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.

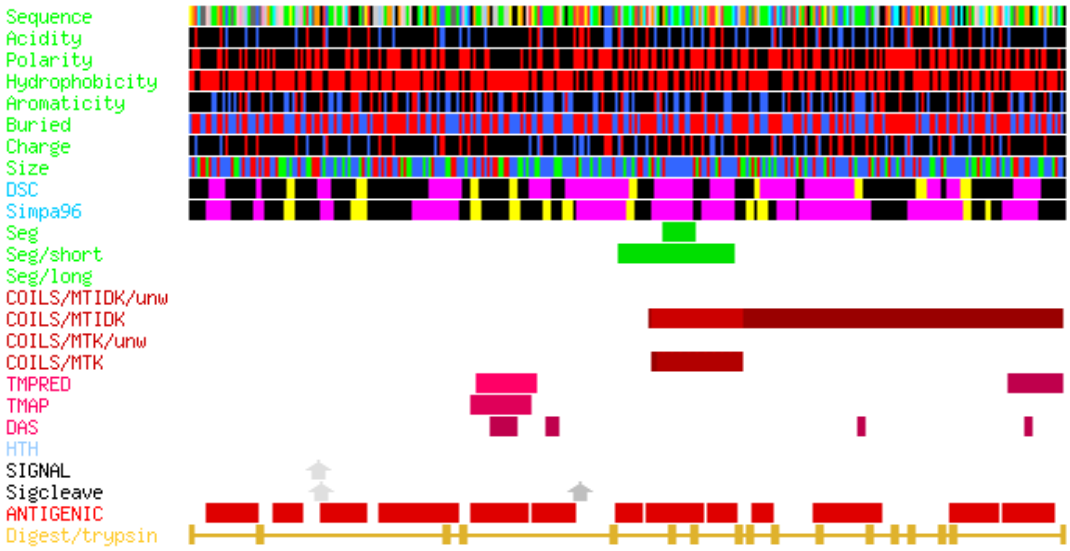


Figure 5.1: PIX display out put showing analysis of the translated coding sequence of dJ222E13.C22.1 (isoform a).
The sequence is displayed in several colour schemes in order to highlight various aspects of the sequence. The key is shown below: letters refer to amino acid symbols.
Sequence: Normal (rasmol) 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

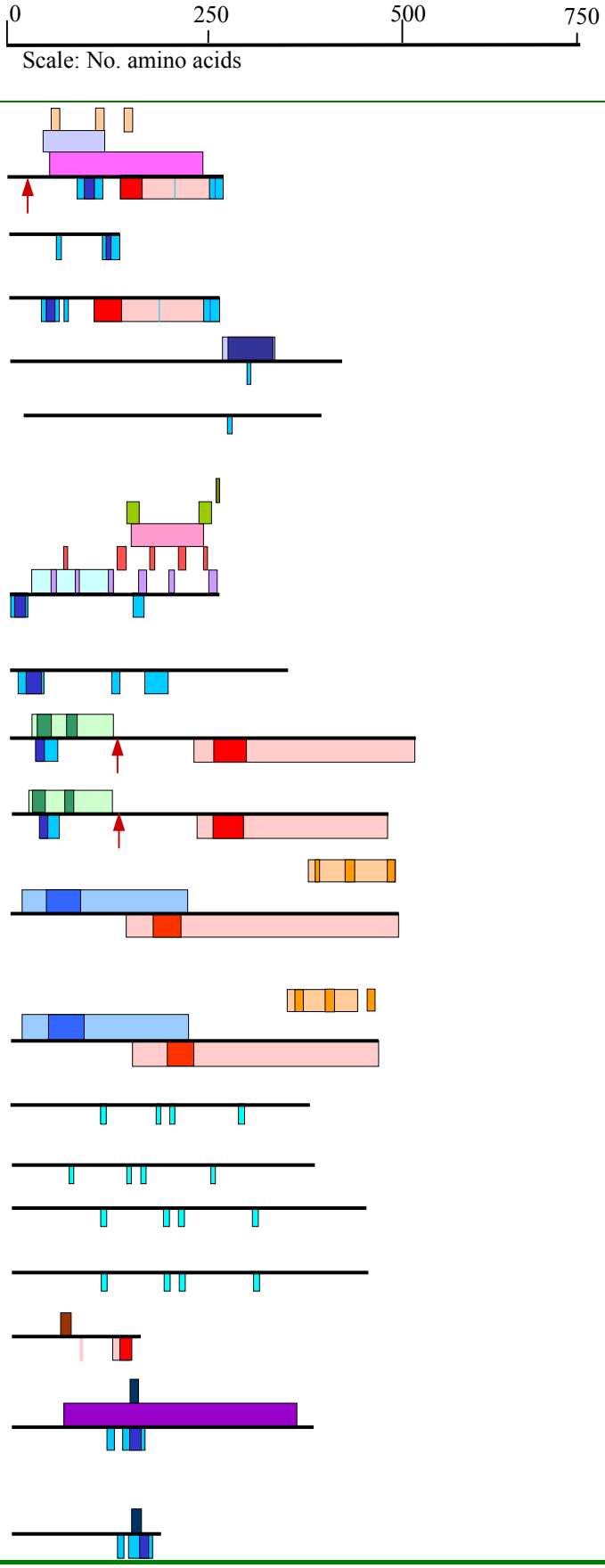
22q13.31. Where possible, InterPro attaches potential functions to the domains. The coding sequences of known alternatively spliced gene structures were included in the search to identify whether the alternative splice altered the domain content of the protein sequence.

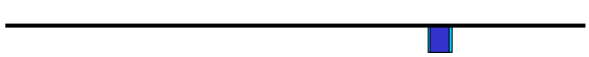
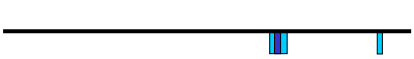
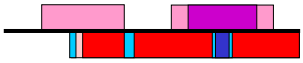
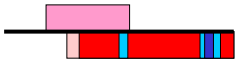
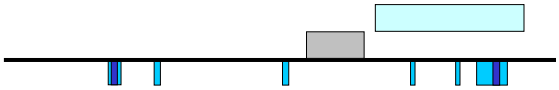
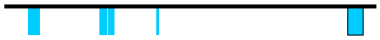
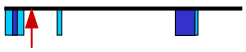
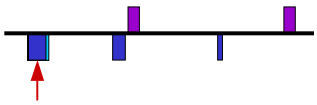
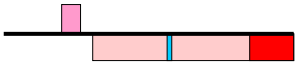
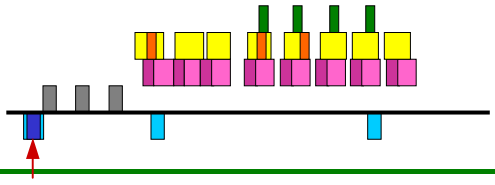
A diagram of each peptide showing InterPro features (depicted above the line in each diagram) and transmembrane, coiled coil regions and potential N-terminal signal peptides (depicted below the line in each diagram), is shown below in figure 5.2. Minimum (dark shades) and maximum (light shades) lengths 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.

