Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Hamid Reza Khorram Khorshid MD, MPH
(Tehran University of Medical Sciences)

Department of Genetics

University of Leicester

November 2003

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Acknowledgements

There is no doubt that throughout my study in Leicester, numerous of people have been involved in the progression of my work and without the help and support of them this project would never have been completed. To list everybody would be lengthy but I would like to thank few people for the major role they played in my study.

First of all I would like to express my sincere thanks and gratitude to my supervisor, Dr. Raymond Dalgleish, for devoting his valuable time and constant guidance throughout my PhD project. I do appreciate his excellent, valuable and patient supervision and extended help at all stages of my work especially during my writing up and its editing. His words and advice not only inspired me to work harder, but also made me realise my potential to work under pressure.

I am grateful to Clare Bennett, Jackie Swallow and other students of our old and lovely Lab G24 for their helpful advice and good friendship during the first year of my study. Thank you to all students and staff of Lab 145 especially Tim, Emma, Lorna, Jo and Julie for their hospitality and friendship during the last couple of years.

I would like to do a general and very special thank you to the whole department for making me feel so welcome and their valuable and constant help over the last four years. Particular thanks go to Prof. Richard Trembath and Dr. Mark Jobling for their continuous assessment of my work and their helpful point of views about my project. Particular thanks to Dr. Howard Pringle and Dr. Wilhelm Schwaeble and the staff and students of their labs for their help and support on the *in situ* hybridisation and isolation of mononuclear cells from blood. Thank you also to Prof. Richard Trembath, Dr. Mark Plumb and Dr. Esther Signer for DNA samples that they donated.

I would like to thank my parents for the upbringing they gave me that enabled me to achieve so much in my life and career.

Finally I would like to thank my government, the Islamic Republic of Iran and the Ministry of Health and Medical Education for funding and supporting my study in the field of genetics in the University of Leicester.

Abstract

Secreted phosphoprotein 24 (spp24) is a member of the cystatin superfamily which was first isolated from acid demineralised bovine cortical bone. Subsequently, human cDNA and genomic DNA clones were isolated and the structure and transcription of the human gene (*SPP2*) were characterised.

This study identifies a rare single-amino acid polymorphism (p.S38F) of human spp24 and its importance has been assessed by comparing the sequence of human spp24 with that of eight other species.

The gene encoding spp24 in mouse (*Spp2*), like its human ortholog, comprises eight exons with an apparently TATA-less promoter. The exon/intron structure is identical in mouse and human and the size and location of intron 1 is conserved between many species. Using several strategies, the gene encoding spp24 in mouse has been mapped to 88832387–88853226 bp of the mouse chromosome 1. To study the function of the spp24 protein, four quantitative trait loci (QTLs) were identified in the vicinity of the mouse *Spp2* gene. Association between each of these QTLs and the mouse *Spp2* gene was investigated, but no association could be demonstrated.

An extensive expression study was carried out on the mouse and human genes encoding spp24. These studies indicated that the gene has an expression pattern of a tissue-specific and cell-specific nature, being expressed predominantly in liver and its expression is down-regulated by lipopolysaccharide (LPS) and tumour necrosis factor α (TNFα).

In an attempt to elucidate the function of the spp24 protein in mouse, a pooled-tissues cDNA library was constructed in a yeast two-hybrid vector. Two different constructs comprising the entire spp24 protein and the C-terminal non-cystatin like domain of the protein were used individually as baits in the yeast two-hybrid system to screen the constructed library. Seven potential interacting proteins were identified including granulin precursor also known as acrogranulin/epithelin (*Grn*), tissue specific transplantation antigen P35B (*Tsta3*), keratin complex 1, acidic, gene 18 (*Krt1-18*), keratin complex 1, acidic, gene 13 (*Krt1-13*), vimentin (*Vim*), similar to protein phosphatase 1, regulatory (inhibitor) subunit 12C (no gene symbol yet assigned) or myosin binding subunit 85 and alpha-actinin-4 (*Actn4*).

Table of Contents

APPENDICES

Chapter 1
Introduction

1.1 General Introduction

Secreted phosphoprotein 24 (spp24) is a 24-kDa novel protein that is encoded by the human *SPP2* gene and which was first co-extracted with MGP (matrix Gla-protein) from an acid demineralised extract of bovine cortical bone (Hu *et al.*, 1995). Because the present study is concerned with this protein and its function, the essential aspects about it will be discussed in more detail, based on the above study in the following sections.
Comparison with existing protein sequences indicated that the N-terminal 107 residues of the bovine protein are related in sequence to the cystatin superfamily of thiol protease inhibitors, but the C-terminal domain from 108-182 did not show any homology with any known protein. Therefore it was suggested that spp24 might inhibit thiol protease activity during bone turnover (Hu *et al.*, 1995). Co-isolation of matrix Gla-protein (MGP, γ-carboxy glutamic acid-containing protein, bone Gla-protein) and spp24 immediately suggested similar properties for spp24 and MGP and a potential role for spp24 in regulatory processes in bone. It was also suggested that any such function might contribute to the phenotype of bone mineral density (BMD) and also to diseases characterised by alterations in BMD such as osteoporosis.

As well as the demonstrated function of cystatin superfamily members in the inhibition of the action of members of the papain superfamily cysteine proteinases, several other roles have been suggested for cystatin superfamily members in health and disease conditions, including:

- Defending the body against infection
- Prevention of formation and metastasis of some tumours
- Protection from autoimmune diseases and inflammatory reactions
- Protection from destruction of cartilage and collagen
- Prevention of some neurological disorders through decreasing the degradation of myelin

Because these roles will be used as a guide for further study about the function of spp24, they will be discussed in more detail in Section 1.6.4

None of the members of the cystatin superfamily that are the most homologous to the N-terminal part of spp24 exhibit typical cystatin anti-protease activity. These include fetuins, kininogens, bradykinin and neutophil antibiotic peptides (Hu *et al.*, 1995) for which many other possible roles have been speculated (Takagaki *et al.*, 1985; Brown *et al.*, 1992). It is therefore probable that spp24 has a totally novel function compared with that of other cystatin superfamily members. Also, it is quite possible that it has multiple roles in different biological processes.

Consequently the main aim of this project was to begin to elucidate some possible functions for the spp24 protein, using different strategies. The following section describes each part in more detail, the reasons for using the selected strategy in each experiment, the results achieved and their analysis in order to develop the best plan for further experiment design and ongoing research.

1.2 Secreted phosphoprotein 24 (spp24)

Secreted phosphoprotein 24 (spp24) is a 24-kDa novel non-collagenous protein and was first isolated from an acid demineralised extract of bovine cortical bone (Hu *et al.*, 1995). Because the present study is concerned with this protein, I will try to explain the essential facts about it in summary, based on the above study.

1.2.1 Purification of spp24 and determination of its amino acid and nucleotide sequence of complete cDNA

Spp24 was initially discovered in the course of developing improved methods for the isolation of MGP (matrix Gla-protein, γ-carboxy glutamic acid-containing protein, bone Gla- protein) from ground bovine bone. Co-isolation of MGP and spp24 during the first step of isolation is an indication of the similar properties of the two proteins. Both proteins were released from ground bovine bone by demineralisation in 10% formic acid. Both proteins were found in the neutral pH-insoluble extract and subsequently purified by application to a Sephacryl S-100 HR column followed by further purification using reverse phase HPLC. The purified spp24 protein was shown to be homogenous by electrophoresis on an SDS polyacrylamide gel (showing a protein of 24 kDa molecular weight) and by N-terminal sequencing of the purified protein and internal peptides released by cleavage (Hu *et al.*, 1995).

The cDNA sequence of bovine spp24 was determined using a combination of RT-PCR (reverse transcription polymerase chain reaction), 3'RACE (rapid amplification of cDNA end) and screening of a λgt11 cDNA library. Based on the N-terminal amino acid sequence of the isolated spp24 protein, degenerate primers were designed and RT-PCR was carried out on bovine liver and bone periosteum RNA preparations. A 380-bp PCR-generated fragment was cloned and sequenced and shown to be identical in both liver and bone. The plasmid cDNA inserts were digested by *EcoR*I and because of an internal *EcoR*I site in the insert, this digestion produced two fragments of 312 and 68 bp. The 312 bp fragment was used as a probe to screen a bovine liver λgt11 cDNA library, which generated a clone covering the 5'-end of the cDNA. In order to determine the 3'-end terminal sequence of the cDNA, 3'RACE was used on both liver and bone RNA preparations. The result indicated the identical 3'-end sequence in both tissues.

Figure 1.1 presents the complete nucleotide sequence of the bovine spp24 cDNA and the deduced amino acid sequence of the protein, some of which was confirmed by N-terminal sequencing.

The bovine spp24 cDNA is 815 bp in length. The coding region of the protein is terminated by a single TGA stop codon at nucleotides 691-693 and an in-frame TAA triplet follows at nucleotides 700-702. The 3'-untranslated region consists of 126 nucleotides with a polyadenylation signal (AATAAA) at nucleotides 790-795. The ATG at nucleotides 91-93 were considered to be the initiation codon based on the rules for translation initiation described by Kozak, 1989. There is a 20-amino acid signal peptide with a potential cleavage site following amino acid residue 20. The N-terminal amino acid sequence of the mature peptide, determined by N-terminal sequencing, is located 20 residues downstream from the presumed initiation methionine. The deduced spp24 mature peptide contains a total of 180 amino acid residues and has a calculated molecular weight of 20,458 in rough agreement with the 24 kDa size of spp24 determined by SDS-polyacrylamide gel electrophoresis. The apparent discrepancy is accounted for by the fact that spp24 is a phosphorylated protein (Section 1.2.3).

1.2.2 Tissue distribution and expression of spp24 mRNA

Hu *et al* (1995) determined the expression of spp24 mRNA in various bovine tissues (bone, liver, heart, lung, kidney and spleen) by northern blot analysis of total RNA using the 312-bp cDNA probe. A single transcript band at 1000-1100 nucleotides was detected which agrees with the length of the spp24 cDNA. Spp24 mRNA was detected in bone and liver tissues but not in bovine heart, kidney, spleen or lung. Therefore Hu *et al.* (1995) suggested tissue-specific expression of spp24 and speculated that the presence of spp24 in bone indicates a possible role in bone turnover.

1.2.3 The identification of phosphoserine and the extent of phosphorylation in the serine rich sequence of spp24

As can be seen in Figure 1.1, there is a serine-rich region between residues 128-136 of bovine spp24 which contains several potential residues of serine phosphorylation. To determine the possible phosphorylation of these serine residues, the spp24 protein was first cleaved at tryptophan 127 using BNSP-Skatole. The peptide corresponding to residues 128-190 was then purified by gel filtration over Sephacryl S-100 HR and transferred to a poly(vinylidene

```
  1  ACAGTCTGAT CTGCCAAGTG CATTATACCA ATATCTCATT AATTCTCCCC

 51  AAACCTCTGA ACGGAAATTG TTCTTCCCAT AATGGAGAAG ATGGCGATGA
                                                M  A  M
101  AGATGTTGGT GATATTTGTC CTTGGAATGA ACCACTGGAC TTGTACAGGT
     K  M  L  V   I  F  V   L  G  M   N  H  W  T   C  T  G
151  TTCCCGGTGT ATGACTATGA CCCGGCTTCC CTGAAGGAGG CTCTCAGCGC
      F  P  V   Y  D  Y  D   P  A  S   L  K  E   A  L  S  A
201  CTCTGTGGCA AAAGTGAATT CCCAGTCACT GAGCCCCTAT CTGTTTCGGG
       S  V  A   K  V  N   S  Q  S  L   S  P  Y   L  F  R
251  CGTTTAGAAG CTCAGTTAAA AGAGTCAACG CCCTGGACGA GGACAGCTTG
     A  F  R  S   S  V  K   R  V  N   A  L  D  E   D  S  L
301  ACCATGGACT TAGAGTTCAG GATTCAAGAG ACGACGTGCA GGAGGGAATC
      T  M  D   L  E  F  R   I  Q  E   T  T  C   R  R  E  S
351  TGAGGCAGAC CCCGCCACCT GTGACTTCCA GAGGGGCTAC CACGTGCCCG
       E  A  D   P  A  T   C  D  F  Q   R  G  Y   H  V  P
401  TGGCCGTTTG CAGAAGCACC GTGCGGATGT CTGCTGAACA GGTGCAGAAC
     V  A  V  C   R  S  T   V  R  M   S  A  E  Q   V  Q  N
451  GTGTGGGTTC GCTGCCACTG GTCCTCCAGC TCTGGGTCCA GCAGCAGTGA
      V  W  V   R  C  H  W   S  S  S   S  G  S   S  S  S  E
501  AGAGATGTTT TTTGGGGATA TCTTGGGATC CTCTACATCA AGAAACAGTT
       E  M  F   F  G  D   I  L  G  S   S  T  S   R  N  S
551  ACCTGCTTGG CCTCACTCCT GACAGATCCA GAGGTGAACC ACTTTATGAA
     Y  L  L  G   L  T  P   D  R  S   R  G  E  P   L  Y  E
601  CCATCACGTG AGATGAGAAG AAACTTTCCT CTTGGAAATA GAAGGTACTC
      P  S  R   E  M  R  R   N  F  P   L  G  N  R   R  Y  S
651  GAACCCGTGG CCCAGAGCAA GAGTAAACCC TGGCTTTGAG TGACAGCCTT
       N  P  W   P  R  A   R  V  N  P   G  F  E
701  AAGCAAAATG CACTGGAAGG AATAGAAGTT CCAATGAAGA AAGATACCTT

751  ATGAATTGTG TAATTTTCTT TTGATCAATT GCAGTCCCTA ATAAATGGCT

801  TACTTTTCCT CTTTCAAAAA AAAAAAAAA
```

Figure 1.1. The bovine cDNA and amino acid sequences (Hu *et al.*, 1995)

The bovine cDNA sequence for spp24 was deduced as explained in section 1.2.1 by Hu *et al.* (1995). The cDNA sequence is shown here with the amino acid sequence underneath. The signal peptide residues are shown in blue and the characteristic cysteine residues, initiation and stop codons are shown in red. The accession numbers of the cDNA and protein are U03872 and Q27967 respectively.

fluoride) membrane and the phosphoserine residues converted to *S*-ethylcysteine (by reaction with ethanethiol) and subjected to N-terminal sequencing. The percentage phosphorylation of each serine residue was determined by the relative amounts serine and *S*-ethylcysteine at the sequenced position (Table 1.1). The presence of phosphoserine in spp24 was confirmed by acid hydrolysis and amino acid analysis.

1.2.4 The structure of bovine spp24 protein and relationship with other known proteins

To evaluate the possible relationship between spp24 and other known proteins, Hu *et al* (1995) compared the complete 200-residue spp24 with known proteins in the non-redundant protein database of NLM using the BLAST search program. This search revealed similarity between the N-terminal region of the bovine spp24 protein and cystatin domains 1 and 3 of human kininogen and the precursor to the bovine neutrophil antibiotic peptide bactenecin. The bovine spp24 was aligned with bovine bactenecin precursor, cystatin domains 1 and 3 of kininogen and two other closely related sequences; porcine cathelin and chicken egg white cystatin (Figure 1.2). As can be seen, the bactenecin precursor and cathelin are more closely related to spp24 than to cystatin domains 1 and 3 of kininogen or to chicken cystatin. Additionally, chicken egg white cystatin and domains 1 and 3 of kininogen are more closely related to spp24 than to cathelin or bactenecin precursor. Based on these results, Hu *et al* (1995) suggested that spp24 was an evolutionary intermediate between kininogen and cystatin and cathelin and bactenecin precursor.

The precursor to the bovine neutrophil peptide bactenecin falls into a group of proteins known as cathelins or cathelicidins. Cathelins are a family of antimicrobial peptide precursors that have a highly conserved N-terminal prepro sequence, followed by a highly variable C-terminal sequence that is the antibacterial peptide. In the mature protein, the precursor is cleaved to release the C-terminal antimicrobial region (Zanetti *et al.*, 1995). The propeptides of cathelins loosely resemble a cystatin domain in that they contain four cysteine residues which are thought to form two disulphide bonds. Although, these propeptides are highly conserved between each other, they show little sequence homology to other cystatin superfamily members, and thus are not included in the cystatin superfamily (Hu *et al.*, 1995). The structure of the proteolytic cystatin region has been determined using crystallographic studies of the 108-amino acid chicken cystatin and is a structure with a 5-stranded β-sheet wrapped around a 5-turn α-helix (Bode *et al.*, 1998). The sequence identities that have been

Amino acid and position in bovine spp24	Degree of phosphorylation (%)
Serine 128	5
Serine 129	63
Serine 130	70
Serine 131	81
Glycine 132	-
Serine 133	82
Serine 134	83
Serine 135	81
Serine 136	78

Table 1.1. Level of phosphorylation in the serine-rich sequence of purified bovine spp24

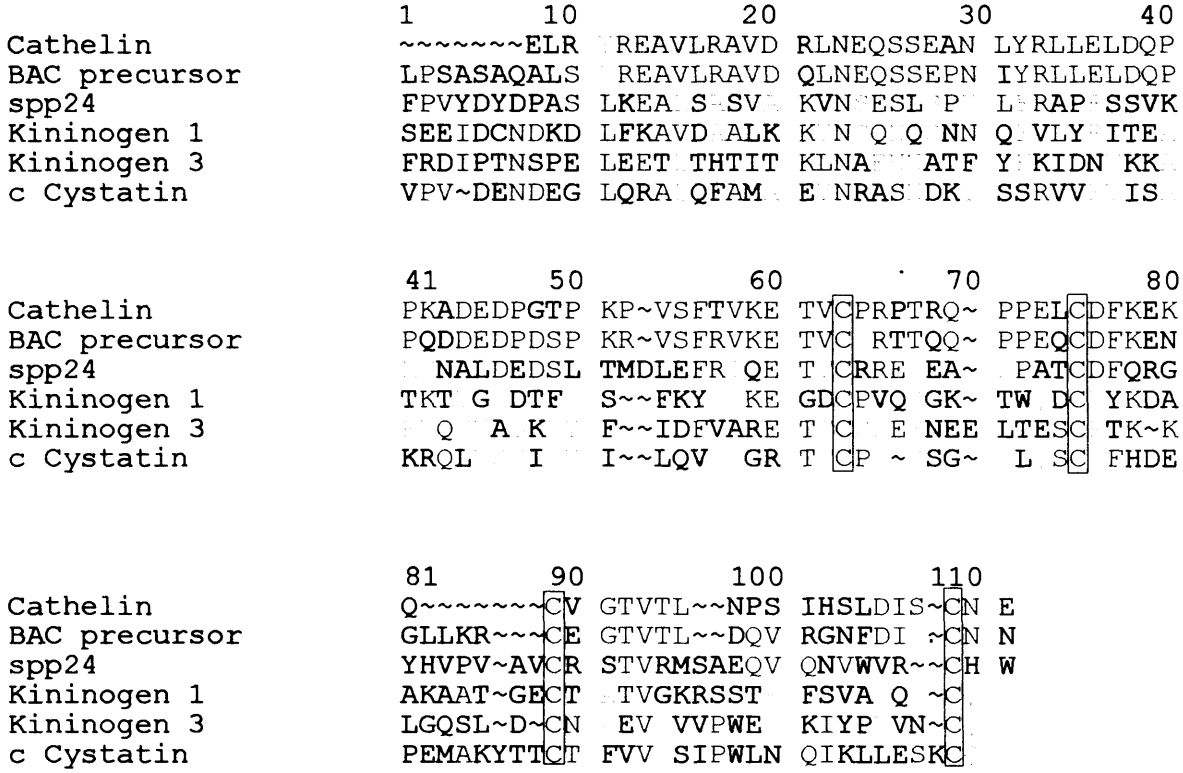


Figure 1.2. Amino acid sequence homologies between spp24 and cathelin, bovine bactenecin precursor, cystatin domains 1 and 3 of human kininogen and chicken cystatin

Residue numbers refer to the sequence position in mature bovine spp24. The other related sequences are residues 1-92 of cathelin, 24-126 of bactenecin precursor, 23-127 of kininogen domain 1, 268-371 of kinininogen domain 3 and 12-116 of chicken cystatin. Identical amino acid sequences are in the same colour in each column and highly conserved cysteine residues are boxed (adapted from Hu *et al.*, 1995).

observed between the 107 residues at the N-terminus of the mature spp24 and the 108-residue chicken cystatin indicate that the entire 107-residue domain between the N-terminus of the mature spp24 and the 11-residue phosphoserine-rich sequence is folded into a cystatin-like tertiary structure. Since the four cysteine residues in the cystatin region of spp24 lie at the sequence positions known to be involved in disulphide bonds in other members of cystatin superfamily, it is probable that the four cysteine residues in the spp24 protein are similarly involved in disulphide bond formation. These disulphide bonds join cys-63 with cys-74 and cys-87 with cys-105 in the mature spp24 protein. Residues 108-180 at the C-terminal end of the mature spp24 protein showed no homology to any known protein. Figure 1.3 illustrates a schematic representation of the bovine spp24 protein.

1.2.5 Speculated functions of spp24

Among the proteins most closely related to spp24, chicken cystatin, cathelin and domain 3 of kininogen have been shown to inhibit protease activity (Salvesen *et al.*, 1986). If spp24 is in fact a cysteine protease inhibitor, the presence of this protein in bone suggests that the target thiol protease could also be found in bone. There are several thiol proteases known to be expressed by osteoclasts to digest collagen and various non-collagen bone proteins under the acidic conditions of osteoclast-mediated bone resorption (Delaissé *et al.*, 1980). According to these results Hu *et al.* (1995) suggested that spp24 might inhibit thiol proteases, as is a feature of most proteins with a cystatin domain.

A second possible role for spp24 was suggested based on the observation that both the cystatin domain 3 of kininogen and bovine neutrophil antibiotic precursor (the two proteins most closely related to spp24) have a cystatin domain that lie to the immediate N-terminus of a peptide which, when released by protease action, has potent biological activity. The common location of these peptides to the immediate C-terminus of the cystatin region suggests that the proteolytic activities which release the active and functional peptides may involve a common mechanism of substrate recognition that is partly based on the presence of the adjacent cystatin region. However, it should be considered that these released peptides show no sequence homology and so must exert their biological activity on different target binding sites and through different pathways.

The third possible function of spp24 was suggested based on the similarities between fetuin and the spp24 protein. Both proteins are synthesised by bone and liver and accumulate in the

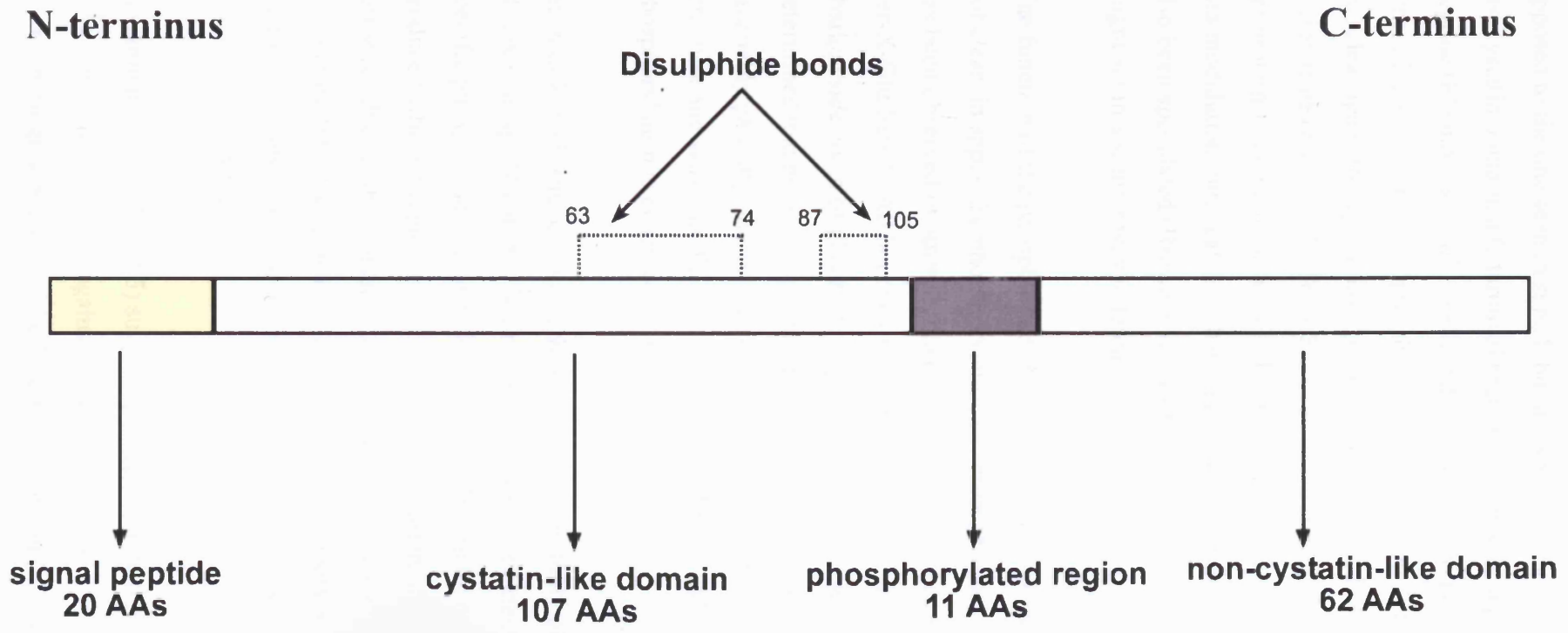


Figure 1.3. A schematic representation of bovine spp24

extracellular matrix of bone (Ohnishi *et al.*, 1993). Fetuin has two cystatin domains as opposed to the one seen in spp24, but it also has an extended C-terminal region following the last cystatin domain, a C-terminal region that could be a precursor to a biologically active peptide (Elzanowski *et al.*, 1988) and finally both proteins contain phosphoserine. The human form of fetuin, α_2HS-glycoprotein, circulates in blood as a cleaved two-chain molecule and this cleavage is thought to take place in the C-terminal sequence following the second cystatin region (Lee *et al.*, 1987). There has been much speculation about the function of fetuins. They are proteins thought to be involved in the acute phase response, have a role in bone formation and modulation, bind calcium, have a role in immunosuppression and many other ideas have also been speculated (Brown *et al.*, 1992). Accordingly, Hu *et al.* (1995) suggested that spp24 might act in a similar way to fetuin.

The function of the phosphorylated serine domain after the cystatin-like domain of spp24 is not clear. In spp24 the phosphorylation of serine residues follows the recognition motif that has been observed in other secreted phosphoproteins (such as osteopontin and MGP), that is Ser-X-Glu/Ser(P). All but one of the phosphorylated serine residues have the negatively charged side chain of glutamate or phosphoserine in the n+2 position. It has also previously determined that phosphoproteins secreted into the extracellular space of cells tend to be invariably partially phosphorylated at serine residues while those phosphoproteins secreted into milk and saliva are fully phosphorylated (Price *et al.*, 1994). The pattern of partial serine phosphorylation is observed in spp24 (Table 1.1).

Hu *et al.* (1995) suggested that the clustering of partially phosphorylated serine residues observed in spp24 could play a role in regulating the extent of phosphoprotein activity by a specific phosphatase or protein kinase and the potential negative charge in this domain could produce sufficient repulsion to prevent the formation of secondary structure. Consequently the serine residues at this domain could act as an anionic spacer region between the cystatin-like and non-cystatin-like domains. The extent of phosphorylation could regulate the separation of the two domains and consequently modulate the activity of spp24 or its susceptibility to proteolytic cleavage.

In summary Hu *et al.* (1995) suggested three possible functions for spp24:

- inhibitory activity against cysteine proteases
- biological activity of the C-terminal domain (following cleavage) on a specific target
- a fetuin-like plasma protein function

Finally it was suggested that spp24 could have a function in bone in which the extent of phosphorylation in the clustered region of serine residues may act to modulate its activity.

1.3 Normal human bone structure and remodelling

Because spp24 was first extracted from bone, a possible role in bone turn over was suggested (Hu *et al.,* 1995). To determine its function we should look at human bone structure and its remodelling, to give us an understanding of the potential relevance of spp24 in the context of bone biology and bone mineral density.

Bone is the hardest tissue of the human body. As the main component of the skeleton, it supports fleshy structures and protects vital organs such as those in the thoracic and cranial cavities. It is the site of haematopoiesis where blood cells are formed. Bone also serves as a reservoir for vital ions like calcium, phosphate and others which can be mobilised if constant concentrations of these important ions in body fluids are required. Bones form a system of levers that multiply the forces generated by muscles, enabling locomotion of the body. There are two main forms of tissue within each bone. Cortical bone (which forms the hard and dense plates of bone) and trabecular bone (inside the cortex, which is a meshwork of bone and is oriented to oppose external pressure). This structure gives bone a high strength to weight ratio and provides a large area for the remodelling process. Bone is a specialised connective tissue composed from extracellular matrix which includes organic and mineral phases and three different cell types: osteocytes which are found in the cavities of matrix, osteoblasts, which synthesise the organic phase of the matrix and the osteoclasts which are involved in resorption and remodelling of bone (Junqueira, 1992).

Bone remodelling is a co-ordinated process of cellular activity, which involves both formation and resorption of bone and is responsible for renewal and repair of damaged bone throughout adult life. Understanding the mechanism of this process is very important, because abnormalities of bone remodelling underlie virtually all metabolic bone disorders. Different studies have shown that bone remodelling begins with activation of osteoclast precursors to the site that is to be remodelled, where they differentiate into mature osteoclasts (Figure 1.4). It is probable that the remodelling process is triggered by mechanical stimuli or release of chemotactic factors from micro fractures of damaged bone (Ralston, 1997). Several haematopoietic growth factors such as interleukin-3, macrophage colony stimulating factor (MCFS), tyrosine kinase c-fos are able to initiate and affect this process and tumour necrosis

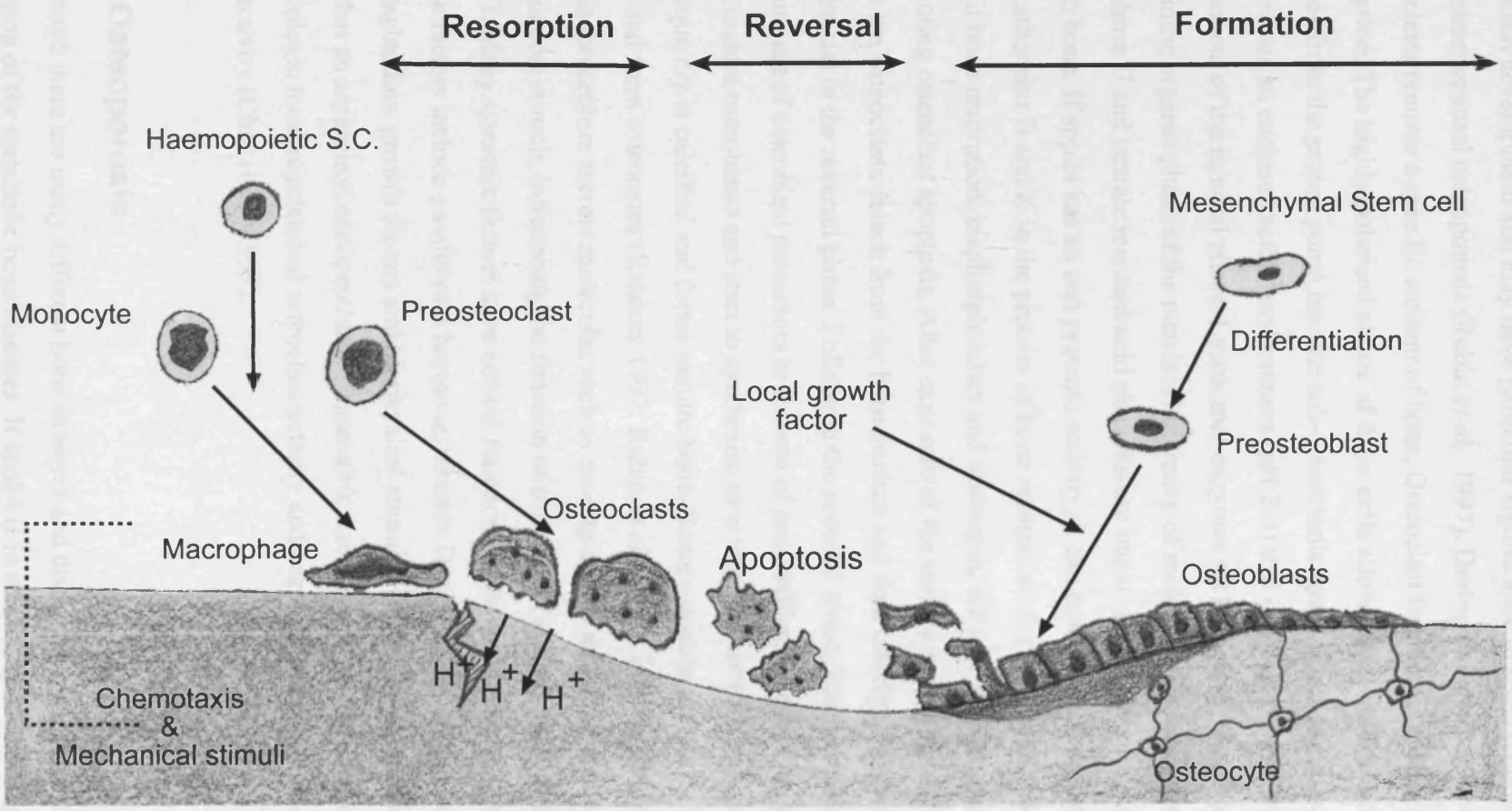


Figure 1.4. A schematic representation of bone remodeling (adapted from Ralston, 1997)

This figure illustrates the processes involved in bone remodelling. Mineralised bone surface is uncovered by resting osteoblasts, osteoclasts remove bone and osteoblasts form new bone. All cells are labelled accordingly.

factor (TNF), systemic factors, 1,25 $(OH)_2D_3$ and PTH can enhance the development of osteoclasts from progenitor cells. Cytokines such as interleukin-6 and -11 also influence osteoclast development and may have particular significance in some pathological states such as postmenopausal osteoporosis (Rubin *et al.*, 1997). During the phase of bone resorption, osteoclasts remove a specific amount of bone. Osteoclast function is specialised for bone resorption. The highly polarised nature of these cells allows them unidirectional secretion of protons from the proton pump into the sub-osteoclastic space where bone resorption occurs and creates an enclosed acidic environment (pH 2-4) that is sufficient to dissolve the mineral component of the mineral phase. Lysosomal enzymes such as cathepsins B and K digest the remaining organic phase of the matrix. Deficiency of molecules such as cathepsin K, carbonic anhydrase II and tartrate resistant acid phosphatase impair the ability of mature osteoclasts to resorb bone. If spp24 has an anti protease activity, it can modulate the function of proteases like cathepsins B and K in the process of bone resorption. Several drugs and hormones that inhibit bone resorption, bisphosphonates and oestrogen, are now thought to act through promoting osteoclast apoptosis. After exposure of the surface receptors to these hormones and drugs the osteoclasts detach from the bone surface and then undergo programmed cell death (apoptosis) in the reversal phase. Following the reversal phase, bone formation begins with recruitment of osteoblast precursors to the site of remodelling. These cells, after maturation, form mature osteoblasts and start to synthesise new bone matrix (osteoid), which subsequently is calcified and forms mature bone. Some osteoblasts become buried in the new bone and form osteocytes (Ralston, 1997; Rubin *et al.*, 1997). It is possible that osteocytes as mechanoreceptors secrete molecules such as prostaglandins and nitric oxide in response to mechanical stimuli, influencing the function of osteoblasts and osteoclasts (Klein-Nulend, 1995). Many systemic factors have critical functions in the regulation of bone metabolism. These factors include parathyroid hormone, vitamin D, oestrogens and androgens, prostaglandins, growth factors and biophysical stimuli (Rubin, 1997). A recent study suggests that that an asparginyl endopeptidase (legumain), a member of thiol protease family, may be a physiologic local regulator of osteoclast activity and negatively regulate osteoclast formation and activity (Choi *et al.*, 1999).

1.4 Osteoporosis

Although there are many different bone diseases and disorders, osteoporosis is the most common of the metabolic bone diseases. If spp24 is in some way involved in bone biology

and metabolism, it could have effects on bone mineral density (BMD) and play a potential role in osteoporosis.

Osteoporosis is a condition in which both cortical and trabecular bone are involved and leads to a generalised reduction in BMD accompanied by deterioration of the micro architecture of bone. These events can lead to a low peak bone mass, increased bone loss or both and consequently an increased risk of fracture. Osteoporosis is defined to exist when the BMD value falls more than 2.5 standard deviations below the young adult mean (Stewart and Ralston, 2000). However, a decrease of just one standard deviation within the normal BMD range is sufficient to increase the risk of osteoporotic fracture by 2.3-3.4 times (Giguere and Rousseau, 2000). BMD decreases with age and therefore osteoporosis affects elderly people in particular. Fractures related to osteoporosis occur in about 150,000 people annually in the United Kingdom and as the age of the population increases, osteoporotic fracture will become more common (Ralston, 1997). In the USA, osteoporosis affects more than 25 million men and women (Melton, 1995), leading to 1.5 million fractures annually (Devoto *et al.*, 1998).

Osteoporosis is a multi-factorial disorder. There are many factors that are thought to increase the chance of osteoporosis, including old age, diet, exercise, age at menarche, age at menopause, body mass index and some drugs. The first direct evidence of a genetic role in BMD was identified by comparing the BMD of monozygotic twins to that of dizygotic twins (Smith *et al.*, 1973). This study indicated closer concordance among the monozygotic twins. Evidence from family and twin studies have indicated that genetic factors play an important role in the regulation of bone mineral density and other related skeletal phenotypes relevant to the pathogenesis of fragility fracture. It has been estimated that heritability of BMD at the spine and hip lies between 70-85% with a value of 50-60% for wrist BMD. So far a variety of different genes (such as the genes encoding type I collagen, oestrogen receptor α, vitamin D receptor gene and transforming growth factor beta gene) have been studied and are thought to be associated with osteoporosis (Ralston, 2002). It appears that BMD may involve different genes, all with relatively small effects individually, but contributing to a high degree of heritability when their effects are considered together.

1.5 Proteases

1.5.1 Introduction

Due to the fact that proteases have important role in bone remodelling and the bovine spp24 protein was first isolated from this organ and may affect the function of proteases, they are explained in more detail. Peptidases or proteases are enzymes which degrade proteins by hydrolysis of peptide bonds. In the cells of higher animals, lysosomes contain endo-peptidases and exo-peptidases, the former degrading proteins into oligopeptides and the latter degrading these oligopeptides completely (Barrett, 1987).

Among the four classes of proteases, the aspartic, the serine, the metallo, and cysteine proteinases, most attention has been paid to the cysteine group of mammalian origin (Turk *et al.*, 1997). The reason for this growing interest is that uncontrolled proteolysis can lead to irreversible damage such as chronic inflammation or tumour metastasis. Cysteine (thiol) proteinases are small proteins with molecular weights varying from 23 to 24 kDa with two catalytic residues, Cys25 and His159, which are involved in the hydrolytic reaction and are most active under reducing and mildly acidic conditions, pH 5–6.5 (Henskens *et al.*, 1996).

Cysteine peptidases, on the basis of their molecular structure, can be divided into about 30 separate families. Three families of cysteine proteinases are known to be present in mammals. The most numerous of these are the papain family, C1 (Chen *et al.*, 1997). This family includes cathepsins B, H, S, L and K that exist inside and around mammalian cells (Alvarez-Fernandez *et al.*, 1999). In spite of high structural similarity among the members of the cathepsin family, cathepsins have discrete expression patterns. Cathepsins B, H and L have ubiquitous distribution, but S is highly expressed in spleen and lung and cathepsin K is found abundantly in osteoclasts (Itoh *et al.*, 1999). These are predominantly lysosomal enzymes, responsible for proteolysis in the endoplasmal and lysosomal systems. Cysteine proteinases are responsible for much of the intracellular proteolysis and therefore they are abundant throughout the body. They are also responsible for the processing of pro-hormones and pro-enzymes (Taugner *et al.*, 1985; Marks *et al.*, 1986), some aspects of bone remodelling (bone resorption) and lysis of collagen (Delaissé *et al.*, 1984). In the cytosolic part of the cell there are members of the other two families of thiol endopeptidases: the calpain family (C2) and caspase family (C14). These peptidases mediate limited proteolysis of cytosolic substances.

Another new family (legumain family, C13) can be also added to the list of mammalian cysteine endopeptidases. Legumain is a cysteine endopeptidase that shows strict specificity for hydrolysis of asparaginyl bonds. The enzyme has been isolated from mammals, plants and blood fluke, Schistosoma mansoni. And belongs to peptidase family C13, and is thus unrelated to the better-known cysteine peptidases of the papain family, C1 (Chen *et al.*, 1997).

1.5.2 Cathepsin K

Because spp24 first was extracted from an acid demineralised extract of bovine cortical bone and because of its similarity to the cysteine protease inhibitor superfamily (Section 1.2), Hu *et al* (1995) speculated that spp24 might have a role in bone turn over and prevention of bone resorption.

Cathepsin K is an important enzyme in the process of bone degradation and turnover and it has been identified as member of the cysteine peptidase superfamily. In recent years several research laboratories have successfully cloned the cathepsin K cDNA from rabbit and human cDNA libraries and its role in bone resorption has been closely studied (Bromme *et al.*, 1995; Inaoka *et al.*, 1995; Li *et al.*, 1995; Shi *et al.*, 1995).

