGAACACACACAAGCC	130
stRPCI23381K9SP6	TGCCATTCTCACATTTCCAA	ATCTGCCTGCTTACACCCC	150
stRPCI23381K9T7	CTGGCCTGGCTGTGAGAT	TTAAAAGCCATACCAACCGC	169
stRPCI23390H1SP6	AGGGCTGTCAGTTCTCTGGA	GGGACTAAACTCCCAAGAAAGA	127
stRPCI23390H1T7	GAAGTGGGCTCAGGACAGAG	AGGCTAAGGTTGCCCTGC	80
stRPCI23397G15SP6	GTGATGGAGACGCCACCTAT	CTTTTGGGTTGTGCTTACTTCC	137
stRPCI23402G11SP6	CTCAGAGCACATTCCAAAAGG	CATCACCTGTCTGAGCTCCA	186
stRPCI23402G11T7	TGTCAGTGTTGCCATATTCCA	ATCCTCGATGACTGGAGGG	157
stRPCI23402I10SP6	TTCAGTGCAATTGATGCAGA	CATCTCAATATGACTGCCTTCG	127
stRPCI23402I10T7	TGTGCTGTACCTGGCATTTT	ATCATTGGGATTTGAAAGTGC	190
stRPCI23403B19T7	AGGGCTCAGTGGGTAGAGGT	ACAAGGCCTCAGTGTGTGTG	130
stRPCI23405F6SP6	TCCCTGTCCTCTTGGGTTC	CAAATGAGGCAGAACTTCAGC	138
stRPCI23405F6T7	TAGTGAATGGTTTCCCCAGTG	TGTAACAGAAGTGGGAGGTGG	138
stRPCI23407C24SP6	TGCCAAACTTCCCAGAATTC	GACAGAATGGCAGAAGCCTC	153
stRPCI23407C24T7	GCAGAGCCTATGAAACCAGC	GCAAAACATTCATGGCTGC	133
stRPCI23412D17SP6	GGATTTCACCTCTGACTTCACC	ACAGCAAGATCACCCAATCC	146
stRPCI23412D17T7	CTAGTGGGAAAGTACCAGCCC	TCTGGGTAATGCATCTGTTCC	183
stRPCI23415K6SP6	ACAGCAGGTTCAGAGATTGACA	AAAGGGGATTTCCAAGCG	167
stRPCI23415K6T7	CTATCCAGCTTGTGTATCTGCG	CTTCTCCTCTTTGCCCCAG	138

stRPCI23437H2SP6	TCGGGACTATGATATTCTTGCC	GGACCCGCTGAACACATC	127	
stRPCI23437H2T7	GCAGGATGACTTCATCTGCA	TGTGCCCCAGTAATCTACTTCC	139	
stRPCI23452B7SP6	AAAGTGTGGACACTTTGCCC	ATCCCTGGATATGGCAATCA	125	
stRPCI23452B7T7	CTGAATACTCCCAGCCTCATG	TCCAGTGGTTCAGCTAAAAATG	88	
stRPCI23452L15SP6	TTCAGCAATCTGCTGTCCTG	GCTCCTTCTGGACAGGTCTG	137	
stRPCI23452L15T7	GGGGAATGTTGTCAGCAAGT	CAGGGTCATGGGGGTTATC	151	
stRPCI23453H13SP6	AATATCACCAAACACCAGGTCC	CCTCTTTCTCCATCTTGTTTGG	122	
stRPCI23454M13T7	GGACAGCAAGAACACAGCAA	CTGCCTCTTCCTCCTGTTTG	153	
stRPCI23458I23T7	CCACCAACCACAAGGACAG	TGTTCCTCTTTTGCCTGCTT	140	
stRPCI23458I24T7	ATCTTGCCTCCTGGGATCTT	GCTATGTAGACCAGGCTGCC	141	
stRPCI23466A19SP6	ACCAACAAGGCAGACAAACC	CTTGTTGTTTCGTCCCGG	129	
stRPCI23466A19T7	TTCTTTTTACCGTGCGTTCC	TCGCTTTCGTGGTTTTCTG	184	
stRPCI23470C10SP6	TCCTCTACTGGGGGCTCC	TTGACTCTGAAGAGGCTTTTCC	176	
stRPCI23470C10T7	GGGATCTGACACCCTCACAC	CCTCAGCCTCTTGAGGTCC	178	
stRPCI23476C13SP6	GTTTCTCAGACCCTCACCCA	AGATAGGAAGGTGGGCCCTA	130	
stRPCI23476C13T7	ACATCAAGATGGTCCTTCGTG	TTTCAGGACCGTCATAGTTGC	124	
stRPCI23478N15SP6	TTGAGCAAACATTTTTGAGAGG	TTTCTCTTCTTCCCTGCTTCC	133	
stSG1759	TCACAGAAATGCAGCCTCC	TCTTGGTGCTGTATTTGAAACG	153	
stSG28458	GGTAGACCTTGCTGTGGAGC	TAAGGAGACATGACTGGGTGG	121	
stSG30720	TGTTGGGTGTATGCATAAGAGC	TGTCAATCACTCTGTTGAGCG	177	
stSG34947	AATGTAACTGTGTGGTTTGCC	TTTGTCAAATAAGGGTTGTGC	124	
stSG35345	AAAGCACGTTGCAAACAAA	GCTTAAATGCAATGACCCC	122	
stSG35376	GTGCGATTCCACTGGTGT	AGGAGTCATTCTCTCCTGGG	168	
stSG41672	CTGAGCTATCCAGTGGAGGC	CCCCAGGCAGTTTAGGTGTA	176	
stSG43419	TCATTCAAAACCACATTTTTGG	CTTGCTTTTCAGAAATTACCGA	142	
stSG45455	TCATTCAAAACCACATTTTTGG	CTTGCTTTTCAGAAATTACCGA	142	
stSG46763	GGGTGGGGGTTTCTTTAAAA	CCCACCCCATCCTTCTTC	128	
stSG49686	AATCCCTGCCCTTTGCTCGTGG	GACATCTTCCGGATCAGATCATG	163	
stSG50904	ACATCAGCCTTTTGTGGGAG	CTTCCTATTTTTCCCATTTCCC	165	
stSG52068	TTCTCAGCTCCCGGTCAG	GACACCCTCAACATCGCC	145	
stSG52305	AAAGAACCCATTTCTTTCTCCC	TCAGGATTTCGCTGCTCC	103	
stSG58518	CAATGGGGAGTGTCACAGG	GGGGAGCTTCTCTGACTGTG	166	
stSG58579	GATGTGCCTGAGAACAGCAA	CCCAGTCTGTTTCCACCAGT	132	
stSG60712	TTAGCTCCCAGCCAGTGTG	TCATTGGTTGCTGTGCTCTC	137	
stSG63111	AGAGAGTCGCTTAGGGAAACG	GCGTACTGCTAACAGACCTGG	168	
stSG64851	TCAGCAATTTCAGCATCAGG	GGCAGGGATGATGAAGAAGA	102	
stSG66348	GTCTGAATGCAGATCACCTCTG	CACAGGACTCAGTGGGGG	135	
stSG88706	GCTCACAGCTGACGATGG	GCCTGAGATGTGGTTACTTGTTC		1634
stSG88707	CAGGTGTCCGACCATGAGC	TCCTCAGGAGGACAGAGGG		1231
stSG88708	CCCCGCAGGATGAAGAAG	GATGCCGCCTTTCTTGCT		1113
stSG88709	GATGGCGGACATCTCCCT	ATAAGCTTTGCCTTGGGGAA		1249
stSG88710	TCCCTCAGCCTTTGAAAAA	AGCTGCACTCGGAAAGGTG		1665
stSG88711	GCCGCCAGAGGAGAAATGT	TAAAAAGGCATCATGAGAAAACA		632
stSG88712	GATGCCTCCGCTCTGGGC	AGGGTTTTCCTGCAGTTGG		995
stSG88713	AATGTCTGTCACATATGATGAT	AGGAACACCATGAAGCCAAG		1590
stSG88714	ATGGCCCACCTGGAGCTGCT	CGACTAAGGTTTCTCTTAAAGGGT		3689