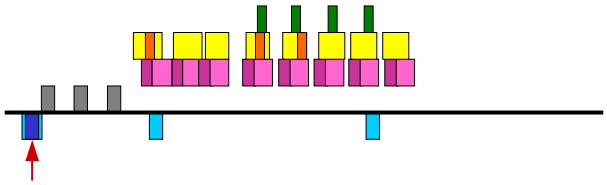
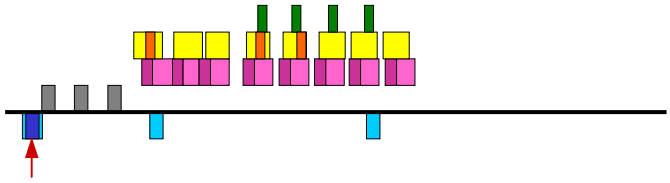
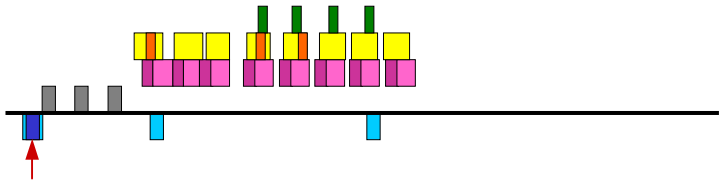
Table 5.2 shows that overall 16 of the 27 protein coding (ignoring alternative splice forms) contained a domain or other InterPro feature. Six of these were identified as multidomain genes. Interestingly, the alternative splice forms of dJ222E13.C22.1, dJ222E13.C22.3 and bK1191B2.C22.3 contained different 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.

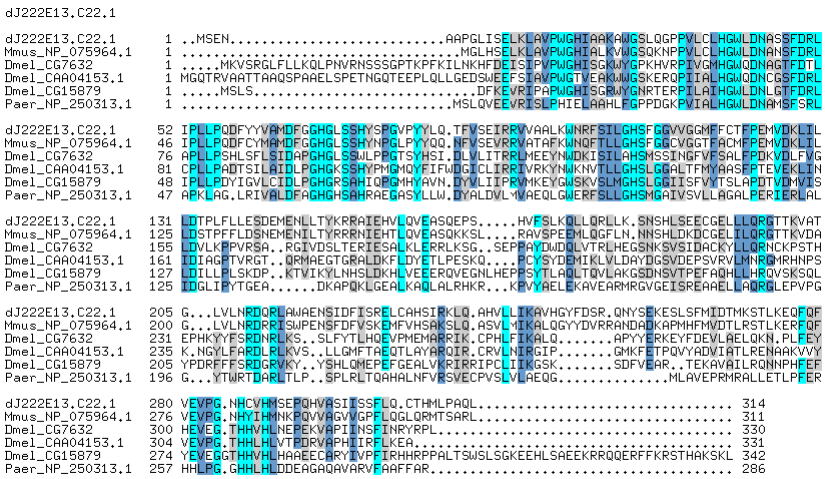


Figure 5.3: Clustalw alignment of the amino acid sequence of dJ222E13.C22.1 against five homologous protein sequences identified from a BLASTP search of the NCBI nonredundant protein sequence database.

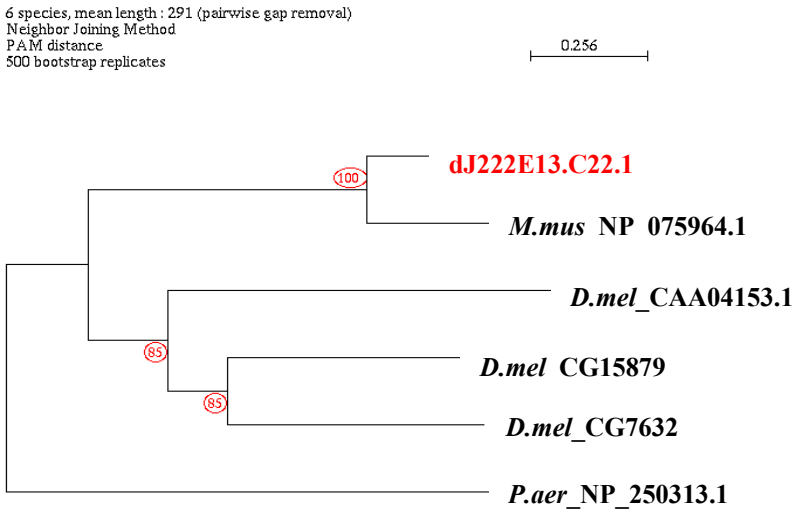


Figure 5.4: Phylogenetic tree derived from the above alignment using the Phylowin package (Galtier *et al.*, 1996). The human protein dJ222E13.C22.1 from chromosome 22q13.31 is highlighted in red. The distance-based tree making method used was the Neighbour-joining method (Saitou & Nei, 1987). The numbers 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.