The rabbit ortholog of cathepsin K, rabbit OC-2, was first identified in a rabbit osteoclast cDNA library and various methods confirmed its expression in osteoclasts and demonstrated that it was highly expressed in these cells compared to other tissues (including liver, spleen, kidney and lung) (Tezuka *et al.,* 1994). Human cathepsin K is highly and selectively expressed in osteoclasts and its level of expression is much higher than that of cathepsins B, L and S (Darke *et al.*, 1996). Also, it was shown that cathepsin K mRNA is expressed at a high level in osteoclastoma (a benign osteoclast-origin tumour which consists of multinucleated giant cells that cause bone degradation), in osteoporotic hip bone, colon and ovary and at lower levels, or not at all, in other tissues (Li *et al.,* 1995). Saneshige *et al.* (1995) demonstrated that retinoic acid, a vitamin A metabolite, can up-regulate the expression of cathepsin K in osteoclasts and increases osteoclast-mediated bone resorption.

Cathepsin K has a very potent proteolytic activity against type I collagen, fluorogenic peptides (synthetic fluorogenic peptides or substrates are used to detect the activity of proteases in living cells) (Boonacker and Van Noorden, 2001) and osteonectin (Bromme *et al.*, 1995; Bossard *et al.*, 1996) and in fact is the only cysteine protease known to be capable

of cleaving collagen at multiple sites both within and outside the helical region (Garnero *et al.*, 1998). Whereas other types of cysteine peptidases can attach and digest only the extra-helical regions at either side of the helical region, cathepsin K has proteolytic activity for the triple helical region of this molecule. It has been shown that cathepsin K can be inhibited by a non-specific cysteine protease inhibitor, E-64, and leupeptin, which is characteristic of thiol proteinases (Bossard, 1995). Also, in rat osteoclast pit formation assays it has been shown that specific inhibition of cathepsin K leads to a decrease in bone resorption (Xia *et al.*, 1999). Cathepsin K antisense phosphothiorate oligodeoxynucleotide (ODN) can inhibit the bone resorption activity of osteoclasts and the inhibitory effect of the ODN is almost equal to that of E-64 (Inui *et al.*, 1997). Together, these data demonstrate that inhibition of cathepsin K should result in a diminution of osteoclast-mediated bone resorption.

1.5.3 Clinical importance of cysteine proteases

Since it is thought that initial pit formation is triggered and mediated by the demineralisation of bone by acidification, cysteine proteases most likely contribute in the later stages of bone resorption. It was shown that optimum concentrations of bafilomycin A1, a specific vacuolar H^+-ATPase inhibitor, almost completely inhibited bone resorption and degradation by blocking osteoclast proton transport (Sundquist *et al.*, 1990; Ohba *et al.*, 1996). Also, the results from another study indicate that cysteine proteases inhibitors, or the reduction of cathepsin K, cannot completely inhibit pit formation and bone resorption, but they impair continuation of the process of cavitation, without significantly affecting the initial pit formation. Also, this study clearly suggested that degradation of the protein component in bone matrix is primarily mediated by the action of cathepsin K rather than cathepsins L and B (Inui *et al.*, 1997). Cathepsin K has been implicated as an effective factor in osteoporosis. Although no natural inhibitor of cathepsin K has been identified so far, synthetic inhibitors of cathepsin K have been shown to reduce osteoclast resorption *in vivo* and *in vitro*. Because of this property of the cathepsin K inhibitors, many pharmaceutical companies are currently studying them with a view to developing new therapeutic agents for disorders which are characterised by bone resorption and degradation. Cysteine proteinases including cathepsin K have also been implicated in other pathological conditions such as metastasis and tumour invasion (Sloane and Honn, 1984) and in infection by different micro organisms (Barrett, 1984) which will be explained in more detail in the following sections. Specific inhibitors of cathepsin K could be expected to therapeutically counteract these pathological states.

Another well known pathological disorder which is caused by abnormality in the function of cathepsin K, and which has helped in further defining the role of this protein, is sclerosing osteochondrodysplasia pycnodysostosis. Pycnodysostosis is a rare autosomal recessive inherited disease (MIM 265800) that was first mapped to the same region as cathepsin K, human chromosome 1q21 (Gelb *et al.*, 1996a). This genetic disease is characterised by osteosclerosis (increased bone density), frequent bone fractures (bone fragility), short stature, acro-osteolysis of distal phalanges, clavicular dysplasia and skull deformities with wide cranial sutures (Maroteaux and Lamy, 1964). From the pathological and ultra structural point of view the number of osteoclasts and their ruffle borders are normal in these patients, but the region of demineralised bone which surrounds each individual osteoclast is increased (Evert *et al.*, 1985). The bone demineralisation in this region is normal, but as is expected the organic matrix is not degraded adequately and bone collagen fibrils are observed. These findings suggest that osteoclasts function normally in demineralising bone but do not adequately degrade organic matrix, leaving an increased amount of demineralised bone at the sites of resorption, which provides further evidence of the essential role of cathepsin K in the degradation of bone organic matrix. Several mutations of cathepsin K have been associated with pycnodysostosis. These mutations include X330W (stop codon suppression), G146R (missense mutation), R241X (nonsense mutation), A277V (missense mutation) and A277E (missense mutation) (Gelb *et al.*, 1996b; 1998; Johnson, 1996).

Since both cathepsin K and spp24 are known to be found in bone and no natural inhibitor has identified for cathepsin K so far, spp24 could be a potential candidate protein as an inhibitor for cathepsin K and could have a role in pathogenesis of diseases that are characterised by loss of bone mineral density such as osteoporosis.

1.6 Cystatins

1.6.1 Introduction

Because a considerable amount of our current knowledge about spp24 has been derived from the original study (Hu *et al.*, 1995) and from inferences made from comparing the spp24 protein with the other members of the cystatin superfamily and it may have a cystatin-like function, this aspect is explained in more detail.

Without any doubt, the activities of all degradative proteases need to be controlled. Cystatin or thiol protease inhibitors are a group of proteins which have this controlling function and regulate the activity of cysteine proteases. The name cystatin was first given to an inhibitor of papain and other related cysteine endopeptidases that had been identified for some years in chicken egg white (Barrett, 1981) and also had been shown to inhibit papain (a plant enzyme), and cathepsins B and C (Fossum and Whitaker, 1968; Keilová and Tomášek, 1974). Cystatins are also known as thiol protease inhibitors or cysteine proteinase inhibitors. The cystatin superfamily (group of proteins for which there is evidence of evolutionary relationship) is a group of proteins that contain one or more cystatin-like sequences and most of them are active cysteine protease inhibitors, which show very high affinity and competitive binding to their target enzymes (Barrett, 1987; Turke and Bode, 1991; Henskins *et al.*, 1996, Brown and Dzeigieleweska, 1997). Cystatins inhibit thiol proteases by occupying the active sites of thiol proteases, where they are hydrolysed far more slowly than the natural substrate and hence limit the processing of substrate molecules by the thiol proteases (Baggio *et al.*, 1996). Many proteins have been identified that are similar in their function and structure to chicken egg white cystatin and thus have been categorised and placed as members in the cystatin superfamily.

1.6.2 Classification

Based on the homology of the different members of the cystatin superfamily to the first-discovered cystatin from chicken egg white, (Barrett, 1981), the size and the complexity of the polypeptide chains, this group has been grouped into 3 further subfamilies (Barrett *et al.*, 1986 a, b), type 1, type 2 and type 3 cystatins. Figure 1.5 shows a schematic representation of these three cystatins subfamilies (adapted from Barrett, 1987). Since the late 1980s, when it was largely accepted that there were three different types of cysteine protease inhibitors,

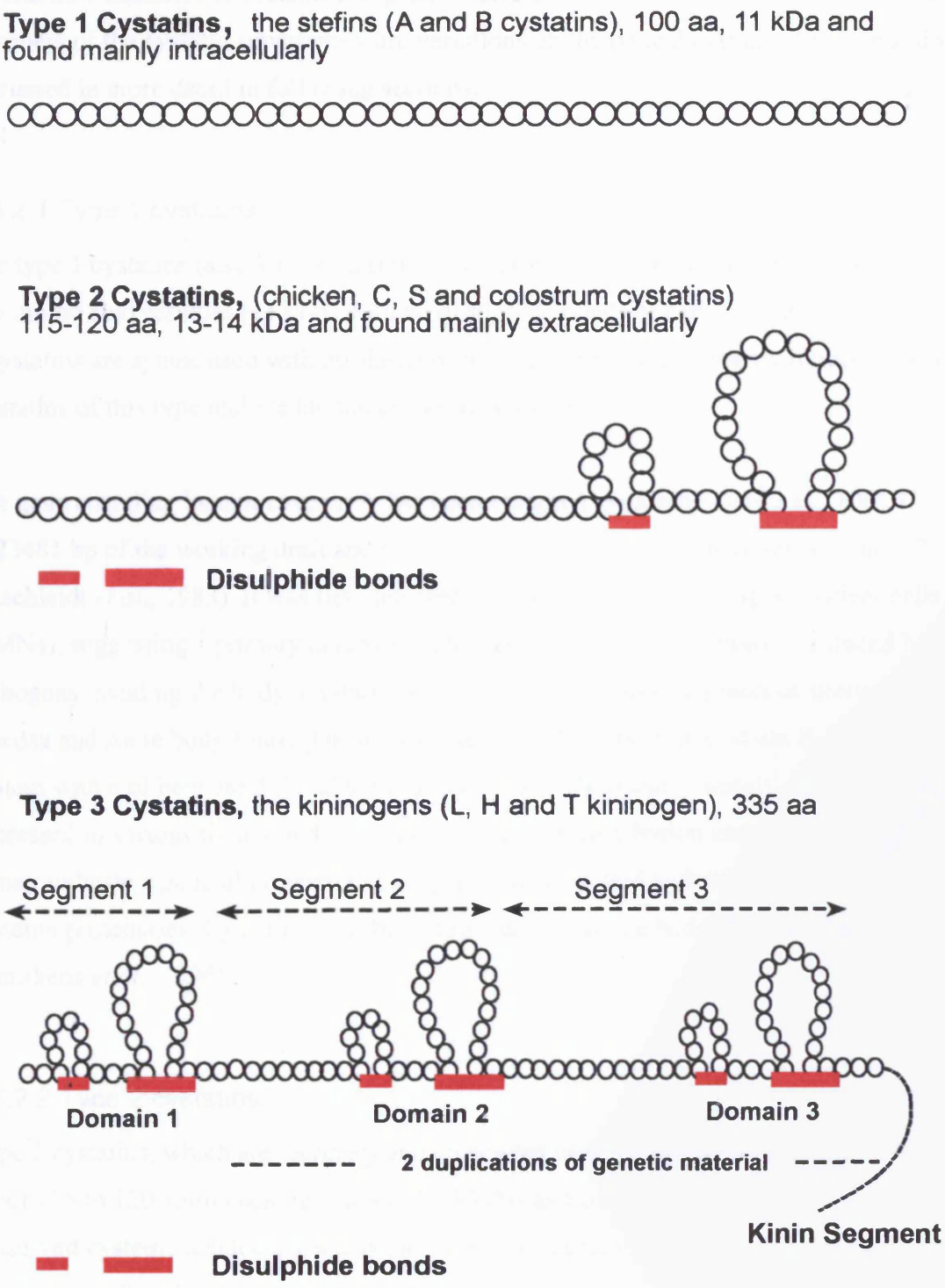


Figure 1.5. A schematic representation of the different sub-types of the cystatin superfamily

Type 1 cystatins, or stefins, have no disulphide bonds and are represented in this figure as a tandem string of amino acids residues. Type 2 cystatins have a single cystatin domain with two disulphide bonds. Finally type 3 cystatins or kininogens have three cystatin domains that each comprise two disulphide bonds plus a non-cystatin-like C-terminal segment (the kinin segment) that is released as a biologically active peptide (not drawn to scale).

several new members or cystatin-like protein have been discovered. All these more recent members of the cystatin superfamily are variations on the typical cystatin structure and will be discussed in more detail in following sections.

1.6.2.1 Type 1 cystatins

The type 1 cystatins (also known as stefins) are proteins whose molecules consist of about 100 amino acid residues (11 kDa) with no disulphide bonds or carbohydrate groups. The type 1 cystatins are synthesised with no signal peptide and are found primarily intracellularly. Cystatins of this type include human cystatins A and B.

The gene encoding human cystatin A has been assigned to chromosome 3 at 1106967-1123481 bp of the working draft and is an acidic protein with a pI between 4.5 and 4.7 (Machleidt *et al.*, 1983). It was first detected in epithelial and polymorpho-nuclear cells (PMNs), suggesting a primary defensive role against cysteine proteinases produced by pathogens invading the body. Cystatin A is found in liver, spleen, placenta, uterus, oral mucosa and some body fluids (Davies and Barrett, 1984). Human cystatin B is a more neutral protein with a pI between 5.7 and 6.3 (Ritonja *et al.*, 1985) and is widely distributed and expressed in various tissues and cells. This ubiquitous distribution and expression of cystatin B may indicate a general protective role against uncontrolled activities of host lysosomal cysteine proteinases. Cystatin B has been detected in just one body fluid, seminal plasma (Henskens *et al.*, 1996).

1.6.2.2 Type 2 cystatins

Type 2 cystatins, which are secretory and somewhat more complex, comprise molecules of about 115 to 120 amino acid residues (13-14 kDa) and contain two disulphide bonds (between conserved cysteine residues) towards the carboxyl end terminus. Type 2 cystatins are mainly found extracellularly in body fluids, but can also be found in different tissues. Examples of this group are chicken egg white cystatin, human cystatins C, D, S, SA, SN, E/M and F. The gene encoding human cystatin types 2 (except cystatin E/M that has been assigned to chromosome 11) are clustered on a 1.2 Mb segment on chromosome 20 (Henskens *et al.*, 1996).

Human cystatin C was first detected as a trace protein in cerebrospinal fluid but not in serum (Clausen, 1961). This protein was later also detected in human biological fluids like plasma,

saliva, seminal plasma, urine and in tissues and cells such as adenohypophysis and brain cortical neurons (Grub *et al.*, 1983). In 1984 it was showed that this protein has high amino acid sequence homology with a cysteine proteinase inhibitor isolated from sera of patients suffering from autoimmune disease and an inhibitory activity towards papain, cathepsins B, H and L were determined (Barrett *et al.*, 1984). Because of the widespread extracellular distribution of this protein, it has been suggested to play a defensive role against host or exogenous cysteine proteinases present in body fluids.

Human cystatins S, SA and SN were first isolated from saliva as acidic and neutral unglycosylated proteins with affinity for hydroxyapatite (Juriaanase and Booij, 1979). Screening of other secretory fluids demonstrated that tears and seminal plasma contained these proteins as well. Cystatin S appears to consist of three isoforms with different phosphorylated serine content, cystatin S1 with no phosphorylated serine, cystatin S2 which is monophosphorylated (Ser^3) and cystatin S3 which is diphosphorylated (Ser^1 and Ser^3) (Isemura *et al.*, 1991). Cystatins S, SA and SN share 90% amino acid sequence homology with each other (Isemura *et al.*, 1987). The biological function of these proteins is not yet established, but it is hypothesised that cystatins S, SA, and SN play a role in the protection of the oral cavity and eyes against the proteolytic activity of cysteine proteases of bacteria, viruses and host inflammatory cells.

Human cystatin D, a neutral cystatin species, was cloned from a genomic DNA library using a cystatin C cDNA probe. Cystatin D expression was identified only in parotid gland tissue and not in other tissues (seminal vesicle, liver and placenta) (Freije *et al.*, 1991). Cystatin D was also found in whole saliva and tears, but its concentrations in seminal plasma, blood plasma, milk and cerebrospinal fluid were below detectable levels (Freije *et al.*, 1993). These results show that the distribution of cystatin D is more tissue-restricted than that of cystatin C.

Cystatin E was identified as an additional human cystatin superfamily member by sequencing ESTs from an epithelial cell-derived cDNA library (Ni *et al.*, 1997). The same cystatin was independently identified by differential display experiments as an mRNA species down-regulated in highly metastatic breast tumour tissue and reported under the name cystatin M (Sotiropoulou *et al.*, 1997). Cystatin E/M is an atypical secreted low molecular weight cystatin (glycoprotein) and shows only 30-35% sequence identity with the human family 2 cystatins.

Cystatin F was identified on analysis of EST sequences obtained from human cDNA libraries. Unlike other members of the human type 2 cystatins, cystatin F has two additional cysteine residues, indicating the presence of an extra disulphide bond stabilising the N-terminal region of the molecule. Like cystatin E/M, cystatin F is a glycoprotein, carrying two N-linked carbohydrate chains. Northern blot analysis and searching different human cDNA libraries revealed that cystatin F gene is primarily expressed in peripheral blood cells, spleen, dendritic and T cells, but no clone was found in several B-cell libraries and in more than 600 libraries from other human tissues and cells (Ni *et al.*, 1998).

Cystatin-like metastasis-associated protein (CMAP) was identified with a differential display system in murine carcinoma cells showing a high rate of metastasis to liver. A human homologue was also found using a PCR-based strategy. This protein shows 22-28% sequence homology to human family 2 cystatins. The human homologue of *Cmap* was found to be expressed in various human cancer cell lines established from malignant tumours. Transfection of *CMAP* antisense DNA into highly metastatic liver cells greatly decreased their metastatic potential and *CMAP* expression, indicating that this protein is involved in the metastasis of malignant liver cells to other locations (Morita *et al.*, 1999).

The cystatin-related epididymal spermatogenic (CRES) proteins are a new subgroup within the human family 2 cystatins. Members of this subgroup are predominantly expressed in human reproductive tissues and lack critical cystatin active-site sequences implying divergent function. The CRES subgroup has six members (*Cres*, *Cres2*, *Cres3*, cystatin T, testatin and cystatin SC) all sharing the unique features of divergent cystatin inhibitory sequences and predominant expression in the reproductive system (Hsia *et al.*, 2003).

Cystatin-like molecule (CLM) is a new member of the cystatin superfamily. It was cloned from a human bone marrow stromal cell cDNA library. The putative CLM protein contains 159 residues with a 29-residue signal peptide. This protein is highly homologous to family 2 cystatins, especially human and mouse testatin. The CLM gene has 2 exons and was mapped on chromosome 20 among the cystatin superfamily gene cluster. Northern blot analysis revealed that CLM is ubiquitously expressed in normal tissues. The results indicate that the secreted CLM protein might play roles in haematopoietic differentiation or inflammation (Sun *et al.*, 2003).

1.6.2.3 Type 3 cystatins

The type 3 cystatins, the kininogens, are the most complex. Each molecule has about 355 amino acids residues and contains three type-2-like domains (D1, D2 and D3) that have clearly resulted from two duplications of genetic material. Domains 2 and 3 are functionally active and have an inhibitory effect on thiol proteases, but domain 1 has no inhibitory activity. At their carboxyl termini, they have an additional polypeptide unrelated to the cystatins and contain the bradykinin sequence that can be released by the action of kalikrein (Barrett, 1987). Bradykinin plays an important role in the intrinsic blood coagulation pathway. In mammals there are three types of kininogens, low-molecular-weight and high-molecular-weight kininogens and T kininogens (Henskens *et al.*, 1996). The kininogen gene has been assigned to the long arm of chromosome 3 at 3q26-qter (Fong *et al.*, 1991). Kininogens are synthesised in liver and then secreted into the bloodstream. Their concentration is highest in blood plasma and synovial fluid, whereas in other body fluids such as tears, seminal plasma, cerebrospinal fluid and colostrums only trace amounts are found. The kininogens and α2-macroglobulin are together the major inhibitors of cysteine proteases in blood plasma (Abrahamson *et al.*, 1986).

Figure 1.6 illustrates the evolutionary relationship between the well known human cystatin domains that are able to inhibit the cysteine proteases (adapted from Ni *et al.*, 1998).

1.6.3 The mechanism of action and stability of cystatins

The hydrolysis reaction of a peptide by a cysteine protease enzyme takes place in two phases. In the first phase, the enzyme interacts with the peptide chain and cleaves a peptide bond. An acyl-enzyme complex is formed between a part of the cleaved substrate and the cysteine protease. The other free part of the cleaved substrate is released as a leaving group. During the second phase, water reacts with the acyl-enzyme. This reaction results in the thiol protease enzyme being regenerated and a second peptide being released (Baggio *et al.*, 1996).

Cysteine protease inhibitors, by replacing the substrate peptide in the formation of the acyl-enzyme complex, can inhibit thiol proteases. The cystatin then hydrolyses very slowly and sometimes not at all. As a result, while the reactive site of the thiol protease is occupied by cystatin, no further substrate molecule can be digested. Figure 1.7 illustrates the hydrolysis of a peptide by a cysteine protease.

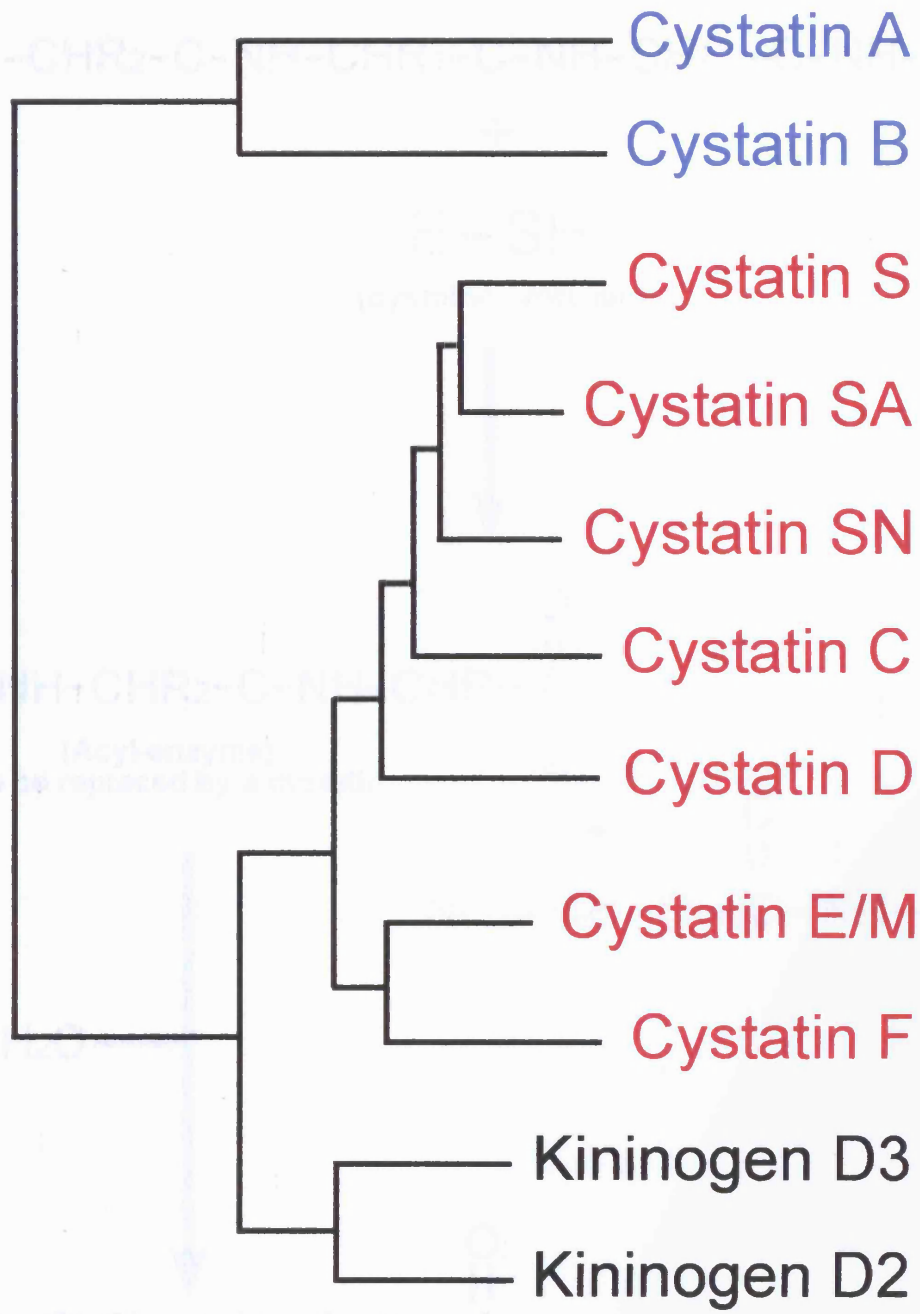


Figure 1.6. Schematic illustration of the evolutionary relationship of members of the cystatin superfamily

This figure shows the evolutionary relationship between all well known human members of human type1, type 2 and type3 cystatins subfamilies (adapted from Ni *et al.*, 1998).

Type1 members are shown in blue, type2 in red and domains 2 and 3 of the type3 are shown in black.

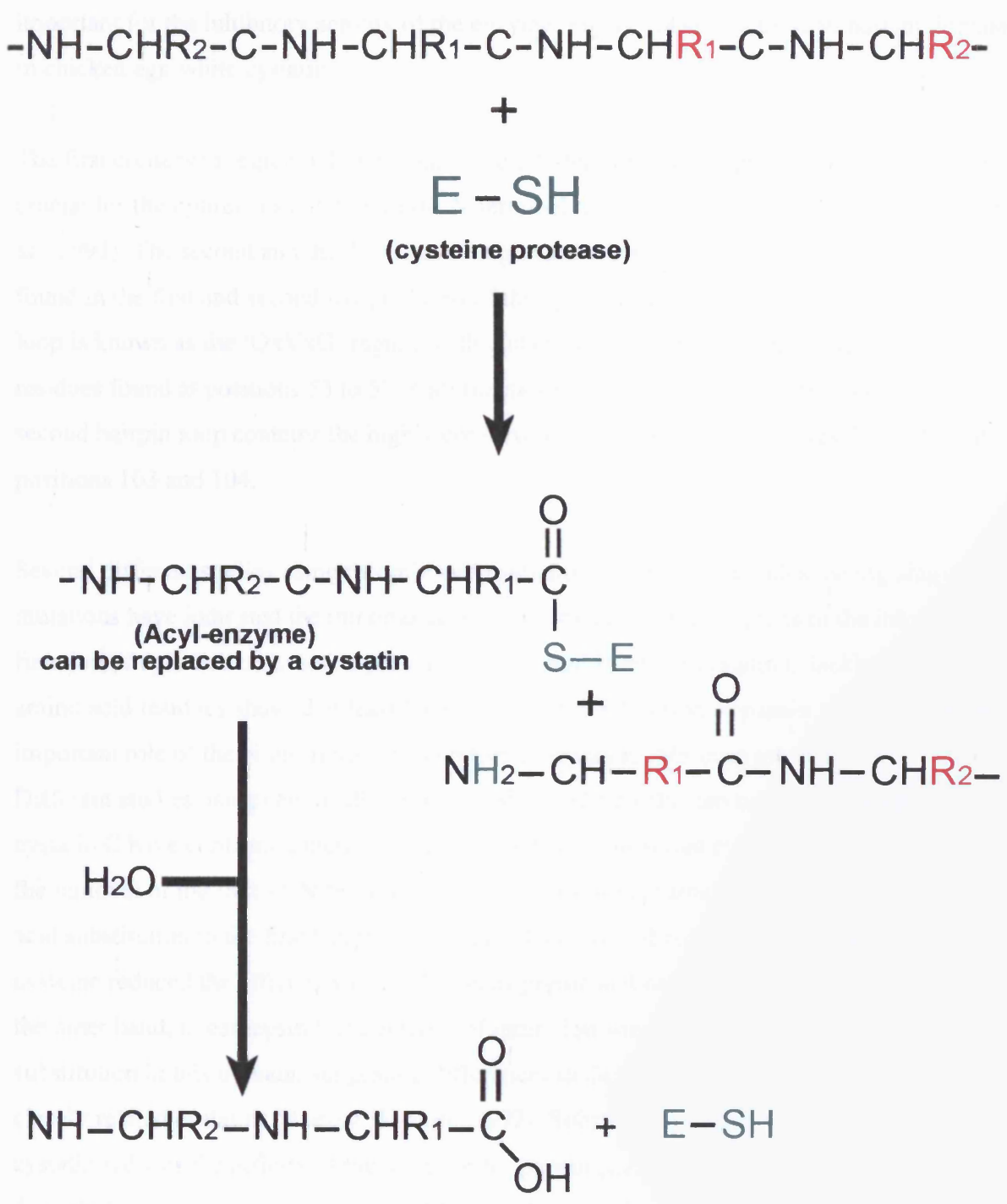


Figure 1.7. Figure illustrating the mode of action of a cysteine protease (adapted from Baggio *et al.*, 1996)

The amino acid sequences of cystatins are highly conserved in three main domains which are important for the inhibitory activity of the enzyme. Figure 1.8 shows these important domains in chicken egg white cystatin.

The first conserved region, Gly9, is found at the N-terminal end of protein and it seems to be crucial for the optimal orientation of the N-terminal region toward the thiol protease (Hall *et al.*, 1993). The second and third conserved regions involved in the inhibition are residues found in the first and second hairpin loops of the cystatin structure. The region in the first loop is known as the 'QxVxG' region, with glutamine (Q), valine (V) and glycine (G) residues found at positions 53 to 57 in all functional cystatins (x can be any residue). The second hairpin loop contains the highly conserved proline and tryptophan residues (PW) at positions 103 and 104.

Several different studies using recombinant cystatins or synthetic peptides having single mutations have indicated the importance of these three conserved regions in the inhibitory function of cystatins on cysteine proteases. Truncated forms of cystatin C lacking the first 11 amino acid residues showed at least 1000-fold weaker inhibition of papain which indicates the important role of the N-terminus for inhibitory interaction (Abrahamson *et al.*, 1987). Different studies using chemically or genetically modified chicken egg white cystatin or cystatin C have confirmed these findings. It has been shown that cystatin A is inactivated after the removal of the first 15 N-terminal amino acid residues (Samejima *et al.*, 1986). Amino acid substitution in the first hairpin loop region (QxVxG) of recombinant chicken egg white cysteine reduced the efficiency of inhibition of papain and cathepsin B by 10 to 1000 fold. On the other hand, in cathepsin L the activity of inhibition was unaffected by amino acid substitution in this domain, suggesting differences in the inhibitory interaction between closely related cystatins (Auerswald *et al.*, 1992). Substitution of Trp104 in chicken egg white cystatin reduces the affinity of the inhibitor for papain (Lidhal *et al.*, 1988). Reduction of the disulphide bond between Cys71 and 81 has no effect on the inhibitory activity of chicken egg white cystatin but, in contrast, the disulphide bond between Cys95 and 115 has an important role for maintaining the native and effective conformation (Björk and Ylinnenjärvi, 1992).

1.6.4 Proposed functions of cystatins

In addition to the established role of inhibition of members of the papain superfamily, several other functions in health and disease conditions have been suggested for cystatins which

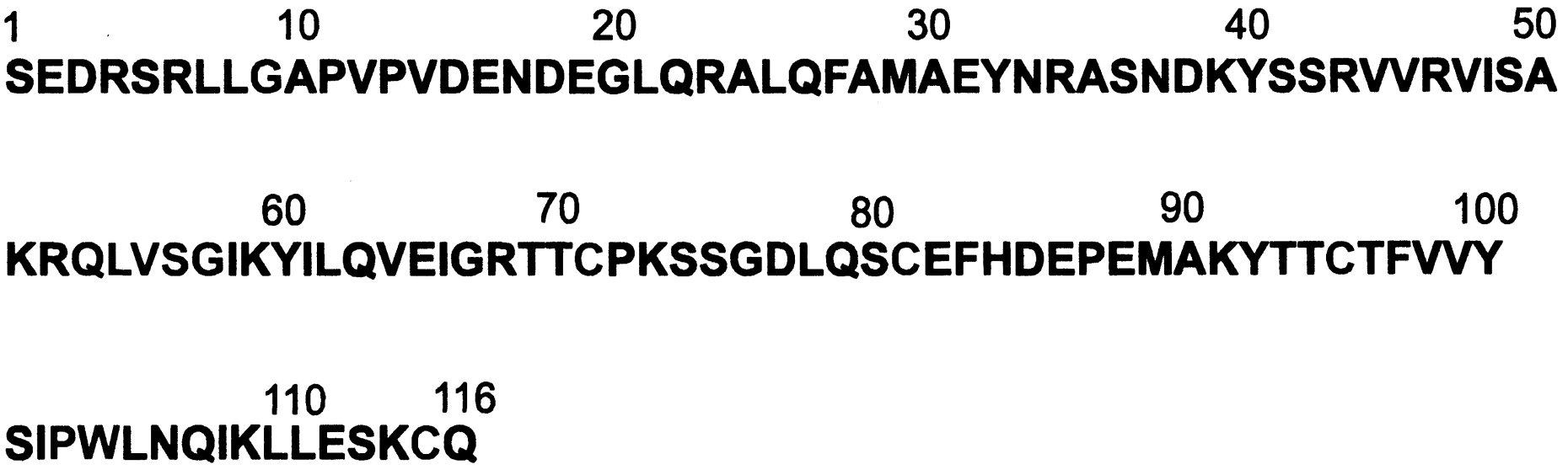


Figure 1.8. The highly conserved residues of cystatins thought to be functionally important

The highly conserved residues are shown in red. The indicated sequence is that of chicken egg white cystatin.
There is a highly conserved glycine residue at position 9, 'QxVxG' sequence at position 53-57 which consists of a glutamine, valine and glycine with 'x' being any amino acid residue and finally the highly conserved proline and tryptophan sequence at positions 103 and 104. The cysteine residues are shown in blue.

include: defending the body against infections, protection from inflammatory reactions, protection from destruction of cartilage and collagen, prevention of some neurological disorders thorough decreasing the degradation of myelin and finally prevention of formation and metastasis of some tumours.

Increased activity of some cysteine proteases has been linked to the formation and progression of different types of malignancies. Different investigations have indicated that an imbalance between cathepsins and cystatins, associated with the metastatic tumour cell phenotype, may facilitate tumour cell invasion and metastasis and may be responsible for early relapse of the disease after the initial removal of the primary tumour (Lah *et al.*, 1998). Cathepsin B can degrade laminin, fibronectin and type IV collagen and facilitate the local dissolution of basement membranes observed during tumour cell extravasation (Sloane, 1990). For example, plasma-membrane fractions of several human tumours such as breast, colon, ovary, bladder and adrenal gland contain higher levels of cathepsin B-like and cathepsin L-like activities compared to normal tissues (Sloane *et al.*, 1987; Rozhin *et al.*; 1989; Chauhan *et al.*; 1991). It was originally thought that either an under expression of cystatins or over expression of cysteine proteases in the cancer cells facilitates the progression of cancers and metastasis. However Collella *et al.* (1993) suggested an opposite effect of cystatins in the process of malignancy; that is cystatins could inhibit the proteolytic activity of cysteine proteases on malignant cells by suppressing the inflammatory responses. For example, cystatin C can inhibit the proteolytic effects of cathepsins on the malignant cells by suppressing the host inflammatory responses and in this way increasing the oncogenicity of cancer cells (Nishida *et al.*, 1984). The different results of clinical investigations on cysteine proteases and their endogenous inhibitors in lung, brain and head & neck tumours, as well as in body fluids of ovarian and uterine carcinoma-, melanoma- and colorectal carcinoma-bearing patients, have indicated that these molecules are highly predictive for the survival time and may be used for risk assessment of relapse and death of patients. Their application for diagnosis, follow-up and anticancer therapy has also been suggested (Kos *et al.*, 1998).

A possible role in inflammatory disorders has also been suggested for cystatins. Due to its epidermal origin, cystatin A has been extensively studied in skin diseases. For example, increased amounts of cystatin A have been identified in psoriatic epidermis and inflammatory skin samples. On the other hand, there is evidence to indicate that a cystatin (probably cystatin A) isolated from psoriatic patients was less active and stable towards papain compared to the inhibitor found in normal cells (Järvinen *et al.*, 1987; Othani *et al.*, 1982).

Periodontal inflammation is a disease which affects the tissues supporting the teeth and is thought to be associated with high levels of cathepsin B, H and L in gingival tissues and gingival cervicular fluids. Salivary cystatins could play a role in protection of periodontal tissues against the effects of exogenous and endogenous cysteine protease sources. It has been shown that cystatin activity of whole saliva is highest in patients suffering from periodontitis when compared to healthy persons (Henskens *et al.*, 1993).

Since synovial fluid of patients suffering from inflammatory joint disease contains high levels of cathepsin B, it is thought that cysteine proteases might play a role in destruction of articular cartilage and collagen. The synovial fluid of patients suffering from rheumatoid arthritis also contains the highest levels of cystatin C (Lenarcic *et al.*, 1988). A possible function for cystatin C in bone resorption was suggested by Lerner and Grubb (1992). Using recombinant cystatin C they showed a significant reduction in the release of ^{45}Ca and ^{3}H in parathyroid hormone-stimulated release of these elements from pre-labelled mouse calvarial bone.

It is thought that cystatins might have a role in defending the body against bacterial and viral infections. For example, some viruses (polio virus, rhino virus, herpes simplex virus) in order to replicate and produce new virus particles require the presence of cysteine proteases in the cytoplasm of the infected cells. Hence, the presence and activity of cystatins could prevent the replication of the virus. It has been shown that chicken egg white cystatin can reduce the replication of polio virus in infected cells and also result in the absence of viral protein synthesis when the cells were exposed to cystatin prior to viral infection (Korant *et al.*, 1985). Cysteine proteases also play a role in the penetration of normal tissues by bacteria such as group A *Streptococci*. A synthesised peptide that mimics the proteinase binding site of cystatin C inhibited the growth of all group A *Streptococci in vitro* and *in vivo* (Björk *et al.*, 1989). It has also been shown that rat cystatins S and A can successfully inhibit the growth of *Porophyromonas gingivalis* and *Staphylococcus aureus* respectively (Naito *et al.*, 1995; Takahashi *et al.*, 1994).

Cystatins have also been implicated in neurological diseases, especially cystatin C, which was first detected in cerebrospinal fluid, and has been implicated to have a function in the aetiology of multiple sclerosis (MS). MS is a neurological disorder that involves demyelination, which might be the result of proteolytic enzymes. It has been suggested that the predominance of macrophages in the lesions of demyelinating diseases like MS could imply a role for cysteine proteinases. Also, a significantly lower level of cystatin C in the CSF

(cerebro spinal fluid) of a large group of MS patients and absence of any correlation between age and cystatin C were demonstrated. Consequently, it was suggested that the decreased level of cystatin C could lead to an increased level of proteinase activity and therefore the degradation of myelin (Bollengier, 1987).

1.7 Aims and Objectives

With the completion of the first draft of the human genome sequence, increasing numbers of new genes are now being identified, most of which have no known function. Current investigation has shown that the human genome contains about 30-40,000 genes, though numbers as low as 28,000 and as high as 140,000 have been suggested (Roest *et al.*, 2000). Once a new gene has been discovered, the question of its function has to be addressed, so the major challenges now, and in the future, are large-scale or global investigation of gene function (functional genomics), deciphering the transcriptome (the total collection of RNA transcripts in a specific type of cell or tissue) and proteome (the total collection of proteins expressed in a specific type of cell or tissue). The new science of proteomics is concerned with the study of protein expression (using high-throughput expression assays), protein function and systematic study of protein-protein interactions using the large-scale application of methods such as yeast two-hybrid screening. The project presented here in this thesis is a typical example of using some of these approaches to find a function for a novel gene.

Secreted phosphoprotein 24 (spp24) was first co-isolated with MGP (matrix Gla protein) from an acid demineralised extract of bovine cortical bone (Hu *et al.*, 1995). This protein is encoded by a novel gene *SPP2*. The main challenge after discovery of this protein and gene are to find their structures and functions. Our group has discovered the structure of the human gene and highlighted the difficulties of determining protein function so far (Bennett, 2002; Bennett *et al.*, manuscript submitted). This thesis attempts to describe the different strategies and methods that have been chosen to clarify and delineate some aspects of these functions.

Fortunately the human gene encoding spp24 was sequenced as a part of the Human Genome Project before the start of this work. This achievement and the other work done before and during my study has been a great influence and help to my project, because they have enabled me to chose more detailed and better aims and objectives. The aims and objective of the work presented in this thesis are as follows:

- **Scanning and screening of the human *SPP2* gene to find any new variations.**

  Finding new variations or mutations in or near the human gene encoding spp24 will be very useful as markers for future linkage or association studies, in particular for

genetic or multifactorial disorders if spp24 is suspected to be involved. Spp24 has already been speculated to have role in bone turn over. Because ninety DNA samples from patients with different and altered bone mineral density are available (David Hosking, Nottingham City Hospital) these single nucleotide polymorphisms (SNPs) can be used in a preliminary study to find any meaningful difference between normal individuals and these patients and, if any is found, further investigations could be carried out. The importance of any detected polymorphisms can be assessed in terms of the sequence conservation between species and also by a consideration of the potential effects of an altered amino acid on the structure of the protein. This work is presented in Chapter 3.

- **To determine and analyse the sequence of the mouse *Spp2* gene and compare the spp24 protein in different species.**

  In order to analyse the complete genomic sequence of the mouse gene encoding the spp24 protein, the whole sequence will be determined and analysed using the HGMP Nix analysis environment. In this part of my study I will try to confirm the sequence of spp24 in cattle and also determine the spp24 protein sequence in other species (sheep and marmoset). This study will enable us to highlight highly conserved residues that are likely to be critical to the function of the protein in more species. This work is presented in Chapter 4.