stSG88715	ATGGAGGCCGAGCGGGGTCCCGAG	AATTCTCCATGCCAGGAACA		2257
stSHGC13585	GGTGGGCCTGACAAACAG	AATCCTTTCCCCCTGTTTTG	132	
stSMC1L2	CATGTTCAGCCGAGTTTTGA	GGACTCTCCGTGTCTCTTGC	85	
stSMC1L2b	CCGATGATGCAGGTGAAC	AGCTGCTGCTTGTGGAAAAT	82	
stSMC1L2na	CCGCGGGCGCTTGATAAC	AGTGAGCCACAGCCTCTTTGG		1397
stSMC1L2r	GAAGTGGATGCAGCCCTAGA	CGTCGGCTCTGGAATAGAAC		127
stSULTX3.3btag	GGCCGATATCGGCGGCATGGCGGAGAGCGAGGCC	GGCCGCTAGCTAAATAAAAGTCAAACGTGAGG		781
stSULTX3.3tag	GGCCGATATCGGCGGCATGGCGGAGAGC	GGCCGCTAGCCCAACTCAAGAAGATC		855
stSULTX3.5btag	GGCCGCGGCCGCCATGGCGGAGAGCGAGGCC	GGCCTCTAGATTATAAATAAAAGTCAAACGTGAGG		781
stSULTX3.5tag	GGCCGCGGCCGCCATGGCGGAGAGC	GGCCTCTAGATCACCTACTCAAGAAGATC		855
stSULTX3na	GGCTGCGAGCCGGG	CTCCCTCCGCTCACGC		1050
stSULTX3nb	CATGGCGGAGAGCGAGG	GACTGTCTGGGTATTGTGAGC		899
stSULTX3r	AGATTCCTGGGGGTGTCCT	ACGTGAGGTCACACTTTCCC		205
stT81609	ACGTACCCTTCATTGATGCC	CTGAGCTTGGTGATGTCCAC	159	
stTTLL1.3tag	GGCCGAATTCTGGATTATGGCAGGGAAAG	GGCCGCTAGCCTTCCAGGTGGTGAGG		1272
stTTLL1.5tag	GGCCGCGGCCGCTATGGCAGGGAAAG	GGCCTCTAGATCACTTCCAGGTGGTGAGG		1272
stTTLL1n	CAGTGACGTCAGCGAGACC	CCAAAGAAGTGCCTTCG		1612
stTTLL1r	TCATCGACGACAAGCTGAAG	GCAGTCTGGGATTTCACCAT		149
stU73200	GAGCAGAGCCAGAAAGAAGC	TGTAGCCTGCAGCACATAGC	120	
stUPK3.nest1	GGGCGATGCCTCCGCTCT	AAGATTTTATTAAGGGTTTTCCTGC		1011
stUPK3na	CTGCTCGCTGGACCGC	GATTTTATTAAGGGTTTTCCTGC		1052
stUPK3nb	CGATGCCTCCGCTCTG	GGGACAAGCCCTGTTTCACCTTC		976
stUPK3r	GCCTCTCTGCATGTTTGACA	TGTCTTGCACTGAGGCATTC		107
stW57231.2	GGCCTGGATAGCTGTCATGT	TTAAGCCTCCCATCATCTGC	143	
stW61968	CTGTAAGCAGCAGTGCCGT	CAGTTATTGGCTGCCTGTGA	127	
stW77006	ATGACAAGGGCTTTGTGGAC	CCTGGTAGACCAGTAGCCCA	156	
stW77006	CCTGGTAGACCAGTAGCCCA	CCTGGTAGACCAGTAGCCCA	156	
stWI13406	CTGTTTCAGACTTTTATTACGTTGC	GCCGAAAATACACACTCTGTTT	150	
stWI14034	TACCGAGCTCATACAAATTTATCTG	AAAAATAGCATCGTGTTTCAGTT	150	
stWI16313	AGGTCCATGTGGCGCTCTAG	CACGTGGGACTGGGAGAAT	101	
stWI17470	CTGACACGTCCCTGTGTGC	GGAAGGCTGATGGTATTTCC	150	
stWI17858	CCCTGCAATCTGGAAAGAGG	AAGCTGTCCCCAGTGGGAT	150	
stX04405.2	GCCCACACTCTCTCCTTTTG	GGGTCTCTTCCTCTTACCCG	150	
stX12944	ACCATGCCCAAAAGAAAGG	TTCCCCTTCTTCCCCTTG	179	
stX53247	ACTCAGCCAATGTGATGGTG	TCTGTGGGTAGGAGAGTGGC	105	
stX56826.2	CACGCATATACCCGCTACCT	CCAGAGTGTTCATTCGAGCA	175	
stX61506	TTCTCTTCGCTCTGCTCTCC	GGTTACGCTGCTCCTTGAAG	171	
stX70854	ACATGCATCAATACAGAGGGC	AAATAGAAGCCAGCCTTGCA	200	
stX85124	CACAGTACATGGAGGGCATG	CCTGCTCCAGTTCTCGGTAG	159	
stX85124.2	AAATATGCCTGCTCAGGTGG	CTGTCTCTGAAATGAGGCCC	131	
stX87671.2	AAATATGCCTGCTCAGGTGG	CTGTCTCTGAAATGAGGCCC	131	
stX89986	ATGACCACTGCCCTGGAG	CTAAACACAGGCCACAGTTAACC	218	