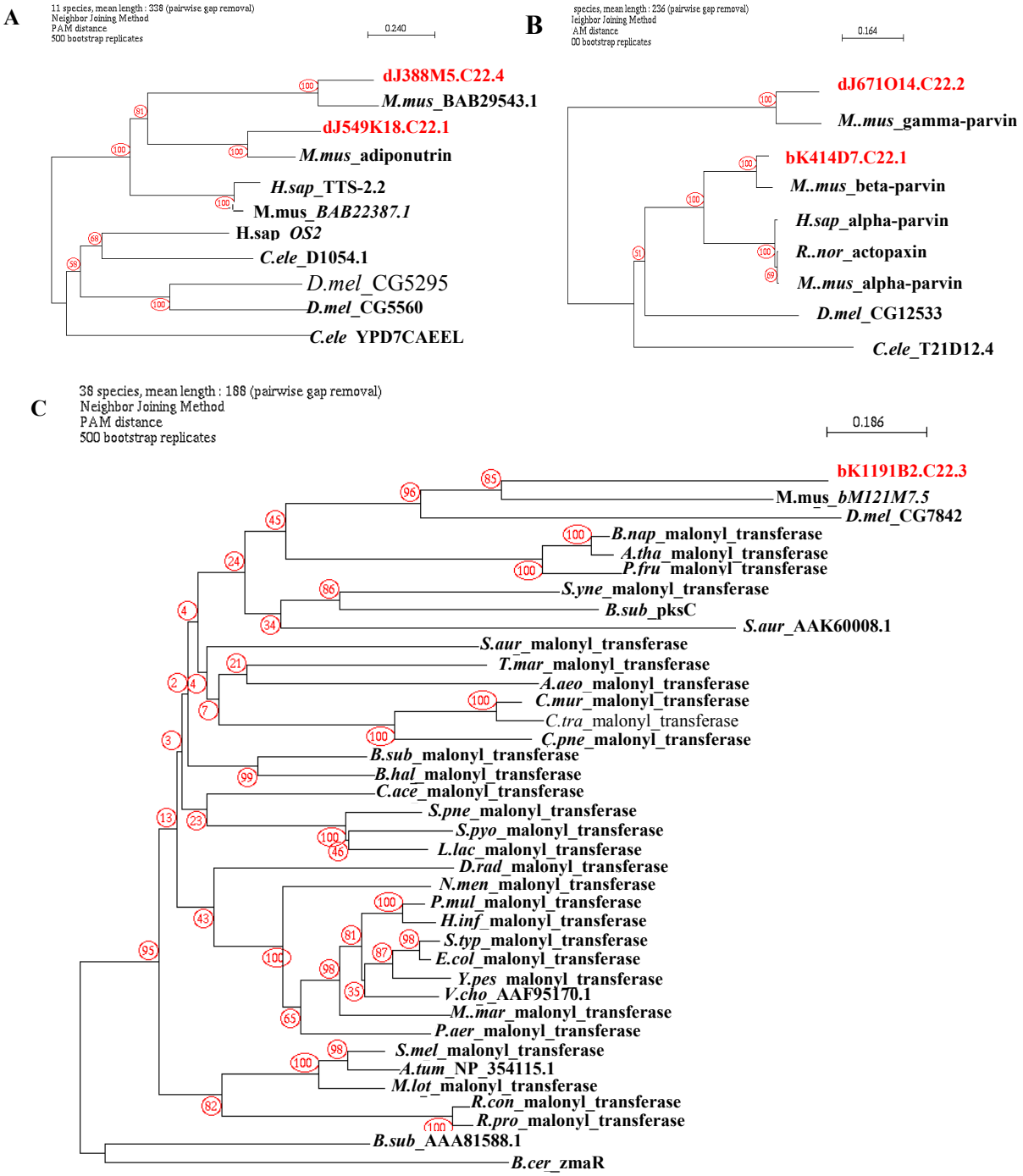


Figure 5.5: Phylogenetic trees derived using NJ methodology from clustalw protein alignments (Phylowin, Galtier *et al.* 1996). Proteins from 22q13.31 are highlighted in red. Other protein identifiers are listed in appendix 8.

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

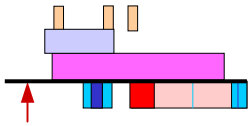
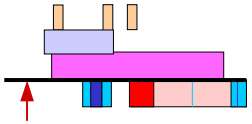
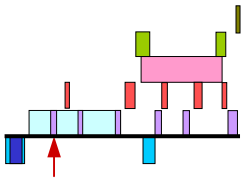
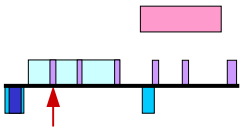
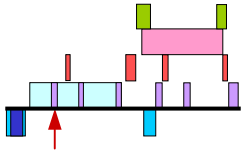
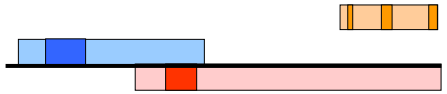
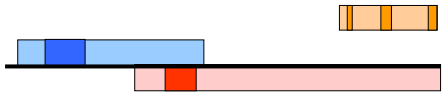
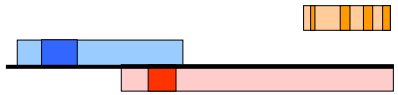
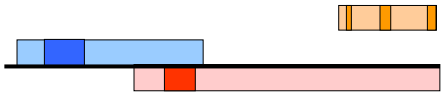
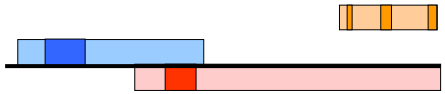
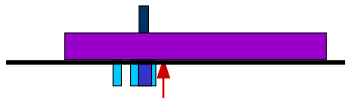
Table 5.4: Potential orthologues of proteins from 22q13.31 identified by phylogenetic analysis. Sequence identifiers are provided in appendix 8.

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’.

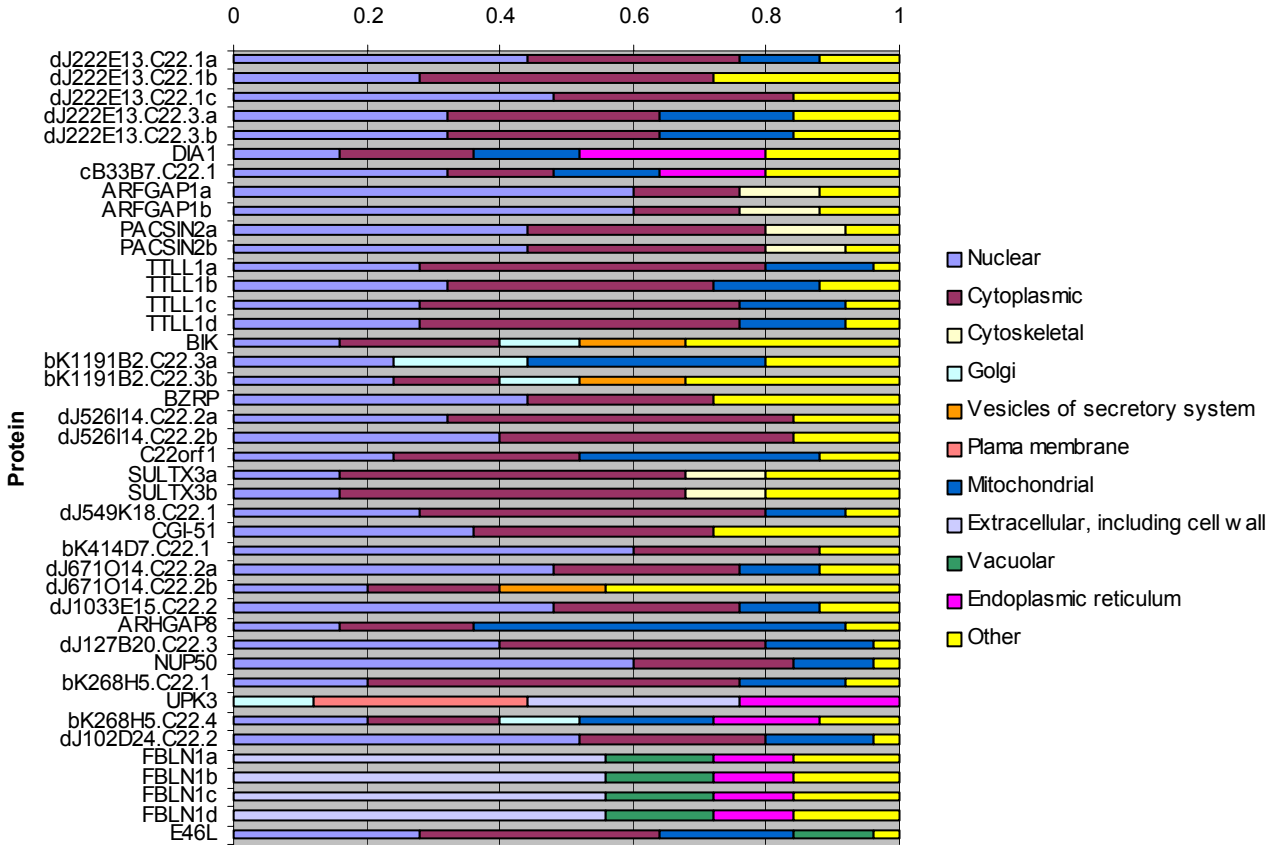


Figure 5.7: Predicted subcellular localisation (PSORT, Nakai & Horton, 1999). The length of each coloured horizontal bar depicts the probability of localisation at a particular cellular location.