- **Mapping of the mouse *Spp2* gene and analysis of mouse QTLs which map close to this gene.**

  The mapping of genes to a specific location on a chromosome is a central focus of medical genetics. One goal of the human genome project is to map all our genes to specific chromosomal locations. As the work progresses, our understanding of the biological basis of genetic diseases will also progress. When a disease gene's location is pinpointed, it is often possible to provide a more accurate prognosis for individuals at risk of a genetic disease. Human-mouse phenotypic similarities (homologies) provide particularly valuable clues for identifying human disease genes for several reasons. Mutations in orthologous genes are likely to produce similar phenotypes in mice and humans. Mouse phenotypic information, deletion maps and mutation maps often translate readily into positional candidate information. Backcross and radiation hybrid mapping allow quick and accurate mapping in the mouse, so most mouse

mutants have been mapped and there is considerable conservation of synteny between mice and humans. Before this study, the human *SPP2* gene was assigned to 2q37→qter by our group (Swallow *et al.*, 1998). One part of this study concerns the assignment of the mouse *Spp2* gene using *in silico* (to predict the likely location of mouse *Spp2* based on the location of human *SPP2*) and radiation hybrid methods. All quantitative trait loci (QTLs) in the vicinity of the mouse *Spp2* gene are to be identified and associations between these QTLs and *Spp2* investigated. In doing so, it might be possible to identify a role for spp24 in multifactorial diseases. This work is presented in Chapters 5 and 6.

- **To identify the mouse *Spp2* gene expression pattern in mouse liver tissue using non-isotopic *in situ* hybridisation method.**

  High resolution spatial expression patterns of RNA in tissues and groups of cells can be obtained by tissue *in situ* hybridisation. In order to maximise sensitivity, a labelled antisense RNA probe (which can be synthesized by *in vitro* transcription from a cDNA cloned in a suitable vector) is used. In this part of my work this strategy is used to determine the expression pattern of the mouse *Spp2* gene in mouse liver tissue. This work is presented in Chapter 7.

- **To characterise the expression profiles of the human, mouse and chicken genes encoding the spp24 protein.**

  The temporal and spatial expression patterns of a gene can provide useful information that may elucidate the potential functions of a protein. In this part of my work the expression pattern information is obtained in different ways, including data from ESTs, microarrays and RT-PCRs. This work is presented in Chapter 7.

- **Construction of a mouse pooled tissues cDNA library and use of the yeast two-hybrid system to find potential proteins interacting with the mouse spp24 protein.**

  In an attempt to identify potential proteins that may interact with the mouse spp24 protein, the whole mature protein and the non-cystatin domain will be expressed as baits in a yeast two-hybrid system. In order to carry out this experiment I need a cDNA clone library. Because spp24 is secreted, it can interact with other proteins,

synthesized in distant organs. To obtain comprehensive results a mouse pooled-tissues cDNA library will be constructed. This constructed library will be screened to try and identify protein motifs that may interact with the two spp24 baits. Protein databases can then be searched to identify proteins containing these motifs. This work is presented in Chapter 8.

Chapter 2
Materials and Methods

2.1 Safety Issues and Regulations for Genetic Modification

All work was carried out observing good laboratory practice. All hazardous chemical substances were handled and in accordance with COSHH (Control of Substances Hazardous to Health) safety regulations. The genetic manipulation procedures in this study were carried out in accordance with the Genetically Modified Organism Regulations and were reviewed and approved by the University of Leicester Genetic Manipulation Safety Committee and classified as a Class 1 activity (group 1 organism in a type A operation). All human blood and animal tissues were handled, used in experiments, and disposed of according to university health and safety regulations. All radioactive substances were handled, used, and disposed of in compliance with the Ionising Radiations Regulations 1999.

2.2 Centrifugation

Unless otherwise stated, all small volume samples (up to 1.5 ml contained in microfuge tubes) were centrifuged in a MSE Micro Centaur microcentrifuge at 6,500 or 13,000 rpm. Larger volumes were centrifuged in Sorvall RT6000D centrifuge with H1100B rotor (up to 3,600 rpm), Rotina 46R centrifuge with 4624 rotor (up to 4,000 rpm) and Sorvall RC-5B with HB4 or SS-34 rotors (up to 13,000 rpm).

2.3 Use of Restriction Endonucleases

Unless otherwise stated, all digestions with restriction endonucleases (restriction enzymes) were carried out in a total reaction volume of 10–20 µl.

Five units of each required restriction enzyme and 2 µl of appropriate 10× Reaction buffer were used per 20 µl of final reaction volume. In double digestions (the use of two restriction enzymes in one reaction), the reaction buffer was chosen to allow maximum activity of both restriction enzymes, according to the manufacturer's guidance. Reactions were incubated at the recommended temperatures (usually 37°C) for 60–90 minutes (single digestion) or 90 minutes (double digestion). Unless otherwise stated, most of the enzymes used were supplied

by Invitrogen. For digests in which two enzymes were added consecutively, reactions were heated at 65°C for 10 minutes before adding the second enzyme, to inactivate enzymes which are sensitive to heat.

2.4 Agarose Gel Electrophoresis

Agarose gel electrophoresis was used to analyse the identity, purity, and quantity of DNA samples or purify fragment of interests such as genomic DNA, plasmid DNA and PCR product. Based on the desired resolution, gels of 0.7–2% (w/v) standard agarose, SeaKem LE agarose (Cambrex) or 1-4% (w/v) MetaPhor agarose (Cambrex) were used to separate DNA fragments of various anticipated sizes. If DNA was to be recovered from a gel then low melting agarose (SeaPlaque) or SeaKem LE agarose of less than 1% was used. The gel solvent and tank buffer was 1× TBE (89 mM Tris base, 89 mM boric acid, 2 mM EDTA) for standard agarose and 1× TAE (40 mM Tris base, 40 mM acetic acid, 2 mM EDTA) for low melting agarose. Staining of DNA was carried out with ethidium bromide which acts as intercalating agent and added to the tank buffer and the molten agarose to a final concentration of 0.5 $\mu g.ml^{-1}$. Gels were cast in 40 ml, 100 ml or 200 ml volume UV-transparent Perspex trays with a plastic comb to create wells, and run in an electrophoresis tank at up to 120 V for DNA samples and up to 250 V for RNA samples. Each well was loaded with up to 22 µl of sample flanked by 200 ng λ/*Hin*dIII and/or 100 ng ΦX174RF/*Hae*III marker DNA fragments, based on the anticipated fragment sizes, using a 1/10 volume of 10× loading dye (Ficoll 400 (20% w/v), 0.1 M EDTA, pH 8.0, Orange G 0.1% w/v). The λ/*Hin*dIII marker DNA was heated at 65°C for 5 minutes and then kept on ice until loading. DNA or RNA fragments were visualised by UV illumination (302 nm, Ultra-Violet Products) photographed, analysed, fragment size and quantity determined using AlphaEase v.4.0, v.5.5 (Alpha Innotech), on an AlphaImager 2000 system. Gel images were stored as bitmaps in uncompressed TIFF format.

2.5 Recovery of DNA from Agarose Gel

2.5.1 Recovery of DNA from an agarose gel using phenol/chloroform extraction

This method was taken from Sambrook *et al.*, 1982.

DNA was electrophoresed in a low-melting temperature, 1% (or less) LE agarose gel, excised and placed in 1 × TAE. The gel fragment containing the desired DNA was excised from gel using a scalpel blade and placed in a microcentrifuge tube (to prevent the degradation and damaging of DNA the DNA was visualised using a Clare Chemical Research Dark Reader (a blue-light transilluminator). The gel slices (not exceeding 200 mg) then were placed in a microcentrifuge tube at 67°C for 10 minutes to melt the agarose. The appropriate volume of pre-warmed (67°C) 1 × TE (10 mM Tris-HCl pH 7.6, 1mM EDTA) was added so that the final concentration of agarose was less than 0.5%. An equal volume of liquified and equilibrated (with 10 mM Tris-HCl pH 7.5, 1mM EDTA) phenol was added to the tube and vortexed for 15 seconds to mix well. The tube was centrifuged at 13,000 rpm for 3 minutes and the upper aqueous layer was removed and placed in a clean microfuge tube. This phenol extraction was repeated once. Next a third extraction was carried out with an equal volume phenol:chloroform:isoamyl alcohol (25:24:1). This was followed by extraction with an equal volume of chloroform. The resulting upper aqueous phase was chilled on ice for 15 minutes and centrifuged for 15 minutes at 4°C. The supernate was removed and placed in a clean microfuge tube and the DNA precipitated with ethanol (Section 2.6).

2.5.2 Recovery of DNA from an agarose gel using the QIAquick gel extraction kit

The QIAquick gel extraction kit from QIAGEN was used and extraction carried out according manufacturer's protocol.

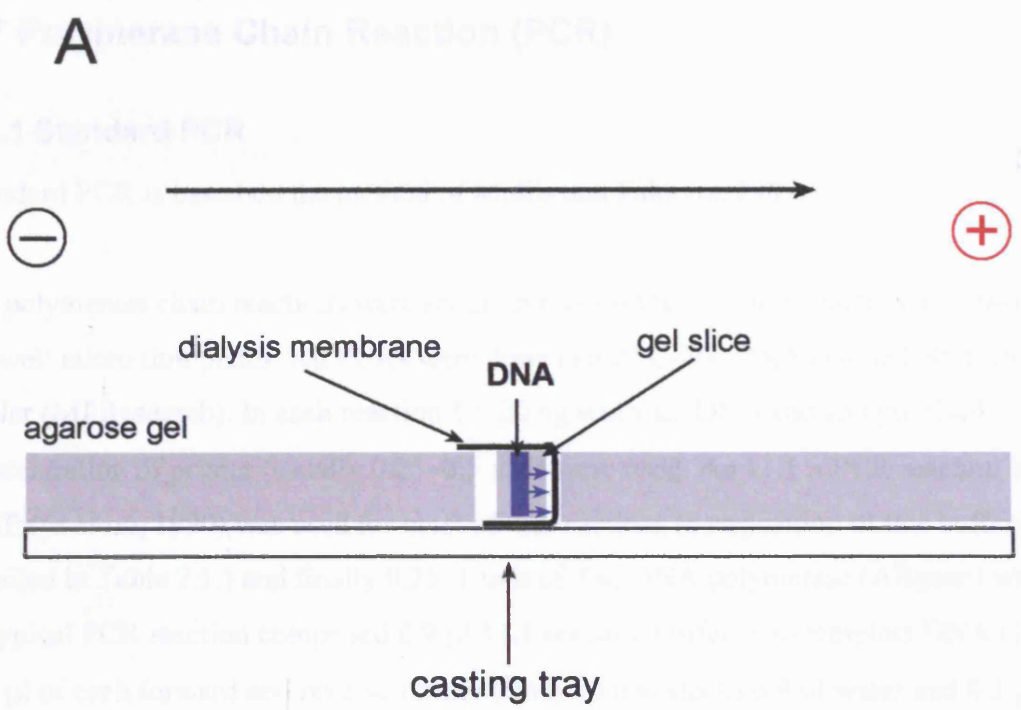
2.5.3 Recovery of DNA fragments from agarose gel by electro elution onto dialysis membrane

The DNA sample (PCR product) was electrophoresed in 1% (w/v) LE agarose gel, excised and placed in a microfuge tube as described in Section 2.5.1. Then a 1% (w/v) LE agarose gel without slots (wells) was prepared (the size of gel depends on the number of DNA fragments and 10–20 DNA fragments can be recovered simultaneously on a 20 × 20 cm gel). The holes were cut in the gel to accommodate the gel fragments and the membrane onto which the DNA was to be eluted. Dialysis membrane (14.3 mm, 12,000–14,000 Daltons, Medicell International Ltd) was cut to the desired size and shape and was boiled in 1 × TE solution (10 mM Tris-HCl pH 7.6, 1mM EDTA). After separation of the two layers of the membrane they were rinsed in water and stored in electrophoresis buffer such as 1 × TBE or 1 × TAE. Then

the gel slice containing the PCR product band and a sheet of dialysis membrane wrapped around the slice (except the cathodic surface) was inserted into the hole (the membrane projected 5 mm on each side of the slice and curved below and above the slice and had a lug to aid removal). Then electrophoresis was carried out to run the DNA onto the membrane at 150 volts. During electrophoresis, the movement and loading of DNA onto the membrane was monitored using a hand held UV lamp (302 nm, Ultra-Violet Products), but exposure was kept to a minimum level to minimise DNA damage. After full loading the DNA onto the membrane, the voltage was reduced to 100 volts and the membrane was transferred quickly to a microcentrifuge tube and the lug trapped in the lid (Figure 2.1 schematically illustrates the recovery of DNA fragments from agarose gel by electro elution onto dialysis membrane). The tube was centrifuged for 10 seconds, flicked to re-wet the membrane and centrifuged for 1 minute and the liquid was transferred into a fresh microcentrifuge tube. The DNA was precipitated by adding 0.1 volume 2 M Na acetate (pH 7.0) plus 2.5 volumes of pure ethanol. The sample was mixed and chilled at -80°C for 3 minutes and centrifuged for 3 minutes. The sample was once more chilled at -80°C for 3 minutes and centrifuged for 3 minutes. The supernate was removed and the pellet washed with 80% (v/v) ethanol. The tube was then centrifuged for 10 minutes. The supernate was discarded and the pellet dried for 5-10 minutes to evaporate the last traces of ethanol. The pellet was then redissolved in the desired volume of 1×TE (10 mM Tris-HCl, pH 7.5, 1mM EDTA) or water. The yield of DNA recovered from the gel using whatever method, was determined by quantitave analysis (by comparison with known amount of size marker DNA), using AlphaImager v.4.0 or v.5.5 software.

2.6 Precipitation of DNA

Unless otherwise stated, all ethanol precipitations of DNA were carried out by adding 2.5 volumes of pure ethanol and 0.1 volume of 3 M sodium acetate (pH 4.5–5.5). The sample was mixed well and then stored at -70 to -80°C for at least 15 minutes. To recover the DNA, the solution was centrifuged at 13000 rpm for 30 minutes. The supernate was discarded and the pellet was washed by adding 70% (v/v) ethanol. The tube was centrifuged once again for 15 minutes. The supernate was discarded and the pellet was dried for 5–10 minutes to evaporate the remaining ethanol. The pellet then was redissolved in the desired volume of water or 1 × TE.



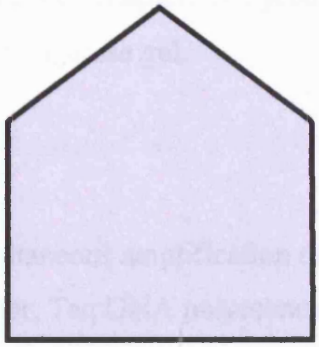


Figure 2.1. Figure illustrating the purification of DNA from agarose gel by electroelution onto dialaysis membrane.

A: This figure schematically shows the direction of the electrical current and the situation of the agarose gel, gel slice, DNA fragment and dialysis membrane to each other.

B: This figure schematically illustrates the shape of the dialysis membrane after cutting and before placing into the well.

2.7 Polymerase Chain Reaction (PCR)

2.7.1 Standard PCR

Standard PCR is based on the method of Mullis and Faloona, 1987.

All polymerase chain reactions were set up in a 10-µl total volume reaction in 200-µl tubes or 96-well micro titre plates. All PCRs were done in a PTC-200 DNA engine Peltier thermal cycler (MJ Research). In each reaction 10–20 ng template DNA and an optimised concentration of primer (usually 0.25–0.5 µM) were used. An 11.1 × PCR reaction buffer (Jeffreys *et al.*, 1990) was used for most of the reactions (composition of this buffer is detailed in Table 2.1.) and finally 0.25–1 unit of Taq DNA polymerase (ABgene) was added. A typical PCR reaction comprised 0.9 µl 11.1 reaction buffer, 1 µl template DNA (10–20 ng) 0.5 µl of each forward and reverse primer (from 10 µM stock) 6.9 µl water and 0.2 µl Taq DNA polymerase (5 units.μl^{-1}).

The concentration of Mg^{2+} (different batches of buffer prepared with different of Mg^{2+} were used here) and the cycling conditions for each PCR reaction were optimised individually and these are detailed accordingly in each chapter. The products of PCR reactions were analysed by electrophoresing 2–5 µl on an agarose gel.

2.7.2 Multiplex PCR

This method was used for simultaneous amplification of different fragments with each other. The concentration of each primer, Taq DNA polymerase and PCR reaction buffer composition were similar to the single PCR method, but the concentration of Mg^{2+} and cycling conditions for each PCR reaction were optimised individually and these are detailed accordingly in the relevant chapters.

2.7.3 RT-PCR

The reverse transcription is based on the methods of Temin and Mizutani, 1970. The PCR is based on the method of Mullis and Faloona, 1987.

Two and half to four micrograms of total RNA and 2.5 µl of reverse primer or 15-mer oligo dT (1 pmol.μl^{-1}) were mixed on ice. To denature the RNA, the sample was incubated at

Constituent	Concentration in PCR reaction
11.1× reaction buffer	
• Tris-HCl pH8.8	45 mM
• Ammonium sulphate	11 mM
• $MgCl_2$	4.5 mM
• 2-mercaptoethanol	6.7 mM
• EDTA pH8.8	4.4 μM
• BSA	113 $\mu g.ml^{-1}$
dNTPs	
• dATP	1 mM
• dCTP	1 mM
• dGTP	1 mM
• dTTP	1 mM
Primers	
• Forward	0.25-0.5 μM
• Reverse	0.25-0.5 μM

Table 2.1 Composition and concentration of PCR reaction buffer and primers

70°C in a water bath for 10 minutes and then, for annealing the primer, it was snapped cooled on ice for 1 minute. Four microlitres of 5× first strand buffer (Invitrogen), 2 µl of 0.1 mm DTT, 1 µl 10 mM dNTP mixture and 0.25 µl of RNasin (Promega, 40 units.μl^{-1}) then were added to the RNA and primer mixture to give a total volume of 20 µl.

The mixture was incubated at 42°C for 2 minutes and then 1 µl of SuperScript II (Invitrogen) was added and incubated at 42°C for a further 50 minutes. Then, to inactivate the enzyme, the mixture was heated at 70°C for 15 minutes and snapped cooled on ice. Two microlitres of cDNA from the reverse transcription reaction were used as the DNA template for a PCR reaction. The PCR reaction was carried out as described in Section 2.7.1, but scaled up to a 40–100 µl total reaction volume. For further analysis, 5–10 µl of the PCR product was electrophoresed on an agarose gel.

2.7.4 Purification of PCR products

To separate any unused primers, dNTPs, Taq DNA polymerase and any unwanted DNA or any component that might inhibit subsequent reactions and experiments such as sequencing and ligation, PCR products were purified using one of two methods:

2.7.4.1 Purification of PCR products using QIAquick PCR purification kit

This method was carried out according the manufacturer's protocol and guidelines (Qiagen).

Five volumes of buffer PB were added to 1 volume of the PCR product and the sample was mixed. The sample then was applied to a QIAquick spin column, placed in a 2 ml collection tube, and centrifuged at 13,000 rpm for 1 min. The flow-through was discarded and the column placed back in to the same collection tube. To wash the DNA, 0.75 ml of buffer PE was applied to the spin column and the sample was centrifuged as previously.

The spin column was then placed in a 1.5 ml microcentrifuge tube. To elute the DNA, 30–50 µl of buffer EB was applied to the centre of the membrane of the spin column and incubated at room temperature for 1–2 minutes. The column was then centrifuged as before and the DNA sample was collected and transferred to a new microcentrifuge tube and stored at -20°C until required.

2.7.4.2 Purification of PCR products using the recovery of DNA from agarose gel

Alternatives versions of this method were explained in detail in Section 2.5.

2.8 Polyacrylamide Gel Elecrophoresis

Gels were prepared 1-24 hours prior to use.

2.8.1 Preparing the plates

Two glass plates (Invitrogen, 33.5 cm × 39 cm and 33.5 cm × 42 cm) were washed with diluted Decon 90 (approximately 1 in 5 dilution) and then washed with distilled water. The plates were then dried with paper towels, washed and cleaned with pure ethanol and then left to dry on the bench at room temperature. The shorter plate was coated with Gel Slick, dimethyldichlorosilane solution, (Flowgen) and left to dry on the bench.
The plastic spacers (vinyl 0.36 mm spacers) were then applied between the two glass plates at the edges. The glass plates were then inserted into an S2 casting boot (Invitrogen).

2.8.2 Pouring the gel

To make a 6% polyacrylamide gel, Sequagel solutions (National Diagnostics) were used. For this purpose 14.4 ml Sequagel concentrate (237.5 $g.l^{-1}$ acrylamide, 12.5 $g.l^{-1}$ methylene bisacrylamide, 8.3 M urea), 39.6 ml Sequagel diluent (8.3 M urea) and 6 ml of Sequagel buffer (50% urea (8.3 M) in 1 M tris-borate, 20 mM EDTA buffer) were added and mixed in a glass beaker. To this mixture, 72 µl of TEMED (N, N, N′ N′-tetramethylenediamine, Sigma) and 336 µl of freshly-prepared 10% (w/v) ammonium persulphate were added and mixed well.

To pour the gel solution between the two glass plates at about 45–50° angle and to avoid air bubbles forming a 60 ml syringe was used. Then a 0.36 mm (thickness) and 28 cm (width) Mylar shark's tooth comb (5.7 mm point to point) was inserted between the two glass plates at the top of the gel (flat edge first) and held in place with bull dog clips clamped at the top of the glass plates. To polymerise the gel it was left for at least 1 hour at room temperature prior to use.

2.8.3 Gel electrophoresis

The glass plates were removed from the rubber casting boot and the comb removed from top of the gel. The comb was then inserted the other way up between the two glass plates so that the teeth protruded about 2–3 mm into the gel. The gel apparatus (Invitrogen, model S2) was assembled and 1 × TBE buffer (89 mM Tris- base, 89 mM boric acid, 2 mM EDTA) was added to the top and bottom chambers. The gel (without loading any sample) was pre-run until the temperature of the gel rose above 50°C and then samples were loaded and the electrophoresis ran at a constant current of 50 mA.

2.8.4 Post-electrophoresis

The plates were removed from the gel running-apparatus and the spacers and comb removed as well. The two plates were then separated from each other, using a thin spatula to leave the gel sticking to the longer plate which had not been silanised with Gel Slick.
To remove the urea, the gel was soaked in 5% (v/v) acetic acid, 15% (v/v) methanol for 5-10 minutes. Then the gel was transferred from the glass plate onto Whatman 3MM paper and dried at 80°C under vacuum for 2 hours on a gel drier (BioRad, model 583).

2.8.5 Autoradiography

The dried gel was placed in an X-ray cassette. In a dark room, a sheet of radiography film (Kodak XAR) was placed over the dried gel within the cassette. Then the cassette was put at room temperature for an appropriate time (1-7 days) and then processed manually. Film was developed by immersion for 5 minutes in Kodak developer (developer LX24), a rinse in stop solution (1 in 200 dilution of glacial acetic acid), and 5-10 minutes in Kodak fixer (containing fixer AL4 and fixer hardener HX40). The film was then washed extensively in running water at 20°C and hung up to dry at room temperature.

2.9 Preparation of labelled DNA size marker

Two hundred nanograms of λ/*Hin*dIII in a 20 µl total volume was added to a microfuge tube. Then 80 µCi of [α-^{32}P] dATP (NEN, 10.0 mCi.ml^{-1}, 111 TBq/mmol) and 1 µl of 5 mM dCTP, dGTP and dTTP were added to the mixture. After mixing, 1 unit of the Klenow fragment (USB Corporation, 1 unit.µl$^{-1)}$ was added and the mixture incubated for 15 minutes at 30°C. To end the reaction, 1 µl 0.5 M EDTA was added.

2.10 DNA sequencing

2.10.1 Automated sequencing using Big Dye Terminator method

DNA fragments (PCR products) were recovered from agarose gel using electroelution onto dialysis membrane (Section 2.5.3). After quantitation of the recovered PCR DNA or plasmid DNA, 10–20 ng/kb of the recovered PCR product or 1 μg of the plasmid DNA/10kb was mixed with 1 μl (3.3 pmole. μl^{-1}) of the desired primer in a total volume of 10 μl. This mixture was amplified using the following cycling conditions:

- Denaturation, 96°C for 10 seconds
- Annealing, about 10°C lower than the optimised annealing temperature in a standard PCR
- Extension, 60°C for 4 minutes
- Number of cycles, 25–30

The sequencing reaction was then transferred to a fresh microcentrifuge tube and the precipitation was carried out according the protocol described in Section 2.6 and the precipitated DNA was submitted to PNACL (The Protein and Nucleic Acid Chemistry Laboratory at the University of Leicester) for electrophoresis. PNACL analysed the products using an ABI-PRISM 377 automated sequencer (Perkin-Elmer).

2.10.2 Direct DNA sequencing of PCR products (manual sequencing)

The method used was modified from Cohen 1999, and Amersham Life Sciences, T7 sequenase version 2.0 DNA sequencing kit protocol using the chain termination method, described by Sanger *et al.*, 1977.

2.10.2.1 PCR reaction and purification of product

The purification of the PCR product was carried out using the QIAquick PCR purification kit according the protocol described in Section 2.7. To achieve good results at least 0.1 pmol of purified product was used (*i.e.* ~60 ng of a 1-kb fragment). The yield of PCR product was determined by quantitative analysis on an agarose gel, using known amounts of DNA markers as a reference and AlphaImager v.4.0 or v.5.5 software.

2.10.2.2 Ethanol precipitation of DNA sample

In cases where the DNA was not concentrated enough (less than 0.1 pmol of DNA in 7 µl 0.1× TE), ethanol precipitation was carried out according the protocol described in Section 2.6.

2.10.2.3 Preparing double-stranded PCR product and setting up dideoxy sequencing reaction.

For each sequencing reaction, a single annealing reaction was set up in a total reaction volume of 10 µl as follows. At least 0.1 pmol of purified PCR product was adjusted to 7 µl with 0.1× TE buffer, pH 8.0. The template was denatured by heating for 2 minutes at 95°C and quickly cooled for 1 minute in a dry ice/ethanol bath. Then 1µl sequencing primer (0.5–1 pmol) and 2 µl of 5× annealing buffer (200 mM Tris-HCl pH 7.5, 100 mM $MgCl_2$, 250 mM NaCl) were added to the frozen DNA samples (in all reactions about a 5 fold molar excess of primer over template was used). The tubes were then thawed and the contents mixed briefly by pipetting. During annealing, four microcentrifuge tubes were individually filled with 2.5 µl of each termination mixture [ddT (80 µM dGTP, 80 µM dATP, 80 µM dCTP, 80 µM dTTP, 8 µM ddTTP and 50 mM NaCl), ddA (as ddT,but with 8 µM ddATP instead of ddTTP), ddC (as ddT, but with 8 µM ddCTP instead of ddTTP) and ddG (as ddT, but with 8 µM ddGTP instead of ddTTP)]. These mixtures were then pre-warmed at 37°C and the labelling mixture (7.5 µM dTTP, 7.5 µM dCTP and 7.5 µM dGTP) was diluted 5-fold with water to its working concentration.

After cooling the annealing mixture, it was centrifuged briefly at 13,000 rpm in Micro Centaur centrifuge and chilled on ice. Then to the 10 µl ice-cooled annealed mixture, 1 µl of DDT (0.1M), 2 µl of diluted labelling mixture, 0.5 µl of [α-^{35}S] dATP (NEN, 12.5 mCi.ml^{-1}, 46.25 TBq/mmol) and 2 µl of diluted (8-fold dilution with dilution buffer supplied in kit) Sequenase polymerase were added. The reactions were mixed and incubated on the bench at room temperature for 2-5 minutes.

To terminate the reactions, 3.5 µl of the above labelling reaction was transferred to each of the pre-warmed termination mixes. These reactions were then incubated at 37°C for 5 minutes. The termination reactions were then stopped by adding 4 µl of stop solution (95% (v/v) formamide, 20 mM EDTA, 0.05% (w/v) bromophenol blue and 0.05% (w/v) xylene cyanol

FF). These final reactions were then mixed and stored on ice or –20°C until ready to analyse by gel electrophoresis.

2.10.2.4 Polyacrylamide denaturing gel electrophoresis

The completed sequencing reactions were heated at 75–80°C for 2 minutes and then snap cooled on ice. Then 3 µl of each sequencing reaction was loaded in each well. Sometimes in order to achieve longer sequencing reads, a second loading was carried out (when the xylene cyanol FF of the first samples travelled half of the length of the gel). The gel electrophoresis and autoradiography were carried as explained in Sections 2.8.1 to 2.8.5.

2.11 Conformation Sensitive Gel Electrophoresis (CSGE)

Conformation sensitive gel electrophoresis method (CSGE) has been developed for scanning PCR products for the presence of single base and larger base mismatches in DNA. For carrying out this method to find any variation, the published protocol with a few changes and modifications was used (Korkko *et al.*, 1998).

- As cross linker, instead of BAP (1, 4-Bis acryolyl piperazine), piperazine diacrylamide, PDA (Biorad) was used at the same concentration (Table 2.2).
- Two concentrations, 20 and 15% of polyacrylamide gel were used (based on the size of fragments).
- Plates, spacers and comb of sequencing procedures that was described in Section 2.8 were used. Therefore the thickness of the gel was 0.36 mm.
- During experiments it was found that the suggested glycerol-based loading dye results in fuzziness of bands, therefore a 6× sucrose based loading dye (xylene cyanol FF 0.25% (w/v), bromophenol blue 0.25% (w/v), sucrose 40% (w/v)) was used.
- It was found that using constant voltage (1300-1500 V) instead of constant power increases the sharpness of bands.

Reagents	Final Conc.	Storage condition
Polyacrylamide: piperazine diacrylamide (PDA) (99:1) 40% (v/v)	Polyacrylamide 15 and 20% (w/v)	4°C
ethylene glycol	10% (v/v)	Room temperature
formamide	15% (v/v)	-20°C (stored in single-use aliquots, 10ml)
20 × TTE (Tris-HCl 44.4 mM pH 9.0, taurine 14.25 mM, EDTA 0.1 mM)	0.5 ×	Room temperature
ammonium persulphate 10% (w/v)	0.1% (v/v)	4°C
TEMED	0.07% (v/v), 42 µl for 60 ml gel	Room temperature

Table 2.2. Reagents required for making denaturing CSGE gels

Following electrophoresis the gel was stained by pouring 1 $\mu g.ml^{-1}$ ethidium bromide on to the centre of the gel for 2 minutes. Bands were then visualised by hand-held UV lamp in a darkened room. The relevant portions of the gel were cut out by scalpel and transferred on to a piece of Whatman 3 MM paper and stained again with ethidium bromide for 10 minutes. After destaining the gel pieces in distilled water for 10 minutes, the pieces of gel were released onto the surface of a transilluminator by wetting the Whatman 3 MM paper surface with water. Finally DNA bands were visualised by UV illumination and photographed. Gel images were stored as bitmaps in TIFF format.

2.12 Culture, Storage and Manipulation of *Escherichia coli* (*E.coli*)

2.12.1 Media and culture

Liquid cultures of *E. coli* bacteria were grown in the rich medium, Luria-Bertani broth (LUB) (10 $g.l^{-1}$ tryptone, 5 $g.l^{-1}$ yeast extract, 10 $g.l^{-1}$ NaCl, pH 7.5, autoclaved) at 37°C on a G10 Gyratory shaker (New Brunswick Scientific) at 230–300 rpm. Liquid cultures of bacteria in LUB were grown in either glass culture tubes or flasks filled to more than 20% of their capacity. Bacteria were inoculated as colonies using sterile wooden tooth picks, or by dilution of existing liquid culture.

For solid cultures, *E. coli* bacteria were streaked or plated on LB agar plates (LUA) (15 $g.l^{-1}$ agar added to LUB medium and autoclaved). After slow melting of LUA in a microwave oven, molten LUA was poured into 100-mm plates and left at room temperature to set. Agar plates were either streaked, using an inoculating loop or by adding a small volume of liquid bacteria culture (usually 200 µl) on to the plates and spread evenly over the surface using a glass spreader. All bacteria were cultured at 37°C for 12-24 hours. Density of bacteria in liquid cultures was determined and monitored using disposable 1 cm cuvettes at 600 nm.

2.12.2 Storage

Plate cultures were kept upside down at 4°C and when long-term storage was required were sealed with parafilm. For long-term storage, *E. coli* were transferred to a medium supplemented with 1 × HMFM (Hogness Modified Freezing Medium: 3.6 mM K_2HPO_4, 1.3 mM KH_2PO_4, 2 mM sodium citrate, 1 mM $MgSO_4$, 4.4% v/v glycerol) in cryotubes at -80°C.

2.12.3 *E. coli* strains

E. coli DH5α and DH10B strains (Invitrogen) were used. The genotypes of these strains are:

DH5α: φ80d*lacZ*ΔM15, rec*A1*, *endA1*, *gyrA96*, *thi-1*, *hsdR17*(r_k^-,m_k^+), *supE44*, *relA1*, *deoR*, Δ(*lacZYA-argF*) U169.

DH10B: F$^-$ *mcrA*, Δ (*mrr*$^-$*hsd*RMS$^-$*mcr*BC), φ80*lacZ*ΔM15, Δ*lacX*74, *deo*R, *rec*A1, *end*A1, *ara*Δ139, Δ (*ara, leu*) 7697 *gal*U, *gal*K λ$^-$, *rps*L, *nup*G λ$^-$, *ton*A.

2.12.4 Antibiotics

All antibiotics were prepared and stored at a 100–1000× concentration at -20°C. The details of each antibiotic have been summarised in Table 2.3.

2.12.5 Preparation and transformation of chemically competent *E. coli*

This method for transformation of chemically competent *E. coli* cells, using calcium chloride procedure is based on the method of Hutchinson and Halvorson, 1980.

2.12.5.1 Preparation of chemical competent cells

A 2.5–3 ml culture of *E. coli* with appropriate antibiotic was grown overnight at 37°C. To inoculate in a larger volume, 1.5 ml of this overnight culture was added to 75 ml of a fresh and pre-warmed LUB medium and allowed to grow to an optical density of 0.36–0.44 at 600 nm, then cooled on ice. The cells were then harvested by centrifugation at 3000 rpm for 5 minutes at 4°C and then resuspended in 20 ml of cold 50 mM $CaCl_2$, and incubated on ice for 15 minutes. The cells were then harvested as before and resuspended in 5 ml of cold 50 mM $CaCl_2$, 5% glycerol. Finally, the cells were split into 200-µl aliquots and snap frozen in capped microcentrifuge tubes in a dry ice/IMS (Industrial Methylated Spirit) bath. The aliquots of competent cells were stored at -70°C until use.

2.12.5.2 Chemical transformation of competent cells

A 200-µl aliquot of frozen competent cells was thawed on ice, and 10 ng DNA in 100 µl of buffer (10 mM Tris-HCl, pH 7.4, 10 mM $MgCl_2$, 10 mM $CaCl_2$) was added to this aliquot. The diluted DNA and thawed competent cells were mixed and placed on ice for 25 minutes. The mixture was then heated (heat shock) at 37°C for 1.5 minutes and left on the bench at

Antibiotics	Solvent	Stock Conc.	Working Conc.	Storage
Ampicillin	Water	5 $mg.ml^{-1}$	50-200 $\mu g.ml^{-1}$	-20°C
Kanamycin	Water	5 $mg.ml^{-1}$	50 $\mu g.ml^{-1}$	-20°C
Chloramphenicol	50% (v/v) Ethanol/Water	3 $mg.ml^{-1}$	30 $\mu g.ml^{-1}$	-20°C Protected from light
Gentamicin	Water	10 $mg.ml^{-1}$	10 $\mu g.ml^{-1}$	4°C

Table 2.3. Antibiotics

Summary of all antibiotics used. Stock concentration refers to the storage concentration of each antibiotic and working concentration refers to concentration used in culture media. All above antibiotics were filter sterilised (0.45 μm) and stored at the indicated temperature.

room temperature for 10 minutes. To help in the expression of the antibiotic resistance of the cells, 1 ml of LUB was added to the cells and they were incubated at 37°C for 1 hour. Different volumes of transformed cells (2 µl, 20 µl, and 200 µl) were spread onto selective plates and incubated at 37°C overnight.

2.12.6 Preparation and transformation of electro-competent *E. coli*

This method is based on the methods of Chang and Lee, 1994 and Siguret *et al.*, 1994 and was used for transformations that required a high efficiency such as in the construction of cDNA clone libraries.

2.12.6.1 Preparation of electro-competent cells

A single colony of *E. coli* was inoculated into 3 ml of LUB broth with antibiotic if appropriate and grown overnight at 37°C with vigorous shaking. To obtain a larger volume of culture, 1 ml of the overnight culture was inoculated into 100 ml of fresh pre-warmed LUB broth. The culture was then grown at 37°C with vigorous shaking to a cell density of approximately 0.6 to 0.65 at OD_{600} and then cooled on ice for 30 minutes.

The cells were then harvested by centrifugation at 4°C, 4000 rpm for 15 minutes and then resuspended and washed twice in 50 ml of ice-cold 10% glycerol at 4°C, 4000 rpm for 15 minutes. The cells were finally resuspended in 0.2 ml ice-cold GYT (10% glycerol, 0.125% yeast extract and 0.25% tryptone) and then divided into 40-µl aliquots and frozen in microcentrifuge tubes in a dry ice/IMS bath and stored at -70°C until use.

2.12.6.2 Electroporation procedure

Purification of the plasmid DNA was carried out using QIAGEN Plasmid Midi Kit. The DNA to be transformed (approximately 50 ng) was in 1 to 5 µl of a low-conductivity medium such as 1 × TE or distilled water to prevent arcing, which could damage the pulse generator. The DNA was mixed with 40 µl of electro-competent cells on ice and transferred to a chilled cuvette. A pulse generator (BioRad Genepulser) comprising a pulse controller and sample carriage was used and the capacitance and resistance were set at 25 µF and 100 Ω respectively. A test pulse at 2.5 kV was carried out without inserting a cuvette. The time

constant was checked to ensure that it was about 2.2–2.4 ms. On ice, 50 ng of plasmid DNA was added and mixed with 40 µl of cells and transferred to a pre-cooled cuvette and placed in the sample carriage of the electroporator and a pulse delivered (2.5 kV, 25 µF). The cuvette was then removed and 1 ml of SOC medium (2% bactotryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 20 mM $MgCl_2$, 20 mM Mg_2SO_4, 20 mM glucose (Hanahan, 1983)) was added, mixed well and transferred to into a 15 ml polypropylene tube to allow the cells to recover. The time constant was checked after each pulse to ensure that it was within range 2.2–2.4 ms. The electroporated cells in SOC medium were then shaken at 37°C for 60 minutes. Aliquots of different volumes were plated out onto the appropriate selective plates and incubated overnight at 37°C.

2.12.7 Selection of transformants

To select transformant cells harbouring a plasmid, appropriate antibiotics were added to LUA plates. When vectors capable of α-complementation were used, indication of plasmids harbouring inserts was carried out using the blue/white colour screening system (Horwitz *et al.*, 1964; Ullmann *et al.*, 1967). For this purpose X-gal (5-bromo-4-chloro-3-indolyl-β-d-galactosidase, 40 µg/ml-1 in N, N dimethylformamide) and IPTG (Isopropyl thio-β-D-galactosidase, 0.1 mM) were added to LUA plates. Colonies containing non-recombinant plasmid gave blue colonies and colonies containing recombinant plasmid produced white colonies.

2.12.8 Isolation of plasmid DNA from *E. coli*

2.12.8.1 BAC and PAC mini- preps using Alkaline-SDS lysis

The method used to isolate plasmid DNA is modified from the method of Ish-Horwicz and Burke, 1981.

A small culture of bacteria in LUB (3 ml) was grown overnight with the appropriate antibiotic for plasmid selection. To harvest the cells, 1.5 ml of the overnight culture was centrifuged at 13,000 rpm for 1 minute. The broth was discarded and the tubes were centrifuged again briefly. The remaining broth was pipetted off. The pellet was resuspended in 200 µl of solution I (50 mM glucose, 25 mM Tris-HCl pH 8.0, 10 mM EDTA) and incubated at room temperature for 5 minutes. Then, 200 µl of solution II (0.2 N NaOH, 1% (w/v) SDS) was

added and the tube mixed gently by inversion several times. The tube was then incubated on ice for 5 minutes. Next, 200 µl of solution III (5 M potassium acetate, pH 5.5) was added and the tube was again gently mixed by inversion and then placed on ice for a further 5 minutes.

The tubes were then centrifuged for 5 minutes at 4°C and 800 µl of the clear supernate transferred to a fresh microcentrifuge tube and centrifuged again to ensure total removal of any white precipitate. The supernate was then transferred into a tube containing 800 µl of ice-cold propan-2-ol (iso-propanol), mixed well and placed on ice for at least 5 minutes. Next, the DNA was pelleted by centrifuging for 15 minutes at 4°C and the supernate again discarded. The pellet was washed in 500 µl 70% (v/v) ethanol, centrifuged and the supernate discarded. The tube was then centrifuged again and the last of the ethanol removed with a pipette. The pellet was air dried for about 15 minutes and then resuspended in 40 µl of 1× TE and placed on ice until dissolved. Because BAC and PAC DNA are high molecular weight and will not dissolved as readily as plasmid DNA, the tube was flicked from time to time to mix the contents to help dissolve the DNA. Plasmid DNA was stored at -20°C. To analyse the product by gel electrophoresis 10 µl of the sample was digested with restriction enzyme in the presence of 2 $\mu g.ml^{-1}$ RNase A in a 20-µl reaction volume.

2.12.8.2 BAC and PAC preps

The method used to isolate plasmid DNA for end sequencing is based on a method modified from the Genome Sequencing Centre of Washington University (http://genome.wustl.edu/tools/protocols/lib_core/BAC.pdf). This method is based on a modified alkaline lysis method and has been scaled up to 50 ml of LUB broth overnight culture.