Appendix 2a: Sequence databases (including latest dates/versions) used during the course of this project

Sequence type	Database	Description	Version/ date	Reference	URL
Nucleotide	22ace	cDNA data resulting from this project	-	-	-
	Incyte	Human ESTs	16.11.00	J. Seilhamer, Incyte, personal communication	http://www.incyte.com/index.shtml
	Unigene	Clustered EST sequences	6.11.01	Bogusket *et al.*, 1993	http://www.ncbi.nlm.nih.gov/UniGene/
	human_MGC	cDNA clones	21.03.01	Strausberg *et al.*, 1999	http://mgc.nci.nih.gov
	Trofatter_exons	Exon trap sequence		Trofatter *et al.*, 1995	
	eos	Predicted exon sequences that have been tested for expression by microarray hybridisation	29.05.01	R. Glynne, Eosbiotech, personal communication	
	embl_vertrna	EMBL 68 vertebrate RNA	68	Baker *et al.*, 2000	http://www.ebi.ac.uk/
	embl_htg	EMBL 68 HTG entries	68	Baker *et al.*, 2000	http://www.ebi.ac.uk/
	embl_htg2	EMBL 68 HTG entries, part 2	68	Baker *et al.*, 2000	http://www.ebi.ac.uk/
	embl_other	EMBL 68 (except EST, STS, GSS, HTG, vertrna)	68	Baker *et al.*, 2000	http://www.ebi.ac.uk/
	emnew_vertrna	EMBL 68 new/changed vertebrate RNA	68.9	Baker *et al.*, 2000	http://www.ebi.ac.uk/
	emnew_htg	EMBL 68 new/changed (HTG)	68.9	Baker *et al.*, 2000	http://www.ebi.ac.uk/
	emnew_htg2	EMBL 68 new/changed (HTG)	68.9	Baker *et al.*, 2000	http://www.ebi.ac.uk/
	emnew_other	EMBL 68 new/changed (except EST,STS,GSS,HTG,vertrna)	68.9	Baker *et al.*, 2000	http://www.ebi.ac.uk/
	EMBL_hum	EMBL 44 human entries (in pri.dat) only	44	Baker *et al.*, 2000	http://www.ebi.ac.uk/
	vertrna	vertebrate RNA (non-EST) in EMBL 51 (Jun-97	51	Baker *et al.*, 2000	http://www.ebi.ac.uk/
	worm34	worm34 C. elegans entries in EMBL 34	34	Baker *et al.*, 2000	http://www.ebi.ac.uk/
	dbEST	dbEST EST database	102701	Boguski *et al.*, 1993	http://www.ncbi.nlm.nih.gov/dbEST/
	dbSTS	dbSTS STS database	102701	Olson *et al.*, 1989	http://www.ncbi.nlm.nih.gov/dbSTS/
	dbGSS	dbGSS Genome Survey database	102701	-	http://www.ncbi.nlm.nih.gov/dbGSS/
	fugu	Fugu sequence from HGMP	2	Elgar *et al.*, 1999	http://fugu.hgmp.mrc.ac.uk/
	non_C.elegans_nematode_ESTs	nematode ESTs (excluding C.elegans) from EMBL	46	Baker *et al.*, 2000	http://www.ebi.ac.uk/
	fantom	Mouse full length cDNAs		Kawai *et al.*, 2001	http://genome.rtc.riken.go.jp/
	blastn_mus	Mouse genomic sequence resulting from this project		-	
	WGS mouse	Whole genome shotgun mouse sequence		MSC, unpublished	http://www.ncbi.nlm.nih.gov/genome/seq/MmHome.html
	nr_RNA	NCBI structural RNA database			http://www.ncbi.nlm.nih.gov
	gbrna	Structural RNA in Genbank 94	108		