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.

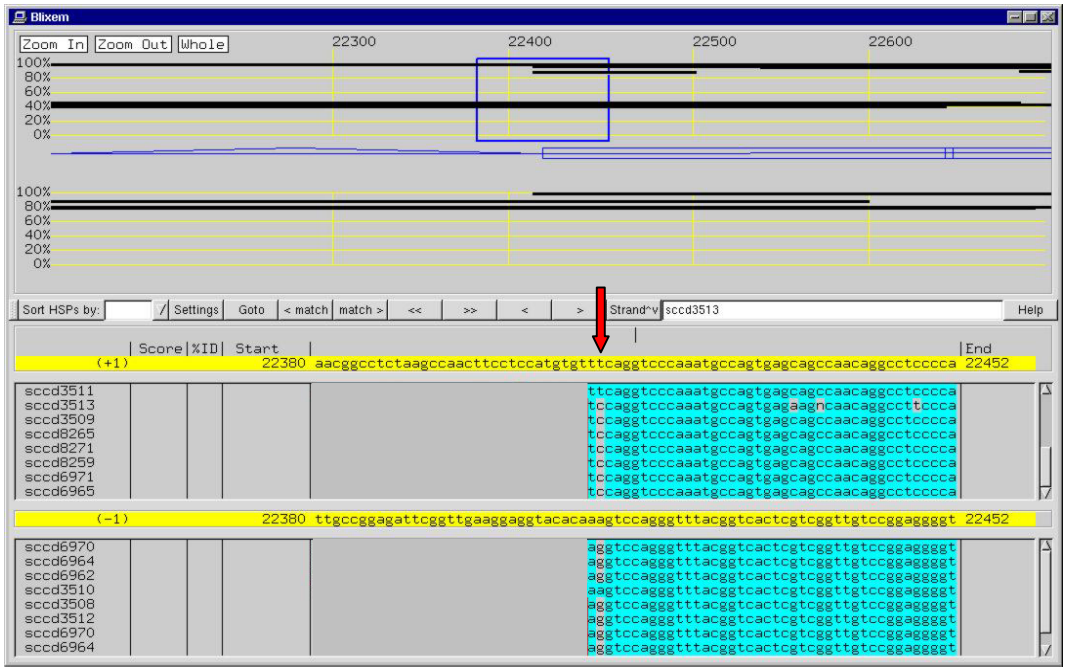


Figure 5.8: Blixem alignment of dJ549K18.C22.1 cDNA clone sequencing reads. The arrow indicates a discrepancy between the cDNA and genomic sequence.

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).

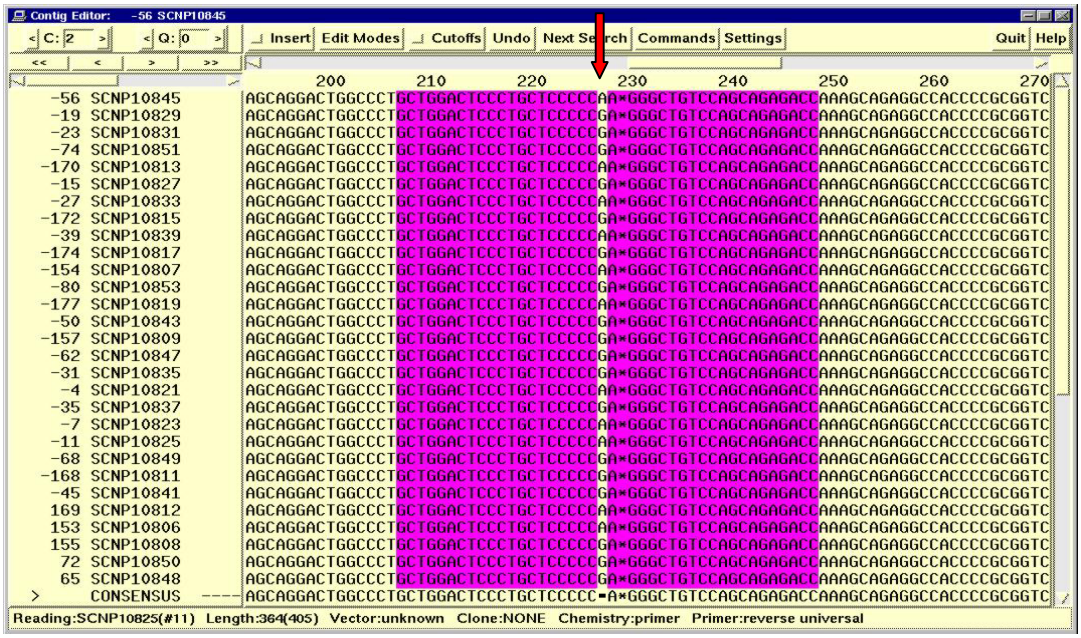


Figure 5.9: Visual display from Gap4 database. The red arrow indicates a potential SNP within the cDNA sequence of dJ549K18.C22.1.

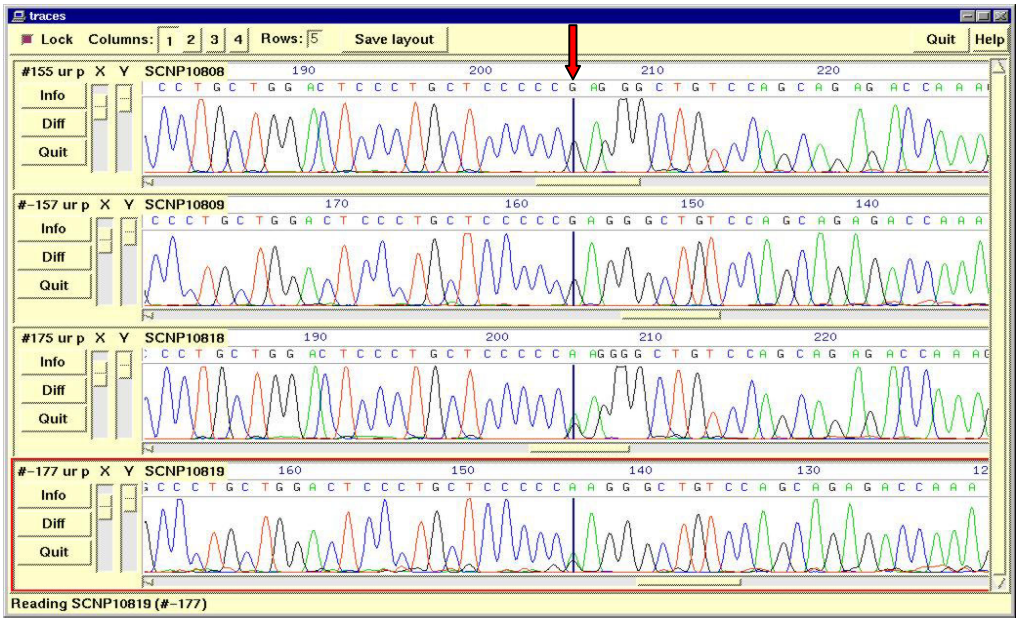


Figure 5.10: Inspection of the forward and reverse traces from two (of 24) individuals. The red arrow indicates a candidate variation in the cDNA sequence from dJ549K18.C22.1.