2.12.8.3 Isolation of plasmid DNA using Qiagen kits

The method used to isolate plasmid DNA from *E.coli* for further analysis was carried out according to manufacturer's protocol.

2.13 Culture, Storage and Handling of yeast (*Saccharomyces cerevisiae*)

2.13.1 Storage of yeast

Yeast plates were kept at 4°C and sealed with parafilm if medium-term storage was required. For long-term storage of stock liquid yeast cultures, they were frozen at -80°C by adding an equal volume of glycerol freezing solution (glycerol 65% (v/v), 25 mM Tris-HCl pH 8.0 and 100 mM $MgSO_4$).

2.13.2 Media

S. cerevisiae containing no plasmids were grown in YPD medium. YPD medium (Yeast extract, Peptone, Dextrose) is an autoclaved blend of 20 $g.l^{-1}$ Difco Peptone, 10 $g.l^{-1}$ yeast extract, pH 6.5. Prior to use, dextrose (glucose) to a final concentration of 2% (w/v) and adenine hemisulfate (to a final concentration of 0.003% (w/v)) were added to the medium. Yeast containing no plasmids were plated or streaked out onto YPD agar (YPDA) medium plates (20$g.l^{-1}$ of agar added to YPD medium).

SD medium (Synthetic dextrose) was used to grow yeast containing plasmid. SD medium was made by adding a minimal SD base (0.67 % (w/v) Yeast nitrogen base (YNB) without amino acids, 2% (w/v) agar (for plates only)) to a stock of Dropout (DO) solution (a solution containing specific amounts of amino acids). Dropout solution does not contain leucine, tryptophan, histidine and adenine. These amino acids and adenine can then be added, or not, as required. Although uracil is not an amino acid, it too was always added to DO solution. SD medium was prepared as described below. To prepare 1 litre of SD medium, 100 ml of 10 × dropout solution, dextrose (glucose) to a final concentration of 2% (w/v) and any required amino acid were added to the required volume of SD base solution up to 1 litre. Any required amino acids were then also added. Solutions of these required amino acids were prepared separately allowing them to be omitted if necessary. Tables 2.4 and 2.5 show the composition of the dropout solution and amount of each individual amino acid solution respectively. All amino acids powders were dissolved in sterile water, autoclaved and for long term use stored at -20°C. Dropout and amino acids solutions were stored at 4°C when in use.

Amino acid	Concentration $mg.l^{-1}$
L-Isoleucine	300
L-Valine	1500
L-Arginine HCL	200
L-Lysine HCL	300
L-Methionine	200
L-Threonine	2000
L-Thyrosine	300
L-Phenylalanine	500
L-Uracil	200

Table 2.4. 10 × dropout solution

The amino acids constituents of 10× dropout solution. It should be considered that amino acids Serine, Aspartic acid and Glutamic acid were not added to 10 × dropout solution because they cause the medium become too acidic and the yeast strain used is prototrophic for these amino acids.

Amino acid	Concentration $g.l^{-1}$
200× Leucine L-Leucine	20
200× Tryptophan L-Tryptophan	4
200× Histidine L-Histidine HCl monohydrate	4
200× Adenine L-Adenine hemisulphate	4

Table 2.5. Individual amino acids

This table lists the details of the amino acids that are prepared separately to enable the appropriate amino acid to be omitted if required.

SD agar plates were prepared by adding 20g.l^{-1} of agar to SD medium and then autoclaving. These SD plates were used to plate or streak out the yeasts containing plasmids.

2.13.3 Strains

Throughout the yeast two-hybrid study, the yeast strain *Saccharomyces cerevisiae* AH109 was used (James *et al.*, 1996). Biological properties and details of this strain are as follows.

Strain AH109

Complete Genotype:	MATa, trp1-901, leu2-3, 112, ura3-52, his3-200, gal4Δ, gal80Δ, LYS::GAL1$_{UAS}$-GAL1TATA-HIS3, GAL2$_{UAS}$-GAL2$_{TATA}$-ADE2, URA3::MEL1$_{UAS}$-MEL1$_{TATA}$-lacZ
Reporter Genes:	HIS3, ADE2, lacZ, MEL1
Transformation Markers:	trp1, leu2

2.13.4 Preparing yeast competent cells

Yeast cells were transformed using the lithium acetate method (Li Ac) according to a modified version of the procedure of Gietz *et al.* (1992).

A 50 ml culture of AH109 cells was grown overnight in YPD medium or appropriate SD medium. The culture was started using 1 ml YPD into which several 2-3 week-old colonies of AH109, 2-3 mm in diameter, had been resuspended. The culture was then grown in an incubator with shaking (250 rpm) at 30°C overnight. If the SD media was used, the duration of incubation was extended to 20–24 hours. After overnight incubation to ensure the culture had reached the stationary phase (OD600>1.5), the density of the cell culture was measured. To produce a cell density of 0.2-0.3 at 600 nm, enough volume of the 50 ml culture was then transferred to a 300 ml YPD or appropriate SD medium. To obtain a cell density of 0.5 ± 0.1 at 600 nm, the 300 ml culture was then grown with shaking at 30°C for about 3 hours. Yeast cells were then harvested by centrifugation at 1000 × g for 5 minutes at room temperature. The pellet was then resuspended in 1.5 ml of freshly prepared sterile 1 × TE/LiAc (10 mM Tris-HCl, 1mM EDTA, 0.1 M Li acetate, pH 7.5).

2.13.5 Transformation of yeast competent cells

The yeast competent cells were transformed using two different scales, small and large. Table 2.6 shows each step and the amount of each component used for each scale of transformation.

2.13.6 Selection for transformants

After transformation of the yeast cells, they were plated on appropriate SD medium. For small-scale and all control transformations 100 µl was plated on each 90 mm plate. For library screening, the large scale transformation was used and the following plates were used to spread yeast cells.

- 100 µl of 1:1000, 1:100 and 1:10 dilutions spread on SD/−Trp/−Leu 90 mm plates as transformation efficiency controls.

- 1 µl (diluted in 100 µl dH_2O or 1 × TE) spread on SD/−Trp and SD/−Leu 90 mm plates to check the transformation efficiency of each bait plasmid construct.

- The remaining suspension was spread on SD/−Leu/−Trp/−His or SD/−Leu/−trp/−His/−Ade 140 mm plates (approximately 250 µl on each plate). All plates were incubated inverted for up to 10 days at 30°C.

2.13.7 Isolation of the plasmid DNA from yeast cells

2.13.7.1 Isolation of plasmid DNA from yeast cells using lyticase and phenol/chloroform extraction

This method has been modified from Ling *et al.*, (1995).

A 5-ml culture was grown for 2 days in the appropriate SD medium to select for the plasmid of interest. The yeast cells were then harvested by centrifuging at 1000 × g for 10 minutes. The supernate was carefully discarded and the pellet resuspended in the residual liquid (about 50 µl) and transferred to a 1.5 ml microfuge tube. Next,10 µl lyticase (Sigma, 5 units. μl^{-1}) solution was added and the contents of tubes were thoroughly mixed by vortexing and then incubated at 37°C overnight with shaking at 200 rpm.

Protocol	Small scale	Large scale
The following amount was added to an appropriate size tube	1.5-ml microfuge tubes	50-ml Falcon tubes
• DNA-BD/bait[a]	0.1 µg	20-100 µg
• AD/library[a]	0.1 µg	10-50 µg
• Herring testes carrier DNA	0.1mg	2 mg
Yeast competent cells were added to each tube and mixed well	0.1 ml	1 ml
Sterile PEG/LiAc solution (10 mM tris-HCl, 1 mM EDTA, 0.1 M LiAc, pH 7.5, 40% (w/v) PEG 4000) was added to each tube and vortexed at high speed to mix	0.6 ml	6ml
The mixture was incubated at 30°C for 30 minutes with shaking (200 rpm)		
DMSO was added to 10% and inverted gently to mix	70 µl	700 µl
The cells were heat shocked in a 42°C water bath for 15 minutes and the cells were swirled occasionally to mix		The cells were swirled occasionally to mix
The cells were chilled on ice for at least 1-2 minutes		
Cells were pelleted by centrifugation at room temperature	5 sec at 14000 rpm	5 min at 1000 × g
The supernate was removed		
Cells were resuspended in 1 × TE[b]	0.5 ml	1.0 or 10 ml[c]

Table 2.6. Yeast transformation protocol

This table shows the details of the yeast transformation method on small and large scales with the amounts to be used for each.

a: For sequential transformation, the DNA-BD vector construct or the AD vector construct (not both) was added.
b: When the high stringency selection method was used, the yeast cells were resuspended in YPD medium. This medium will aid the yeast in recovery from the heat shock of transformation, but will not adversely affect screening.
c: Used 1.0 ml for simultaneous co-transformation and 10 ml used for the second transformation in sequential transformation protocol.

After the overnight incubation, 10 µl of 20% (w/v) SDS was added and the mixture was mixed thoroughly by vortexing vigorously at high speed for 1 minute. The mixture was then put through one freeze and thaw cycle at -20°C. After thawing, the solution was mixed again by vortexing. The volume of the solution was then increased to 200 µl by adding 1 × TE. An equal volume of phenol:chloroform:isoamyl alcohol (25:24:1) was added and the sample was vortexed at high speed for 5 minutes. The sample was then centrifuged for 10 minutes at 13,000 rpm and about 200 µl of aqueous phase transferred to a clean tube. Then, 8 µl of 10 M ammonium acetate and 500 µl of absolute ethanol were added and the sample placed at -70°C for 1 hour.

The sample was then centrifuged as previously for 10 minutes and the supernate poured off. The pellet was washed twice in 70% (v/v) ethanol, centrifuged and supernate discarded. The pellet was air dried for about 5 minutes and then resuspended in 20 µl of sterile water. The DNA was then stored at -20°C. To transform into *E. coli,* using the electroporation method (Section 2.12.6.2), 4 µl of DNA was used with 50 µl of electro-competent cells.

2.13.7.2 Yeast plasmid miniprep using glass bead

A 5-ml culture was grown for 2 days in appropriate SD medium to select the desired plasmid. The yeast cells were then harvested by centrifuging at 1000 × g for 10 minutes. The supernate was carefully discarded and the pellet resuspended in the residual liquid (about 50 µl) and transferred to a 1.5 ml microfuge tube.

Next, 200 µl plasmid rescue solution (2% (v/v) Triton X-100, 1% (w/v) SDS, 0.1 M NaCl, 10mM Tris-HCl pH 8, 1 mM EDTA), 100 µl phenol:chloroform:isoamyl alcohol (25:24:1) and 0.3 g glass beads (425-600 µm, Sigma) were added. The content of tube was vortexed at high speed for 2 minutes and then centrifuged for 5 minutes at 13,000. About 200 µl of aqueous (upper) phase was transferred to a clean tube and 20 µl of 3 M sodium acetate and 440 µl of absolute ethanol were added and then centrifuged as previously for 20 minutes and the supernate poured off. The pellet was washed twice in 70% (v/v) ethanol, centrifuged and supernate discarded. The pellet was air dried for about 5 minutes and then resuspended in 20 µl of sterile water. The DNA was then stored at -20°C. In order to transform into *E.coli,* using electeroporation method, 4 µl of DNA was used with 50 µl of electro-competent cells.

2.14 Isolation of Mononuclear Lymphocytes and Monocytes from Whole Fresh Blood

This method uses Ficoll-Paque (Lymphoprep, Pharmacia Biotech) and was used to isolate mononuclear cells from whole fresh blood and was carried out with modifications according to the manufacturer's protocol.

In a 50-ml polypropylene, screw-cap centrifuge tube the anticoagulant-treated (heparinised) fresh whole blood was diluted with an equal volume of the sterile, cold 1× PBS, phosphate buffered saline (130 mM sodium chloride, 7 mM sodium dihydrogen orthophosphate, 3 mM di-sodium hydrogen orthophosphate) and mixed well by pipetting. The diluted blood was layered carefully to another 50-ml polypropylene tube containing an equal volume of Ficoll-Paque (equal volume to the blood and 1× PBS). To avoid mixing the Ficoll-Paque and the diluted blood sample, the tube containing the Ficoll-Paque was held in a 45° position and the blood sample added very slowly. The sample was then centrifuged at 800× g with no brake for 30 minutes at 18–20°C. After centrifugation the interface layer which contained mononuclear cells was transferred to a clean 50-ml polypropylene tube using a Pasteur pipette. The mononuclear cells were then centrifuged (washed) twice with 20 ml sterile and cold 1× PBS and 2% FCS (foetal calf serum) and centrifuged at 400× g for 10 minutes. Total RNA isolation from mononuclear cells with no further treatment was carried out at this step

For further culture and treatment of mononuclear cells in a suspension medium, the washed mononuclear cells were suspended in an appropriate volume of RPMI 1640 medium (L-glutamine 2mM, penicillin 200U.ml^{-1}, streptomycin 200 μg.ml^{-1}, 10% FCS and 1× non-essential amino acids (Biochrom) and divided in aliquots (2 ml) for further treatment.

Monocyte cells stick to the surface of polycarbonate and grow as monolayers surface, so this biological property of monocytes was used to isolate them from mononuclear lymphocytes. After resuspending the whole mononuclear cells in RPMI 1640 with 10% FCS they were cultured in a 9-cm polycarbonate dish and incubated at 37°C with 5% CO_2 overnight. On the following day, the liquid medium containing mononuclear lymphocytes was discarded. Then 5 ml RPMI 1640/FCS 10% was added to the plate and monocytes were scraped and removed using a cell scraper.

To study the effects of different substances on gene expression, 100 ng.ml^{-1} lipopolysaccharide, LPS (isolated from *Salmonella minnesota*, Sigma) and 100 U.ml^{-1} tumour

necrosis factor alpha, TNFα (Sigma) were added to samples. Based on the nature and aim of the subsequent experiments, the samples were incubated in a culture incubator for 1 to 24 hours at 37°C with 5% CO_2.

2.15 Genomic DNA Extraction from Fresh Whole Blood

2.15.1 Genomic DNA extraction from frozen whole blood

Blood samples (10 ml) in 10-ml EDTA tubes were thawed and poured into 50-ml polypropylene, screw cap centrifuge tubes. Ice-cold distilled water was added to a final volume of 50 ml and the sample mixed well by shaking.

The tubes were then centrifuged at 3,000 rpm for 10 minutes at 4°C. The supernate was discarded and the pellet resuspended in 25 ml of ice-cold sucrose lysis buffer (0.32 M sucrose, 5 mM $MgCl_2$, 1% (v/v) Triton X-100, 10 mM Tris-HCl pH 7.5) and allowed to stand on ice for up to 30 minutes. The samples were then centrifuged at 3,000 rpm for 10 minutes at 4°C and the supernate was poured off, the pellet resuspended in 3 ml CVS buffer (10 mM Tris-HCl pH 10.5, EDTA 1 mM, NaCl 15 mM, SDS 0.05% (w/v)) and incubated at 60°C for 2-3 hours.

After incubation, 2 ml of 5 M NaCl was added to the samples which were centrifuged at 3,500 rpm for 10 minutes at 4°C with the brake off. The supernate was then collected in a 15 ml tube, filled with absolute ethanol at room temperature and inverted 3 times to mix and precipitate the DNA. The DNA was then lifted out with a blue pipette tip and placed in a microfuge tube containing 1 ml 70% ethanol. The samples were centrifuged and ethanol carefully removed. The pellet was then air dried and redissolved overnight in 1 ml of 1× TE and stored at -20°C.

2.15.2 Isolation of genomic DNA from whole blood and animal tissues using Promega Genomic DNA Purification kit

This kit was used for purification of 300 µl to 3 ml fresh or frozen whole blood and 10–20 mg fresh or thawed animal tissues. Purification of DNA was carried out according to the manufacturer's protocol (Wizard Genomic DNA Purification Kit).

2.16 Extraction of Total RNA from Mammalian Tissues or Cells

2.16.1 Total RNA extraction using RNAzol B kit

This method uses RNAzol B Kit (AMS Biotechnology (Europe) Ltd) and was used to purify total RNA from animal tissues (more than 200 mg animal tissues) and carried out according to the manufacturer's protocol.

The tissue sample was homogenised using a portable homogeniser (Status × 120, Philip Harris Scientific) in RNAzol B (2 ml per 0.1 g tissue). To every 2 ml RNAzol B, 0.2 ml of chloroform was added and the sample was shaken vigorously for 15 seconds. Next, the sample was chilled on ice for 5 minutes and then centrifuged at 10,000 rpm for 15 minutes at 4°C. The aqueous phase (upper layer) was transferred to a clean tube and an equal volume of isopropanol was added. (In tissues like liver that are rich in glycogen, before adding the isopropanol, the pellet was washed twice with 4 M lithium chloride by vortexing and subsequent centrifuging as previously for 8 minutes). The sample was incubated on ice for 15 minutes and then centrifuged as previously for 15 minutes. The supernate was discarded and the pellet washed in 1 ml of 75% (v/v) ethanol by vortexing with subsequent centrifugation as previously for 8 minutes. The pellet was then resuspended in dH_2O or 1 mM EDTA, pH 7.0 (both solutions were pre-treated with diethylpyrocarbonate, DEPC). To make DEPC-treated water or solution, DEPC was added to ultra pure water or solutions to a final concentration of 0.1% (v/v). To dissolve the DEPC, the containing bottle was shaken well, allowed to stand overnight in a fume cupboard and autoclaved the following day. Tris is destroyed by DEPC, therefore solutions containing this reagent was made in RNase-free glassware using water that had been previously treated with DEPC.

2.16.2 Total RNA extraction, using Promega kit

This method was used to isolate total RNA from 5-10 ml whole fresh blood and carried out according to the manufacture's protocol (SV Total RNA Isolation System).

2.16.3 Total RNA extraction, using QIAGEN kit (RNeasy Mini Kit)

This method was used to purify total RNA from small amount of animal tissues (20-30 mg) and cell cultures (1×10^7 cells grown in suspension (lymphocytes) or in monolayers

(endothelial and monocyte cells)) and was carried out according to the manufacturer's protocol. To homogenise tissues or cells a portable homogeniser was used.

For more efficient use of the total RNA in subsequent procedures (RT-PCR or mRNA isolation from total RNA) further DNA removal was carried out using RNase-free DNase (QIAGEN) according to the manufacturer's protocol.

2.16.2 Purification of mRNA from total RNA, using QIAGEN (Oligotex kit)

This method was used to purify polyA$^+$ mRNA from up to 250 mg total RNA per spin-column of the Oligotex mini kit and all procedures were carried out according to the manufacturer's protocol.

2.16.5 Precipitation of total RNA and poly A$^+$ mRNA with TouchDown Precipitation Reagent

TouchDown Precipitation Reagent (Active Motif) is a novel reagent that enables the precipitation of nucleic acids without addition of salt or any carrier. This method is ideal for precipitating and increasing the concentration small amounts of mRNA. Furthermore this method can be carried out at 4°C, eliminating the need to freeze the sample at -70°C and the yield is about 100%. This method was carried out according to the manufacturer's protocol.

2.17 Characterisation of Size and the Quality of RNA Using Denaturing Agarose Gel Electrophoresis

This method was used to determine the size and the quality of isolated total RNA and *in vitro* transcribed RNA and was modified from Jones *et al.*, 1994, Brown *et al.*, 1999 and Ambion TechNotes.(http://www.ambion.com/techlib/tn/73/737.html).

The appropriate amount of agarose was dissolved in water and cooled to 60°C in a water bath. One percent agarose is suitable for RNA molecules 500 bp to 10 kb in size. A higher-percentage gel (1 to 2%) was used to resolve smaller molecules and a lower-percentage gel (0.7 to 1%) for larger molecules. In a fume hood, the appropriate amount of 1× MOPS buffer (20 mM MOPS, 5 mM sodium acetate, 0.5 mM EDTA, pH 7.0) and formaldehyde (6.5 final concentration) were added. After removing the comb, the gel was placed in the gel tank and 1× MOPS and 6.5% formaldehyde used as running buffer.

Samples were prepared in a total volume of 18 µl by addition of 3× sample buffer to a final concentration of 1× (0.7× MOPS buffer, 9.2% formaldehyde, EtBr 30 $\mu g.ml^{-1}$). The RNA molecular weight marker was treated in exactly the same manner as samples. The samples were then denatured by heating in a boiling water bath for 5 minutes and snap cooling on ice and a one-tenth volume of 10× loading buffer (50% glycerol, 1 mM EDTA pH 8.0, 0.25% bromophenol blue, 0.25% xylene cyanol FF) was added to each sample. Electrophoresis was carried out at 5V/cM until the bromophenol blue dye migrated to 2/3 the length of the gel.

2.18 Preparing and Fixation of Tissues for *In Situ* Hybridisation

Mouse tissues were obtained by dissection and were immediately fixed using two alternative methods.

To prepare wax-embedded tissue (formalin fixation method) the desired tissues were cut into small pieces (less than 0.4 × 0.4 × 0.3 cm), immersed into 4% paraformaldehyde, pH 7.2–7.4 and submitted to the Department of Pathology, University of Leicester for paraffin wax processing.

To prepare frozen tissue sections, a thin layer (5 mm) of OCT compound (H&E Company) was spread onto the surface of a cork mount (5 mm × 5 mm × 2 mm). Using a liquid nitrogen Dewar, a glass beaker containing 25 ml isopentane was cooled until the isopentane began to freeze (-150°C). Using a new scalpel, the freshly excised tissues were trimmed to the less than 3 mm pieces and each piece was oriented and laid down onto the surface of the cork mount. Using an artery forceps, the mount was quickly lowered into the beaker until the tissue submerged below the surface of the cold isopentane. The cork mounts were left there until the OCT compound was frozen and had turned white. The frozen corks with tissues attached were then transferred to labelled Nunc cryotubes and stored in liquid nitrogen vapour. The tissues frozen in this way can be kept for RNA *in situ* hybridization for six mounts to one year. All tissue slides from frozen section or paraffin wax embedded tissues were prepared by the Department of Pathology, University of Leicester.

2.19 Labelling of Oligonucleotide Probes

This method was used to label the 3′ end of an anti-sense oligonucleotides cocktail of mouse mitochondrial P47 gene RNA.

In a microfuge tube, 2 µl of P47 oligonucleotides cocktail (five different oligonucleotides of the mouse mitochondrial P47 gene RNA, 0.5 $\mu g.\mu l^{-1}$), 4 µl 5× TdT (terminal deoxynucleotidyl transferase) reaction buffer (Roche), 4 µl cobalt chloride (25 mM), 1.7 µl 1 mm Digoxigenin-11-dUTP (Roche), 1.7 µl 5 mM dATP and 3 µl TdT enzyme (15 $units.\mu l^{-1}$, Invitrogen) were added in a total volume of 20 µl and mixed by pipetting. The reaction was incubated at 37°C for 15 minutes and, to stop the TdT activity, 1 µl of 0.5 M EDTA pH 8.0 was added to the mixture. To bring the final concentration of oligonucleotides to 20 $ng.\mu l^{-1}$, 30 µl of sterile water was added and the sample was divided in 200 ng aliquots and stored at -20°C.

2.20 Haematoxylin-Eosin Staining (H&E Staining)

To study the structure of the tissues of interest before *in situ* hybridisation studies, H&E staining was carried out.

The sections were de-waxed and rehydrated by passing the slides rack through a series of staining dishes containing the following solvents: 2 changes of Xylol for 3 minute, 99% IMS for 1 minute, 2 changes of 99% IMS for 1 minute and running tap water for 1 minute. Following the de-waxing step, the slide rack was placed in Haematoxylin for 5 minutes and then the slides were rinsed in running tap water for 1-2 minutes. Finally, the slides rack was placed in Eosin for 1 minute and then washed in running tap water for 30 seconds.

To dehydrate and clear the sections, the slides rack was passed through a series of staining dishes as follow: 95% IMS for 15 seconds, 99% IMS for 1 minute, 99% IMS 1 minute, Xylol (for mounting) for 3 minutes. Slides were mounted by placing DPX mountant (H&E Company) onto each cover slip using an orange stick. The slide was lowered onto the cover slip with the section facing downwards, allowing the DPX to spread out, and then the cover slip was pressed gently to remove any air bubbles. Finally, the slides were left to air dry in a slide tray.

2.21 mRNA *In Situ* Hybridisation Using Digoxigenin Labelled Olgonucleotides

2.21.1 Pre-treatment of sections

The sections were de-waxed and rehydrated by passing the slides rack through a series of staining dishes containing the following solvents: Xylol for 5 minutes, xylol for 2 minutes, 2 changes of 99% IMS for 1 minute, 95% IMS for 1 minute and finally the slides were immersed twice in DEPC treated water, each time for 10 minutes. The slides were then placed on an incubation tray and sections completely covered with 200 µl proteinase K solution (50 mM Tris-HCl pH 7.6 and 5-20 $\mu g.ml^{-1}$ proteinase K) and incubated at 37°C for the required time (the optimal concentration of proteinase K and duration of incubation were determined experimentally). After incubation, the slides were then immersed in DEPC-treated water twice each for 5 minutes.

2.21.2 Hybridisation

The slides were placed on an incubation tray and 100 µl prehybridisation solution (0.6 M NaCl, 1× PE (50 mM Tris-HCl pH 7.5, 0.1% (w/v) sodium pyrophosphate, 0.2% (w/v) polyvinylpyrrolidone, 0.2% (w/v) Ficoll 400, 5 mM EDTA), 10% dextran sulphate, 150 $\mu g.ml^{-1}$ ssDNA and 30% (v/v) formamide) was added to each slide, followed by incubation at 37°C for 1 hour. During the prehybridisation step, the probe solution (hybridisation solution) was prepared by adding the probe to an aliquot of prehybridisation solution to the required concentration. After the incubation period, the prehybridisation solution was removed from slides and a minimum volume of hybridisation solution sufficient to completely cover the section was added. The coverslip was replaced and the slide incubated at 37°C overnight.

After an overnight incubation, the coverslips were removed (by allowing them to slide from the microscope slide into a beaker) in a fume cupboard and then the slides were immersed in two changes of 2× SSC for 10 minutes each. The slides were then immersed in TBSBT solution (50 mM Tris-HCl pH 7.4, 100 mM NaCl, 3% (w/v) BSA, 0.1 % v/v) Triton X-100, prepared on the day required and filtered before use) for 5 minutes.

2.21.3 Detection

The slides were placed on an incubation tray and the section covered completely with 100 µl anti-digoxigenin conjugated antibody diluted in 1 in 600 with TBSBT solution and incubated at room temperature for 30 minutes. The slides were immersed in two changes of TBS (50 mM Tris-HCl pH 7.6, 100 mM NaCl) for 5 minutes and then washed by immersing in two changes of DEPC-treated water for 5 minutes. The slides were then placed on a covered incubation tray and covered completely with 200-400 µl of fresh filtered NBT/BCIP solution (0.4 $mg.ml^{-1}$ nitro blue tetrazolium chloride, 0.19 $mg.ml^{-1}$ 5-bromo-4-chloro-3-indolyl phosphate, 100 mM Tris-HCl pH 9.5, 50 mM $MgSO_4$). The slides were then incubated in the covered tray at room temperature and monitored microscopically to determine the appropriate time of incubation (usually 2-6 hours). After the incubation period, the coverslips were drained into a beaker, the slides washed in running tap water and after air drying they were mounted (Section 2.20).

2.22 DNA Cloning Procedures

2.22.1 Dephosphorylation

This procedure was carried out according to the manufacturer's protocol (Amersham Pharmacia Biotech).

The DNA (the fragment of interest or cut plasmid) to be dephosphorylated was incubated with shrimp alkaline phosphatase buffer (20 mM Tris-HCl pH 8.0, 10 mm $MgCl_2$) and 0.1 unit per pmol of 5'-protruding DNA termini of shrimp alkaline phosphatase for 1 hour at 37°C in a total reaction volume of 10 µl. The phosphatase enzyme was then inactivated by heating the reaction at 65°C for 15 minutes.

2.22.2 Ligation

This method was based on the method described by Sambrook *et al.* (1989).

The DNA of interest was ligated using T4 DNA ligase (Invitrogen) in the recommended buffer (50 mM Tris-HCl pH7.6, 10 mM $MgCl_2$, 1mM ATP, 1 mM DTT, 5% (w/v) polyethylene glycol-8000). A molar ratio of 1:3, vector: insert, was used in a total reaction volume of 10 µl. The reaction was incubated at 16°C for at least 16 hours.

2.22.3 Site specific-recombination

To carry out this procedure the appropriate protocol in the Invitrogen Gateway Cloning Technology manual was followed. The site specific-recombination cloning was carried out in four steps.

1 PCR amplification of the fragment of interest using attB-flanked forward and reverse primers.
2 Purification of the attB-PCR product.
3 Creation of an entry clone from attB-flanked PCR product via the BP reaction.
4 Creation of an expression clone via the LR reaction.

All the above procedures were carried out according the manufacturer's protocol except for one modification. In step 2 for purification of the attB-PCR product, the QIAGEN PCR purification kit was used (Section 2.10.2.1).

2.23 *In Vitro* Transcription

To transcribe the sense and antisense (negative control) RNA and compare unlabelled (cold transcript) and labelled transcripts (hot transcript), this procedure was carried out in two steps. All reagents and enzymes were supplied by Invitrogen.

2.23.1 Unlabelled *in vitro* transcription

This method was used to transcribe unlabelled RNA from linearised DNA templates using T3 and T7 RNA polymerases.

In a 1.5 ml microfuge tube, 2 µg linearised DNA, 20 µl 5× transcription buffer (0.2 M Tris-HCl pH 8.0, 40 mM $MgCl_2$, 10 mM spermidine-$(HCl)_3$, 125 mM NaCl), 10 µl 100mm DTT, 2.5 µl RNasin (40 units.μl^{-1}), 2.3 µl 100 mM NTPs, 50 units RNA polymerase T3 or T7 were added in a total reaction volume of 100 µl and mixed well by pipetting. The reaction was incubated at 37°C for 2 hours and then the reaction was stopped by heating at 65°C for 15 minutes.

2.23.2 Labelled *in vitro* transcription (Digoxigenin-labelled RNA probes).

This method was used to transcribe the sense RNA (using RNA polymerase T3) and the antisense RNA (as the negative control, using RNA polymerase T7) using the linearized DNA template with labelling. To label the RNA strands during transcription, DIG RNA labelling mixture (Roche) was used according to the manufacturer's protocol (Roche).

In a 1.5 ml microfuge tube, 1 µg linearised DNA, 2 µl 10× Dig RNA labelling mixture (10 mM ATP, 10 mM CTP, 10 mM GTP, 6.5 mM TTP, 3.5 mM DIG-11-UTP, pH 7.5), 2 µl 10 × transcription buffer (0.4 M Tris-HCl pH 8.0, 60 mM $MgCl_2$, 20 mM spermidine), 0.5 µl RNasin (40 units.μl^{-1}) were added in a total reaction of 20 µl and mixed well by pipetting. The reaction was incubated at 37°C for 2 hours and then the reaction was stopped by heating at 65°C for 15 minutes.

2.23.3 Ethanol precipitation of RNA transcripts

The RNA transcripts were purified from unincorporated NTPs using the lithium chloride/ethanol procedures and carried out according to the manufacturer's protocol (Roche).

2.24 Construction of a Mouse Pooled Tissues cDNA Library (Yeast Two-hybrid Library)

To construct a mouse pooled tissues cDNA library, the Superscript Plasmid System with Gateway Technology for cDNA Synthesis and Cloning from Invitrogen, was used. All the following methods, except for a few modifications, were carried out according to the manufacture's protocols and are based on the method described by Okayama and Berg (1982) with modifications by Gubler and Hoffman (1983).

2.24.1 First strand synthesis

This step was carried out in a 20 µl total volume reaction and has been designed to convert up to 5 µg of mRNA into first strand cDNA.

Two microlitres of *Not* I primer-adaptor were added to a 1.5 ml sterile RNase-free microfuge tube. Then 5 µg mRNA in 5 µl DEPC-treated water was added and mixed. The mixture was heated at 70°C and then snap cooled on ice. The contents of the tube were collected by a brief centrifugation and then 4 µl 5× first strand buffer, 2 µl 0.1 mM DTT, 1 µl 10 mM dNTPs

mixture and 1 µl [α-^{32}P]dCTP (NEN, 10.0 mCi.ml^{-1}, 111 TBq/mmol) were added and mixed by gentle vortexing and collected by brief centrifugation. The tube was incubated at a 37°C for 2 minutes to equilibrate the temperature and then 5 µl of SuperScript II reverse transcriptase (200 units.µl^{-1}) was added, mixed and incubated at 37°C for a further 1 hour (The amount of SuperScript II reverse transcriptase depends upon the amount of starting mRNA and 200 units of enzyme is used for ≤1 µg of mRNA). The tube was then placed on ice and 2µl of the reaction was removed and added to a microfuge tube containing 43 µl of 20 mM EDTA pH 7.5 and 5 µl of yeast tRNA. This mixture was used to measure the quantity and quality of the first strand yield. The remaining 18 µl of the first reaction was used to synthesise the second strand of cDNA.

2.24.2 Determination the quantity of the first strand cDNA

Ten-microlitre aliquots from the diluted first strand cDNA sample were spotted in duplicate onto two pieces (1 cm^2) of Whatman DE-81 filter. Both filters were dried at room temperature. One of these filters with no further treatment, was used to determine the specific activity of the dCTP in the reaction. The other was washed three times, for 5 minutes each, in 50 ml of fresh, ice cold 10% (w/v) trichloroacetic acid (TCA) containing 1% (w/v) sodium pyrophosphate. The filter was then washed with 50 ml of 95% ethanol at room temperature. The filter was air dried and used to determine the yield of the first strand cDNA. Both filters were counted in 3 ml standard scintillant to determine the total amount of ^{32}P in the reaction, as well as the amount of ^{32}P that was incorporated into cDNA.

The remaining 30 µl sample from the reaction was precipitated by adding 15 µl (half volume) of 7.5 M acetate ammonium, followed by 90 µl of cold (-20°C) absolute ethanol. The sample was vortexed thoroughly and centrifuged at room temperature for 20 minutes at 14,000 × g. The supernate was then removed carefully and the pellet washed with 0.5 ml of 70% ethanol (-20°C) for 2 minutes at 14,000× g. The supernate was removed and the pellet (cDNA) was dried at 37°C for 10 minutes to evaporate the residual ethanol. This cDNA was used to determine the quality of the first strand cDNA, using alkaline agarose gel electrophoresis.

2.24.3 Gel analysis

To estimate the size range of the synthesized cDNA, alkaline agarose gel electrophoresis was used. The ethanol-precipitated first strand sample was dissolved in 10 µl 1× alkaline agarose gel sample buffer (30 mM NaOH, 1 mM EDTA, 10% (v/v) glycerol, 0.01% bromophenol

blue). The 1.4% (w/v) agarose gel was cast in 100 ml of 30 mM NaCl, 2 mM EDTA and then equilibrated for 2 to 3 hours in alkaline electrophoresis buffer (30 mM NaOH, 2 mM EDTA) before loading the sample. Size marker prepared according the method described in Section 2.9. After loading the samples, electrophoresis was carried out for 5 to 6 hours at 50 V. The gel was dehydrated using a hair dryer and then exposed to X-ray film overnight at room temperature.

2.24.4 Second strand synthesis

This step describes the synthesis of the second strand of cDNA was carried out using the cDNA product from the first step Section 2.24.1).

On ice, the following reagents were added to the remaining 18 µl of the first reaction: 93 µl of DEPC-treated water, 30 µl of 5× second strand buffer, 3 µl of 10 mM dNTP mixture, 1 µl of *E.coli* DNA ligase (10 units.µl^{-1}), 4 µl of *E.coli* DNA polymerase I (10 units.µl^{-1}), 1 µl of *E.coli* RNase H (2 units.µl^{-1}). The tube was gently vortexed and incubated at 16°C for 2 hours and then 2 µl of T4 DNA polymerase (5 units.µl^{-1}) was added, mixed and incubated at 16°C for a further 5 minutes. To terminate the reaction, the tube was placed on ice and 10 µl of 0.5 M EDTA was added. Then 150 µl (equal volume of the reaction) of phenol:chlorophorm:isoamyl alcohol (25:24:1) was added to the sample, vortexed thoroughly and centrifuged at room temperature for 5 minutes at 14,000× *g* to separate the phases. Carefully 140 µl of the upper aqueous layer was removed and transferred to a fresh 1.5 ml microfuge tube and ethanol precipitation was carried out (Section 2.24.2).

2.24.5 *Sal* I adaptor addition

The tube containing the precipitated second strand cDNA was placed on ice and 25 µl DEPC-treated water, 10 µl 5× T4 DNA ligase buffer, 10 µl *Sal* I adaptors, 5 µl T4 DNA ligase (1 unit. µl^{-1}) were added to the tube. The sample was mixed gently and incubated at 16°C for a minimum of 16 hours. The reaction was again phenol extracted and ethanol precipitated (Sections 1.24.2 and 2.24.4).

2.24.6 *Not* I digestion

The precipitated cDNA (Section 2.24.5) was resuspended in 41 µl DEPC-treated water and 5 µl of manufacturer recommended restriction enzyme buffer (React 3 buffer) and 4 µl of *Not*I

(15 units.μl^{-1}) were added to the tube. The reaction was mixed gently and incubated at 37°C for 2 hours. After incubation, the phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation were carried out and the pellet was dried.

2.24.7 Column chromatography

Column chromatography (Invitrogen) was carried out to remove residual adaptors, small fragment of DNA released by *Not*I digestion, and to select appropriate size cDNA fragments. This step was carried out according to the manufacturer's protocol. The amount of double stranded cDNA in each fraction was determined by Cerenkov counting.

2.24.8 Ligation of the cDNA to the vector

After selection of fractions of cDNA for ligation, 4 µl 5× T4 DNA ligase buffer, 2 µl pEXP-AD502 *Sal*I-*Not*I-cut vector (50 ng.μl^{-1}), 30 ng cDNA and enough DEPC-treated water for a 19 µl total volume were added to a microfuge tube. Then 1 µl of T4 DNA ligase (1 unit.μl^{-1}) was added and mixed by pipetting. The reaction was incubated at 4°C overnight. Following incubation, 5 µl of yeast tRNA (1µg. μl^{-1}) was added to the reaction, ethanol precipitation was carried out and the pellet was dried. Next 5 µl of sterile distilled water was added to the dried pellet and 1 µl samples were used to carry out individual electroporations.

2.24.9 Introduction of ligated cDNA into *E. coli* by electroporation

One microlitre aliquots of ligated cDNA were added to 50 µl DH10B electro competent cells (Section 2.12.6.1) and transformation was carried out at 2.5 kV at settings of 100 ohms and 25 µF (Section 2.12.6.2).

2.24.10 Expansion of plasmid cDNA library

To expand the plasmid library, approximately 50,000 colonies of primary cDNA transformants (colonies of original library) were plated on 140 mm LUA plates plus ampicillin (200 µg/ml) and enough plates used to cover at least 6 times the number of independent clones in the original library. Plates were incubated at 30°C for two nights or until the colonies were nearly confluent. The colonies were then scraped using a cell scraper into 15 ml LUB and 1× HMFM. The scraped colonies were mixed well by pipetting and subdivide into 1-ml aliquots and stored at -80°C.

2.25 Yeast Two-hybrid Library Screens

All the following protocols with some modifications were taken from the Clonetech MATCHMAKER *GAL4* Two-Hybrid System 3 and libraries user manual.

2.25.2 Library titering and DNA preparation

An aliquot of library was thawed and a serial dilution from 10^{-3} to 10^{-6} made in LUB broth. Fifty-microlitre and 100 µl aliquots from each dilution were spread on LA/ampicillin plates. Plates were then incubated at 37°C overnight and the resulting colonies counted and used to calculate the library titre in cfu.ml^{-1} (colony forming unit).

Approximately 200,000 colonies of the yeast two-hybrid library were plated on 140 mm plates and enough plates used to cover at least 6 times the number of independent clones in the library. Plates were incubated at 30°C for two nights or until the colonies were nearly confluent. The colonies were then scraped using a cell scraper into 15 ml LUB. The scraped colonies were then transferred into 1 litre of LUB/ampicillin broth and incubated at 37°C for 2-3 hours with shaking. A maxi-scale plasmid DNA isolation was then carried out using a Qiagen kit as described in Section 2.12.8.3.

2.25.3 Screening the library

The yeast two-hybrid library screens were carried out by sequential transformation. An initial transformation was carried out to transform the plasmids containing the bait constructs into the yeast AH109. Transformation of different baits was carried out using small-scale transformations as described in Section 2.13.5. The yeast AH109 harbouring individual baits was then grown on SD/-leucine medium plates and a large-scale transformation carried out to introduce the library plasmids (prey) into the strain already containing the bait plasmid. These yeast, containing both the bait and prey plasmids, were plated onto SD/-leucine -tryptophan medium plates.

The library screen plates were then incubated at 30°C for up to 2 weeks and any positive clone was patched and transferred onto SD/-leucine -tryptophan -histidine medium plates. Growing colonies were then replica plated onto SD/-leucine -tryptophan -histidine -adenine medium plates to eliminate false positives.

2.26 Southern Blotting

Southern blotting, a technique developed by Ed Southern (Southern, 1975), and modified by Dalgleish, 1987 was used to analyse mouse DNA BAC clones.

2.26.1 Digestion of BAC DNA

Restriction enzyme-digested DNA samples were electrophoresed on a 0.7% LE agarose gel and photographed as detailed in Section 2.4.