PROTEIN	swir	Non-redundant Swissprot39+, SPTREMBL15, Wormpep24	23	Rice, P and Sonnhammer, E (unpublished)	
	swiss	Swissprot 39	39	Bairoch *et al.*, 2000	http://www.expasy.ch/sprot/sprot-top.html http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html
	swnew	Swissprot 40 new/changed entries	40.6	Bairoch *et al.*, 2000	http://www.expasy.ch/sprot/sprot-top.html http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html
	wormpep_current	Translated inhouse worm genes, rel. 25 (Nov 00)	25		http://www.sanger.ac.uk/Projects/C_elegans/wormpep/
	prodom	Clustered Swissprot 28	28	Corpet *et al.*, 2000	http://protein.toulouse.inra.fr/prodom/doc/prodom.html
	sbase	SBASE protein domain library (S.Pongor, ICGEB)	3	Murvail *et al.*, unpublished	http://www3.icgeb.trieste.it/~sbasesrv/
	hum24	hum24 Human proteins in Swissprot 24	24	Bairoch *et al.*, 2000	http://www.ebi.ac.uk/swissprot/Information/information.html
	sptrembl	Swissprot translated EMBL	15.0	Bairoch *et al.*, 2000	http://www.ebi.ac.uk/swissprot/Information/information.html
	trnew	Translated EMBL new entries listed by PID	15.9	Bairoch *et al.*, 2000	http://www.ebi.ac.uk/swissprot/Information/information.html
	swall	sp_tr_nrdb	110201		http://www.ebi.ac.uk/swissprot/

Appendix 2b: sequence analysis programs

Type of Analysis	Analysis	Reference	URL
Search	BLAST suite of programs	Altshul *et al*., 1990	
	Exonerate	Slater, unpublished	http://www.ensembl.org/Docs/wiki/html/EnsemblDocs/Exonerate.html
	BLAT	Kent, unpublished	http://genome.ucsc.edu/cgi-bin/hgBlat?db=hg7
Sequence content	RepeatMasker	Smit & Green, unpublished	http://ftp.genome.washington.edu/RM/RepeatMasker.html
	GC-profile	Durham, unpublished	
	CPGFIND	Micklem, unpublished	
	TRNASCAN	Finchant and Burks, 1991	
Exon prediction	XGRAIL	Uberbacher and Mural, 1991	http://compbio.ornl.gov/Grail-bin/EmptyGrailForm
	exofish	Crollius *et al*., 2000	http://www.genoscope.cns.fr/proxy/cgi-bin/exofish.cgi
	Hexon	Solovyev *et al*., 1994	http://dot.imgen/bcm/tmc.edu:9331/seq-search/gene-search.html
	xpound	Kamb *et al*., 1995	
	fexh	Solovyev *et al*., 1994	http://dot.imgen.bcm.tmc.edu:9331/seq-search/gene-search.html
Gene prediction	genscan	Burge and Karlin, 1997	http://genomic.stanford.edu/GENSCANW.html
	genewise	Birney, unpublished	http://www.sanger.ac.uk/Software/Wise2/genewiseform.shtml
	fgenes	Solovyev, unpublished	http://dot.imgen.bcm.tmc.edu:9331/seq-search/gene-search.html
	fgenesh	Solovyev, unpublished	http://dot.imgen.bcm.tmc.edu:9331/seq-search/gene-search.html
U12 intron prediction		Levine and Durbin, 2001	
Promoter prediction	Eponine	Down, unpublished	
	PromoterInspector	Scherf *et al*., 2000	http://genomatix.gsf.de/cgi-bin/promoterinspector/promoterinspector.pl

Appendix 4: Gene data

Locus name	Alternative names	Gene structure (22ace)	Genomic size(bp)	Transcript size (bp)	Coding sequence	5' UTR	3' UTR	No. exons	Type	Accession number
dJ222E13.C22.1		dJ222E13.C22.1.mRNA	20489	1399	942	127	330	12	Full gene	AL589866, AL590118, AL590120
dJ222E13.C22.2		dJ222E13.C22.2.mRNA	7936					7	Pseudogene	
dJ222E13.C22.3		dJ222E13.C22.3a.mRNA	31241	3440	1263	104	2073	9	Full gene	AL160111, Al160112
dJ222E13.C22.5		dJ222E13.C22.5	312					1	Pseudogene	
dJ222E13.C22.7		dJ222E13.C22.7	149	149				1	snRNA	J04119
DIA1		dJ222E13.C22.4.mRNA	30591	1954	903	82	969	9	Full gene	M16462
cB33B7.C22.1	A14GALT	cB33B7.C22.1.mRNA	3513	2020	1059	239	772	2	Full gene	AB037883
dJ47A17.C22.1		dJ47A17.C22.1	894					1	Pseudogene	
dJ47A17.C22.2		dJ47A17.C22.2.mRNA	2626					7	Pseudogene	
ARFGAP1		dJ437M21.C22.1.mRNA	60743	2699	1548	84	1067	16	Full gene	AL159143, AF111847
dJ437M21.C22.4		dJ437M21.C22.4	684					1	Pseudogene	
PACSIN2		dJ323M22.C22.1.mRNA	145373	3247	1458	192	1597	11	Full gene	AAD41781, AL136845
TTLL1		dJ323M22.C22.2.b.mRNA	145373	1618	1182	301	135	12	Full gene	AL589867, AL096886, AL096886, AF104927
bK1191B2.C22.1		bK1191B2.C22.1	365					1	Pseudogene	
BIK	NBK	bK1191B2.C22.2.mRNA	19111	1099	480	209	410	5	Full gene	X89986, U34584
bK1191B2.C22.3		bK1191B2.C22.3a.mRNA	11181	2048	1170	38	840	4	Full gene	AL359401, AL359403
BZRP		dJ526I14.C22.1b.mRNA	11698	850	507	89	254	4	Full gene	M36035
dJ526I14.C22.2		dJ526I14.C22.2.mRNA	20480	3353	1932	32	1389	14	Full gene	AL590888, D63487
dJ526I14.C22.3		dJ526I14.C22.3.mRNA	139477	3179	2802	60	317	22	Unpub. partial gene	
dJ100N22.C22.5		dJ100N22.C22.5.mRNA	2848					1	Rejected (PolyA)	AL442096
dJ754E20A.C22.4		dJ745E20A.C22.4	8897	951	189	217	545	3	Unpub. partial gene	