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.

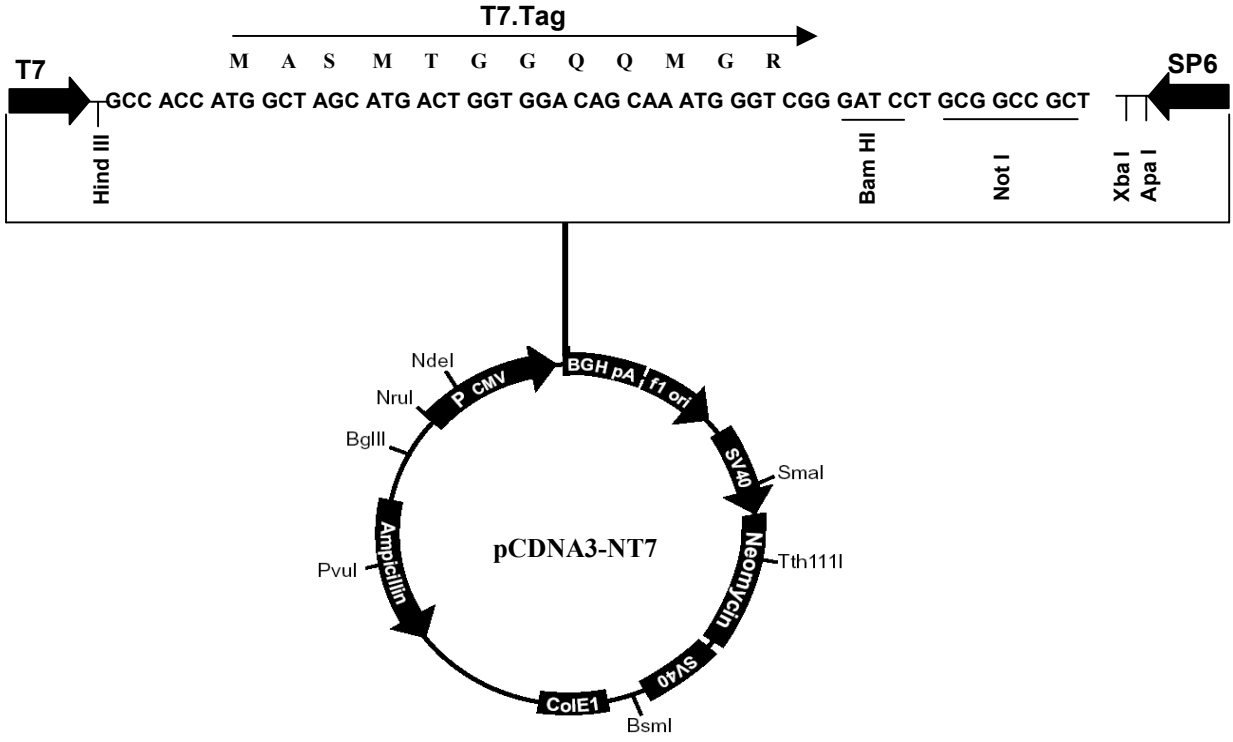


Figure 5.11: Schematic of the mammalian cell expression vector pCDNA3-T7-C. The PCDNA3 vector (Invitrogen) polylinker site was modified as described in the text.

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.

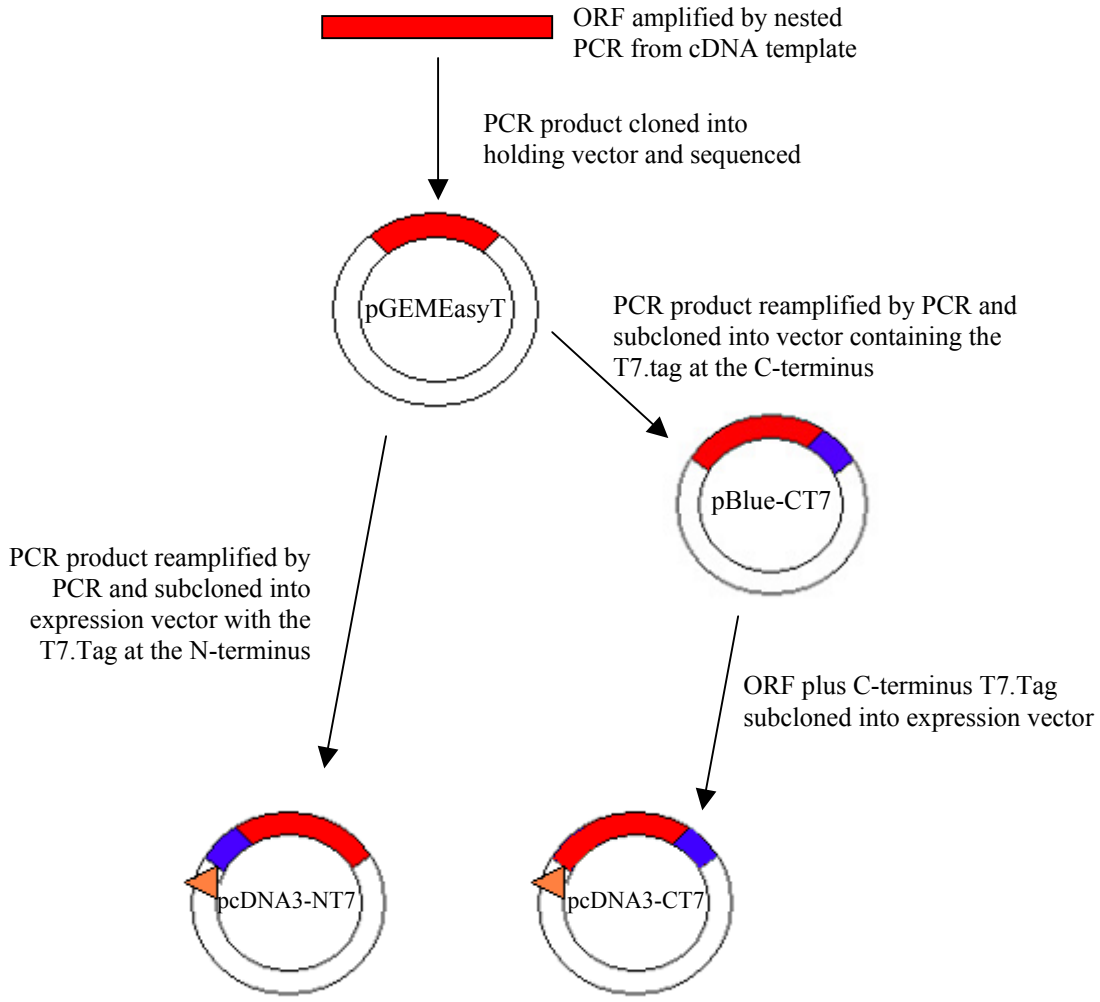


Figure 5.12: Schematic showing strategy used to generate N- and C- terminally T7-tagged clones. The ORF of the gene under investigation (shown in red) was amplified from a cDNA template by nested PCR. The PCR product was then cloned into a holding vector, pGEMEasyT (Promega) and the insert sequenced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