2.26.2 Pre treatment of the gel

Prior to blotting, the gel (100 ml) was pre-treated by washing for 7 minutes in depurinating solution (0.25 M HCl), rinsing briefly in distilled water and washing for 30 minutes in denaturing solution (0.5 M NaOH, 1.5 M NaCl). The gel was then rinsed in distilled water again and washed for 30 minutes in neutralising solution (3 M NaCl, 0.5 M Tris-HCl, pH 7.4). Each of the washes was done in a volume of 250–300 ml on a shaking platform (Stoval Life Science Inc).

2.26.3 Blotting the gel

The blotting apparatus was set up in the conventional method as shown in Figure 4.2 of Dalgleish (1978). A glass plate was placed across a suitably-sized plastic tray. A sheet of Whatman 3 MM paper was then placed over the glass plate with the ends hanging down into the tray. About 250-300 ml 20× SSC (3 M NaCl, 0.3 M sodium citrate, pH 7.0) was then poured into the tray until the ends of the paper dipped into this solution. The neutralised gel was then placed onto the blotting apparatus. An exactly gel-sized nylon membrane (Hybond-N, Amersham Pharmacia Biotech) was placed over the surface of the gel, followed by a gel-sized Whatman 3 MM paper above. Both the membrane and paper were pre-washed in a 3× SSC. A stack of paper towels was placed on the top of the gel, followed by a glass plate and a weight of about 250 g. The blot was then left overnight with wet paper towels being replaced with dry ones every 5 minutes during the first 30 minutes.

After completion of blotting, the apparatus was dismantled. The gel origin was marked on the membrane and the bottom right corner was cut off for orientation purposes. The membrane

was then left on Whatman 3MM paper to air dry and was then cross linked in a UV cross-linker (Amersham Life Science model RPN 2500/2501, 10-15 seconds at 70,000 $\mu J/cm^2$).

2.26.4 Preparation of oligo-labelling buffer (OLB)

The OLB is made from solutions A, B, and C, which are mixed together in the ratio of 2:5:3 (Dalgleish, 1987). Solution A is made from 625 µl 2 M Tris-HCl, pH 8.0, 25 µl 5 M $MgCl_2$, 350 µl distilled water, 18 µl 2-mercaptoethanol, 5 µl dTTP, 5 µl dGTP, 5 µl dATP (each dNTP was dissolved in 3 mM Tris-HCl, 0.2 mM EDTA, pH 7.0 at a concentration of 0.1 M). Solution B is 2 M HEPES, pH 6.6 and solution C is random hexadeoxyribonucleic acids (Amersham Pharmacia Biotech) at 90 OD units.ml^{-1}.

2.26.5 Preparation of the probe

Probes for hybridisation to Southern blots and in library screening were radiolabelled with ^{32}P using the oligo-labelling method. DNA to be labelled as a probe was boiled for 5 minutes and then chilled on ice unless the sample also contained agarose, in which case the sample was kept at room temperature. The labelling reaction was set up at room temperature by adding these components in the following order to give a final volume of 15 µl: distilled water (to a total volume of 15 µl), 3 µl of oligo-labelling buffer (OLB), 0.6 µl of BSA bovine serum albumin, 10 $mg.ml^{-1}$), 5-10 ng of DNA, 1 µl [α-^{32}P] dCTP (NEN, 10 $mCi.ml^{-1}$, 111 TBq/mmole) and 0.6 µl Klenow fragment DNA polymerase I (USB Corporation, 1 $unit.\mu l^{-1}$). The oligo-labelling reaction was left to proceed over night at room temperature. The reaction was then stopped with 85 µl of oligo-labelling stop solution (20 mM NaCl, 20 mM Tris-HCl pH7.5, 2mM EDTA, 0.25% (w/v) SDS).

2.26.6 Test of incorporation of radionucleotide (dCTP) into the probe

This method was adapted from (Sambrook *et al.*, 1989).

One microlitre of labelled probe (after adding oligo-labelling stop solution) was taken and diluted with 11 µl of distilled water. Five microlitres of this diluted labelled probe was spotted onto each of two pieces of Whatman DE-81 paper (1.5 cm × 1.5 cm). One filter was then washed 6 times (for 5 minutes per wash) in 0.5 M NaH_2PO_4, followed by two washes in water and two in ethanol, each for one minute per wash. The filter was left on the bench to dry and

then the counts on the washed and unwashed filters compared. The percentage of the radiolabel incorporated dCTP into the probe was calculated.

2.26.7 Pre-hybridisation wash and hybridisation of the probe

After the UV cross-linking of the DNA onto the nylon membrane, it was washed for 2 hours in 25 ml pre-hybridisation solution (1.5× SSPE (0.27 M NaCl, 15 mM Na_2PO_4 (pH 7.7), 1.5 mM EDTA), 0.5% (w/v) Marvel low fat dried milk (Premier Beverages), 1% (w/v) SDS, 6% (w/v) polyethylene glycol 8000) at 65°C in a Hybaid hybridisation oven (Hybaid Limited). The labelled probe DNA was denatured by boiling for 5 minutes and then snap cooled on ice. The probe was then added and mixed to the same pre-hybridisation buffer and the hybridisation allowed to proceed at 65°C overnight.

2.26.8 Post-hybridisation washes

After overnight incubation, the hybridisation buffer was discarded and the nylon membrane was then rinsed 3 times briefly and washed twice for 10 minutes each in 25 ml of 3× SSC, 0.1% (w/v) SDS at 65°C. To increase the stringency, the membrane was then washed four times in 25 ml 0.5× SSC, 0.1% (w/v) SDS for 10 minutes per wash at 65°C.

2.26.9 Autoradiography

Prior to autoradiography, the membrane was blotted dry on filter paper and then placed and wrapped onto a sheet of card with plastic film (Saran wrap). The wrapped card and membrane was then placed into a radiography cassette with an intensifying screen, and a sheet of X-ray film (Kodak XAR) was applied over the membrane within the cassette with intensifying screens.

The cassette was then put at -70°C for an appropriate time and then processed manually (Section 2.8.5).

2.27 Northern Hybridisation (Northern Blot)

This transfer of denatured or non-denatured RNA to Hyband-N^+ nylon membrane was adapted from Dalgleish, 1987 and Sambrook *et al.*, 1989. All solutions and reagents used were RNase free and prepared using DEPC-treated water.

2.27.1 Transfer of non-denatured RNA to Hybond-N^+ nylon membrane filter

Transfer of denatured or non-denatured RNA to Hybond-N^+ nylon membrane (Amersham Pharmacia Biotech) filter was carried out according to the method in Section 2.26.3.

2.27.2 Detection procedure

After transferring the RNA onto a nylon filter, the filter was washed in 100 ml washing buffer (0.3% (v/v) Tween 20, 0.1 M maleic acid and 0.15 mM NaCl, pH 7.5) for 1 minute. The filter was then soaked and washed in 1× block solution (10× block solution (Amersham Pharmacia Biotech) diluted in 0.1 M maleic acid and 0.15 M NaCl, pH 7.5, on a shaking platform for 30 minutes. In the next step, 1/5000 dilution of Anti-Digoxigenin-Ab (Amersham Pharmacia Biotech) was added to the 1× block solution and shaken on a shaking platform for at least 30 minutes. The filter was then washed twice in washing buffer for 15 minutes and transferred to a 100 ml detection buffer (0.1 M Tris-HCl, 0.1 M NaCl pH 9.5) and washed for 5 minutes. Then a 1/250 volume of CDP-Star (Amersham Pharmacia Biotech) detection buffer was added and shaken for 5 minutes.

Finally, the filter was semi-dried at room temperature and placed and wrapped onto a sheet of card with plastic film (Saran wrap). The wrapped card and membrane was then placed into a radiography cassette, and a sheet of X-ray film (Kodak XAR) was applied over the membrane within the cassette. The cassette was then put at room temperature for 5 minutes and the film processed manually (Section 2.8.5).

2.28 RNA Dot Blotting

Dilutions RNA were manually dot blotted onto a piece of Hybond-N^+ nylon membrane (Amersham Pharmacia Biotech) filter. The dot blots were allowed to be air dried completely. The filter was then wrapped in a Saran plastic film and UV cross-linked in a UV cross-linker (Amersham Life Science model RPN 2500/2501, 10-15 seconds at 70,000 $\mu J/cm^2$). All solutions and reagents used were RNase free and prepared using DEPC-treated water.

2.28.1 Pre-hybridisation and hybridisation washes and conditions

After the UV cross-linking of the RNA, the nylon membrane was washed for 1 hour in 20 ml pre-hybridisation solution with (0.1–0.6 M NaCl, 1× PE solution (50 mM Tris-HCl, pH 7.5,

1% SDS, 0.2% Ficoll 400, 5 mM EDTA, 0.1% tetra-sodium pyrophosphate, 0.2% polyvinyl pyrolidone), 150 μg/ml ssDNA (<500 bp), 30–60% formamide, 2% block solution (Amersham Pharmacia Biotech)) at 37–42°C for 1 hour. Then 5 ng/ml Dig-labelled riboprobes was added to pre-hybridisation solution and hybridisation carried out overnight. The concentration of NaCl, formamide and the temperature were modified in different experiments to obtain the best stringency (better sensitivity and specificity) for hybridisation.

2.28.2 Post-hybridisation washes

After an overnight incubation in hybridisation solution, the filter was washed twice in 20 ml post-hybridisation solution (0.2-0.5 × SSC and 30-60% formamide) each time for 10 minutes at 37°C. The concentrations of SSC and formamide were modified in different experiments to obtain the best stringency (better sensitivity and specificity).

2.28.3 Detection

In this step, all procedures were carried out at room temperature. Following the post-hybridisation washes the filter was treated with the same procedures as described in Section 2.27.2, to detect hybridisation between the riboprobes and the target RNA target.

2.29 Computer Hardware, Software and Internet Sites

The following list summarises the different computing facilities have been used in this project.

2.29.1 Computer facilities (hardware)

Analyses were carried out on IBM compatible microcomputers running various of the Microsoft Windows operating system or on a Silicon Graphics Origin 200 computer running the IRIX v.6.5 operating system.

2.29.2 Software

- Wisconsin Package v.10.0, v.10.1, v.10.2 and v.10.3 for IRIX (Accelerys)
- Chromas v.1.45 (Technelysium)
- AlphaEase v.3.24I (Alpha Innotech Corporation)

- Microsoft Internet Explorer v.5.5
- Hummingbird eXceed v.6.1 and v.6.2
- Adobe Photoshop v.3.0 LE
- Microsoft Office 97, 2000 and XP professional
- EndNote v.4.0 (Thomson ISI Researchsoft)
- Serif PhotoPlus v.6.0
- Macromedia FreeHand v.8.0
- Fluorchem v2.01 (Alpha Innotech Corporation)

2.29.3 Primer design and internet sites used

All primers were designed using the program Primer3 (Rozen and Skaletsky, 2000). This program is available at: http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi.

Other internet sites that were used frequently are as follow:

- Ensembl human genome (http://www.ensembl.org).
- Human Genome Mapping Project Resource Centre (http://www.hgmp.mrc.ac.uk/).
- National Centre for Biotechnology Information (http:/www.ncbi.nlm.nih.gov/).
- Network Protein Sequence Analysis (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_server.html).
- The Restriction Enzyme Database (http://rebase.neb.com/rebase/rebase.html).
- Washington University Genome Sequencing Centre (http://genome.wustl.edu/est/).

Chapter 3
Scanning of the human *SPP2* gene for new variants and comparison of the spp24 protein sequence in eight other species

3.1 Introduction

Genomes are dynamic entities that alter continuously over time due to the effects of mutation (small-scale sequence changes) or recombination (large-scale sequence changes). Both mutation and recombination can have great impacts on the cell (and subsequently the organism) in which they occur. The product of a new mutation may have decreased or no function (loss of function mutation) and often cause recessive phenotypes or the product may perform some new abnormal function (gain of function). Also, mutations may cause a synonymous change, the new codon encodes the same amino acid as the unmutated codon. Therefore this kind of mutation has no impact (silent mutation) on the coding function of the gene, but it should be considered that some synonymous changes can create cryptic splice sites or can disrupt exonic splice enhancers (ESEs) and not all mutations occur within genes.

As the main goal of the Human Genome Project changes from listing genes to understanding their biological functions, the study of molecular pathology has moved to the centre stage. The main aim of molecular pathology is to clarify why a given genetic change causes a specific phenotype. For most inherited diseases where the relevant gene has been identified, many different mutations are known. We are not able to do experiments on humans or breed them to determine the exact effect of a mutation but humans provide unique opportunities to observe the clinical effects of many different changes in a particular gene. These types of studies are able to produce hypotheses that must then be tested in animal models. Therefore, studies of naturally occurring human mutations are usually followed and complemented by study of particular mutations in transgenic animals.

As discussed in Chapter 1, bone mineral density has been shown to be under strong genetic control (Gennari *et al.*, 2002). The initial aims of this study did not include any plan for linkage or association studies with respect to osteoporosis or other multifactorial disorders, but aimed to determine the function of the gene encoding the spp24 protein that has been speculated to have a role in the regulation of bone mineral density (Hu *et al.*, 1995). However,

as the study progressed, two variants were discovered, making it feasible to carry out a limited study to determine if any meaningful differences in the frequency of these variants existed between the normal population and osteoporotic patients. DNA samples were provided by Prof David Hosking (Nottingham City Hospital) from patients with low bone mineral density or osteoporosis. Detection of new variants and the characterisation of these variants in the human *SPP2* gene will be valuable for any further large-scale association study. These potentially have greater power of detection than linkage studies, therefore may identify genes contributing to a disease with a small or moderate effect which might otherwise be missed.

This chapter describes the detection and characterisation of two variants in intron 2 and exon 2 of the human *SPP2* gene and their subsequent use in a small-scale association study in the normal population and osteoporotic patients. This chapter also evaluates the importance of the variant in exon 2 by comparing the conservation of the altered amino acid between different species.

3.1.1 Linkage and association studies and bone mineral density (BMD)

As discussed in more detail in Chapter 1, osteoporosis is a condition where both cortical and trabecular bone become thinned and are therefore more susceptible to fracture. The most commonly affected area is the hip. To evaluate the severity of the osteoporosis, bone mineral density (BMD) is measured. BMD is usually measured at the spine and hip by DEXA (Dual Energy X-ray Absorptiometry), scanning that measures the fraction of radiation passing through the bone. BMD is measured in $g.cm^{-2}$, but is usually indicated as a T or Z score. The T score indicates the number of standard deviations relative to an average young person with an optimum bone density. The Z score indicates the number of standard deviations relative to an average person of the same age, sex and body size. For all the data received from Nottingham City Hospital that were used in this study, BMD has been recorded as a T score. All of these samples are from Caucasian, post-menopausal women. T scores have been calculated using the Hologic reference curve (Bennett, 2002). For each of the 90 BMD samples the T-score was provided for 6 different regions. These regions were spine, the total hip and then four regions within the hip and pelvis. Using the WHO criteria described in Table 3.1 each region for each sample was designated normal, low or osteoporotic. In this way each sample had a profile of BMD designations. If a sample's profile contained all normal designations, it was said to be a normal person from a BMD point of view. If a

Category	BMD expressed as a T score
Normal	BMD <1 SD below the mean young adult range
Low BMD	BMD 1–2.5 SD below the mean of young healthy women
Osteoporosis	BMD >2.5 SD below the mean of young healthy women
Severe Osteoporosis	BMD >2.5 SD below the mean of young healthy women and the presence of one or more fragility fracture

Table 3.1. The WHO criteria for low BMD and osteoporosis in women

This table illustrates details of the criteria used to define osteoporosis (WHO Technical report series, 1994).
These criteria described by WHO are based on the patient's T score which refers to the number of standard deviations relative to the mean of BMD in young healthy women with optimum bone strength.

sample's profile had at least one low, but no osteoporotic value, it was considered to be low overall. If a sample had at least one osteoporotic value, it was said to be osteoporotic overall.

As discussed in Chapter 1, BMD has been demonstrated to be under strong genetic control. It has also been demonstrated that there are separate genes that independently control BMD and susceptibility to fracture. Genome-wide scans using linkage analysis have been used in various studies to identify regions of the genome (in different species including human) that may contribute to low BMD. The human *SPP2* gene was initially localised to chromosome band 2q37→qter by *in situ* hybridisation (Swallow *et al.*, 1997). Regions thought to contain genes contributing to low spine and hip BMD were identified on chromosome 2p (Devoto *et al.*, 1998; Niu *et al.*, 1999). The only 2q region that has been identified was 2q13 in a study searching for genes that predispose to distal interphalangeal joint osteoarthritis (Leppävuori *et al.*, 1999). In spite of *SPP2* not mapping to a region of chromosome 2 thought to play a role in determining BMD, it was thought worthwhile to carry out a small pilot study to test for any possible association.

Four polymorphic microsatellites (CA, AG, GTT and AT; Figure 3.1) had previously been identified within or near the human *SPP2* gene (Bennett *et al.*, manuscript submitted). Three of these microsatellites (AG, GTT and AT) were typed in 109 individuals from various ethnic groups about whom no BMD data were available. In addition, there were 90 individuals for whom BMD measurements existed; approximately equal numbers with normal, low, and very low BMD (osteoporotic). The small-scale association study indicated that there were some differences between the genotypes seen in normal individuals and low BMD or osteoporotic samples, with the rare alleles more commonly present in the low BMD or osteoporotic samples. However, analysis showed that these results were statistically insignificant due to the small sample size and the rarity of the rare alleles.

Association studies should be interpreted cautiously because many factors (that will be discussed in Section 3.3) can produce spurious associations between a disease and a variant allele. Therefore, it is desirable to have large sample numbers, adequate controls, robust statistical methods and to obtain reproducible results in repeat studies in other populations.

This chapter presents the detection and characterisation of human *SPP2* gene variants and their use in a small-scale association study. If meaningful association-study differences had

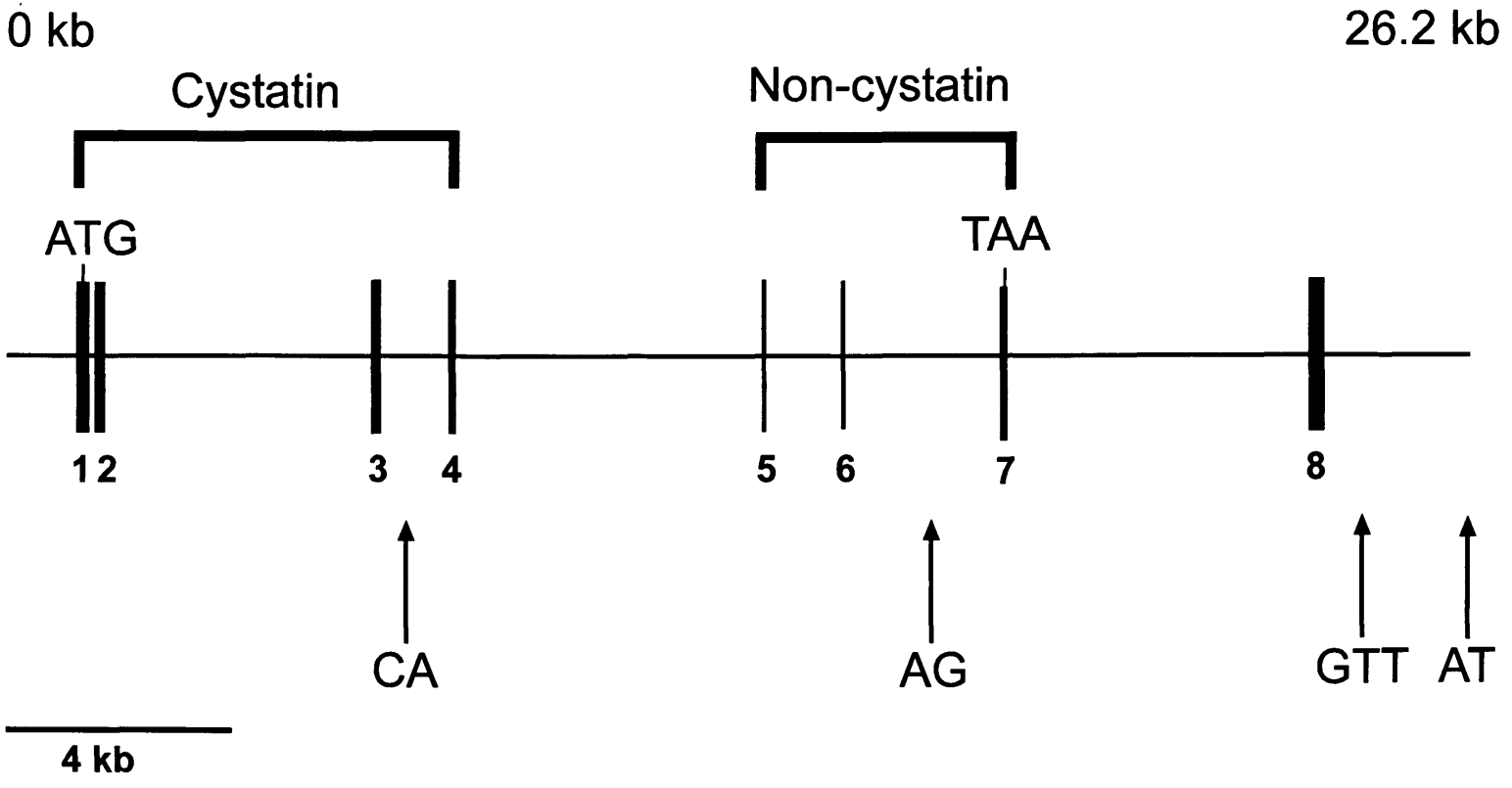


Figure 3.1. The exon/intron structure (map) of the human SPP2 **gene**

This figure illustrates the 26.2 kb region of the genomic DNA segment (accession number AC006037) that contains the human *SPP2* gene. The 'ATG' start codon in exon 1 and the 'TAA' stop codon in exon 7 are indicated. Exons 1-4 encode the cystatin-like region and exons 5-7 the non-cystatin-like region. Also shown on this figure are positions of three di-nucleotide repeats and one tri-nucleotide repeat found in the human *SPP2* gene (The map is not drawn to scale).

been identified, it might have provided justification to go on and complete a larger and more powerful controlled study.

3.1.2 Identification of genomic clones containing the human *SPP2* gene

The bovine spp24 cDNA sequence (accession number U03872) was used to search the human EST database and several ESTs were identified. One of the initial human ESTs was used to screen a human genomic PAC library RPCI1 obtained from the HGMP UK Resource Centre (Bennett *et al.*, manuscript submitted). This library comprised about 120,000 clones each containing an insert of approximately 110 kb in the recombinant P1 vector pCYPAC-2. Four clones were identified to contain most, if not all, of the *SPP2* gene. The clone numbers for these four positive clones were 14 E15, 37 E17, 137 C1 and 318 P19 (the first part of the designation refers to the microtitre plate number and the second to the location of the clone in that plate). Next, by screening a human chromosome-2 library (Gingrich *et al.*, 1996) from HGMP, two further PAC clones were identified (Dalgleish, unpublished) designated 3 N4 and 6 M9 (only one of which was shown to contain the entire gene).

3.1.3 Determination of the human SPP2 gene cDNA sequence and its exon/intron structure

Using the published bovine mRNA sequence (accession number U03872), several human ESTs were identified by FASTA searches of the EST division of the EMBL DNA sequence database. These ESTs, and others comprising the human spp24 UniGene cluster (Hs.12230), were aligned into a consensus cDNA sequence that, initially, was not complete at the 5′ end compared to the bovine cDNA sequence. To obtain a near full-length cDNA sequence, a cDNA library of 2.5×10^6 clones was constructed from normal human liver. About 10,000 cDNAs clones were screened by hybridisation with a part-length cDNA probe that yielded three positive clones whose identities were confirmed by DNA sequencing. Among these cDNAs, a full-length clone was identified (designated pB1.1) that added an additional 253 bp to the 5′ end of the initial consensus cDNA. The sequence of the 1018-bp full-length cDNA was submitted to the EMBL database with the accession number AJ308009 (Bennett *et al.*, manuscript submitted, Figure 3.2).

The *SPP2* gene sequence was determined by the Human Genome Project. The gene falls on the complementary strand of the reported sequence (accession number AC006037). The exon

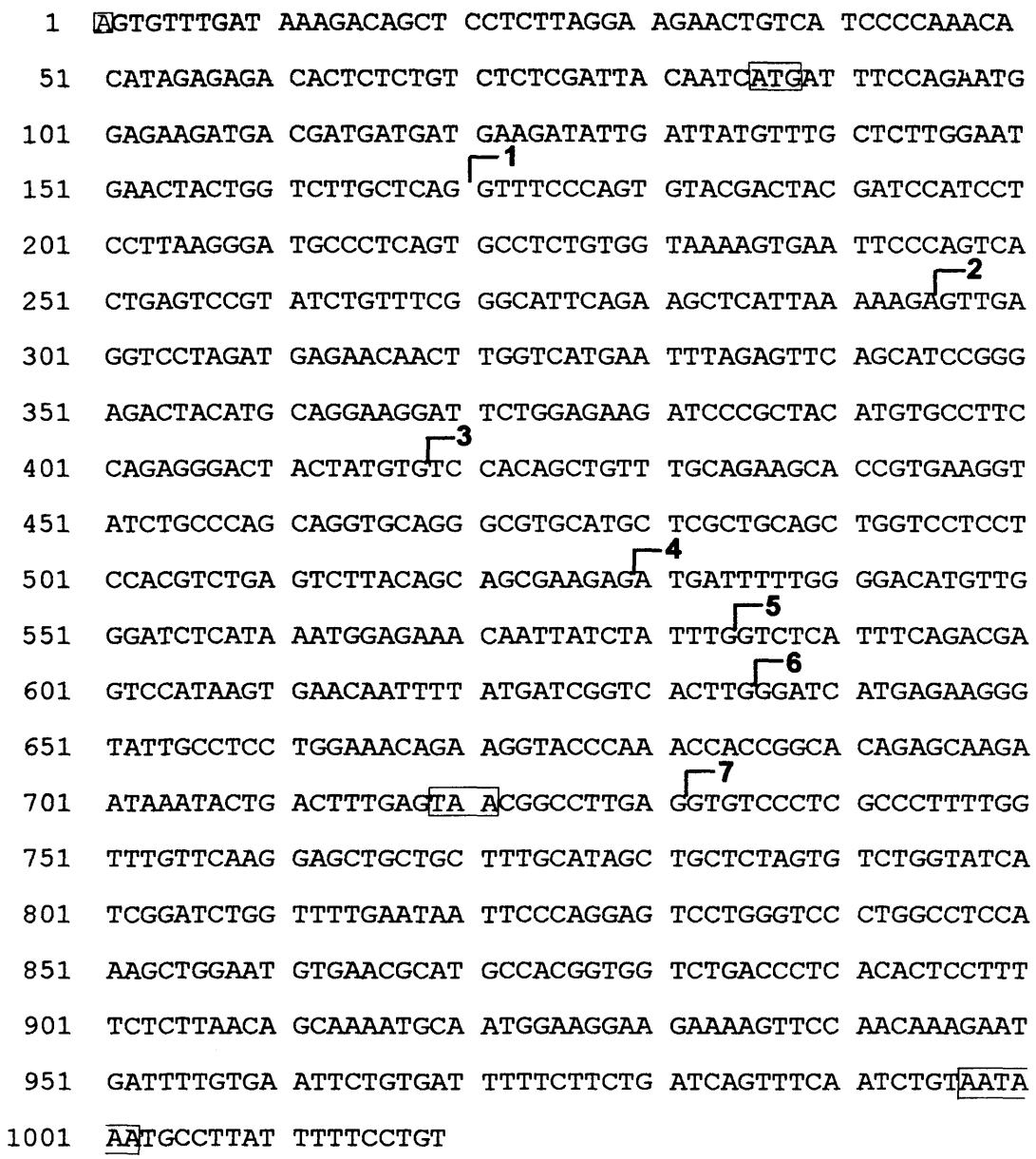


Figure 3.2. The human *SPP2* cDNA

This figure shows the human *SPP2* cDNA. The 'ATG' start codon, the 'TAA' stop codon and the 'AATAAA' polyadenylation signal are boxed. The first base of the cDNA is based on the 5'RACE data and the primary transcription initiation site in liver is marked in red and boxed. This figure also illustrates the exon/intron boundaries of the 8 exons in the human *SPP2* cDNA sequence.

and intron structure, sizes and boundaries were determined by comparing the human spp24 cDNA and the *SPP2* gene. The gene has 8 exons, with all the exon/intron boundaries conforming to the GT/AG consensus at the donor and acceptor splice sites. The ATG start codon is located in the first exon and the TAA stop codon falls in exon 7 (Bennett *et al.*, manuscript submitted) (Figure 3.1, Figure 3.2 and Table 3.2). This location of the stop codon in the penultimate exon is a feature of only 7% of genes, of which 70% encode secreted or cell surface proteins (Nagy and Maquat, 1998), and is consistent with the presence of a putative signal peptide in spp24 (Chapter 1). The sequence of *SPP2* has been annotated and deposited in the EMBL database with the accession number AJ272265.

3.1.4 Sequence variations of *SPP2* gene

Three potential polymorphisms were identified during compilation of the human consensus cDNA sequence. These polymorphisms were at positions 270, 324 and 579 counting from the first base of the start codon (Figure 3.2). Each of these potential polymorphisms falls within the mature peptide-coding region of the cDNA. Two of these (270 and 579) are conservative changes and encode no change in the amino acid sequence, but the other (324) introduces an amino acid change, Asp→ Glu, both of which are acidic residues. Using the GCG map program, the recognition sites of common restriction enzymes site were identified in both the variant and wild type alleles with respect to this sequence dimorphism (Table 3.3). The variant allele contains an additional site for the restriction enzyme *Sca*I. A small-scale study of the polymorphism at position 324 was carried out by Southern blotting (at that time the genomic sequence of human *SPP2* had not been determined and a PCR-based test was not a possibility) using 25 samples of unrelated Caucasian genomic DNAs digested with *Sca*I and probed with a human *SPP2* cDNA. However no polymorphism at this position was found in these samples (Kitchen, 1999).

Exon/Intron number	Exon size (bp)	Intron size (bp)
1	170	99
2	125	7740
3	123	1410
4	111	6053
5	55	636
6	51	2653
7	96	6821
8	288	

Table 3.2 Exon and intron sizes of the human *SPP2* gene

Base number of polymorphism	270	324	579
Base in contig and confirmatory sequence	A Most ESTs	C Most ESTs	A Most ESTs
Alternative base and corresponding sequence	T pB1.1*	G T74678	G T74678
Codon change and corresponding aa residue change (if any)	ACA→ ACT Thr→ Thr	GAC→ GAG Asp→ Glu	GGA→ GGG Gly→ Gly
Restriction sites gained or lost	No change	Gain of *Sca*I, *Tat*I, *Rsa*I, and two *Cje*I sites Loss of *Bsm*FI site	Gain of *Bsa*JI site

Table 3.3. Potential polymorphisms of the human *SPP2* cDNA

* The pB1.1 spp24 cDNA is the longest clone identified in the screen of the human liver cDNA library (Bennett *et al.*, manuscript submitted). T74678 is the accession number of the I.M.A.G.E. cDNA clone number 84837 which encodes spp24.

3.2 Results

3.2.1 Analysis of other previously-identified single nucleotide polymorphisms

dbSNP (Smigielski et al., 2000) is a database which contains descriptions of single nucleotide polymorphisms (SNPs) that have been identified throughout the human genome. The data can be accessed using the Entrez interface at NCBI. The *SPP2* gene sequence (accession number AJ272265), which includes 2319 bp upstream of the transcription initiation site and 1250 bp downstream of the polyadenylation signal, was searched against the data in dbSNP. Thirty seven SNPs were identified and are listed in Table 3.4. One SNP lies in exon 3, with the remaining 36 in introns or 3′ to the polyadenylation signal. None of these 36 SNPs change donor, acceptor or branch splicing sites. The SNP that lies in exon 3 (genomic DNA, rs593668) is identical to the one previously identified as the EST polymorphism at position 270 (in pB1.1 spp24 cDNA clone) (Section 3.1.4). MatInspector (Quandt *et al.*, 1995) was used to determine if any of the introns contained known transcription factor binding sites. Although binding sites were identified, none of the SNPs were located within them. Because none of the 36 non-exonic SNPs lies in an identifiably functional DNA sequence, they have not at present been analysed any further.

3.2.2 PCR amplification of the exons of the human *SPP2* gene

The gene sequence and the exon/intron boundaries within the cDNA and gene have been identified (Bennett *et al.*, manuscript submitted; Figures 3.1 and 3.2; Table 3.2). Therefore, it was possible to design primers for the PCR amplification (Section 2.7.1) of relevant sections (exons 1–7) of the human *SPP2* gene to allow scanning of the amplification products and their eventual sequencing in the event of finding any variants. PCR optimisation was achieved by varying the annealing temperature between 55°C and 65°C in 1°C steps, and by using PCR buffer with $MgCl_2$ concentrations of 1–4.5 mM in 0.5 mM steps. The sequence of each primer, the sizes of fragments after amplification, the optimised annealing temperature and $MgCl_2$ concentration are shown in Table 3.5. The standard cycling conditions used, unless described otherwise, were 30 cycles each comprising 96°C for 30 seconds, annealing temperature (Table 3.5) for 30 seconds, 72 °C for 30 seconds. Figure 3.3 shows an agarose gel of the PCR products from optimised reactions of exons 1–7 of the human *SPP2* gene.

SNP number	Number of Intron/Exon	Base number from the start of the sequence (AJ272265)	Base change & NCBI SNP cluster ID
1	Intron 2	3544	C/A (rs1507516)
2	Intron 2	4034	C/T (rs1507517)
3	Intron 2	4048	A/G (rs1507518)
4	Intron 2	4177	T/C (rs1507519)
5	Intron 2	4409	A/G (rs1354895)
6	Intron 2	4474	T/C (rs1354896)
7	Intron 2	4824	G/T (rs2267900)
8	Intron 2	5264	T/A (rs2267901)
9	Intron 2	5331	C/A (rs3201395)
10	Intron 2	5625	G/A (rs2267903)
11	Intron 2	5805	C/A (rs2267904)
12	Intron 2	5963	T/G (rs617970)
13	Intron 2	6057	G/T (rs633847)
14	Intron 2	6351	G/A (rs1911592)
15	Intron 2	6373	T/C (rs1911593)
16	Intron 2	6415	T/G (rs1911594)
17	Intron 2	6682	C/T (rs1911595)
18	Exon 3	10513	T/A (ACT→ ACA) **Thr→ Thr** (rs593668)
19	Intron 3	10960	C/T (rs689395)
20	Intron 3	11025	C/T (rs2239536)
21	Intron 4	14949	A/C (rs613714)
22	Intron 5	18425	G/A (rs250978)
23	Intron 6	20057	A/G (rs2284294)
24	Intron 6	20575	G/A (rs2284295)
25	Intron 6	21337	G/A (rs250977)
26	Intron 7	21688	C/T (rs2286699)
27	Intron 7	21948	T/G (rs2286698)
28	Intron 7	23852	T/A (rs250975)
29	Intron 7	24035	C/T (rs250976)
30	Intron 7	24512	C/T (rs250974)
31	Intron 7	25725	G/A (rs1507522)
32	Intron 7	25871	G/A (rs1995697)
33	Intron 7	25975	G/A (rs1507522)
34	Intron 7	27805	G/A (rs250971)
35	Intron 7	27835	A/C (rs250970)
36	3′ of the gene	29409	C/G (rs250966)
37	3′ of the gene	29450	A/T (rs1507521)

Table 3.4. SNPs recorded in dbSNP for the human *SPP2* gene

Exon/Exons	Primers 5′→3′	Size of PCR fragment	$[Mg^{+2}]$ (mM)	Ann. T.°C
1+2	Forward: CAGAAATATTGACCCCAGGA Reverse: GACAGCATTGGAAGGAGGAG	511 bp	4.5	61
2	Forward: CTGCTCTGGATCATGCAGAG Reverse: GACAGCATTGGAAGGAGGAG	245 bp	4.5	63
3	Forward: GCTTTCATGGTGGACAATTC Reverse: CATTTCTGGGATGGGTCTC	268 bp	3.5	60
4	Forward: CAATGGAGGCTATCCCTTTCC Reverse: CCTAAGAGGTGGGGTCTGG	217 bp	4	59
5	Forward: TTTCATGTGCTGACACATCC Reverse: AAATGACTCACTAACAAAGAGTTGC	173 bp	4	59
6	Forward: AACATTCTGGAACAGTGAGAGG Reverse: TGATCAGAAAAGGGTCTGGTG	153 bp	4	59
7	Forward: AGAGCCTATGCTTCCCTTTTC Reverse: CAGCAGTTTTAAGGCGTTCAC	199 bp	4	59

Table 3.5. The human *SPP2* primers

The sequences of the primers, the sizes of fragments generated by PCR, the annealing temperature (Ann. T.) and the concentration of magnesium ions in each PCR reaction.

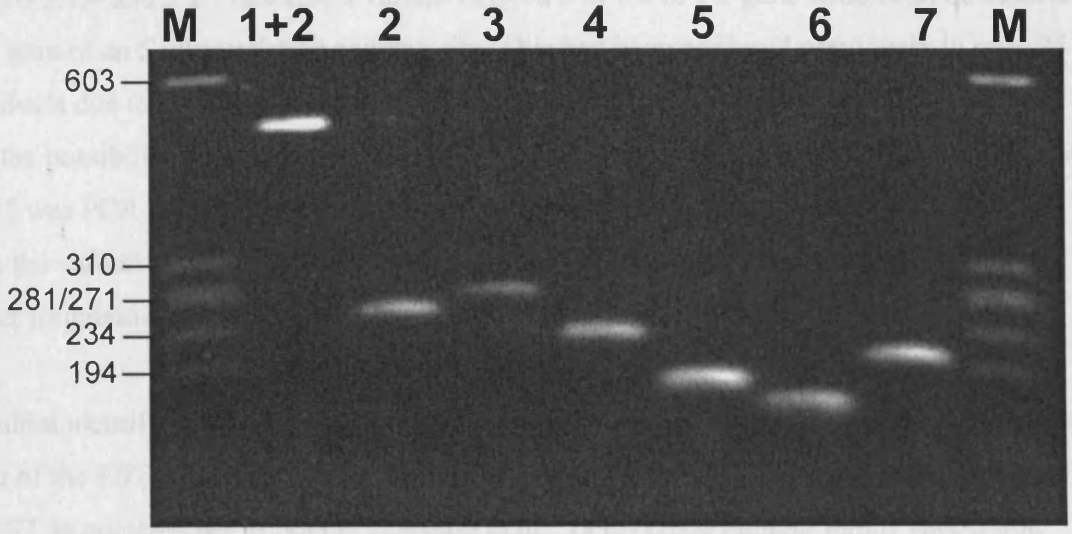


Figure 3.3. PCR amplification of the human *SPP2* exons

The exons of the human *SPP2* gene after PCR amplification and agarose gel electrophoresis.

M = marker (ΦX174RF/*Hae*III)

3.2.3 Investigating the potential variant in exon 3

Sections 3.1.4 and 3.2.1 describe a variant in exon 3 of the *SPP2* gene which can be detected as the gain of an *Sca*I restriction enzyme site. This had been analysed previously in only 25 individuals due to the limitations imposed, at the time, by analysis using Southern blotting. With the possibility of analysis by PCR, larger numbers of individuals could now be screened. Exon 3 was PCR amplified for 100 additional DNA samples from unrelated individuals. When the variant is present, *Sca*I is predicted to cut the PCR product (268 bp) into two smaller fragments (230 and 38 bp). No variant was seen among these 100 samples.

The initial identification of this potential amino acid variant was as a DNA sequence variant of one of the ESTs used to compile the consensus cDNA sequence for spp24 (Section 3.1.4). The EST in question is I.M.A.G.E. (Lennon *et al.*, 1996) clone number 84837 (accession number T74678; sequencing trace yc57h11.r1). Digestion of plasmid DNA prepared from this clone with restriction enzyme *Sca*I confirmed the presence of the variant in accordance with the recommendations of Yang *et al.*, 2000. Hence, it is possible that the G base seen in the cloned sequence is an error arising during the cDNA cloning or that the polymorphism is rare, with an allele frequency less than 1 in 250 (the number of alleles screened in both studies).

3.2.4 Scanning for sequence variations in the human *SPP2* gene

Scanning the exons of the human *SPP2* gene was carried out using the conformation sensitive gel electrophoresis method (CSGE) (Körkkö e*t al.*, 1998) as described in Section 2.11. Exons 1–2, 2 and 3 were PCR amplified and scanned individually, but exons 4, 5, 6 and 7 were PCR amplified (using multiplex PCR amplification) and scanned together. CSGE has advantages over other methods for scanning PCR products which will be discussed in more detail in Section 3.3. Results with respect to specific exons of *SPP2* are presented in the following sections.