C22orf1	239AB	cB13C9.C22.1.mRNA	63348	2323	903	120	1300	4	Full gene	U84894
dJ345P10.C22.4		dJ345P10.C22.4.mRNA	283450	4866	4575	198	133	33	Pub. partial gene	AB051459
dJ345P10.C22.1		dJ345P10.C22.1	1035					1	Pseudogene	
dJ388M5.C22.1		dJ388M5.C22.1	607					1	Pseudogene	
HMG17L-1		dJ388M5.C22.2.mRNA	1460	1159	219	146	794	2	Unpub. partial gene	
SULTX3	SULT4A1	dJ388M5.C22.3.mRNA	37874	2347	852	0	1495	7	Full gene	AF188698, AF115311
dJ388M5.C22.4		dJ388M5.C22.4.mRNA	15878	1837	1218	230		11	Unpub. partial gene	
dJ549K18.C22.1		dJ549K18.C22.1.mRNA	23830	2805	1443	173	1189	9	Full gene	AK025665
CGI-51		dJ796I17.C22.2.mRNA	41112	1716	1410	156	150	15	Full gene	AF151809
dJ796I17.C22.3		dJ796I17.C22.3	322					1	Pseudogene	
bK414D7.C22.1	beta-parvin	bK414D7.C22.1.mRNA	144894	1650	1092	3	555	13	Full gene	AL159142, AF237769
dJ671O14.C22.2	gamma-parvin	dJ671O14.C22.2.mRNA	25679	1503	993	278	232	14	Full gene	AL590887, AL55092, AF237772
dJ671O14.C22.1		dJ671O14.C22.1	353					1	Pseudogene	
dJ671O14.C22.6		dJ671O14.C22.6.mRNA	42074	6440	396		6044	2	Pub. partial gene	AB051431
dJ32I10.C22.9		dJ32I10.C22.9	753					1	Pseudogene	
bK397C4.C22.1		bK397C4.C22.1.mRNA	289					1	Pseudogene	
dJ1033E15.C22.1		dJ1033E15.C22.1.mRNA	618					1	Pub. partial gene	AF086048
dJ1033E15.C22.2		dJ1033E15.C22.2.mRNA	1563	2677	339	195	2143	1	Full gene	AL136553
dJ474I12.C22.5		dJ474I12.C22.5.mRNA	3111	720	339	67	314	4	Unpub. partial gene	
dJ474I12.C22.2		dJ474I12.C22.2.mRNA	19083	817	354		463	5	Unpub. partial gene	
dJ474I12.C22.1		dJ474I12.C22.1	1285					1	Pseudogene	
dJ181C9.C22.1		dJ181C9.C22.1	692					1	Pseudogene	

ARHGAP8		dJ181C9.C22.2.mRNA	160453	2264	1929	224	111	17	Full gene	AL355192
dJ127B20.C22.2		dJ127B20.C22.2	833					1	Pseudogene	
dJ127B20.C22.3		dJ127B20.C22.3.mRNA	91484	3282	1296	63	1923	13	Full gene	BC012187
dJ753M9.C22.4		dJ753M9.C22.4.mRNA	6412					1	Rejected (PolyA)	AB051448
NUP50	NPAP60L	bK217C2.C22.1.mRNA	24114	5172	1404	403	3365	8	Full gene	AF107840
bK268H5.C22.1		bK268H5.C22.1.mRNA	48619	6306	1212	90	5004	10	Full gene	AB023147
UPK3		bK268H5.C22.2.mRNA	10893	1051	861	32	158	6	Full gene	AF085808
bK268H5.C22.3		bK268H5.C22.3	483					1	Pseudogene	
bK268H5.C22.4		bK268H5.C22.4.mRNA	32018	2879	1071	259	1549	9	Full gene	AK000642
SMC1L2		bK268H5.C22.5.mRNA	69557	4253	3705	52	496	25	Unpub. partial gene	
dJ102D24.C22.2		dJ102D24.C22.2.mRNA	18719	1392	927	399	66	7	Full gene	AL442116
FBLN1		bK941F9.C22.1.mRNA	98481	2112	2109			17	Full gene	AF126110, U01244, X53741, X53742, X53743
bK941F9.C22.6		bK941F9.C22.6.mRNA	21982	376	312		64	2	Unpub. partial gene	
E46L	SCA10	bK941F9.C22.2.mRNA	173511	3331	1425	265	1641	12	Full gene	AF119662
dJ37M3.C22.5		dJ37M3.C22.5	1226					1	Pseudogene	

Appendix 5: Mouse sequence clones.