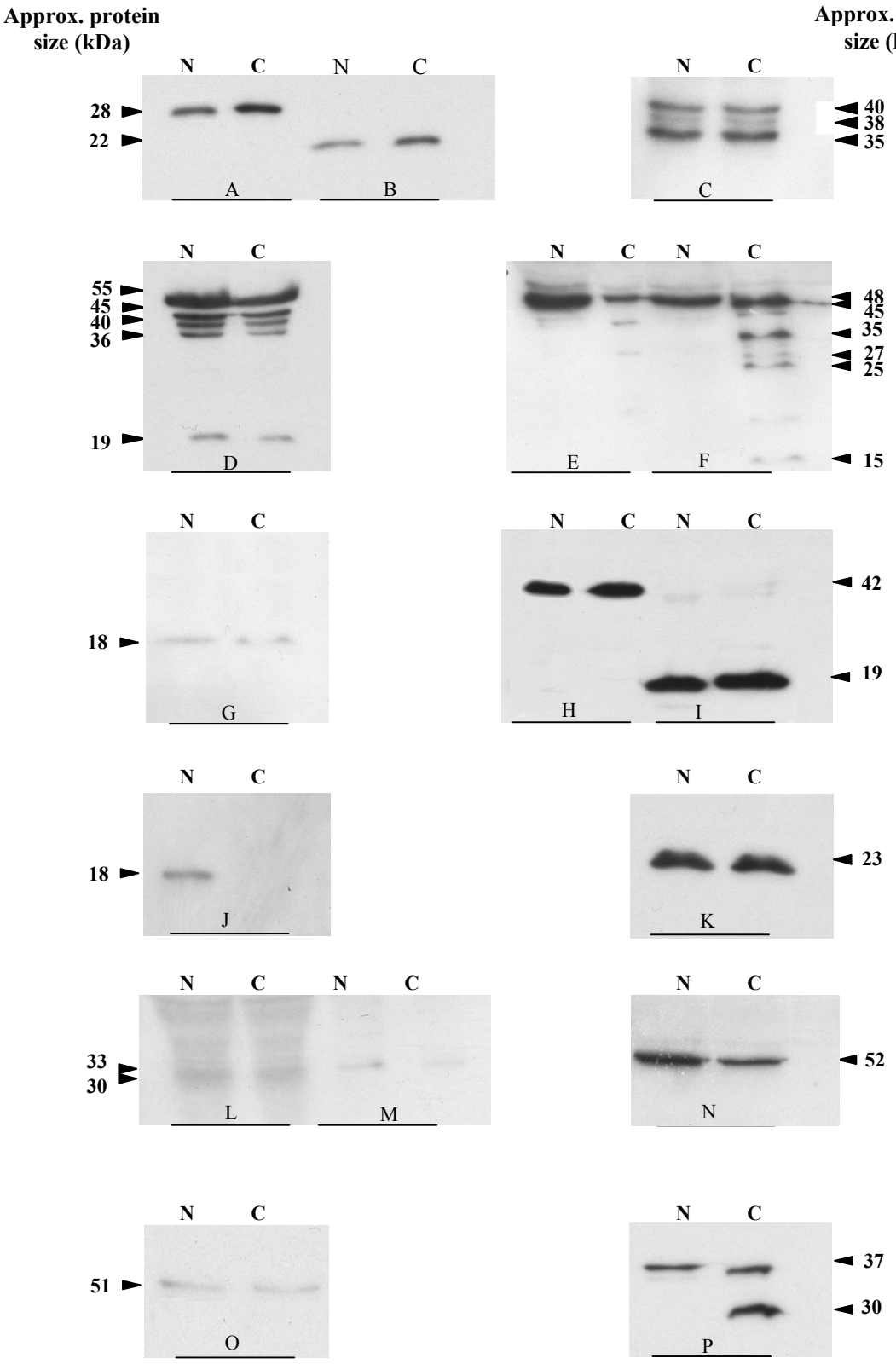


Figure 5.13: Western blot analysis of transiently transfected COS-7 cells. N- and C- terminally tagged constructs are shown. A) dJ222E13.C22.1a; B) dJ222E13.C22.1b; C) cB33B7.C22.1; D) PACSIN2a; E) TTLL1a; F)TTLL1c; G) BIK; H) bK1191B2.C22.3a; I) bK1191B2.C22.3b; J) BZRP; K dJ526I14.C22.2b; L) SULTX3a;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

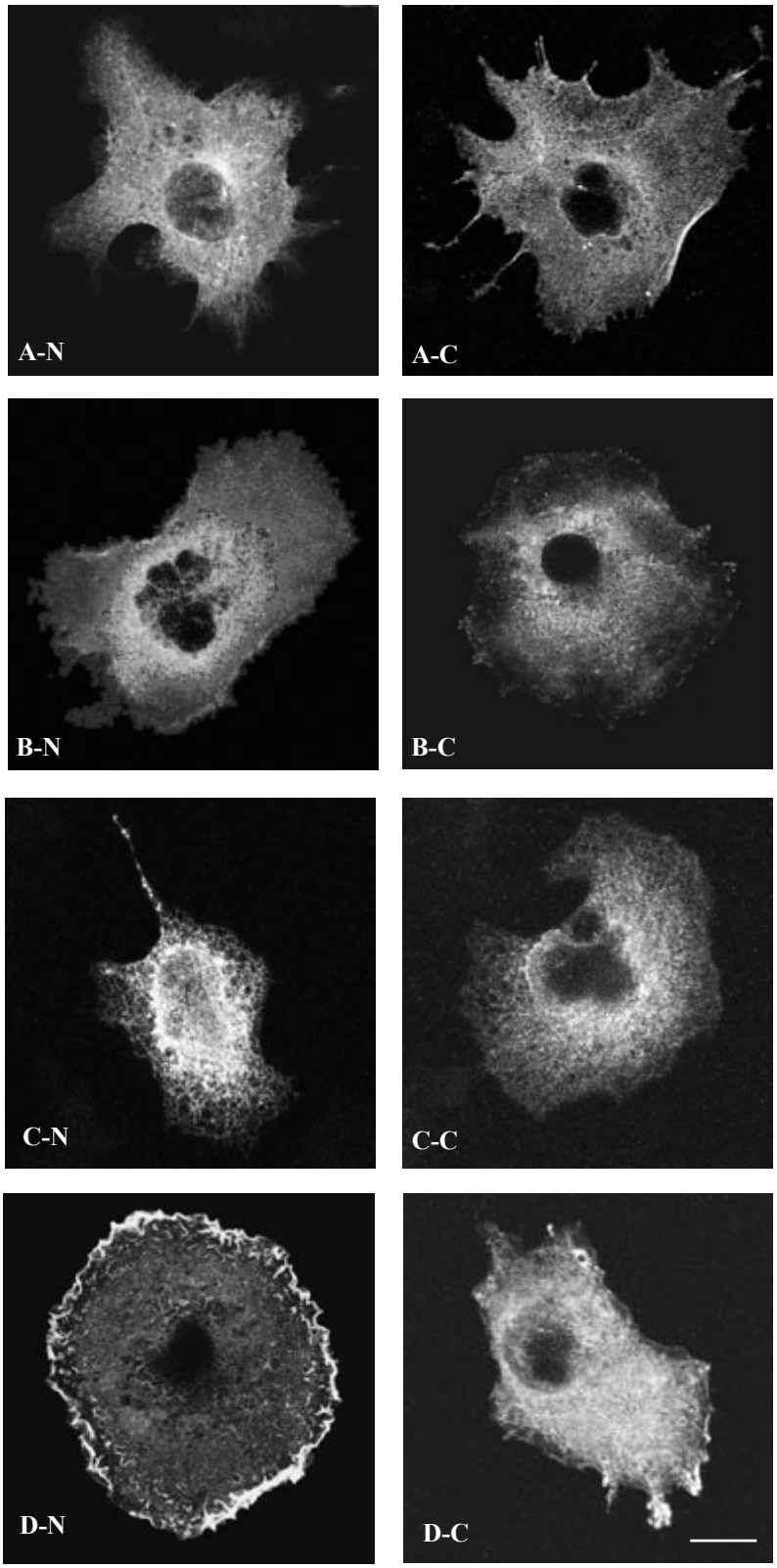
The presence of extra bands, smaller than the expected size of the protein construct, was noted in several cases (figure 5.13.C, D, E, F and P). These bands may be caused by partially degraded copies of the protein construct, or could be the result of post-translational modifications. Interestingly in 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,