3.2.5 Scanning and screening for variations in the exon 3

The first PCR primers to be designed for analysis of *SPP2* were those for exon 3 for the analysis of the potential SNP in that exon (Section 3.2.3). This method has been developed for scanning PCR products for the presence of single-base and larger mismatches in DNA. The assay is based on the assumption that a mildly-denaturing solvent gel in an appropriate buffer can enhance the conformational changes produced by single-base mismatches in double

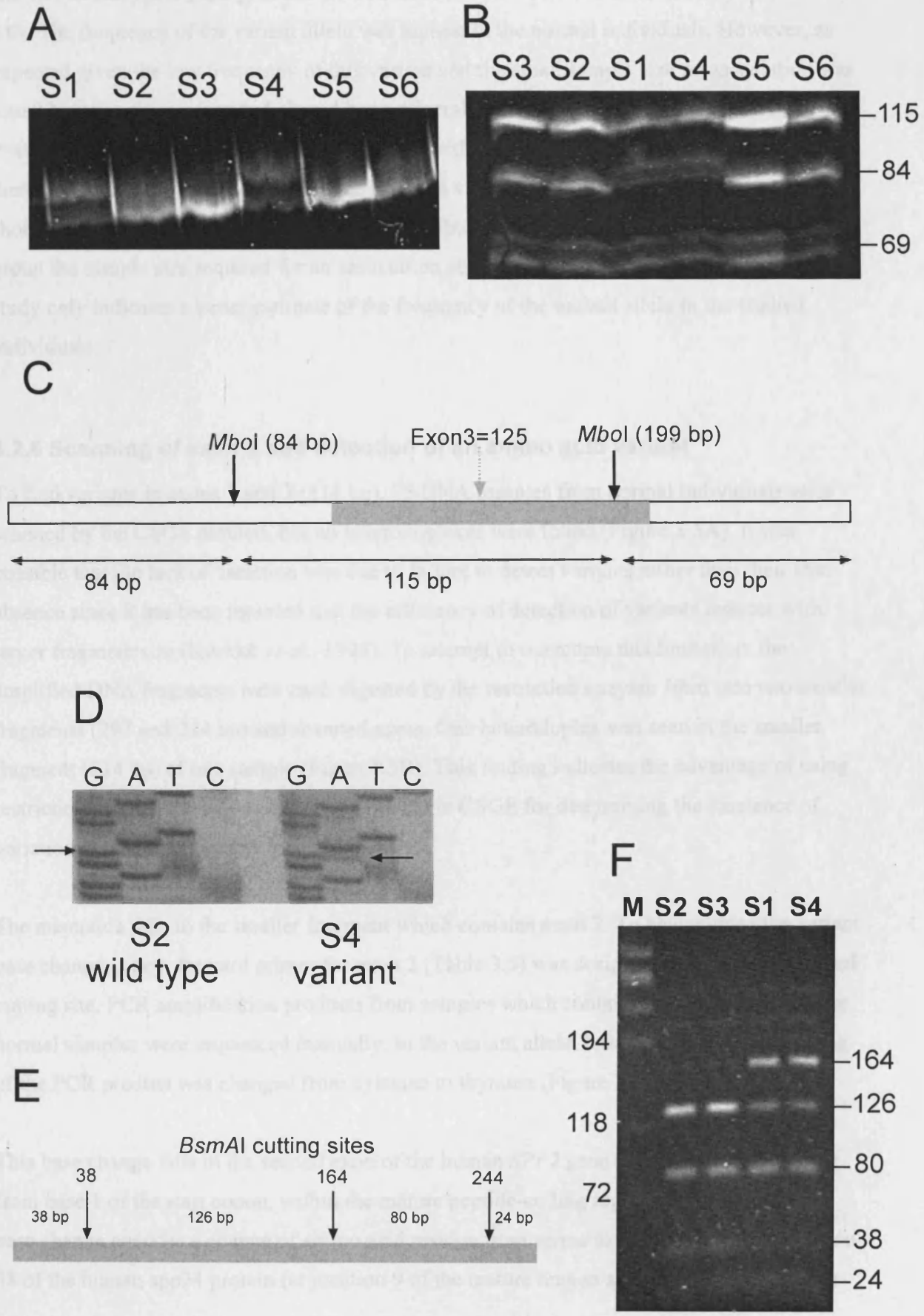
stranded DNA and thereby increase the differential movement and migration in electrophoretic gels of hetero- and homoduplexes (Körkkö *et al.*, 1998; Ganguly *et al.*, 2002). To scan for any additional variants in exon 3, forty two samples from unrelated Caucasians individuals were PCR amplified and scanned by the CSGE method. Among these samples, one common heteroduplex was found (Figure 3.4A, lanes S1 and S4). To determine the approximate site of variation, the *Mbo*I restriction enzyme was used to cut the fragment (268 bp) into three smaller fragments, 84, 115 and 69 bp (Figure 3.4B and C) which were scanned again by CSGE. The variant lies in the 84 bp fragment (second intron of *SPP2* gene) (Figure 3.4B, lanes S1 and S4).

To determine the exact sequence of the variant allele, samples from variant and wild type alleles were sequenced manually (Section 2.10.2). In the variant allele (which is in the second intron) the 46th base from the beginning of the 268 bp fragment is adenine and in the wild type is guanine (Figure 3.4D).

	40 46 52
Wild type	GGTAGA**G**ACAATG
Variant	GGTAGA**A**ACAATG

To determine the frequency of the variant allele, a simpler screening method was needed. By using the map program of GCG, 20 bp of DNA sequence on both sides of the variant allele were selected and restriction enzyme sites were determined. The variant allele does not gain any restriction site, but loses restriction sites for the *Bsm*AI and *Eco*57I restriction enzymes. *Bsm*AI was chosen for screening of the variant allele for reasons of availability. This restriction enzyme cuts the wild type allele PCR amplification product into four fragments (38, 126, 80 and 24 bp), but cuts the variant allele into 164, 80 and 24 bp fragments. Therefore, in homozygotes for the wild-type allele, four fragments (126, 80, 38 and 24 bp) are seen and in heterozygotes five fragments (164, 126, 80 38, and 24 bp) (Figure 3.4E). To determine the variant type allele frequency, 48 DNA samples from unrelated normal individuals were PCR amplified and digested by *Bsm*AI. Seven heterozygote genotypes of the variant allele were seen among these samples, Therefore the allele frequency for variant allele is about 0.073 in the normal population.

Ninety samples from patients with varying bone mineral density were screened for this polymorphism (already described in Section 3.1.1). According to the WHO criteria, 23


A
S1 S2 S3 S4 S5 S6
B
S3 S2 S1 S4 S5 S6
115
84
69
C
MboI (84 bp)
Exon3=125
MboI (199 bp)
84 bp
115 bp
69 bp
D
G A T C G A T C
S2
wild type
S4
variant
E
BsmAI cutting sites
38
164
244
38 bp
126 bp
80 bp
24 bp
F
M S2 S3 S1 S4
194
118
72
164
126
80
38
24

patients were classified as normal, 35 as low BMD and 32 as osteoporotic. Nine heterozygote and two homozygote genotypes for the variant allele were detected in these samples (Table 3.6). The frequency of the variant allele was highest in the normal individuals. However, as expected given the low frequency of this variant and the small sample size no association was found between this variant and altered bone mineral density (1-tailed Fisher exact test, P=0.99). Low BMD is a multifactorial disorder with a continous or quantitative pattern, therefore in any attempt to determine the role of a variant, the distribution of the character should be first determined and based on the distribution of the character in a specific ethnic group the sample size required for an association study should be calculated. Therefore, this study only indicates a better estimate of the frequency of the variant allele in the studied individuals.

3.2.6 Scanning of exon 2 and detection of an amino acid variant

To find variants in exons 1 and 2 (511 bp), 75 DNA samples from normal individuals were scanned by the CSGE method, but no heteroduplexes were found (Figure 3.5A). It was possible that the lack of variation was due to failure to detect variants rather than their true absence since it has been reported that the efficiency of detection of variants reduces with larger fragment size (Körkkö *et al.*, 1995). To attempt to overcome this limitation, the amplified DNA fragments were each digested by the restriction enzyme *Hha*I into two smaller fragments (297 and 214 bp) and scanned again. One heteroduplex was seen in the smaller fragment (214 bp) of one sample (Figure3.5B). This finding indicates the advantage of using restriction enzyme digestion of large fragments in CSGE for determining the existence of variation and its approximate location.

The mismatch falls in the smaller fragment which contains exon 2. To characterise the variant base change, a new forward primer for exon 2 (Table 3.5) was designed upstream of the *Hha*I cutting site. PCR amplification products from samples which contained the variant and other normal samples were sequenced manually. In the variant allele, base 102 from the beginning of the PCR product was changed from cytosine to thymine (Figure 3.5C).

This base change falls in the second exon of the human *SPP2* gene at position 113 counting from base 1 of the start codon, within the mature peptide-coding region of the cDNA. This base change encodes a change of amino acid residue from serine to phenylalanine, at position 38 of the human spp24 protein (at position 9 of the mature human spp24 protein). In bovine

Groups	Homozygote (wild type allele)	Heterozygote	Homozygote (variant allele)
Normal people (n=23)	19	3	1
Patients with low BMD (n=35)	32	2	1
Patients with very low BMD (n=32)	28	4	0
Pooled data for reduced BMD (n=67)	60	6	1

Table 3.6. Number of different genotypes (G/A variation in intron 2) among 90 people with varying BMD

Figure 3.5. Results of scanning and screening of the fragments which contain exons 1 and 2 of the human *SPP2* gene

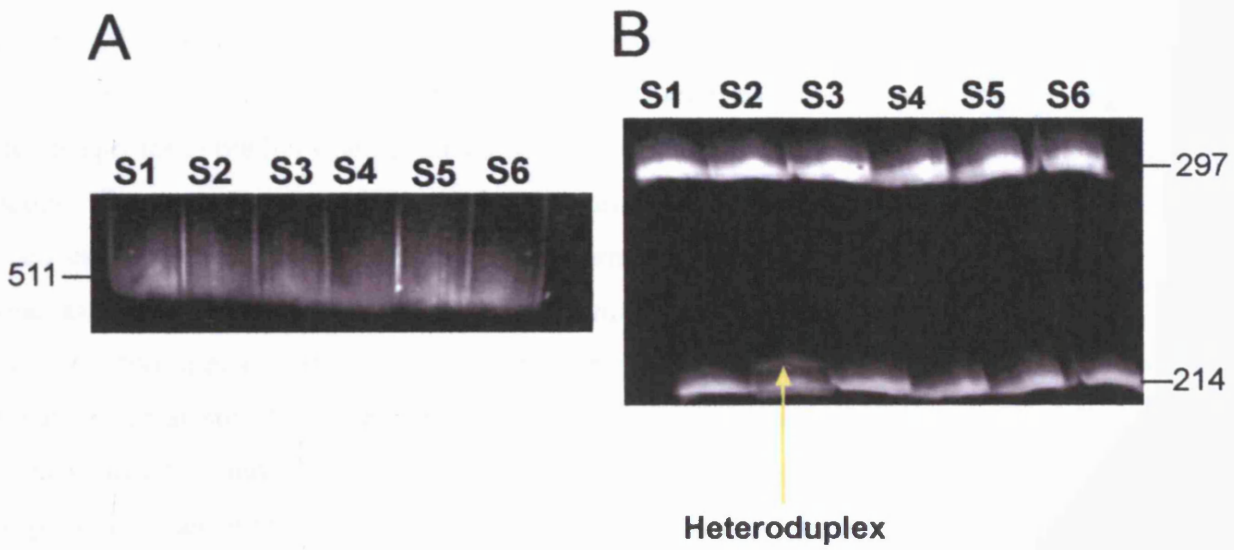
A: Result of CSGE (20% polyacrylamide gel) of the 511 bp PCR fragment which shows no mismatch in samples 1– 6

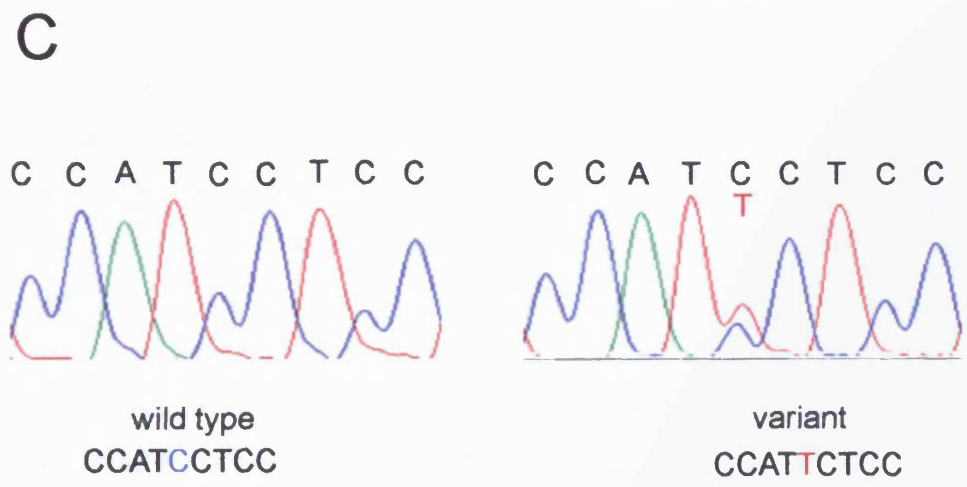
B: Result of CSGE (20% polyacrylamide gel) of the 511 bp PCR fragment after restriction digestion by *Hha*I (297 and 214 bp) which shows a heteroduplex in the 214 bp fragment (which contains exon 2) of sample S2.

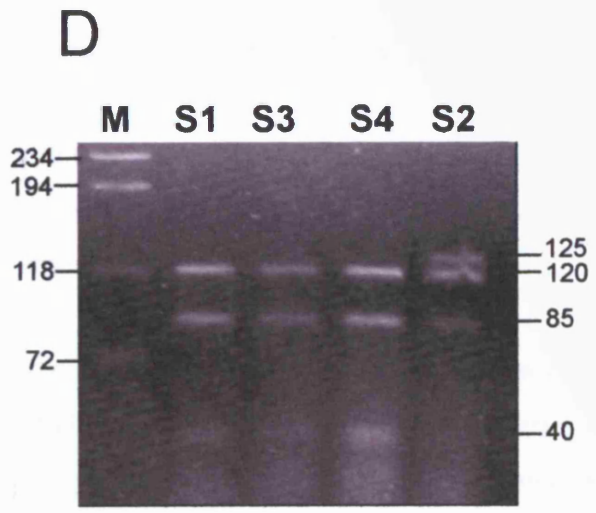
C: Result of automated sequencing of sample which contain variant allele (sample S2) and samples which contain wild type allele.

D: Result of screening of exon 2 by *Fok*I restriction enzyme, which shows the heterozygote genotype of variant allele (sample S2) and homozygote genotype of the wild type allele (other samples).

M = marker (ΦX174RF/*Hae*III)


A
S1 S2 S3 S4 S5 S6
511
B
S1 S2 S3 S4 S5 S6
297
214
Heteroduplex


C
C C A T C C T C C
wild type
CCATCCTCC
C C A T C C T C C
T
variant
CCATTCTCC


D
M S1 S3 S4 S2
234
194
118
72
125
120
85
40

spp24 there is an alanine at this position. Serine has an uncharged polar R group, while alanine and phenylalanine have non-polar R groups. Phenylalanine's R group is bulkier than that of serine or alanine. Due to the fact that amino acid substitutions can potentially affect protein function, the significance of the serine to phenylalanine substitution was assessed using the SIFT (Sorts Intolerant From Tolerant) computer program (Ng and Henikoff, 2001). This program is an amino acid sequence-based tool which uses sequence homology among different species to predict whether an amino acid substitution affects the protein structure and function. The program carries out a search of protein sequence databases for related sequences, and identified spp24 sequences for mouse, rat and pig in addition to those for human and bovine. An alignment of the protein sequences is shown in Figure 3.6. On the basis of the five spp24 protein sequences, SIFT predicated that the substitution at position 9 of the mature human spp24 (or at position 38 of the precursor protein, p.S38F) would induce secondary structure and possible functional changes. The significance of this amino acid change is discussed at much greater length in Chapter 4.

To find the allele frequency of the variant and for larger-scale screening, restriction sites for restriction enzymes in both the wild and variant alleles were determined using the map program of GCG. The variant allele does not gain any restriction site, but loses five restriction sites for the *Fok*I, *Mnl*I, *Eco*NI, *Bsl*I and *Bst*F5I restriction enzymes. *Fok*I restriction enzyme was chosen (because of its availability) for screening of the variant allele among the population. This restriction enzyme cuts the wild type allele fragment into three smaller fragments (85, 40 and 120 bp), but cuts the variant allele fragment into two fragments (125 and 120). Therefore, in the heterozygote genotype four fragments (125, 120, 85, and 40 bp) are seen (Figure 3.5D). To confirm the result of scaning by CSGE the same 75 DNA samples from unrelated normal individuals were were PCR amplified and digested by *Fok*I. Only 1 heterozygote genotype was found and the result of CSGE was confirmed.

To investigate the frequency of the variant in the individuals with varying BMD, ninety samples from people with different bone mineral density (Section 3.2.4) were screened, but no variant allele was found among these people.

Further screening was carried out on 120 samples from unrelated psoriatic Scottish patients (DNA samples provided by Richard Trembath, Department of Genetics, University of Leicester). Two persons heterozygous for the variant allele were found (allele frequency ~0.0083).

A

```
                    96        102       108
Wild type           ATCCATCCTCCTT
Variant             ATCCATTCTCCTT
```

B

```
                        9
Mouse       FPVYDYDPSSLQEALSASVVKVNSQSLSPYLFRAT
Rat         FPVYDYDPSSLQEALSASVVKVNSQSLSPYLFRAT
Human       FPVYDYDPSSLRDALSASVVKVNSQSLSPYLFRAF
Bovine      FPVYDYDPASLKEALSASVVKVNSQSLSPYLFRAF
Variant     FPVYDYDPFSLRDALSASVVKVNSQSLSPYLFRAF
Pig         FPVYDYDPSSLREAVGASVVKVNSQSLSPYLFRAF
```

Figure 3.6. Amino acid change in exon 2 of the human *SPP2* gene

A: This figure illustrates the nucleotide change (C/T) at position 102, counting from the start of the PCR product, which changes the amino acid serine to phenylalanine.

B: The mouse, rat, human, bovine, pig and variant allele of spp24 protein (aa 1-35 of the mature peptide), which indicates the highly conserved regions (boxed sequences) among these species and the variant allele.

If we consider all the samples that were screened for this variation, 285 individuals of Caucasian origin were screened and three individuals heterozygous for the variant were identified, and no homozygotes. Therefore the allele frequency is about 0.0053 in the population and there is an expectation of approximately 1 person in 36,000 being homozygous for the allele if the sample population is in Hardy-Weinberg equilibrium.

3.2.7 Scanning for variation in exons 4, 5, 6 and 7

By designing appropriate primers (Table 3.5) with similar annealing temperatures that would amplify products with reasonable length differences, exons 4, 5, 6 and 7 were amplified simultaneously. Therefore, by multiplex PCR, these four exons can be scanned together (Figure 3.7). PCR amplification was carried out as described in Section 2.7.2.

For scanning these exons, 144 DNA samples from unrelated individuals (mostly northern European), were amplified by the multiplex PCR method. Exon 5 was amplified with variable efficiency from sample to sample (Figure 3.7) even after optimising the Mg^{2+} ion concentration and annealing temperature. Because of this, exon 5 was also amplified and analysed separately as well as in the multiplex reaction. Using the CSGE method, no heteroduplexes or mismatches was found in these samples.

3.2.8 Comparing and confirming the sensitivity and specificity of CSGE to denaturing high performance liquid chromatography (DHPLC)

The fact that only two variants were detected during the analysis of the protein-coding exons of *SPP2* (one in exon 2 and one in intron 2), is cause for concern. It is possible that more variants exist but have not been detected by CSGE. It is certainly the case that individuals homozygous for variants would only be detected very inefficiently (if at all) as CSGE depends on the formation of heteroduplexes formed between DNA strands form different alleles. To an extent, this can be overcome by mixing all tested DNA samples with a reference DNA sample prior to amplification to ensure that heteroduplexes will be formed even when the tested DNA is homozygous for a variant. An alternative to this strategy is to use a mutation detection method that does not rely solely on the formation of heteroduplexes. Denaturing high performance liquid chromatography (DHPLC) can reveal the presence of variation by differential retention of hetero- and homoduplex DNA under partial denaturation during reverse phase column chromatography. In principle, DHPLC is capable of distinguishing homoduplexes that differ only by a single base from one another. The ability of

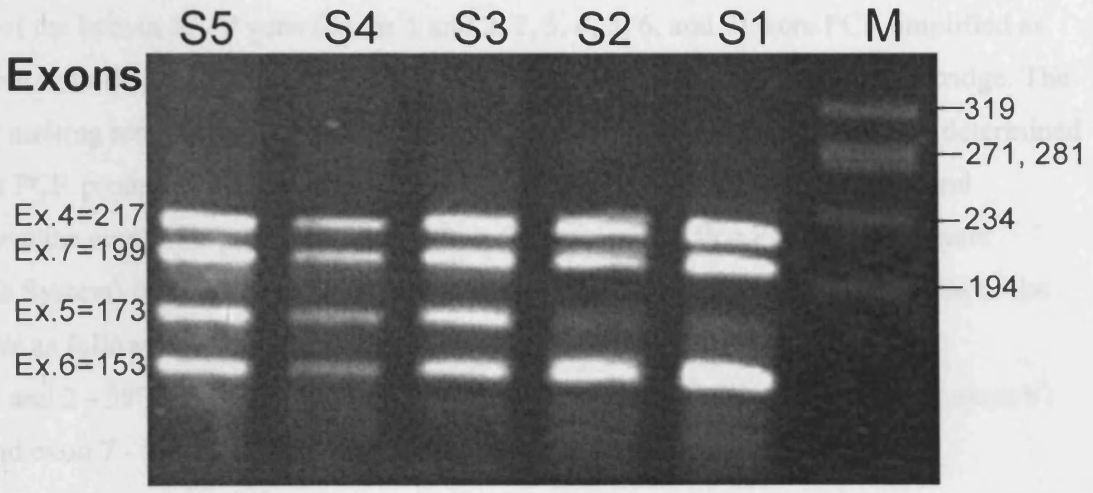


Figure 3.7. Multiplex PCR

This figure shows the multiplex PCR of exons 4, 5, 6 and 7 of the human *SPP2* gene. Marker DNA fragment sizes are shown on the right, and the PCR fragment sizes on the left of the gel image.

M = marker ΦX174RF/*Hae*III

the DHPLC to resolve heteroduplexes from homoduplexes in minutes, and the possibility that it can distinguish nearly-identical homoduplexes, makes it a powerful tool in the field of mutation detection (Cotton, 1997; Underhill *et al.*, 1996). Figure 3.8 schematically illustratest the result of DHPLC for a heterozygote variant.

Samples previously analysed by CSGE were re-analysed by DHPLC. All exons and flanking introns of the human *SPP2* gene (exons 1 and 2, 2, 3, 4, 5, 6, and 7) were PCR amplified as before but with BSA omitted from the reaction as this can damage the DHPLC cartridge. The optimal melting temperature for partial denaturation on the DHPLC apparatus was determined for each PCR product by varying the melting temperature in the range 6°C below and 6°C above the computer-predicted temperature (Transgenomic WAVE DNA Fragment Analysis System) in 1°C steps. Optimal melting temperatures for partial denaturation of the exons are as follows:
exons 1 and 2 - 59°C, exon 2 - 56°C, exon 3 - 56°C, exon 4 - 58°C, exon 5 - 54°C, exon 6 - 57°C and exon 7 - 57°C.

The same 90 samples from individuals with known BMD and 48 samples from individuals with unknown BMD described in Section 3.2.5 were re-analysed with respect to each exon by DHPLC. The results did not reveal any further variants.

3.2.9 The bovine spp24 cDNA sequence confirmation

The amino acid change of serine to phenylalanine in human exon 2 falls in one of the highly conserved regions of the spp24 amino acid sequence in human, mouse, rat, cattle and pig (Figure 3.6A and B). The amino acid serine falls at this position in all these species except cattle, where the amino acid is alanine. It was possible that the presence of alanine in the cattle spp24 protein sequence might have been due to an error in the original cDNA sequencing.

Due to the fact that the size of intron 1 of the *SPP2* gene is 99 bp in human and 100 bp in mouse (see Chapter 4), it was speculated that it might be similarly short in the bovine gene. If so, it would be feasible to design PCR primers based on the bovine cDNA sequence to allow amplification across exons 1 and 2 from bovine genomic DNA. Two primers (For2: 5′ GAGAAGTGGCGATGAAGATG 3′and Rev2: 5′ TGACTGGGAATTCACTTTTGC 3′) were designed in the presumed 5′ end region of the

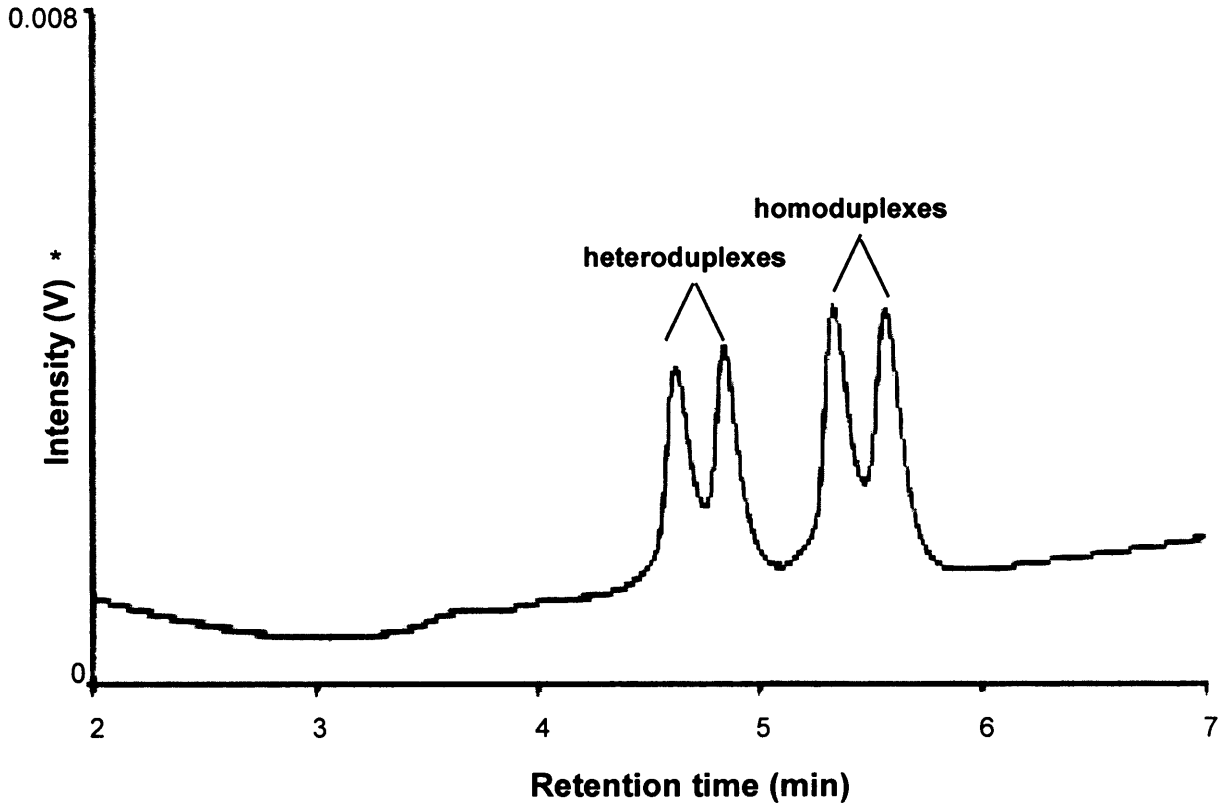


Figure 3.8. Analysis of heteroduplexes and homoduplexes using DHPLC method by the Transgenomic WAVE DNA Fragment Analysis System

This figure schematically illustrate the result of DHPLC for a heterozygote variant. Separation of heteroduplexes from homoduplexes is accomplished under partially denaturing conditions.

*The signal intensity of DHPLC profiles is shown in mV or V at A_{260}.

first bovine exon and the presumed 3′ end region of the second bovine exon using the location of the mouse and human introns as a guide. PCR amplification was carried out as described in Section 2.7.1 for 30 cycles each comprising 96°C 30s, 60°C 30s, 72°C 30s.

The 244-bp amplified PCR product, containing parts of exons 1 and 2 and the entire intron 1, was then purified and sequenced in the forward and reverse directions as described in Sections 2.5.3 and 2.10.1 respectively. It was determined that alanine falls at the position corresponding to the variant human serine/phenylalanine amino acid, thus confirming the original cattle cDNA sequence.

3.2.10 Computer-based analysis of the spp24 protein phosphorylation

Serine is an amino acid that is commonly phosphorylated in proteins. Substitution of the serine at amino acid position 38 (serine 38) might remove a potential phosphorylation site on spp24 and/or might influence the ability of the adjacent serine at amino acid position 39 (serine 39) to act as a phosphorylation site.

The NetPhos protein phosphorylation computer program (Blom *et al.*, 1999; http://www.cbs.dtu.dk/services/NetPhos) produces predictions for serine, threonine and tyrosine phosphorylation sites in eukaryotic proteins. This program predicted 13 phosphorylated serine residues, 4 phosphorylated threonine residues and 3 phosphorylated tyrosine residues in the human spp24 protein. These residues are indicated in Figure 3.9.

Most of these predicted phosphorylated residues fall in the cystatin-like and serine-rich regions of the spp24 protein. The prediction in the serine-rich region is supported by the degree of phosphorylation identified experimentally in the same region of the bovine spp24 protein (Hu *et al.*, 1995). The phosphorylated residues predicted in the non-cystatin-like region of the spp24 protein are not conserved residues between species (see Chapter 4) and so are unlikely to be significant to the function of the protein. However, seven of the nine residues predicted to be phosphorylated in the cystatin-like region of the human spp24 protein are conserved between human, bovine, sheep, mouse, rat and pig, one of these being serine 39. This suggests that they are functionally important residues and thus the extent of phosphorylation could also be important to the protein function. With respect to the variant protein sequence, serine 38 is predicted not to be a substrate for phosphorylation and its substitution by phenylalanine is not predicted to diminish the potential to phosphorylate serine

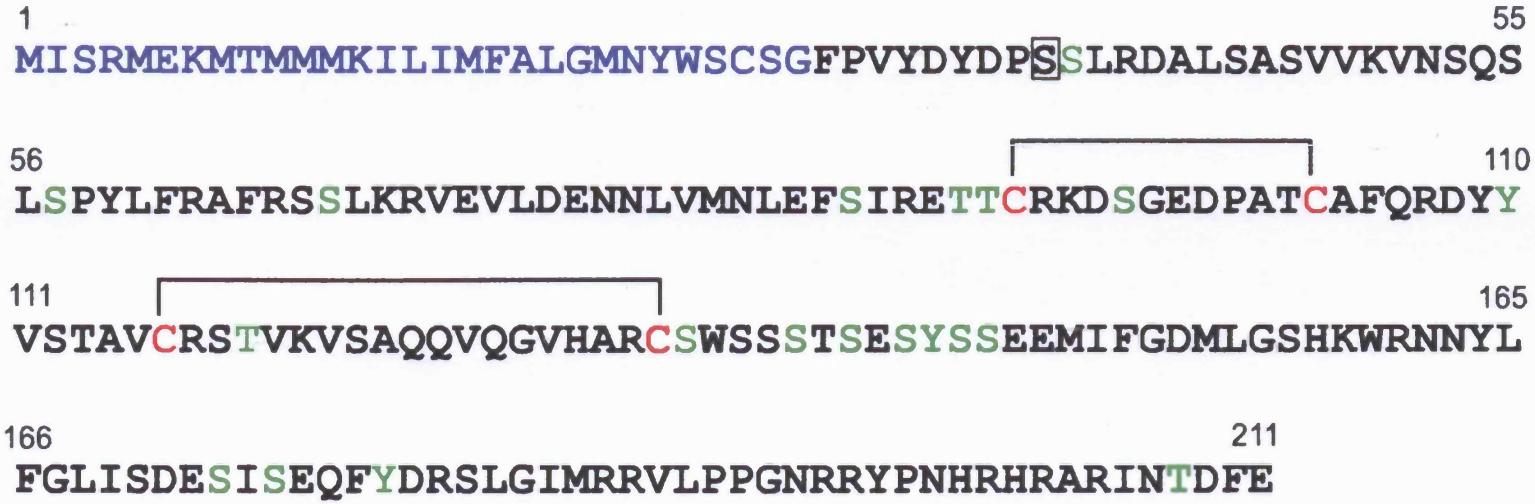


Figure 3.9. Phosphorylation of the human spp24 protein

The phosphorylation prediction program NetPhos predicted 13 phosphorylated serine residues (S), four phosphorylated threonine (T) and three phosphorylated tyrosine residues in the human spp24 protein. Green coloured residues indicate that the program NetPhos predicted that the residue would be phosphorylated. The first 29 amino acid residues (signal peptide) are shown in blue and mature peptide in black. The two square brackets show the disulphid bonds and the cysteine residues are shown in red and the amino acid residue serine at which the substitution to phenylalanine was found (section 3.2.6) is boxed. The cystatin-like domain consists of residues 30 to 136.
The phosphorylated region runs from residues 139 to 146 and the non-cystatin-like domain consists of residues 147 to 211.

39. Hence, if the substitution of serine 38 by phenylalanine is disease-causing, its effects are likely to be through a direct alteration of the three-dimensional structure of spp24 rather than by changes to its phosphorylation.

3.3 Discussion

Several major single-nucleotide polymorphism (SNP) discovery projects have been launched. dbSNP is a database which contains descriptions of SNPs have been found throughout the human genome (the data can be accessed using the Entrez interface in NCBI). So far, few of these SNPs have been validated. Moreover, most validations have been carried by re-sequencing the same cloned DNA sample used for its initial discovery. This approach checks the experimental artefacts in the discovery process, but does not control for DNA cloning artefacts nor does it provide any information about allele frequency in the population. Some genuine SNPs will be very rare in the population and are unlikely to be observed again in repeat samples from that, or another, population (Yang *et al.*, 2000).

The important point that must be considered is that none of the variants of the human *SPP2* gene recorded in dbSNP have been validated. Therefore, to use these variants in genetic studies (if they are real and of sufficiently high enough frequency) in a specific ethnic group, their frequencies should be determined, because if they are rare then we may have difficulty in using them in association studies (because of insufficient statistical power).

The ability to detect single-base changes (mutation detection) is of fundamental importance in molecular genetics. This is especially true in human genetics, where interest in linking mutations of identified genes to particular diseases is most crucial. For the great majority of diseases there is extensive allele heterogeneity, and genetic testing requires a search for mutations anywhere within or near the relevant gene. The biggest current problem in laboratory genetic diagnosis is the lack of any quick, cheap and reliable method for doing this. All the different methods have their advantages and disadvantages, from the point of view of a researcher or a diagnostic laboratory. Sequencing, single-strand conformational polymorphism (SSCP) and heteroduplex analysis have been the methods most frequently used. Most scanning methods for finding a new variation or mutation involve analysis of a heteroduplex formed between a single strand of the DNA being examined and the complementary strand of DNA that does not have any mutation (Cotton, 1997).

After studying different methods of heteroduplex analysis, and using some of them practically, conformation sensitive gel electrophoresis (CSGE), with a few minor modifications (Section 2.11), was finally chosen for experimental use.
The advantages of CSGE include:

- It is very simple and does not need any complex equipment or facilities
- It requires no special preparation of the PCR product
- It can handle a large number of samples
- It can simultaneously scan several DNA fragments from multiplex PCRs
- It does not require the use of radioactive substances
- It is inexpensive
- Its interpretation is very simple

The disadvantages of CSGE include:

- It cannot reveal mismatches which fall in the middle of fragments larger than 400–500 bp (like other heteroduplex analysis methods). This limitation can be overcome using restriction digestion to cut the PCR products into smaller fragments prior to analysis
- It does not reveal the position of any base change

The sensitivity and specificity of DHPLC has previously been studied. Gross *et al.*, 1999 in a study that compared the BRCA1 mutation analysis by direct sequencing, SSCP (Single Strand Conformation Polymorphism) and DHPLC showed that DHPLC could resolve 100% of DNA alterations that had been confirmed by DNA sequencing. In contrast, mutation analysis by SSCP detected only 94% of these variants. In this study, there was complete correspondence between the results obtained by CSGE and DHPLC. Therefore it was concluded that in our study CSGE is as good as DHPLC to scan for new variations, screen for known variants and it provides a powerful, cost-effective way to scan genes with high sensitivity and specificity though this may not be true of other regions of the genome depending on amplicon etc.

Osteoporosis is a multi-factorial disease to which many factors are known to contribute. Valuable information was available for 90 individuals including age, body mass index (BMI) and number of years since cessation of menstruation. The averages of each of these parameters were calculated for each group and it was determined that the average age was higher in patients with osteoporosis or low BMD, the average BMI was lower and the average number of years since cessation of menstruation was higher. All of these are known to be risk factors in osteoporosis. The sample classes have not been equally matched; for example the average age in the normal control group (52.6, SD=3.63) is not matched to the average age of those in the low BMD (53.8, SD=3.26) and osteoporotic group (54.63, SD=3.23). The life style, job, previous medical history, ethnic origin and quality of nutrition in each individual person are very important factors that should also be matched and considered in any

association study. There is no information regarding these important factors. Therefore, for any association observed, it is commonly impossible to deduce unequivocally the parameter with which the variant is associated.

Association studies have limitations that should be considered in any study, but also have the potential to be more powerful than linkage studies. Association studies should be interpreted cautiously because many factors can produce spurious associations between a disease and a potential risk factor or variation. There are three main reasons why an association might be seen in a study. Either the variation itself is causative of the disease, the variation is in linkage disequilibrium with the real cause, or the variation is more common in the studied population (in different ethnic groups, allele frequency can vary considerably). To be able to differentiate between these three reasons, and avoid any bias, the association study must have good statistics and consequently a well-matched control group.

By careful design of a study, it is possible to overcome many potential pitfalls of the association studies. The common errors in association studies are (Cardon and Bell, 2001):

- Small sample size
- Random error
- Control group poorly matched
- Failure to replicate the results
- Failure to detect linkage disequilibrium between adjacent loci
- There is analysis of subgroups or there are multiple tests carried out on the same samples
- Over-interpretation of the results by researchers introducing a positive bias

Little could be deduced from the association study presented in this chapter given the very low frequency of the variant allele and the small sample size. This analysis really only indicates a better estimate of the frequency of the variants in the Caucasian population. To perform a worthwhile study, many more samples would need to be analysed. The rarer the variant to be studied, the larger the number of samples and controls that need to be analysed. For example, if we want to study the effect of the variant allele, which has a frequency of 0.0077 (three heterozygote genotypes in 195 normal individuals), in a polygenic dichotomous character (a character like polydactly, which some people have and others do not, as opposed to a continuous or quantitative character like height) the following numbers of samples are needed:

with an odds ratio of 1.5 ($\alpha = 0.05$, power = 0.8) 10,828 individuals
with an odds ratio of 2 ($\alpha = 0.05$, power = 0.8) 3,360 individuals
with odds ratio 4 ($\alpha = 0.05$, power = 0.8) 665 individuals
To determine the role of the variant allele in any continuous or quantitative character (such as osteoporosis) using an association study, the distribution of the character (mean and standard deviation of BMD) in the normal population of interest and osteoporotic individuals should be first determined (http://www.mc.vanderbilt.edu/prevmed/ps/). To address this, a study will be carried out in collaboration with GlaxoSmithKline (GSK) using a cohort of approximately 3000 DNA samples from peri-menopausal women from Aberdeenshire.

In summary, two variations in the human *SPP2* gene were identified that are potentially valuable in future association studies. Such studies should be carried out cautiously to ensure significant, reproducible results. Secondly, the potential of the amino acid substitution variant in exon 2 to alter the function of the spp24 protein needs to be assessed.

Chapter 4

Sequence Determination and Analysis of the Mouse *Spp2* Gene and Comparison of the spp24 Protein in Different Species

4.1 Introduction

Structural characterisation of a gene can provide valuable information about the encoded protein, the expression of the gene and its regulation. Before performing any study about the function of a gene, such as expression of the protein or mouse knockouts, it is also crucial to know its structure.

The homology of a protein to other known related proteins can provide clues to possible protein function and to identify highly conserved residues that may be functionally important. Homologies between proteins can also enable us to predict the 3D-structure of a protein based on the 3D-model of an existing homologue. In this way, protein homologies can be used to predict the function, functionally important residues and to predict a possible model for the protein and its mode of action. Although these comparisons cannot replace practical experiments, providing conclusive evidence in themselves, they can guide us in the right direction for functional investigation and may form the basis for proposing different theoretical mechanisms.

The original report of spp24 (Hu *et al.*, 1995) presented the bovine cDNA sequence and the deduced protein sequence, but the structure of the bovine gene was not determined. This chapter presents the determination and detailed analysis of the mouse *Spp2* gene and its exon/intron structure. Among the genetically well-explored organisms, mice are one of the closest to human in evolutionary terms therefore the structure of the mouse gene is obviously of importance in determining the role of spp24 in human health and disease. The structure of the mouse gene is also crucial to enable a gene knockout in mice to be made in the future.

4.1.1 The human gene encoding secreted phosphoprotein 24

Details of the isolation of human cDNA and genomic DNA clones and their analyses are described in Section 3.1.3.

Northern blot analysis using an spp24 cDNA showed the presence of spp24 mRNA in bovine bone and liver (Hu *et al.*, 1995). In human, expression of the *SPP2* gene has been confirmed in liver and also detected in kidney (Bennett *et al.*, manuscript submitted). Using the primer extension method, it has been shown that the transcription start sites in liver and kidney are different, with the former having a transcript that is 24 bases longer at the 5′ end. Multiple transcription start sites are common in genes with a TATA-less promoter.