Clone name	Sequenced by	Mapped to genomic location	EMBL accession and version number	Orthologous human region	Status of available sequence (5-10-01)
RP-10K12	Sanger Institute	MMU 15	AL583889.8	22q13.31	Unfinished
RP-121M7	Sanger Institute	MMU 15	AL583887.9	22q13.31	Finished
RP-85M21	Sanger Institute	MMU 15	Al591964.5	22q13.31	Unfinished
RP-150J22	Sanger Institute	MMU 15	AL513354.14	22q13.31	Finished
RP-237G11	Sanger Institute	MMU 15	AL603867.2	22q13.31	Unfinished
RP-23A6	Sanger Institute	MMU 15	AL626769.11	22q13.31	Unfinished
RP-95H5	Sanger Institute	MMU 15	AL603714.4	22q13.31	Unfinished
RP-74I9	Sanger Institute	MMU 15	AL611986.8	22q13.31	Unfinished
RP-320B4	Sanger Institute	MMU 15	AL611987.9	22q13.31	Unfinished
RP-180L12	Sanger Institute	MMU 15	AL513352.11	22q13.31	Finished
RP-98L10	Sanger Institute	MMU 15	AL626761.6	22q13.31	Unfinished
RP-451I21	Sanger Institute	MMU 15	-	22q13.31	-
RP-292L2	Sanger Institute	MMU 15	AL583891.15	22q13.31	Finished
RP-290M7	Sanger Institute	MMU 15	AL591946.9	22q13.1	Unfinished
RP-79F10	Sanger Institute	MMU 15	AL590144.5	22q13.1	Unfinished
RP-189A18	Sanger Institute	MMU 15	AL583893.17	22q13.1	Finished
RP-385C21	Sanger Institute	MMU 15	AL589692.9	22q13.1	Finished
RP-81H23	Sanger Institute	MMU 15	AL589650.13	22q13.1	Unfinished
RP-402G11	Sanger Institute	MMU 15	AL583886.11	22q13.1	Finished
RP-188D8	Sanger Institute	MMU 15	AL592187.4	22q13.1	Unfinished
RP-21H23	Sanger Institute	MMU 15	AL591864.6	22q13.1	Unfinished
RP-89G22	Sanger Institute	MMU 15	-	22q13.1	-
RP-422F22	Sanger Institute	MMU 15	AL591892.2	22q13.1	Unfinished
RP-412D17	Sanger Institute	MMU 15	AL603843.2	22q13.1	Unfinished
RP-175A3	Sanger Institute	MMU 15	-	22q13.1	-
RP-55O11	Sanger Institute	MMU 15	-	22q13.1	-
RP-359D20	Sanger Institute	MMU 15	-	22q13.1	-
RP-77D8	Sanger Institute	MMU 15	-	22q13.1	-
RP-267J18	UOKNOR	MMU 8	AC076974.23	22q13.1	Unfinished
RP-254F2	Sanger Institute	MMU8	AL603837.2	22q13.1	Unfinished
RP-290L7	AECOM	MMU8	AC084823.10	22q13.1	Finished
RP-478N15	Sanger Institute	MMU8	AL603864.3	22q13.1	Unfinished
RP-480M4	Sanger Institute	MMU8	AL603782.5	22q13.1	Finished
RP-477E1	Sanger Institute	MMU8	-	22q13.1	-

Appendix 6: DBA output (Jareborg *et al*., 1999) showing alignments of conserved mouse-human sequences within regions up to 3 kb upstream of four annotated gene transcription start sites.

The alignment is show below as a series of blocks. Each block is in one of four classes, A (65%), B (75%),C (85%),D(95%). Small gaps are permitted in the blocks.

```
Alignments upstream of Gene : TTLL1 (human) bM121M7.1 (mouse).


bM121M7      507     CAGCCCCGATCCTTCTCTTCGCTTTTTCCTTCTCTTCATTGCTTCTTT
                 B   CAGC  CGA CC T TCTTC CT   TC C TC CTTC TT CTTCTT
GoldenPath22262      CAGCAGCGAACCCTTTCTTCCCTCCCTCCCCTCCCTTCCTTCCTTCT-


bM121M7      555     TTCTTG-CTTCTTCT
                 B   TTCTT  CTTCTTCT
GoldenPath22310      TTCTTTCCTTCTTCT


bM121M7      1857    TCTTTTTTTTTTTTTT
                 D   TCTTTTTTTTTTTTTT
GoldenPath22327      TCTTTTTTTTTTTTTT


bM121M7      2286    TTTTTTTG-TT-GTTG-TTTTTTTT
                 D   TTTTTTTG TT GTTG TTTTTTTT
GoldenPath221510     TTTTTTTGTTTGGTTGTTTTTTTTT


bM121M7      2322    TTTTTTGTTTTTTTTTTCGAGACAGGGTTTCTCTGTGTAGCCCTGGCT
                 B   TTTTTT TTTTT TTTTT GA A  GGGT T  CT T T G CCGGCT
GoldenPath221705   TTTTTTCTTTTTTTTTTTAGAAATGGGGTCTTGCTATATTGGCCAGGCT


bM121M7      2371    GTCCCGGAACTC
                 B   G  C GGAACTC
GoldenPath221753     GGTCTGGAACTC


bM121M7      2418   CCGCCTGCTTCTGCCTCCCAAAGTGCTGGGATTAAAGGCATGCGCCACC
                 C   CC CC GC TCTGCCTCCC AAGTGCTGGGATTA AGGC TGGCCACC
GoldenPath222480     CCACCCGCCTCTGCCTCCCAAGTGCTGGGATTACAGGCGTGAGCCACC


bM121M7      2647   CTTTATCCGCTGGCCCTGGCCCTTACAACTCTCATTTCTGGCCCCTGAA
                 B   CTT ATCC  TG  CC G CCCTT   A TCTC TTTCTGG CCC GA
GoldenPath222768     CTTAATCCCGTGCTCCAGACCCTTCTCAATCTCCTTTCTGGCCCCGA-


bM121M7      2694    AGGGTGCGCGG
                 B   AGGGTG  CGG
GoldenPath222817     AGGGTGGACGG


bM121M7      2809   TCAGGAAGCAGTAGCGCCAGCGGTTTTCGCGTTCTCGGTTGCTAGGACA
                 B   TC GGAA  AGT GC  C  CGG   TCGCGT CT GGTTGC AGG C
GoldenPath222910   TCCGGAAATAGTCGCAGCGCCGGCGGTCGCGTCCTGGGTTGCCAGGGCG


bM121M7      2858    CCTCTCCGGAAGTGGAGTGAAGC
                 B   CC C C GGAAGT GAGTG AGC
GoldenPath222959     CCGCCCTGGAAGTAGAGTG-AGC
```