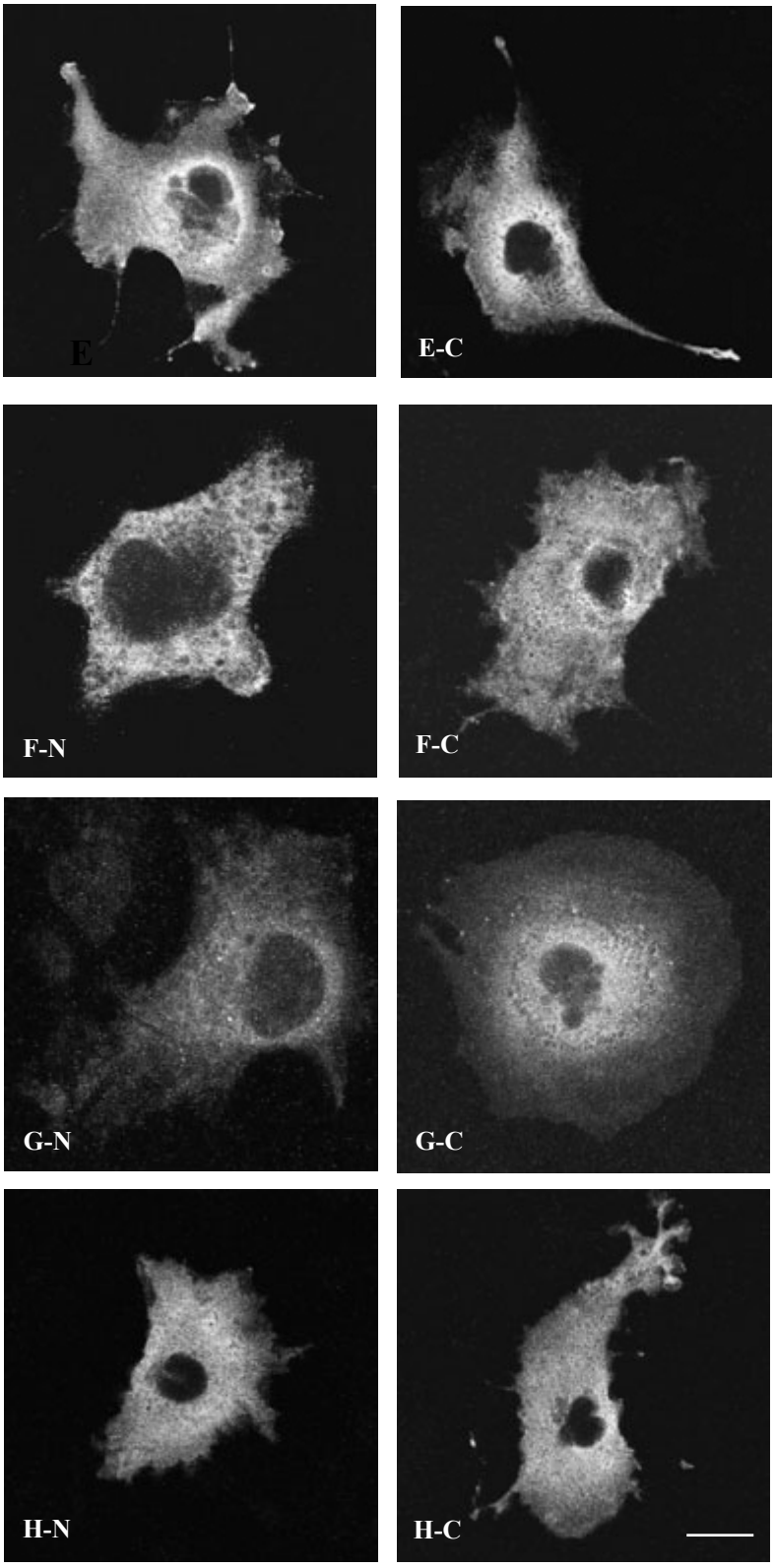
although the use of β-mercaptoethanol in the sample preparation should preclude this. Alternatively, these larger bands could result from glycosylation, or similar post-translational modification, of the expressed protein construct.