No functional elements that might influence gene expression were identified in the 5′-UTR or 3′-UTR of the human mRNA. Using different promoter-identification programs no conventional promoter containing either a CAAT or TATA box was identified. The GC content of the 1 kb upstream to the initiation sequence was very low (about 38%) and no CpG island was identified either. Genes which are ubiquitously expressed usually have promoters with high GC content and a TATA box. The promoters of tissue-specific genes tend to have a lower GC content and sometimes lack a TATA box (Itoh *et al.*, 1999). Hence, these data are compatible with the previous data about the expression of the gene encoding spp24 in different tissues.

4.1.2 The mouse gene encoding secreted phosphoprotein 24

Previous studies in this laboratory identified 57 mouse ESTs that showed strong homology to the bovine spp24 cDNA (accession number U03872). These ESTs were aligned using the GCG program "pileup" and manually edited in the GCG SeqLab editor. The consensus mouse *Spp2* cDNA was submitted to EMBL with the accession number AJ315513 (Figure 4.1). The mouse gene encoding the spp24 protein has been assigned the symbol *Spp2* by the Mouse Genome Database (MGI: 2177623).

A mouse genomic PAC library RPCI21 (Osoegawa *et al.*, 2000) containing 254,217 clones with an insert of average size of 137 kb in the vector pPAC4 was screened for the mouse *Spp2* gene (Manship and Dalgleish, unpublished) and several positive clones were identified carrying the mouse *Spp2* gene. Subsequently, two of the PAC clones known to contain the *Spp2* gene were used to make a small-insert sub-clone library that was screened for inserts containing CA di-nucleotide repeats (to use those subclones in mapping of the mouse *Spp2* gene by backcross method) (Swallow and Dalgleish, unpublished). Some of the exon/intron boundaries of the mouse *Spp2* gene were identified by analysis of preliminary mouse genomic sequence from the NCBI sequencing trace archive (Bennett and Dalgleish, unpublished).

```
  1  ACAAGAATAA GACAGCCACC CTCTGAAAGA GCTGTCATCC AGAAGCCTGG
 51  AGAGAGGCCG TCTCCCTGAC TCTGGGTCGC CATCCTCTCA GTATGGAGCA
101  GGCAATGCTG AAGACGCTGG CTTTGTTGGT GCTGGGCATG CACTACTGGT
151  GTGCCACAGG TTTCCCGGTG TACGACTACG ACCCTTCCTC TCTGCAGGAA
201  GCTCTCAGTG CCTCAGTGGC AAAGGTGAAC TCGCAGTCCC TGAGTCCTTA
251  CCTGTTTCGG GCGACCCGGA GCTCCTTGAA GAGAGTCAAC GTCCTGGATG
301  AAGACACATT GGTCATGAAC TTAGAGTTCA GTGTTCAGGA AACCACATGC
351  CTGAGAGATT CTGGTGATCC CTCCACCTGT GCCTTCCAAA GGGGCTACTC
401  TGTGCCAACA GCTGCTTGCA GGAGCACTGT GCAGATGTCC AAGGGACAGG
451  TAAAGGATGT GTGGGCTCAC TGCCGCTGGG CGTCCTCATC TGAGTCCAAC
501  AGCAGTGAGG AGATGATGTT TGGGGACATG GCAAGATCCC ACAGACGAAG
551  AAATGATTAT CTACTTGGTT TTCTTTCTGA TGAATCCAGA AGTGAACAAT
601  TCCGTGACCG GTCACTTGAA ATCATGAGGA GGGGACAGCC TCCCGCCCAT
651  AGAAGGTTCC TGAACCTCCA TCGCAGAGCA AGAGTAAATT CTGGCTTTGA
701  GTGACATCCT GGAGATTTCA TGAAAGAAAG AGAAGCAGAA GCTGAAATGA
751  AGAAAGGCAT GGAGAATGGT GTCTTTTTCC TTTTTATAAT CTCCACTCTG
801  CAATAAAGAT CTTTCCCTTC CTTTAAAAAA
```

Figure 4.1. The consensus mouse *Spp2* cDNA determined by the alignment of mouse ESTs

The 'ATG' start codon, 'TGA' stop codon, polyadenylation signal and the start of the poly A tail (AAAAAA) are boxed. This figure also illustrates the exon/intron boundaries of the 8 exons in the mouse Spp2 cDNA sequence. This sequence has been submitted to EMBL with the accession number AJ315513.

At that time, the complete genomic mouse *Spp2* sequence was not available. This chapter describes the compilation of the mouse *Spp2* gene sequence from the preliminary Mouse Genome Sequencing Project data, closure of a gap in the sequence data from that project and extensive analysis of the entire mouse *Spp2* gene. The chapter also describes sequence comparisons of the spp24 protein in human, mouse, rat, cattle, sheep, pig, chicken and salmon.

4.2 Results

This section presents the identification of the introns flanking each individual exon and an extensive sequence analysis of the mouse S*pp2* gene allowing the annotation of the DNA sequence. This was submitted to EMBL and allocated the accession number CAAA01218969. The annotations of CAAA01218969 are presented in Appendix A.

4.2.1 Identification of mouse *Spp2* DNA traces containing individual exons and flanking introns

All of the work presented here was carried out using the programs pileup and SeqLab in the GCG molecular biology package and BLAST (Section 2.29.2).

Although the boundaries between the individual exons of the mouse *Spp2* gene had previously been identified, little sequence data for the introns flanking these exons had been compiled. Consequently, there were insufficient data for the design of PCR primers to amplify these exons and their immediate flanking intron sequences. Individual exons of the mouse *Spp2* sequence (Figure 4.1, accession number AJ315513) were used to perform BLAST searches of the mouse genomic sequencing trace archives (NCBI). The traces containing individual exons with 100% identity and their flanking introns were selected. Table 4.1 presents the details of the mouse DNA traces that were aligned to generate individual consensus mouse *Spp2* exons and flanking introns.

Thirty one DNA traces were identified of which 25 showed 100% identity to the individual exons of the mouse *Spp2* gene used to query the trace data. The remaining 6 traces showed similarities (80–90%) to the mouse query sequences, but not to a sufficiently high extent to believe that they represented genuine mouse sequences. These identical DNA traces were aligned using pileup and then viewed and manually edited in SeqLab. All anomalies between traces were checked and edited using the original sequence trace data. These edited traces were aligned to generate individual consensus mouse *Spp2* exons and flanking introns.

4.2.2 PCR amplification and optimisation of the promoter region and exons 1–8 of the mouse *Spp2* gene

Identification of variations in the sequence of the mouse *Spp2* gene and a corresponding change in phenotype may provide information concerning the function of spp24. One method

Exon	Trace archive ID
1	gnl|ti|17208091 G10P617462RB5.T0 gnl|ti|11634955 ml2C-a84g10.q1c gnl|ti|17892499 G10P629193RG12.T0 gnl|ti|13433728 ml2B-a1366d06.q1c gnl|ti|13290377 ml2B-a171e09.p1c
2	gnl|ti|29129761 jli48g09.b1 gnl|ti|13433728 ml2B-a1366d06.q1c
3	gnl|ti|18512459 G10p634443FE7.T0 gnl|ti|19824919 G10P636988RA11.T0 gnl|ti|3468936 G10P69389RG3.T0 gnl|ti|13224112 mk2A-a4827h09.p1c
4	gnl|ti|19426564 G10P637862FC4.T0 gnl|ti|16722548 jlf75e03.g1 gnl|ti|13245908 mk2A-a4838d11.q1c
5	gnl|ti|18599363 G10P625878FB12.T0 gnl|ti|21264180 jrr89d01.g1 gnl|ti|12044598 jil80e03.b1
6	gnl|ti|4890613 G10P62463FH7.T0 gnl|ti|5524319 G14P6216FE9.T0
7	gnl|ti|10926951 G10P617847FD9.T0 gnl|ti|18101168 G10P623963FA8.T0
8	gnl|ti|20840990 G10P636931FD4.T0 gnl|ti|1353022 ml1B-a961g10.p1c gnl|ti|133773409 ml2C-a6796g09.q1c gnl|ti|13482307 ml2B-a1200d10.q1c

Table 4.1. The sequences from the mouse genomic sequencing trace archive (NCBI) that were aligned to generate the consensus sequence for each individual exon and flanking introns of the mouse *Spp2* gene

The mouse 'trace archive' at NCBI was BLAST searched using the individual exons of the mouse *Spp2* gene. A total 25 sequences were identified that contained part or all of an exon and its flanking introns. This table illustrates the details of the trace archive sequences and the mouse exons that they contain.

of doing this is to PCR amplify each of the exons and the promoter region of the mouse *Spp2* gene and scan for sequence differences between different mouse strains with specific phenotypes that might be a consequence of variation in the *Spp2* gene.

By designing appropriate primers (Table 4.2), using the sequences of the flanking introns of each individual exon and the 5′ region of the mouse *Spp2* gene all 8 exons and the 5′ region (about 1000 bp upstream of the transcription initiation site) were PCR amplified. To achieve the best sensitivity and specificity for amplification of each fragment, the PCR conditions were optimised. The following conditions were used for all PCR amplification reactions: 96°C 30s, 60–62°C 30s, 72°C 30–40s × 30 (for fragments larger than 500 bp 40s extension time and for fragments smaller than 500 bp 30s were used). PCR amplification was carried out as described in Section 2.7.1. These PCR conditions were confirmed by amplification and agarose gel electrophoresis (Figure 4.2).

4.2.3 Determination of the entire mouse genomic *Spp2* sequence and confirmation of the exon/intron boundaries

It was at this time that the almost complete sequence of the region of mouse chromosome 1 that contained the nearly whole *Spp2* gene (except a gap for one region in intron 7) became available through the Softberry DNA database (http://softberry.ru/berry.phtml?topic=mousexp). This sequence began 10 bp before the presumed 5′ end of the first exon and ended 10 bp 3′ to the end of exon 8 (Cchr1.fa001017; range 108347804–108366968; length 19165). However, as it mentioned there was a gap in their presentation of seventh intron represented by 123 "N" bases (108364936–108365058).

To complete the whole sequence of the genomic mouse *Spp2* gene, two primers were designed using the sequences of the regions flanking the gap:

Forward 5′ AGCTTGTGCCTAAGAGCAGTG 3′

Reverse 5′ CTATGAGCAGGGCCTCTCTG 3′

After optimisation, the following PCR conditions were used for amplification of this gap fragment: 96°C 30s, 63°C 30s, 72°C 40s × 30. PCR amplification was carried out as described in Section 2.7.1. The PCR conditions were confirmed by amplification and agarose gel electrophoresis. The PCR amplified fragment (about 660 bp) was gel extracted and sequenced in both forward and reverse directions as described in Sections 2.5.3 and 2.10.1

Amplified region	Primers 5′→3′	Size of fragment after PCR	$[Mg^{+2}]$ (mM)	Ann. T (°C)
Promoter 2 500-1000 bp upstream of the 1st exon	Forward GAGACCTGGTGCACTCATACACTC Reverse CTCCAGGCTTCTGGATGACAG	572 bp	4.5	62
Promoter 1 1-500 bp upstream of the 1st exon	Forward ATAGCATGAGCGTGACTGGATG Reverse CATGGTTCAGAAACCACAATGG	596 bp	4.5	62
Exons 1+2	Forward: GTCTCCTGCTGTGGGATAAAG Reverse: AACAAAGTAGGTGAAGTCAAGACG	628 bp	4.5	60
Exon 3	Forward: TGAGTAGGTGACCCTGTTTATGG Reverse: CTCAGGATTTCACTTCCTCAGTC	413 bp	4.5	60
Exon 4	Forward: TTTTGTTGTTTGGTTCTGCAATG Reverse: GGAAAATGAGCCTGTAGTGATTC	428 bp	4.5	60
Exon 5	Forward: CCAGCCTCTTTCTGTCCTACTG Reverse: AAGTCAGGTGTTCGCTTTTTTG	362 bp	4.5	60
Exon 6	Forward: CCTCTGATTAAGGAGGATATCCAC Reverse: AGCCAAGTGGTGACTTTTTAGG	371 bp	4.5	60
Exon 7	Forward: AGTTTCCACAGATGCCTTAAAGAG Reverse: TGGGTACTAAAGGGTCTGGAAAG	461 bp	4.5	60
Exon 8	Forward: CATTTAGATCACTTTCTGAGCAGTG Reverse: TCATGCCTCCTAACATCAATTTC	380 bp	4.5	60

Table 4.2. The mouse *Spp2* primers

Sequences of primers, the sizes of fragments generated by PCR, the annealing temperature (Ann. T.) and the concentration of magnesium ions in individual PCR reactions.

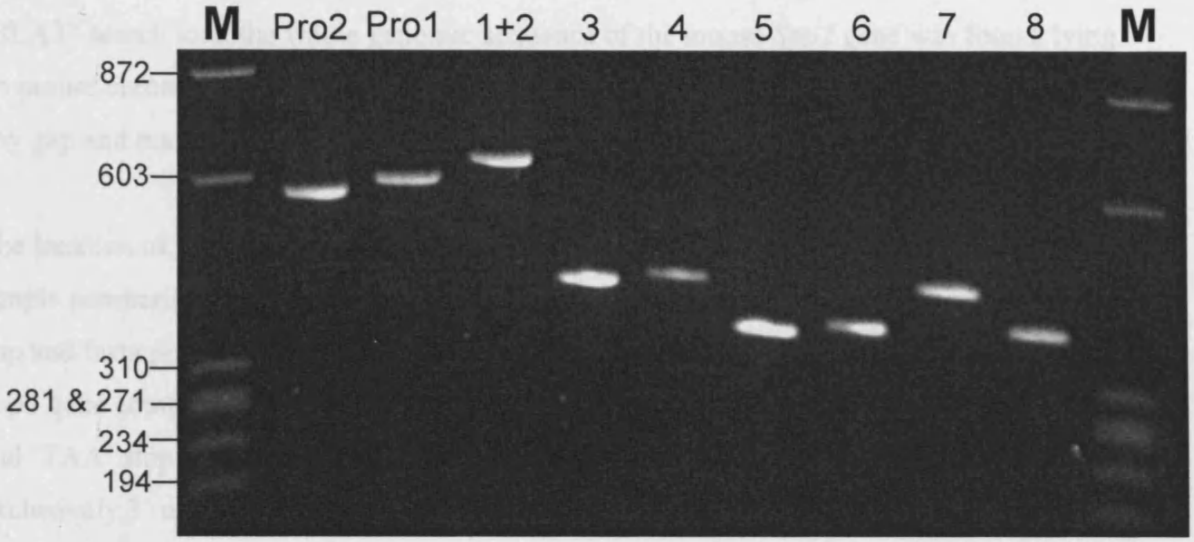


Figure 4.2. PCR amplification of the mouse *Spp2* gene exons and promoter regions

The promoter regions and exons of the mouse *Spp2* gene PCR amplified and analysed by agarose gel electerophoresis.

M = marker (ΦX174RF/*Hae*III)

respectively. The result of the sequencing revealed that the actual size of the gap was 405 bp and using these sequence data, the gap was closed.

At this time a further assembly of the mouse genome sequence data became available (the UCSC Genome Web site, http://genomefff.ucsc.edu). By searching their data with their 'BLAT' search tool, the whole genomic sequence of the mouse *Spp2* gene was found, lying on mouse chromosome 1 between 88832387 and 88853226. This sequence did not contain any gap and matched exactly with our assembled sequence.

The location of the exon/intron boundaries and the sizes of the introns were determined by a simple comparison between the mouse *Spp2* cDNA and the genomic DNA sequence using the gap and fasta programs within the GCG molecular biology package (Section 2.29.2). The *Spp2* gene comprises 8 exons and 7 introns. The 'ATG' start codon is located in the first exon and 'TAA' stop codon is located in the penultimate exon, the final exon containing exclusively 3′ untranslated region. The gene spans approximately 21 kb and is shown schematically in Figure 4.3.

Table 4.3 lists the sizes of each exon and intron and the DNA sequence at the boundaries. All boundaries show the consensus GT/AG sequences although not all junctions conform exactly to the consensus of 'GTRAGT' and 'YYTTYYYYYYNCAG' for the donor and acceptor sites respectively (Senepathy *et al.*, 1990). The size of exon 1 could not be precisely determined as the site of transcription initiation has not been identified for the mouse *Spp2* gene. The size given here for exon 1 is based on a comparison of the gene sequence with the consensus mouse *Spp2* cDNA with accession number AJ315513 (Figure 4.1).

4.2.4 An extensive sequence analysis of the mouse *Spp2* gene

The DNA sequence containing the mouse *Spp2* gene was extensively analysed using the NIX analysis environment at the HGMP website (http://www.hgmp.mrc.ac.uk/NIX/) (Williams *et al.*, 1998). The NIX analysis package comprises the programs listed in Table 4.4 with their respective functions as described. Any sub-group of analyses use several programs to search for consensus features and indicates a likelihood of that feature being real. Therefore the possibility of extra exons or alternative splicing can be determined and identification of likely promoter regions, repetitive elements, polyadenylation signals and open reading frames is possible. This package also contains programs that identify any ESTs or proteins that show

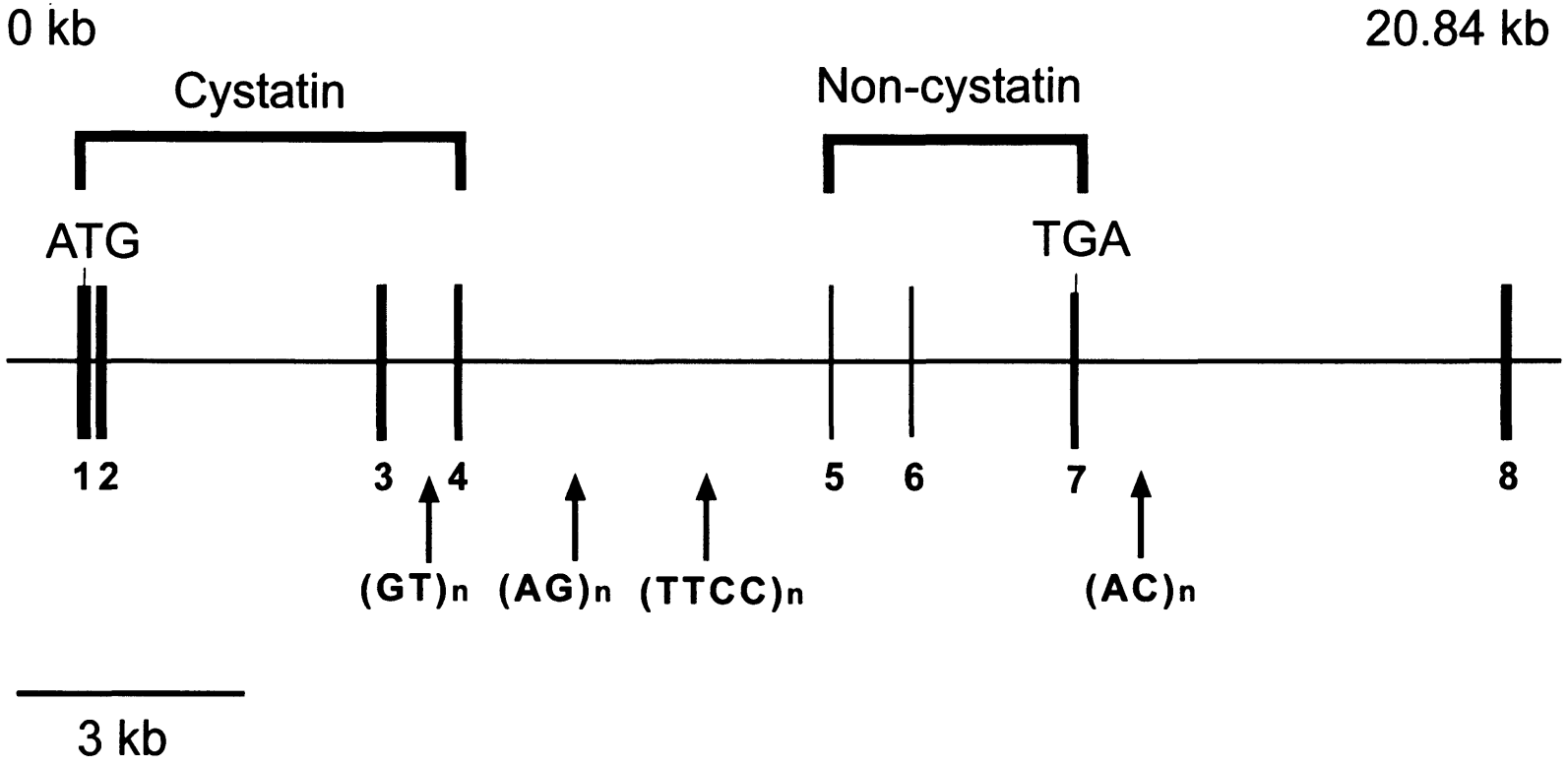


Figure 4.3. The exon/intron structure of the mouse *Spp2* gene

This figure illustrates the 20.84 kb region of the genomic DNA segment (Accession number CAAA01218969) containing the mouse *Spp2* gene. The figure is drawn approximately to scale with the exons labelled 1 to 8. The 'ATG' start codon falls in exon 1 and the 'TGA' stop codon in exon 7. The exons encoding the cystatin-like domain of the spp24 protein and the exons encoding the non-cystatin-like domain of the protein are indicated. The locations of three di-nucleotide repeats and one tetra-nucleotide repeat are indicated.

Exon/intron number	Exon size (bp)	Intron size (bp)	Exon/Intron boundaries sequence
1	159	100	.ACA....CAG**gt**aaag
2	125	3667	tgtttgctgtct**ag**GTT....AGA**gt**aagt *
3	120	1022	ttgcctttgtga**ag**GTC....GTG**gt**aagt *
4	108	4879	tctgtcttttcc**ag**CCA....GAG**gt**atga
5	55	1094	ttaaatttctttagATG....TTG**gt**aagt *
6	51	2145	ctaatgtgttac**ag**GTT....TTG**gt**aagt *
7	96	5616	ccttccatcctt**ag**AAA....GAG**gt**aagg
8	110	-	tctctcttgaat**ag**ATT....TTT

Table 4.3 The exon/intron sizes and boundaries of the mouse *spp2* gene

This table shows the exon/intron boundaries and the sizes of the exons and introns. The consensus gt/ag sequences are shown in bold. A '*' indicates that this junction conforms exactly the consensus of 'GTRAGT' and 'YYTTYYYYYYNCAG' for the donor and acceptor sites respectively (Senepathy *et al.*, 1990). All other junctions are very similar to these consensus sequences, but not identical. Exon sequence is shown in upper case and intron sequence in lower case.

Table 4.4. The programs used in the HGMP NIX analysis package

The NIX analysis package is from the HGMP website (http://www.hgmp.mrc.ac.uk/NIX/). This environment is a world wide web tool that enables analyses of any DNA sequence using analysis programs simultaneously. Therefore it is possible to perform an extensive DNA sequence analysis and search for features that conform to a consensus and give a likelihood of that feature being real.

[1]GRAIL references: Uberbacher *et al.*, (1991), Guan *et al.*, (1991a), Uberbacher and Mural, (1991), Einstein *et al.*, (1991), Mural *et al.*, (1991), Guan *et al.*, (1991b), Guan *et al.*, (1992), Uberbacher *et al.*, (1992), Einstein *et al.*, (1992), Xu *et al.*, (1994a), Xu *et al.*, (1994b), Mural *et al.*, (1993), Xu *et al.*, (1994c), Uberbacher, (1994), Shah *et al.*, (1995), Matis *et al.*, (1996), Uberbacher *et al.*, (1995a), Guan and Uberbacher, (1996), Xu *et al.*, (1995a), Uberbacher *et al.*, (1995b), Xu *et al.*, (1995b), Xu *et al.*, (1995c), Mark *et al.*, (1995), Uberbacher, (1995), Xu and Uberbacher, (1996a), Xu and Uberbacher, (1996b), Shah, (1996).

Program in NIX	Function of program	References
GRAIL/cpg	Predicts CpG islands	See legend [1]
GRAIL/polIIprom	Predicts promoters	See legend [1]
TSSW/Promoter	Predicts promoters	Solovyev and Salmanov (1997)
GENESCAN/Prom	Predicts promoters	Burge and Karlin (1997a) Burge and Karlin (1997b) Burge, (1997) Burset and Guigo (1996)
Fgenes/Prom	Predicts promoters	Slovyev (1995) Solovyev and Lawerence, (1993)
Fex	Predicts exons	Solovyev *et al.*, (1994a) Solovyev *et al.*, (1994b)
Hexon	Predicts exons	Solovyev *et al.*, (1994a) Solovyev *et al.*, (1994b)
MZEF	Predicts exons	Zhang (1997)
Genemark	Predicts exons	Borodovsky and McIninch (1993)
GRAIL/exons	Predicts exons	See legend [1]
GRAIL/gap2	Predicts genes	See legend [1]
Genefinder	Predicts genes	Green (unpublished)
FGene	Predicts genes	Solovyev (1995) Solovyev and Lawerence (1993)
GENSCAN	Predicts genes	Burge and Karlin (1997a) Burge and Karlin (1997b) Burge (1997) Burset and Guigo (1996)
FGenes	Predicts genes	Solovyev (1995) Solovyev and Lawerence (1993)
HMMGene	Predicts genes	Krogh (1997)
BLAST/trembl	Blasts against tremble database	Altschul *et al.*, (1990) Altschul *et al.*, (1997)
BLAST/swissport	Blasts against swissport database	Altschul *et al.*, (1990) Altschul *et al.*, (1997)
BLAST/EST	Blasts against EST database	Altschul *et al.*, (1990) Altschul *et al.*, (1997)
BLAST/Embl-	Blasts against EMBL database	Altschul *et al.*, (1990) Altschul *et al.*, (1997)
GENSCAN/polya	Predicts polyadenylation signals	Burge and Karlin (1997a) Burge and Karlin (1997b) Burge (1997) Burset and Guigo (1996)
FGenes/polya	Predicts polyadenylation signals	Solovyev (1995) Solovyev and Lawerence (1993)
GRAIL/polya	Predicts polyadenylation signals	See legend [1]
BLAST/gss	Predicts if frameshift errors are likely	Altschul *et al.*, (1990) Altschul *et al.*, (1997)
BLAST/sts	Blasts against STS database	Altschul *et al.*, (1990) Altschul *et al.*, (1997)
BLAST/ecoli	Blasts against *E. coli* database	Altschul *et al.*, (1990) Altschul *et al.*, (1997)
BLAST/vector	Blasts against vector database	Altschul *et al.*, (1990) Altschul *et al.*, (1997)
RepeatMasker	Predicts repetitive elements	Smit and Green (unpublished)
tRNAscan-RE	Scan for tRNA	Fichant and Burks (1991) Eddy and Durbin (1994) Pavesi *et al.*, (1994) Lowe and Eddy (1997)

homology to the investigated sequence. Therefore this package enables extensive sequence analyses using several different programs to be performed simultaneously. However, the NIX package does not search sequences for tandem DNA repeats. Such analyses were carried out using Tandem Repeats Finder (http://c3.biomath.mssm.edu/trf/) (Benson, 1999). The NIX analysis results for *Spp2* are shown in Figure 4.4.

Unsurprisingly, the programs GRAIL, Genefinder, FGene, GENESCAN, FGenes and HMMGene all confirm the presence of a gene in the region of the mouse *Spp2* gene sequence.

The programs FEX, HEXON, MZEF and Genemark all predict the location of exons. As expected, all exons predicted by each individual program clustered around the location of the exons previously determined by comparison with the consensus cDNA sequence (Section 4.2.4). There was no other clustering in any other segment of the sequence that would suggest previously unidentified exons.

The programs GRAIL/prom, TSSW/Promoter, GENSCAN/Prom and FGenes/Prom are capable of predicting the location of possible promoters. No promoters were predicted by these programs around the expected region in the vicinity of the transcription initiation site of the mouse *Spp2* gene. This result suggests that the *Spp2* gene has no conventional promoter. The program GRAIL/CpG did not indicate any CpG islands preceding the mouse *Spp2* gene.

The mouse *Spp2* gene has an obvious polyadenylation signal that was predicted by GENSCAN/polya. This signal corresponds to the 'AATAAA' sequence identified at the 3′ of the consensus cDNA sequence (Figure 4.1).

As expected no homology to any vector or *E. coli* DNA sequence was found in the BLAST searches. BLAST searches against the TREMBL, Swissprot, UniGene, EMBL databases showed homology to the expected rat, human, bovine spp24 sequences. The only non-spp24 protein that showed a significant homology was a chicken protein (accession number Q91982) that will be discussed further in Sections 4.2.5 and 4.2.8 and whose expression profile will de discussed in Chapter 8.

The program Repeat Masker predicted the location of many interspersed repetitive sequences (SINEs, LINEs, LTR and DNA elements) within the region of the mouse *Spp2* gene. Interspersed repeats from all of these families were indicated by this program within the

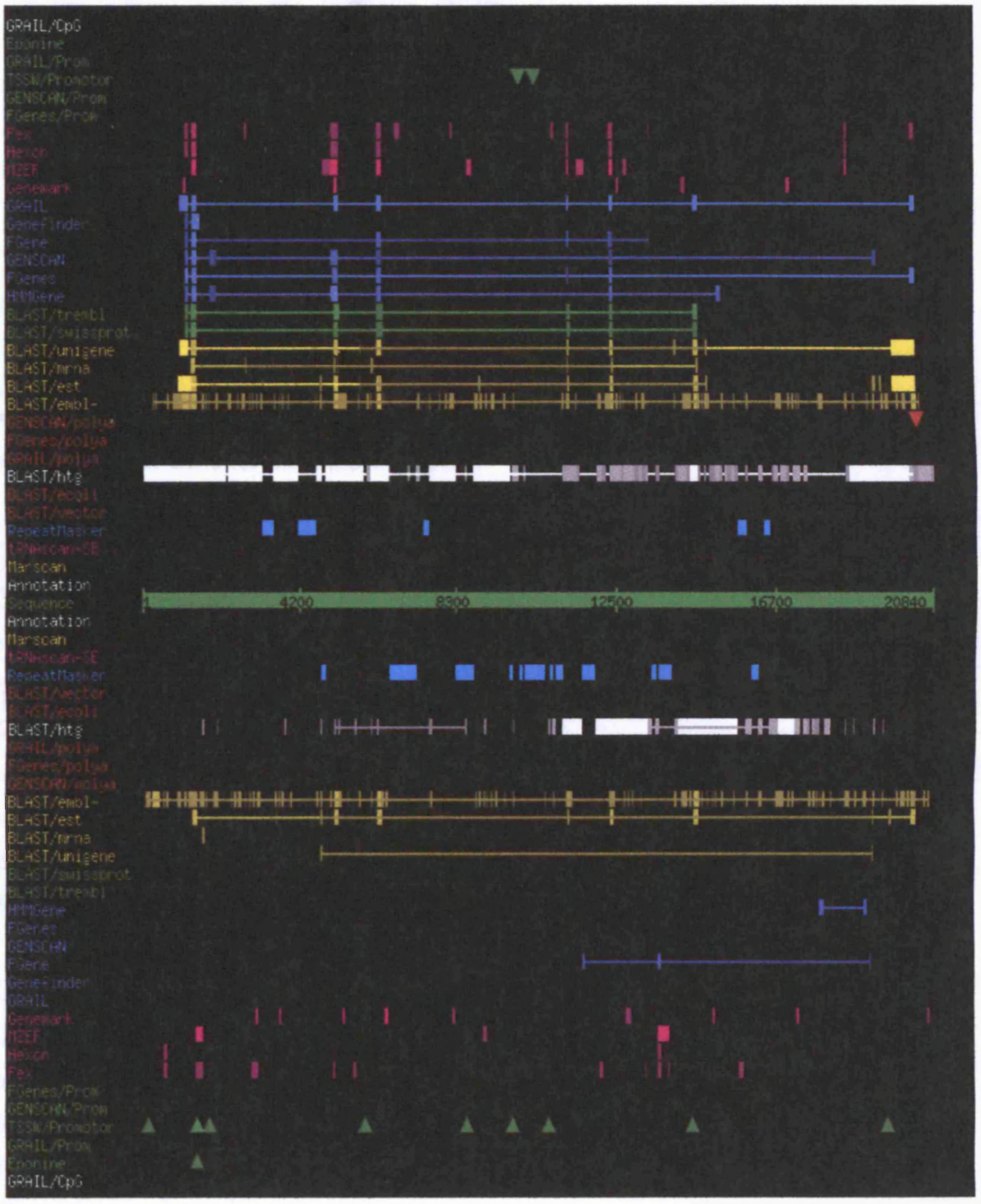


Figure 4.4. NIX analysis of the mouse *Spp2* gene

The complete sequence of the mouse *Spp2* gene was analysed using the NIX analysis environment at the HGMP website (http://www.hgmp.mrc.ac.uk). The individual programs in this environment are described in details in section 4.2.1. The results of each program are shown graphically above. The top half of the figure shows the analysis of the sense strand and the lower half the complementary strand.

mouse *Spp2* gene, but none of them was found in the mRNA coding regions of the gene. Details of the location of these interspersed repetitive elements can be found in Appendix A.

Using the program Tandem Repeat Finder (Benson, 1999), the mouse *Spp2* gene was also searched for tandem repeats. This program identified a variety of tandem repeats such as GT, AG, AC and TTCC. Details of these tandem repeats can be found in Appendix A and their location is indicated in Figure 4.3.

The annotated DNA sequence for the mouse *Spp2* gene was submitted to EMBL DNA database and allocated the accession number CAAA01218969.

4.2.5 Determination of the sheep spp24 cDNA, exons 1 and 2 and intron 1 sequences in Chicken and Marmoset

To confirm the whole sequence of the bovine mature spp24 cDNA two primers (For3: 5′ TCTGAACGGAAATTGTTCTTCC 3′and Rev4: 5′ TGGAACTTCTATTCCTTCCAGTG 3′) were designed in the presumed 5′ end region of the first and the presumed 3′ end region of the eighth exons and the fragment was amplified using PCR amplification of reverse-transcribed bovine liver mRNA. PCR amplification was carried out as described in Section 2.7.1 for 30 cycles each comprising 96°C 30s, 60°C 30s, 72°C 40s. Extraction of total RNA and RT-PCR were carried out as described in Sections 2.16.3 and 2.7.3 respectively. The PCR amplified fragment was purified and sequenced in the forward and reverse directions as described in Sections 2.5.3 and 2.10.1. Using this sequence, the original amino acid sequence for the bovine spp24 protein (accession number Q27967) was confirmed.

Using these two primers, it was also possible to PCR amplify (using reverse-transcribed ovine liver RNA) the corresponding region of the ovine spp24 cDNA with the same annealing temperature and PCR conditions. The PCR amplified fragment was purified and sequenced in forward and reverse directions as described in Sections 2.5.3 and 2.10.1 respectively. The spp24 cDNA sequence for sheep was submitted to the EMBL DNA sequence data base with the accession number AJ544160 (Appendix B).

The chicken spp24-like protein (accession number Q91982) mentioned in Section 4.2.4, was originally described as the product of "growth hormone regulated gene 1 (GHRG-1)" by Agarwal *et al.* (1995). A detailed analysis of the data presented in that paper and comparison

of their predicted protein sequence with that of human spp24 led to the realisation that the cDNA sequence presented in that paper was a chimera of cDNA and genomic DNA sequence (Bennett *et al.*, manuscript submitted). Consequently, a revised sequence for the chicken promoter, exon 1, intron 1 and start of exon2 was compiled and deposited in the EMBL DNA database with the accession number BN000081.

It was speculated that marmoset (*Callithrix jacchus*) might be amplifiable using human the PCR primers for *SPP2* for exons 1 and 2 (Table 3.5). PCR amplification was carried out as described in Section 2.7.1 for 30 cycles comprising 96°C 30s, 55°C 30s, 72°C 30s. The PCR-amplified fragment was purified and sequenced in the forward and reverse directions as described in Sections 2.5.3 and 2.10.1. Figure 4.5 illustrates PCR amplifications of part of exons 1, 2 and intron 1 in marmoset, cattle and sheep using their genomic DNA, and PCR amplification of the mature peptide region of spp24 cDNA using reverse transcribed liver mRNA from cattle and sheep. The results indicate that the cattle primers amplify the sheep genomic DNA and cDNA as efficiently as the corresponding cattle DNAs. The marmoset genomic DNA amplifies efficiently with the human primers, without generating any artefact bands.

4.2.6 Comparison of the first intron of the gene encoding the spp24 protein in six species

In an attempt to find any common transcription factor binding sites in the first intron among different species, it was decided to sequence the first intron in each of 6 species. Introns and non-coding DNA sequences that flank genes are generally highly diverged such that alignment of corresponding regions of orthologous genes from two or more species can be extremely difficult. However, the alignment process is simpler where the intron is short and its size is evolutionarily well conserved, such as turns out to be the case in *SPP2*.

The size, location and sequence of the first intron of the gene encoding the spp24 protein in human and chicken had already been determined (accession numbers AJ272265 and BN000081 respectively). In this study, the corresponding data were determined for cattle, sheep, marmoset and mouse (accession number CAAA01218969). Therefore it was decided to align the first intron for these six species to see if any features could be identified that are common to all six.

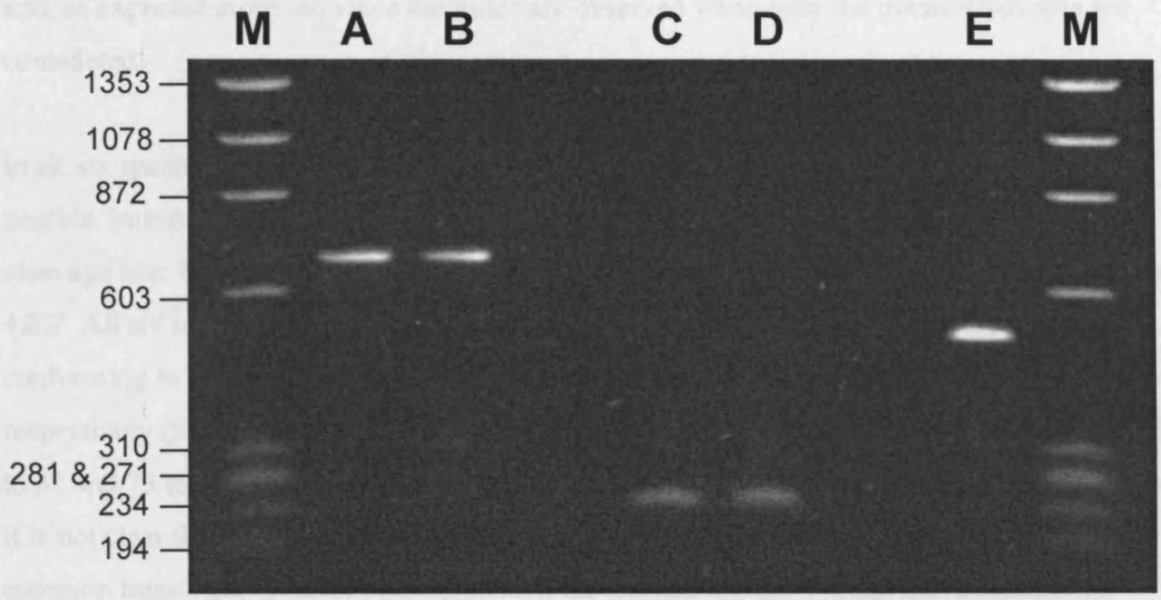


Figure 4.5. PCR amplification of cattle and sheep spp24 cDNA and exons 1, 2 and intron 1 of cattle, sheep and marmoset

A: PCR amplified fragment using reverse-transcribed bovine liver mRNA (678 bp) containing the entire mature spp24 protein cDNA sequence.

B: PCR amplified fragment using reverse-transcribed ovine liver mRNA (678 bp) containing the entire mature spp24 protein cDNA sequence.

C: PCR amplified fragment (using bovine genomic DNA) containing part of exons 1, 2 and intron 1 (244 bp).

D: PCR amplified fragment (using ovine genomic DNA) containing part of exons 1, 2 and intron 1 (241 bp).

E: PCR amplified fragment (using marmoset genomic DNA) containing exons 1, 2 and intron 1 (511 bp).

M = marker (ΦX174RF/*Hae*III)

Figure 4.6 shows the alignment of the first intron of the gene encoding the spp24 protein from human, marmoset, mouse, cattle, sheep and chicken. Beneath the aligned sequence a consensus sequence is given, with sequence identities for all six species depicted in uppercase letters. The chicken first intron sequence is the most diverged from the consensus sequence and, as expected more sequence identities are observed when only the mammalian data are considered.

In all six species, intron 1 lies at the boundary between the signal peptide and the mature peptide, interrupting the codon (in phase 1) encoding the glycine immediately preceding the cleavage site. The significance of the presence of this intron is discussed more fully in Section 4.2.7. All six introns are of the predominant gt/ag type with their donor and acceptor sites conforming to the consensus sequences 'GTRAGT' and 'YYTTYYYYYYNCAG' respectively (Senepathy *et al.*, 1990). Two identical motifs (CCTGC and CTCAC at bases 28 to 32 and 73 to 77 respectively) are observed in all species with the exception of chicken, but it is not clear that this finding is significant. A search was carried out for the presence of common transcription factor binding sites in the aligned introns 1 using the MatInspector program (http://www.genomatix.de/cgi-bin/matinspector_prof/mat_fam.pl). No such sites were found to be common to either all six introns, or to those of the mammalian species. Hence, it is presumed that intron 1 does not harbour any sequences that are important for the control of expression of the gene in all six species.