Alignments upstream of Gene: BIK (human) Biklk (mouse)

```
bK1191B2    981      GGGGTTTCTCCATGTTGGTCAGGCTGG-CTCAAACTC
                C    GGGGTTTC C ATGT    CAGGCTGG CT AAACTC
bM121M7     790      GGGGTTTCACTATGTACCCCAGGCTGGCCTGAAACTC

bK1191B2    1238   CCTTTTGATCAGCATATTGTCTTGGGGATTTTGCAAAGTAAATAAGTAC
                A    CC TTTGATCAGC T TTG CTT GGGAT TTGCA   TAAATA  TC
bM121M7     1819     CCATTTGATCAGCCTGTTGCCTTGGGATCTTGCAGGCTAAATATTTTC

bK1191B2    1286   TGTATTTGCACCCACTCTGCCCTTGAATCATCCAGTGTCCCCAAACGGT
                A     T      C  CC CT  GCCC T A TC  C    TG CC CAAAC G
bM121M7     1868   ACTCACCCCCTCCCCTGGGCCCCTAAGTCCCCATCTGACCTCAAACTGG

bK1191B2    1335     CCTTCTTTCTCCATCTTTTCTGCT
                A    CCT CTT CTCCA CTT TCT CT
bM121M7     1917     CCTCCTTCCTCCACCTTGTCTACT

bK1191B2    1437     TCAGGTCTTTCGGGACCCTGAGACCCCGTGTCATTGCTTTACCTCCCT
                C    TCAGGTCTTTC GGAC   TGAGACCCC TGTCA TGC T  C TCCT
bM121M7     1967     TCAGGTCTTTCGGACTGTTGAGACCCCTTGTCACTGCCTCTCTTCTCT

bK1191B2    1485     GAGTCTCAATTT
                C     AGTCTCAATTT
bM121M7     2015     CAGTCTCAATTT

bK1191B2    1498     TTCATCTGCAAAATGCATTCCCAGAG
                D    TTCATCTGCAAAATG ATT  CAGAG
bM121M7     2035     TTCATCTGCAAAATGTATT--CAGAG

bK1191B2    2233   TAAACAAGCTTTGCCGTGCCCAGGACAATTGTTACTTTGTTATTCCAGG
                A    TAAACAAG TTTG  GT      GA A T GTTACTTTGT ATTC
bM121M7     2419   TAAACAAGTTTTGTTGTTGTATAGATAGTGGTTACTTTGTAATTCGGCC

bK1191B2    2282     AGCGCTCTGCCTTCTCCCACC
                A    AG GCTC   CTT  C CACC
bM121M7     2468     AGTGCTCCATCTTTACTCACC
```

Alignments upstream of Gene BZRP (human) Bzrp (mouse)

```
dJ526I14    685      TTTTTTTTTTTAAA
                D    TTTTTTTTTTTAAA
bM121M7     21       TTTTTTTTTTTAAA

dJ526I14    978      CATGTGTGTCTT-TTTTTATTTATTTTATTTTTTGTTTTGTTTTTTTG
                B    CAT TGT T TT T TTT TT T TTT  TT TTTGTTT  TTTTTTG
bM121M7     1031   CATCTGTTTGTTTTGTTTGTTTTGTTTGTTTGTTTGTTTGTTTTTG

dJ526I14    1027     AGAAAGAGTCTCACTGTGTCACCCAAGCTG
                B    AGA AG GT TC CTGTGT  CCC  GCTG
bM121M7     1078     AGACAGGGTTTCTCTGTGTAGCCCTGGCTG
```

```
dJ526I14     1246    ATCCACCGGCCTCGGCCTCACAAGTGCTGGGATTACAGGTGTGAGCCA
                 C   AT CACC  CCTC GCCTC C AAGTGCTGGGATTA AGGTGTGGCCA
bM121M7      1152   ATTCACCTACCTCTGCCTCCCAAAGTGCTGGGATTAAAGGTGTGCGCCA

dJ526I14     1295    CCACGCCCGGCT
                 C   CCACGCCC GCT
bM121M7      1200    CCACGCCCAGCT

dJ526I14     1623    TTAAAAAAGAAAAAAAAAAAAAAAAACAAA
                 B   TTAAAA A A AAA AAA AAA AAACAAA
bM121M7      2664    TTAAAACAAACAAACAAATAAACAAACAAA
```