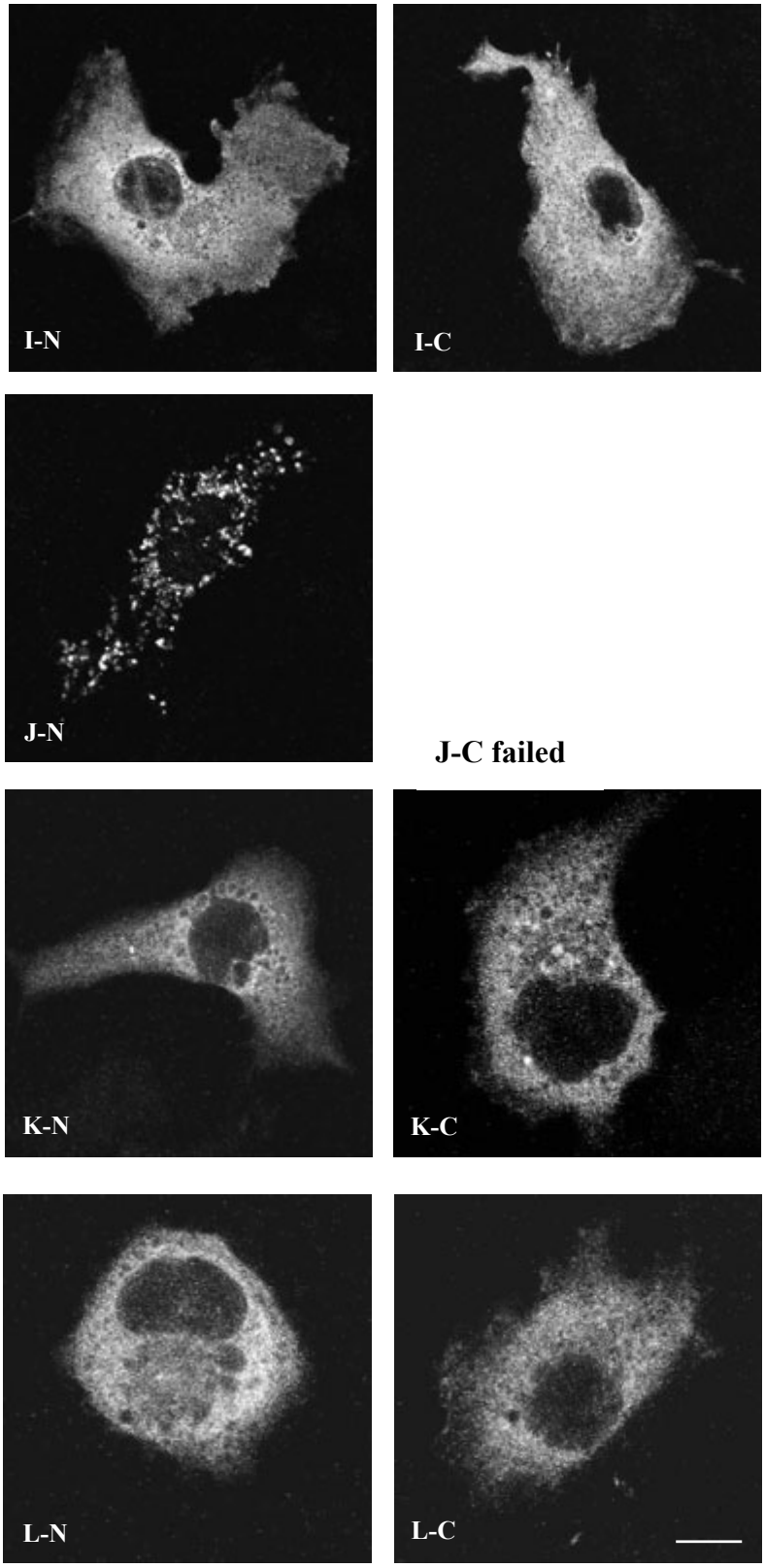
5.4.5 Analysis of T7.Tag protein subcellular location

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

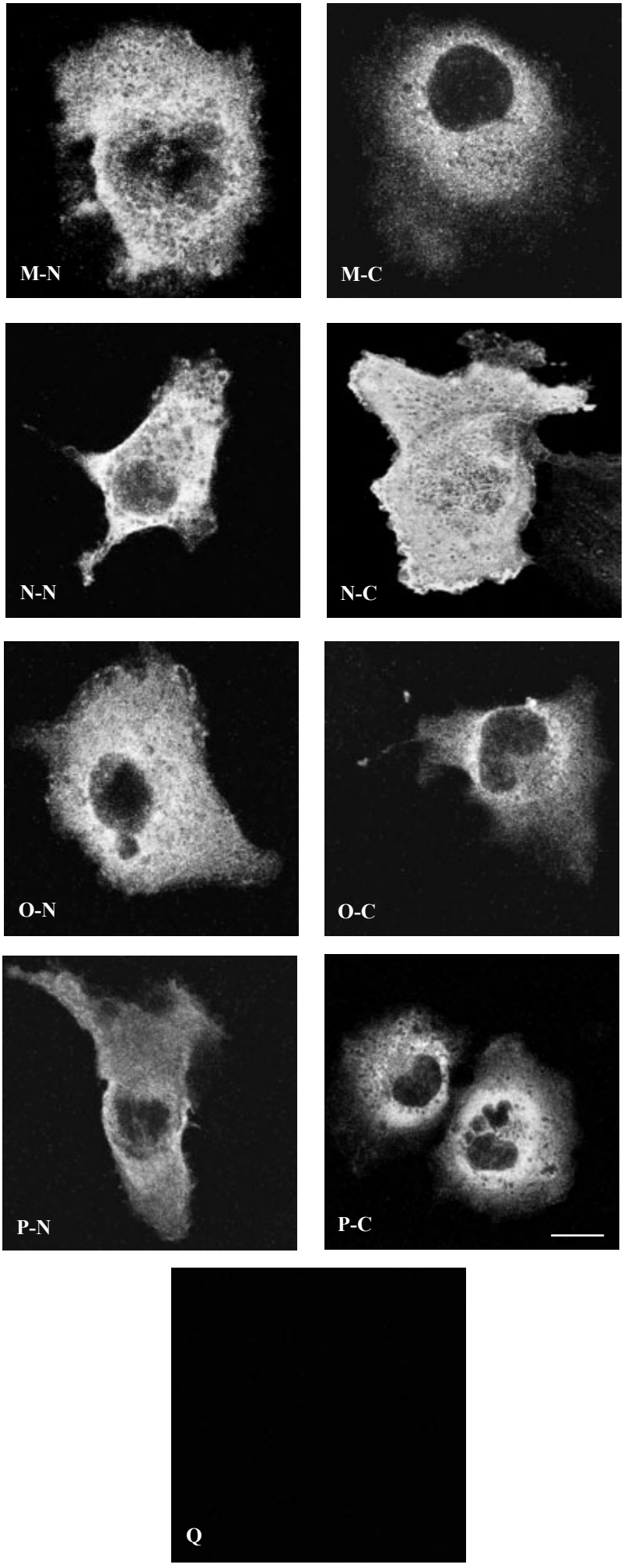

M-N
M-C
N-N
N-C
O-N
O-C
P-N
P-C
Q

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).

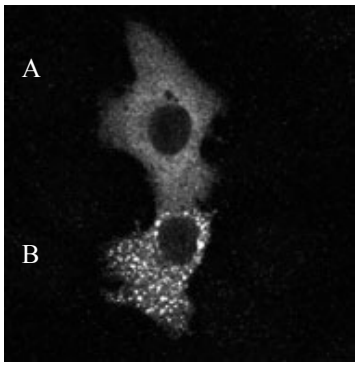


Figure 5.15: An example of possible aggresome formation. COS-7 cells were transfected with the N-terminal T7 tag construct of dJ222E13.C22.1a. The construct displays a cytosolic localisation pattern in cell A, which was also observed in the majority of other transfectants. Possible over-expression of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

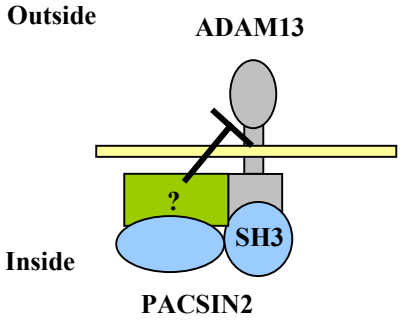


Figure 5.16: Adapted from Cousin *et al.* (2000). Schematic representation of the regulation of ADAM13 by X-PACSIN2. ADAM13 is in grey, the plasma membrane in yellow and X-PACSIN2 in blue. In this model, X-PACSIN2 binds to the ADAM13 cytoplasmic domain through its SH3 domain and to a putative repressor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