Alignments upstream of Gene C22orf1 (human) bM150J22.1 (mouse)

```
GoldenPath221006    TCT-CTGCCTCACCCTCTC-AGT-GCTGGGATTACAGGTG
                 C  TCT CT CCTC CCCTCTC AGT GCT GG TTACAGGTG
bM150J22     1378   TCTCCTCCCTCTCCCTCTCAAGTGGCTTGGCTTACAGGTG

GoldenPath221618   AAGGAGGATATTGCTAATTTATTTCACCTTCTAGGGAGATGATCAAGAT
                 B  AA GAG  TA TGCT  TTTATTT ACCTT TA GG GATGAT  AGA
bM150J22     1600   AAAGAGATTACTGCTTTTTATTTTACCTTGTATGGGGATGATTGAGAA

GoldenPath221666    TT
                 B  TT
bM150J22     1649   TT

GoldenPath221670    AAAAATTAATAACCCATTTCTCCTTGCACATAATTAAAATGTTCTCCA
                 C  AA AATTAATAACCC TTTCTCCTTGCACATAATTAAAATGTTCTCCA
bM150J22     1659   AAGAATTAATAACCCGTTTCTCCTTGGCACATAATTAAAATGTTCTCCA

GoldenPath221719   GTCTCTAATTTTTGTCTTTTTCCTAATCTAATTTGTTTTCTGACTGTGT
                 C  GTCTCTAATTTTTGTC TTTT  TAATCTAATTTGTTTT  GATGTGT
bM150J22     1707   GTCTCTAATTTTTGTCATTTTTATAATCTAATTTGTTTTTGGAGTGTGT

GoldenPath221768   CGATTCTTCTTCCAAGCGCAAAGCAAAGGGGATTTTTCTTCATTTAATG
                 C  CGATTC   TTCC AGC  A A CAA GGG ATTTTTCTTC TTTA T
bM150J22     1756   CGATTCCCTTTCCGAGCATAGAACAAGGGGAATTTTTCTTCCTTTAGTA

GoldenPath221817   TGATTGCGATATGAGTGTCCAGGAATAGTTTAAATGATGTTATTTTCTC
                 C  TG TTG GAT   AGTGTCCAGGAA   TTTAAA A TATTTTCTC
bM150J22     1805   TGTTTGGGATCCCAGTGTCCAGGAACCATTTAAACTACATGATTTTCTC

GoldenPath221866    CTTGGTTAAATACAGCGCAAAAGGAATCGTTGGAGGGTTCTTAA
                 C  CTTGGTTAAATACAG GCAAA GG   C TTG AGGG TCTTAA
bM150J22     1854   CTTGGTTAAATACAGTGCAAATGGGGCCATTGAAGGG-TCTTAA

GoldenPath221939   CTTCAAGGCATTTCCCATTTACACAGTTTAAAAAAATAATTATGAAAAG
                 B  CTTCAAG CA T CC   T ACACAGTT A AAAAAT A TATGAAAG
bM150J22     1989   CTTCAAGTCACTGCCTGCTCACACAGTTAAAAAAATTACTATGAACAG

GoldenPath221987    G
                 B  G
bM150J22     2038   G
```

```
GoldenPath222376    CTGGGAAAATCCCTTCCTCAGGGCACCACTAAAAGATATCTTTAGATGA
                 B  CT GGAAA  CCCTT CTC GGG  C  CTAAAAGAT TCTT A   A
bM150J22    2410    CTTGGAAATCCCCTTCTCTGGGTGCATCTAAAAGATGTCTTGAAGCTA

GoldenPath222424    AATCGATGTCGAGGGAGGAATTTTCGCCCGGCTGTCTCCACCTGCTCA
                 B  AAT  ATGT GA GGAGGAATTTTC  CC  GC GTC CCACCTGC
bM150J22    2459   AATTTATGTTGAAGGAGGAATTTTCATCCCGGCAGTCCCCACCTGCATG

GoldenPath222473    GGCTTGCAGGGGTGTGGGCGTGGGGCCATGTGGGTGTGT
                 B  GGCTTGC GGGGTG GG  GTG GGC  TGTGG T TGT
bM150J22    2507    GGCTTGCTGGGGTGCGGCTGTGTGGCTGTGTGGCTATGT

GoldenPath222541    TGTAGGGAATCCTTCCTGTTCCCCAATTCTGAAAAAGGTAAAACTTAA
                 C  TG AGG GAATCCTTCCTGTTCCCCA  TCTG AAAA   AAAATTAA
bM150J22    2568    TGGAGGGGAATCCTTCCTGTTCCCCATATCTG-AAAAAAAAAATTTAA

GoldenPath222588   CTCACCTGTTTACAAAATACCAGCCATTGTCTTTACCCCAGCTCACCTG
                 C  CT  CCTGTTTACAAAATACCAGCCATTGTCT   CCC AGCTCACTG
bM150J22    2616    CTTCCCTGTTTACAAAATACCAGCCATTGTCT-GTCCCTAGCTCACTG

GoldenPath222635    ---CCCTGTAGGCTC-GGAGGATTTTGTTGAAGGAAAAAAAAATGTCT
                 ??    C CTGTAG CTC  GAG AT TTGTTG AGG  AAAAAAATGTCT
bM150J22    2665   CCCCTCTGTAGACTCCCGAGAATCTTGTTGGAGGGGGAAAAAAATGTCT

GoldenPath222684    TAAGTATTTAAACACGTTGAGCCATGCATGCATCCGTCCA
                 C   AAGTATTTAAACA GTTG  CCATGCA GCATCCGTCCA
bM150J22    2709    GAAGTATTTAAACATGTTGGACCATGCACGCATCCGTCCA

GoldenPath222761   GTCCCCTCCCTCCCTTCCCTTTTCTTTTTTACCAAAGTATATTCATCAA
                 D  GTCCCCTCCCTCC TTC CT TTC TTTTTACAAAGTATATTCATCAA
bM150J22    2777    GTCCCCTCCCTCCGTTCTCTCTTCTTTTTACTAAAGTATATTCATCAA

GoldenPath222809   ACTGCTGAGTTGGAAAGATTTGTAATGAGTTTTTGAGCTTTGTACGACT
                 D  ACTGCTGAG TGGAAAGATTTGTAATGAGTTTTTGAGC  TGTCGACT
bM150J22    2826    ACTGCTGAGATGGAAAGATTTGTAATGAGTTTTTGAGC-TGTGCGACT

GoldenPath222856    GTGTTT
                 D  GTGTTT
bM150J22    2875    GTGTTT

GoldenPath222873   CCCCCCCGCCCTCTCCCTCTTTCTAAATCTTCATCTGACATTAAATAAA
                 D  CCC CCCCCCTCTCCCTCTTTCTAAATCTTCATCTGACATTAAATAAA
bM150J22    2902    CCCCCCTCCCTCTCCCTCTTTCTAAATCTTCATCTGACATTAAATAAA

GoldenPath222921   GCAAATCCCAAACAGATTAACTGTCGCACGGTTCTGCTCCGTCTCCTCA
                 D   CAAATCCCAAACAGATTAACTGTCGCA GGTTCTGCTCCGTCTCTCA
bM150J22    2951   ACAAATCCCAAACAGATTAACTGTCGCATGGTTCTGCTCCGTCTCTTCA
```