4.2.7 Comparison of the spp24 protein and cystatin superfamily proteins

As described in Chapter 1, the spp24 protein has an N-terminal cystatin-like domain, a serine-rich domain and a C-terminal non-cystatin-like domain. The cystatin-like domain of spp24 is encoded by the first four exons of the gene, with exon 1 encoding exclusively the signal peptide and the non-cystatin-like region is encoded by exon 5–7 (Figure 4.3). This gene organisation is completely conserved between mouse and human and there is evidence for the conservation of the location of some introns in other species (Dalgleish, unplublished).

Using the Genbank DNA database, the genomic sequences, exon/intron boundaries and gene organisation of all members of the human cystatin superfamily were determined. The family 1 cystatins (including cystatins A and B) have 3 exons with no signal peptide. The family 2 cystatins (including cystatins C, D, SA, SN and S) and cystatin E/M all have 3 exons with the entire signal peptide and the start of mature peptide encoded by the first exon. Cystatin F

Figure 4.6. The alignment of the first intron of the gene encoding the spp24 protein from six different species

The first introns of the genes encoding the spp24 protein from six different species are aligned. Identical motifs and nucleotides are shown in blue. A consensus sequence has been generated to highlight potentially important motifs or sequences.
Sequences that are identical in all species with the exception of chicken are shown in red.
The last 10 bp of the first exon and the first 10 bp of the second exon in all species are shown in upper case and the phase of splicing is shown by boxing the last bp of the first exon and the first two bp of the second exon.
The branch site that is conserved in all mammals (ctcac) and chicken (ctaat) is underlined.
The donor and acceptor sites are boxed.

(leukocystatin) has 4 exons, but again the entire signal peptide and the start of mature peptide are encoded by the first exon. In the family 3 cystatins, or kininogens, there are 3 cystatin domains and, again, there is no intron between the sequence encoding the signal peptide preceding the first cystatin domain and the start of mature peptide. Hence, introns 2 and 3 of the genes encoding spp24 probably correspond respectively to introns 1 and 2 of the classical cystatins. This is supported by the fact that all are in phase 0 relative to the protein sequence.

Figure 4.7 shows the exon/intron structure and boundaries in the human spp24 gene and consensus for the family type 2 cystatins. It can be seen the cystatin-like domain of spp24 is encoded by the first four exons of the gene, with exon 1 encoding exclusively the signal peptide, but in the family 2 cystatins, exon 1 encodes the signal peptide and the start of the mature peptide as well. Interestingly, intron 1 of *SPP2/Spp2* in all six species is very small, averaging 99 bp. Absence of a corresponding intron in the genes encoding other members of the cystatin superfamily suggest that this intron is "new" and that spp24 may be a relatively young member of the superfamily.

4.2.8 Alignment of the spp24 protein sequence from nine species

Consensus spp24 cDNA sequences were previously compiled for pig, salmon and chicken and were deposited in the EMBL DNA database with the accession numbers AJ315513, AJ308100 and AJ428527 respectively. The amino acid sequences were then deduced from the cDNAs. Figure 4.8 illustrates the alignment of the spp24 protein from rat, mouse, human, cattle, sheep, pig, chicken, salmon and marmoset (the first 60 amino acid residues corresponding to the first and second exons) using GCG pileup and displayed in GCG SeqLab where the alignment was optimised manually. Because in marmoset, no frozen liver tissue was available therefore it was not possible to extract total RNA to carry out cDNA syntesis and sequencing of the whole cDNA (like cattle and sheep). The alignment of the amino acid sequence of the spp24 protein from nine different species provides an opportunity to assess the conservation of amino acid sequence along the whole protein. The results of the spp24 comparison in these species will be discussed in more detail in section 4.3.

The evolutionary relationship of the spp24 protein among these species is also calculated by GCG pileup and displayed as a dendrogram (Figure 4.9) which shows the evolutionary relationship of the spp24 between these species. The relationship, based on spp24 sequences, is as expected from previous evolutionary studies (Futuyma, 1986).

Figure 4.7. A comparison of exon size between the exons of family 2 cystatins and those seen in the cystatin-like domain of the human *SPP2* gene and schematic illustration of the human spp24 protein and family 2 cystatin exon/intron structure

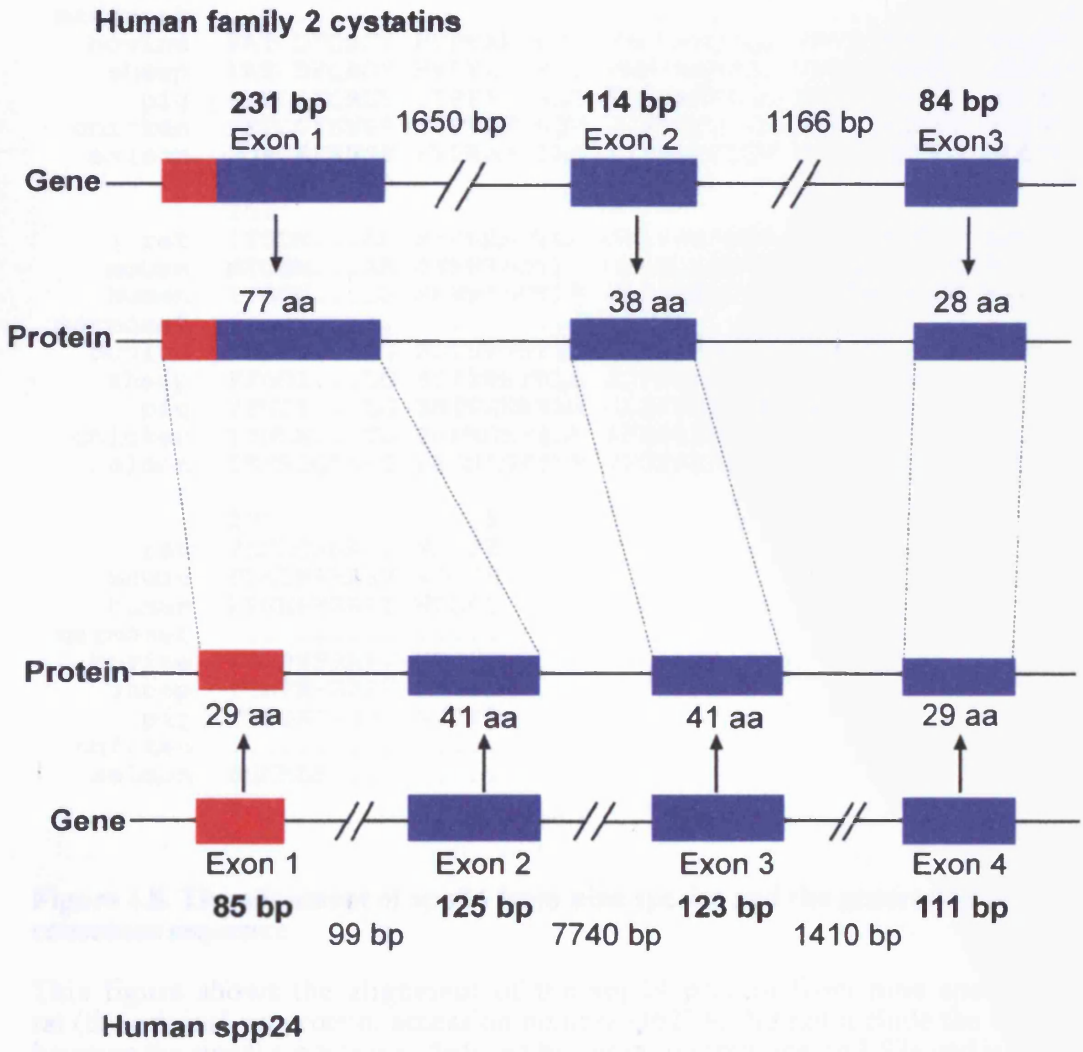
A: The human cystatins C, D, SN, SA and S were used to calculate the average exon and intron sizes in family 2 cystatins. The size of each exon is shown in bp and the number of encoded amino acids in aa. The exons are shown aligned against the corresponding exons of the *SPP2* gene.
The signal peptides are included and, in the case of spp24, approximately 8 amino acid residues in the exon 4 correspond to the phosphorylated region of the protein.

B: The figure shows the exon/intron structure in the human spp24 and family 2 cystatins. The cystatin-like domain of spp24 is encoded by the first four exons of the gene with exon 1 encoding exclusively the signal peptide, but in the family 2 cystatins exon 1 encodes the signal peptide and also the start of the mature peptide. The mature peptides in the family 2 cystatins and spp24 are shown in blue and signal peptides in red.
It should be mentioned that the cystatin-like region of the gene encoding spp24 protein has sequence similarity with cystatin superfamily. The sizes of exons in the cystatin-like region of gene are similar to the corresponding exons in family 2 cystatins as well.

A

Family 2 cystatins		SPP2	
Exon	**Size in base pairs and amino acids**	**Exon**	**Size in base pairs and amino acids**
1	231 bp/77 aa	1	85 bp/29 aa (only signal peptide)
	-	2	125 bp/41 aa
2	114 bp/38 aa	3	123 bp/41 aa
3	84 bp/28 aa	4	111 bp/29 aa (excluding the phosphorylated region)
Intron	**Size in bp**	**Intron**	**Size in bp**
	-	1	99
1	1697	2	7740
2	1202	3	1410

B



```
          1                                                        50
     rat  ......MELA TMKTLVMLVL GMHYWCASGF PVYDYDPSSL QE.ALSASVA
   mouse  ......MEQA MLKTLALLVL GMHYWCATGF PVYDYDPSSL QE.ALSASVA
   human  MISRMEKMTM MMKILIMFAL GMNYWSCSGF PVYDYDPSSL RD.ALSASVV
marmoset  .......... MMKILIMFAL GMNYWSCSGF PVYDYDPSSL SD.ALSASVA
  bovine  .........M AMKMLVIFVL GMNHWTCTGF PVYDYDPASL KE.ALSASVA
   sheep  .........M VMKMLVIFVF GMNHWTCTGF PVYDYDPASL KE.ALSASVA
     pig  ......MEKR AMRMLAMFVL GTSFWSCAGF PVYDYDPSSL RE.AVGASVA
 chicken  MGKTPEDFER HTMRSLIFVL ALSVFTCSGF PVYDYELPVT EE.ALNASIA
  salmon  ....LRHEQK MKWCGVLMVA LLQSLCCSGL PLYQSELAST ADKALVVTMT

          51                                                      100
     rat  KVNSQSLSPY LFRATRSSLK RVNVLDEDTL VMNLEFTVQE TTCLRESG.D
   mouse  KVNSQSLSPY LFRATRSSLK RVNVLDEDTL VMNLEFSVQE TTCLRDSG.D
   human  KVNSQSLSPY LFRAFRSSLK RVEVLDENNL VMNLEFSIRE TTCRKDSGED
marmoset  KVNSQSLSPY LFRALRSSLK R......... .......... ..........
  bovine  KVNSQSLSPY LFRAFRSSVK RVNALDEDSL TMDLEFRIQE TTCRRESEAD
   sheep  KVNSQSLSPY LFRAFRSSIK RVNALDEDSL TMDLEFRIQE TTCRRESEAD
     pig  KVNSQSLSPY LFRAFRSSLK RVNVLGEDSL SMDIEFGIRE TTCKRDSGED
 chicken  RINSQTWGPN LYGVVRSHVR HVDMWNSNDY RLELQLSIRE TECTKASGRD
  salmon  QVNNLYAGLR LYRVSRGSIK RVVPLGLNTY DLIMNFGIKE TDCLKSSGED

          101                                                     150
     rat  PSTCAFQRGY SVPTAACRST VQMSKGQVKD VWAHCRWRS. TSESNSSEEM
   mouse  PSTCAFQRGY SVPTAACRST VQMSKGQVKD VWAHCRWAS. SSESNSSEEM
   human  PATCAFQRDY YVSTAVCRST VKVSAQQVQG VHARCSWSSS TSESYSSEEM
marmoset  .......... .......... .......... .......... ..........
  bovine  PATCDFQRGY HVPVAVCRST VRMSAEQVQN VWVRCHWSS. SSGSSSSEEM
   sheep  PATCDFQRGY HVPVAVCRST VRMSAERVQD VWVRCHWSS. SSGSSSSEEM
     pig  PATCDFQRGY FTPSAICRST VQISAEKVQD VWVRCRWSS. SSESNSSEEM
 chicken  PFTCGFKVGP FVPTAVCKSV VEVSSEQIVN VIVRCHQSTF SSESMSSEEM
  salmon  PQRCAFRVGF FVPAASCTAR VRVTAEFTQV VSLNCGQDSS SSESSSEENF

          151                                                     200
     rat  IFGDM...AR SHRRRNDYLL GFLYDEPKGE QFYDRSIEIT RRGHPPAHRR
   mouse  MFGDM...AR SHRRRNDYLL GFLSDESRSE QFRDRSLEIM RRGQPPAHRR
   human  IFGDM...LG SHKWRNNYLF GLISDESISE QFYDRSLGIM RRVLPPGNRR
marmoset  .......... .......... .......... .......... ..........
  bovine  FFGDI...LG SSTSRNSYLL GLTPDRSRGE PLYEPSREMR RNFPL.GNRR
   sheep  FFGDI...LG SSTSRNSHLL GLTPDRSRGE PLYERSREMR RNFPL.GNRR
     pig  IFGDI...LG SSTSRNNYLR GLIPDVSRTE PLYERSLETM RRFPPPGNRS
 chicken  TYMLM...TD PRKRGSSRSE AFSSRGRGHS NGDWRKPDYT SPGKVE....
  salmon  TRKRQQLNVQ PFGNRGPVLP VPGFSEATRF PSHSFSRQEV EPQPIPRGDS

          201          215
     rat  FLNLQRRARV NSGFE
   mouse  FLNLHRRARV NSGFE
   human  YPNHRHRARI NTDFE
marmoset  .......... .....
  bovine  YSNPWPRARV NPGFE
   sheep  YSNPWPRARV NPGFE
     pig  FPNQWPRART NTGFE
 chicken  .......... .....
  salmon  FGNHLE.... .....
```

Figure 4.8. The alignment of spp24 from nine species and the generation of a consensus sequence

This figure shows the alignment of the spp24 protein from nine species including rat (the original rat protein, accession number Q62740 did not include the signal peptide, however the signal peptide was deduced by our group from the rat ESTs and is shown in this figure), mouse, human, bovine, sheep, pig, chicken, salmon and marmoset. All the signal peptides are shown in blue.

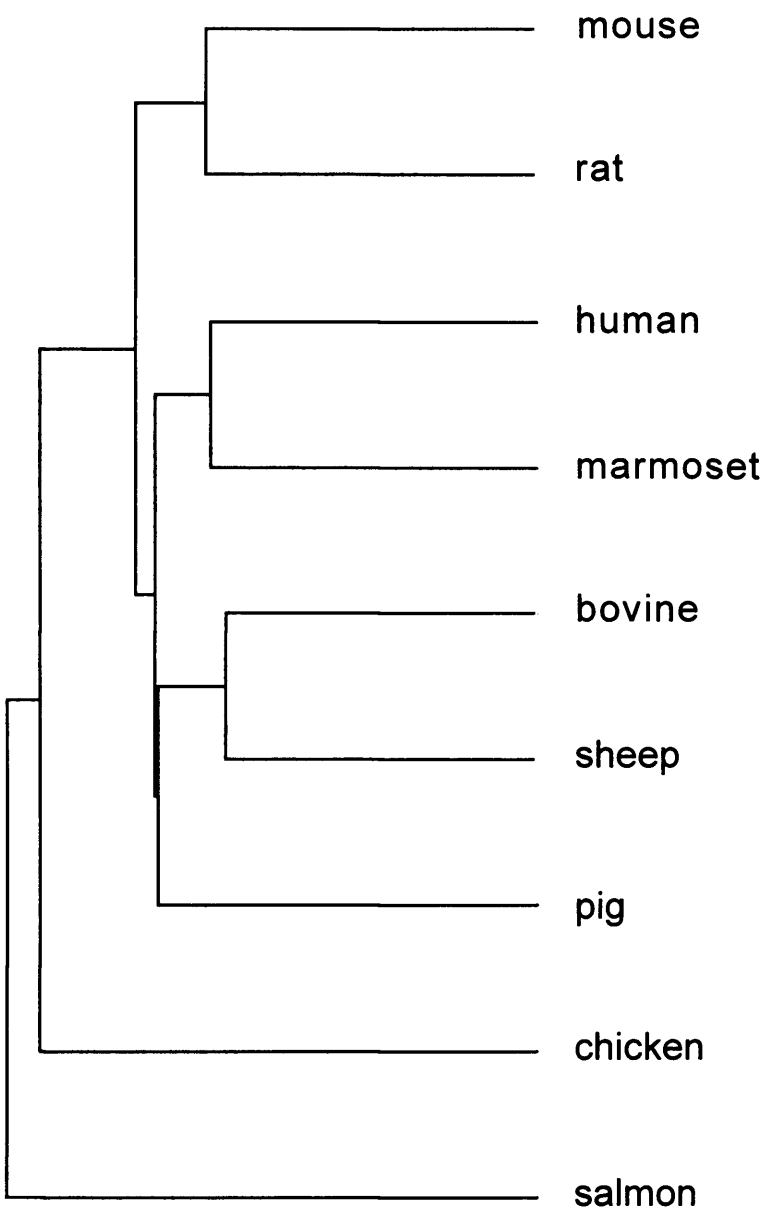


Figure 4.9. Schematic illustration of the evolutionary relationship of the spp24 protein between nine different species

4.2.9 Analysis of the promoter region of the mouse Spp2 gene

The extensive analysis of the mouse *Spp2* gene revealed a gene with no striking characteristics. The ATG start codon is located in the first exon, the TGA stop codon is in the penultimate exon and the final exon codes solely for the 3′ UTR. It is rare to find the termination codon in the penultimate exon, however this is not uncommon in secreted and trans membrane proteins (the more details of these genes can be found at http://rpci.med.buffalo.edu/scientific.report/maquat1.html) (Nagy and Maquat, 1998). The mouse *Spp2* gene does not contain any obvious promoter. Using different promoter prediction programs in the NIX analysis environment, no consensus promoter was determined. The mouse *Spp2* gene does not contain the TATA box, but it is possible to identify an Inr sequence. The consensus Inr sequence is defined as Py Py A^{+1} N T/A Py Py, where N is any nucleotide and Py is a pyrimidine (C or T) (Smale 1997). A more detailed analysis indicated that within this consensus sequence an A at position +1, a T or an A at +3 and a pyrimidine at –1 are the most crucial in determining the strength of the Inr element (Javahery *et al.*, 1994; Lo and Smale, 1996).

Two transcription initiation sites have been determined so far. The first (which was deduced from a full-length cDNA sequence compiled by alignment of the ESTs for spp24 comprising the mouse UniGene cluster Mm.28247 with accession number AJ315513) lies in the sequence 'TGACAAG', with the underlined A being in the +1 position. This sequence matches at two positions thought to be crucial in determining the Inr strength *i.e.* an A at +1 and an A at +3. The second transcription initiation site (which was found in a *Mus musculus* adult male kidney full-length cDNA with the accession number AK002814) lies in the sequence 'CCAGATT', with the A being in the +1 position. This sequence could be an initiation sequence. It matches the consensus sequence at every position thought to be crucial in determining the Inr strength *i.e.* an A at +1, an A at +3 and a C at –1. It also has pyrimidines at positions –2 (C), +3 and +4 (TT), making it a very good candidate for an initiation sequence because it matches the consensus at every position (crucial and non crucial) in determining the Inr strength (Figure 4.10).

The transcription initiation sites were identified in the human gene using primer extension with RNA isolated from liver and kidney (Bennett *et al.* manuscript submitted). The primary liver transcription initiation site lies in the sequence 'CCAGTGT', with the A being in the +1 position. This sequence matches the consensus sequence at every position thought to be crucial in determining the Inr strength and also has pyrimidines in several of the surrounding

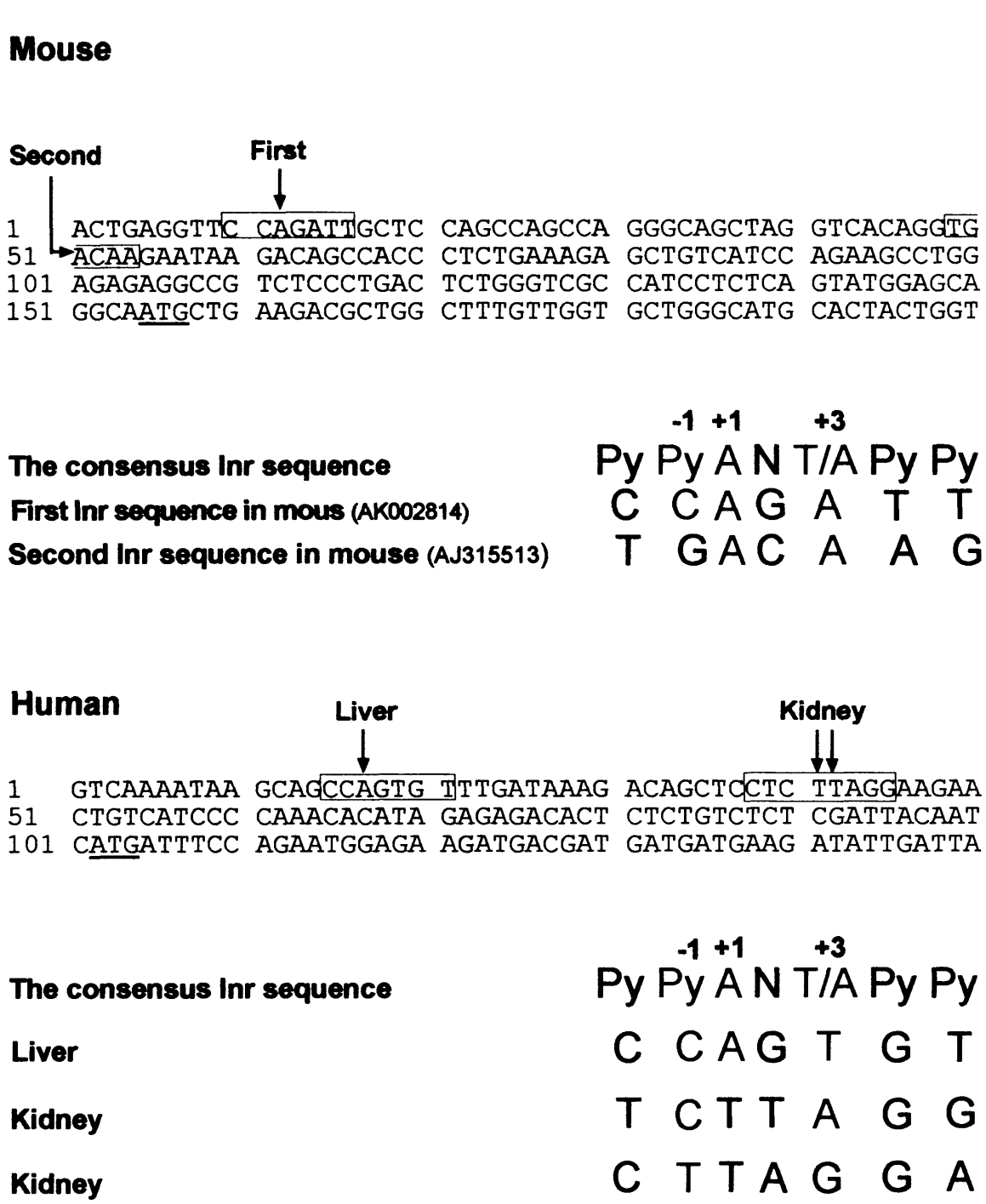


Figure 4.10. The consensus Inr sequence and potential Inr sequences of the gene encoding spp24 protein in mouse and human

The consensus Inr sequence is defined as Py Py A^{+1} N T/A Py Py, where N is any nucleotide and Py is a pyrimidine (C or T) (Smale, 1997). Within this consensus sequence an A at position +1, a T or an A at +3 and a pyrimidine at –1 are the most crucial in determining the strength of the Inr element.

This figure illustrates the two potential Inr elements in mouse and three potential Inr elements in human (in liver and kidney). The most crucial sequences in determining the strength of the Inr element in consensus sequences and matching sequences in potential Inr elements in human and mouse are shown in red. As can be seen, the first potential Inr element in mouse perfectly matches the consensus at every position thought to be crucial and non crucial in determining the Inr strength. In human the first potential Inr element (liver) matches the consensus at every position thought to be crucial in determining the Inr strength and also has pyrimidines in several of the surrounding positions. The major transcription initiation sites in human and mouse are shown by the vertical arrows and the translation start codons are underlined.

positions. The two transcription initiation sites mapped in kidney lie in the sequences 'TCTTAGG' and 'CTTAGGA' respectively with the second T in each sequence being in the +1 position of the Inr consensus. The first sequence matches the consensus Inr sequence more closely than the second, but neither match is as good as the potential Inr site in liver (Figure 4.10).

In an attempt to identify potential Inr elements in the mouse gene, 500 bp around the presumed transcription initiation site (taken from the mouse genomic DNA sequence with accession number CAAA01218969) was searched using the program "findpatterns" in the GCG molecular biology package. One potential transcription initiation site that lies in the sequence 'CCAGATT' with the A being in the +1 position was identified (Figure 4.10). This sequence matches perfectly the consensus sequence at every position thought to be crucial and non crucial in determining the Inr strength. Using the BLAST search of the mouse EST database ten ESTs (UniGene EST database, cluster Mm.28247) were identified whose 5′ end lay just 3′ to this theoretical Inr sequence. These ESTs suggest that this Inr element is a true transcription initiation site for the mouse *Spp2* gene however, it needs to be confirmed by practical experiments such as primer extension.

In an attempt to identify any common potential upstream regulatory regions of the gene encoding spp24 in human, mouse and chicken, approximately 500 bp upstream to the start of transcription in each gene were searched using the program MatInspector V2.2 (http://www.genomatix.de/cgi-bin/matinspector_prof/mat_fam.pl) (Quandt *et al.*, 1995). Search programs of this type can predict the presence of many transcription binding sites as they will pick up sites that only match loosely with consensus binding sites for transcription factors. Therefore this program should only be used as a start point to detect potential binding sites and results should be confirmed by experimental evidence.

MatInspector identified 17 transcription factor binding sites (common to at least two species) in the upstream regions of these genes. However, only two transcription binding sites were common in all three species, including ectopic viral integration site 1 encoded factor (EVI1) (family/matrix, V$EVI1.02) and nuclear factor of activated T-cells (NFAT-1) (family/matrix, V$NFAT/NFAT.01). The consensus binding sequence for viral integration site 1 encoded factor is 'AAGA' and was found at –133 to –129 and –369 to –366 in human, –209 to –206 in mouse and –173 to –170 in chicken. The consensus binding sequence for nuclear factor of

activated T-cells is 'GAAA' and was found at –81 to –78 in human, –93 to –90, –102 to –99 and –383 to –380 in mouse and –193 to –190 in chicken.

4.3 Discussion

The spp24 protein can be divided into two main domains, the cystatin-like domain and non-cystatin-like domain. Figure 4.3 illustrates the exons that encode these two domains of the spp24 protein in mouse. The mouse *Spp2* gene has 8 exons and 7 introns (Table 4.4). Exon 1 encodes the 5′ UTR and signal peptide, exons 2 to 4 encode the cystatin-like domain, exons 5 to 7 the non-cystatin like domain and exon 8 solely encodes the 3′ UTR. In this study it was shown that the gene organisation and exon/intron structure of the gene encoding spp24 is identical in human and mouse and also that the size and location of intron 1 is conserved between several different species (Section 4.2.6).

The classical cystatin genes comprise three exons with the two disulphide-bonded loops encoded in individual exons, but the cystatin-like region of spp24 is encoded by four exons, again with the two disulphide-bonded loops encoded in individual exons. Figure 4.7 compares the sizes of the typical cystatin exons (families 2 of cystatin superfamily) and the exons observed in the cystatin-like region of the human *SPP2* gene (identical in its structure and organisation with the mouse *Spp2* gene).

Exons 3 and 4 of the *SPP2* gene and exons 2 and 3 of the family 2 cystatins are similar in size. However, exon 4 of *SPP2* encodes the phosphorylated serine rich domain of the spp24 protein (not present in the family 2 cystatins) and therefore the cystatin-like part of the protein encoded in exon 4 is actually smaller than that encoded in exon 3 of typical cystatin. Intron 2 of the *SPP2* gene is much larger than its equivalent in the family 2 cystatins, but intron 3 is more similar. This similarity of the exon/intron structure and organisation between typical cystatin and spp24 cystatin-like domain provides further support for spp24 being a member of the cystatin superfamily, as suggested originally by Hu *et al*, (1995).

If spp24 is a member of the cystatin superfamily, it seems that the first exon in the classical cystatin genes has been split into two separate exons by an intron (intron 1) that is not present in the typical cystatin genes. It is interesting that the size of intron that separates the signal peptide from the mature protein in the human *SPP2* is only 99 bp and its size and location are conserved between several different species (Section 4.2.6). From the limited information available for human and mouse, it is interesting to note that the conservation of the size of intron size is more marked than that of the other introns. Altogether, this study may be significant from an evolutionary point of view and may indicate that the spp24 protein is a

relatively new member of cystatin superfamily. However, the differences observed between spp24 and the cystatins suggest that the former may only be a relatively distant member of the cystatin superfamily.

Eukaryotic genes normally have two kinds of promoter, the core promoter which usually lies in the region adjacent to the transcription initiation site and a more distant enhancer region (Roeder, 1991; Tjian and Maniatis, 1994). Within the core promoter the two main elements are TATA box and an initiator sequence (Inr) (Smale and Baltimore, 1989). A core promoter could contain both of these elements, either one or the other, or neither of them.

It has been demonstrated that an Inr element cannot be replaced by a TATA box in terms of a promoter having the transcriptional responses necessary for lineage-specific gene expression (Novina and Roy, 1996). For example, in the *FcγR1b* gene promoter ($TATA^{-}Inr^{+}$), the Inr element is required for myeloid-specific expression and selective interferon-γ (IFN-γ) responsiveness. The artificial replacement of a TATA box, either in place of or in addition to the Inr sequence, results in an increase in gene expression, but a loss of lineage specificity (Eichbaum, 1994). It has been also suggested that Inr elements could be responsible for temporal regulation of gene expression (Novina and Roy, 1996). An example of this is observed in the *Drosophila Adh* gene (alcohol dehydrogenase). The Inr elements in this gene mediate the molecular switch between a distal promoter, preferentially used during embryonic and adult developmental, and the proximal promoter used at other times (Hansen and Tjian, 1995). Inr elements are also thought to be responsible for the control of spatial expression. The *Drosophila Dpp* gene (decapentaplegic) encodes a protein related to transforming growth factor β (TGF-β), which has an important role in dorsal-ventral pattern formation. The promoter of this gene ($TATA^{-}Inr^{+)}$ controls the spatial expression of the gene in the development of the embryo and is resistant to ventral activation, therefore preventing dorsalisation of the embryo (Schwyter *et al.*, 1995).

These findings suggest that the core promoters of the genes encoding spp24 in human and mouse are $TATA^{-}Inr^{+}$, perhaps indicating that these genes have a lineage-specific expression pattern and that the temporal and spatial expression of these genes are under the control of Inr elements.

As explained in Section 4.2.9 two common potential transcription factor binding sites were identified in the upstream regulatory regions of the gene encoding spp24 in human, mouse

and chicken including EVI1 and NFAT-1. Expression of the gene encoding the EVI1 transcription factor is activated in murine myeloid leukaemia by retroviral insertions. Aberrant expression of this gene has been shown to interfere with myeloid differentiation, which is proposed to be the basis for its role in leukaemias (Dewel *et al.*, 1993). NFAT-1 is known to be found specifically in nuclear extracts of activated T cells and transmits signals initiated at the T cell antigen receptor and can play a regulatory role in early T cell gene activation (Shaw *et al.*, 1988). Finding the consensus sequences of these two transcription factors in the upstream regulatory regions of the gene encoding spp24 in human, mouse and chicken may indicate that *SPP2* gene is expressed in lympho-myeloid lineage cells such as lymphocytes and has a role in the immune system (like some members of the cystatin superfamily that have an anti inflammatory and anti infection role in the body which has already been discussed in more detail in Chapter 1).

The study of the mouse *Spp2* gene promoter region, although not conclusive, suggests (when combined with the other evidence and the results of human *SPP2* gene study) that the gene encoding spp24 protein in mouse and human is expressed in a tissue-specific manner. It may be also expressed by cells of a specific lineage that are found in liver or kidney at specific stages of embryonic and adult development. This will be discussed further in Chapter 8.

The alignment of the amino acid sequence of the spp24 protein from nine different species provides an opportunity to assess the conservation of amino acid sequence along the entire protein. There is absolute conservation of the two pairs of cystatin residues (Figure 4.8), characterising the protein as cystatin-like as originally described by Hu *et al*, (1995). The spp24 protein in seven mammals for which we have the amino acid sequence (the first 60 amino acid residues only in marmoset) is of approximately the same length with considerable conservation of sequence. In contrast, the length of the spp24 protein in salmon and chicken are shorter (in the non-cystatin-like domain) and shows less conservation and greater divergence. The N-terminal 41 amino-acid region that is encoded by exon 2 in mouse and human is highly conserved between mammals compared to the rest of the protein sequence. This suggests that perhaps this part of the cystatin-like domain is under greater selective pressure compared to the rest of protein sequence and may play a functional or structural role that is tightly constrained. The first six amino acid residues of the mature protein in chicken are also absolutely conserved with those of the mammals. The region containing phosphorylated serine residues also appears to be important. This is expected if the suggestion by Hu *et al*, (1995) that this region has a regulatory function is correct. The lack of identical

residues in the non-cystatin-like region suggests that perhaps this domain is not functionally important or its function does not depend on strict sequence conservation or that it may have evolved to have specific roles in each species. It is notable that the extreme C-terminal end of the spp24 sequence is also highly conserved between mammals and all eight species have glutamic acid as the C-terminal amino acid which may also indicate a possible functional or structural role for this residue.

Following alignment, an RGD motif (arginine-glycine-aspartate) was observed in the extreme C-terminal domain of the salmon spp24 protein, but not in other species. Osteoclastic bone resorption requires the formation of a tightly sealed compartment between the mineralised bone matrix and osteoclast. This complex acts as an extracellular lysosome which contains proteolytic enzymes and acids. Vitronectin receptors (VnR, integrin alphavbeta3) which display on the surface of osteoclast cells may play a role in the attachment of osteoclasts to the resorption surface (Fisher *et al.*, 1993). It has been shown that VnR is bound to the RGD-containing matrix proteins and it has been reported that soluble peptide containing RGD sequences can block osteoclast attachment to bone and inhibit bone resorption in vitro (Fisher *et al.*, 1993). It has also been shown that infection by field strains of foot-and-mouth disease virus (FMDV) is initiated by binding to certain species of RGD-dependent integrin including alphavbeta3 and the epithelial alphavbeta6 (Jackson *et al.*, 2002). Therefore it is possible that in salmon the spp24 protein can inhibit osteoclast-mediated bone resorption *in vivo* or *in vitro* or prevent infection by agents similar to foot-and-mouth disease virus. It is perhaps notable that the RGD motif is conserved in trout which is closely related to salmon (Dalgleish, unpublished).

The amino acid alignment provides an opportunity to evaluate the importance of the amino acid substitution found in human at amino acid 38 (amino acid 9 of the mature protein, Section 3.2.5). This alignment reveals that the amino acid serine at this position is not absolutely conserved across all species. Human, marmoset, mouse, rat and pig have serine at this position whereas cattle and sheep have alanine and chicken and salmon (the more evolutionary distant species) have alanine and proline respectively. It is notable that the amino acid change of serine to phenylalanine falls in one of the most highly conserved regions of the spp24 amino acid sequence, especially amongst mammals. Phenylalanine is an aromatic amino acid and is bulkier than serine, alanine and proline and would probably not be as easily accommodated as the other residues into the three dimensional structure of the spp24 protein. This will be discussed further in Chapter 9.

Chapter 5
Assignment (Mapping) of the Mouse *Spp2* Gene

5.1 Introduction

The mapping of genes to specific locations on chromosomes is a central focus of medical genetics. Dramatic advances in molecular genetic technology and important developments in the statistical analysis of genetic data have greatly increased the rate at which genes are being mapped. One of the most important goals of the Human Genome Project is to map all our genes to specific chromosome locations. As this aim progresses, our understanding of the biological basis of genetic diseases will also progress. Gene mapping is an important step in the understanding, diagnosis and eventually better treatment or management of genetic diseases. Therefore gene mapping contributes directly to many of the primary goals of medical genetics.

Mouse genetics is one of the most powerful systems for the study of many aspects of vertebrate biology, including development, physiology and pathobiology. Human and mouse phenotypic homologies provide valuable clues toward identifying human disease genes for several reasons. Among the genetically well-explored organisms, mice are one of the closest to human in evolutionary terms. Therefore mutations in orthologous genes are likely to induce similar phenotypes in mice and humans. Mouse phenotypic information usually translates readily into positional candidate genes. Backcross or radiation hybrid mapping allows quick and accurate mapping in the mouse. Exon sequences and arrangements are usually well conserved between orthologous genes in human and mouse, therefore once a human or mouse gene is isolated, primers or probes of one species could be used to screen DNA libraries from the other species to identify the orthologous gene.

The human *SPP2* gene, encoding the spp24 protein has previously been assigned to the tip of the long arm of chromosome 2 (2q37-qter) using *in situ* hybridisation (Swallow *et al.*, 1997). When this study started there was no available data about the assignment of the *Spp2* gene in mouse. This chapter presents the assignment of the gene orthologous to the human *SPP2* gene in the mouse genome using methods including *in silico* and radiation hybrids for gross mapping, and eventually the Ensembl Genome Browser for an exact and detailed assignment.

5.1.1 General aspects of genome organisation and gene distribution in human and mouse

Cytogenetic analyses reveal very different chromosome organisations between human and mouse. The mouse has 20 pairs of acrocentric chromosomes whereas there are 23 pairs of human chromosomes, most of which are metacentric or submetacentric. Nevertheless, comparison of high resolution human and mouse chromosome maps has revealed that orthologous chromosomal segments are located in regions where there is considerable similarity of the cytogenetic banding pattern (albeit relatively small chromosomal regions) between human and mouse (Sawyer and Hozier, 1986). Gene order has not been generally well conserved between mouse and human chromosomes. However, again, there is a conservation of gene order over small to moderately sized sub-chromosomal segments between mouse and human. This type of partial conservation of synteny (a group of linked genes in one species is paralleled comparably by a linkage group between the orthologous genes in the other species) is very useful in assignment of the gene of interest based on the gene order in other species.

5.1.2 Techniques for the mapping of specific gene

There are numerous approaches by which localisation of a gene of interest can be identified and the method selected in each case is largely dependent on the resource available for study. Two major types of gene mapping can be distinguished. In genetic mapping (linkage analysis) the frequency of meiotic crossover between loci is used to estimate inter-locus distances. Genetic linkage analysis is reliant on the existence of pedigrees in which the phenotype of interest is known to segregate. This method also requires that a polymorphic marker exists in sufficiently close proximity to the locus involved to enable linkage to be identified. Linkage analysis allows us to determine the relative distance between loci, but it does not assign specific locations to markers or disease genes. Physical mapping which involves the use of cytogenetic and molecular techniques to determine the actual physical locations of genes on chromosomes accomplishes this goal.

1 A gene could be physically mapped by associating cytogenetically observable variations (heteromorphism, translocations, deletions, and duplications) with the presence of a genetic disease.

2 *In situ* hybridisation is a physical mapping technique in which a labelled probe is hybridised to fixed metaphase chromosomes to determine the chromosomal location of the DNA represented by the probe (the gene of interest).

3 Different strains of mice have different alleles at many polymorphic loci, making it easy to recognise the origin of a marker allele. This property is exploited to construct a marker framework map or mapping a new phenotype or gene. In the backcross method any marker or cloned gene can be assigned rapidly to a small chromosomal segment defined by two recombination breakpoints in the collection of backcrossed mice.

4 Somatic cell hybridisation (SCH) is another physical technique in which human (or other species) and rodent cells are hybridised, resulting in cells that have a reduced number of donor chromosomes. In this method, somatic cells from different species (for example hamster and mouse) co-cultured in the presence of agents such as polyethylene glycol or Sendai virus, will sometimes fuse together to form hybrid cells (somatic cell hybridisation). The resulting hybrid cells contain the chromosomes of both species. The cells are exposed to selective media such as HAT (hypoxanthine, aminopterin and thymidine) to eliminate those cells that did not fuse. The hybrid cells are then allowed to replicate. These cells will begin to lose some of their mouse chromosomes as they undergo mitosis. Finally, cells remain that have a full set of hamster chromosomes but only one or a few mouse chromosomes. These cells are then karyotyped to determine which mouse chromosomes remain. Panels of such cells are used to correlate the presence of a gene with the consistent presence of one chromosome (or when translocations are available, a specific segment of a chromosome).

5 Radiation hybrid mapping (RH) is a powerful and widely used technique that will be described in more detail in next section.

5.1.3 Radiation hybrid (RH) analysis

A natural progression from SCH mapping was the development of methods by which assignment could be made at a sub-chromosomal level. The discovery of radiation hybrid mapping satisfied this need and was developed by Cox *et al.* (1990). The methodology is similar to that of the SCH mapping except that, prior to the fusion event, donor chromosomes are exposed to a measured and lethal dose of radiation, from an X-ray source. This radiation causes each donor chromosome to be broken into a series of random smaller fragments, which themselves are not able to be propagate (Goss and Harris, 1975). Fragments are rescued by

fusion with a recipient cell line, typically of rodent origin. A subsequent selection stage, comparable to that used in the generation of SCH panels, ensures that only hybrid cells survive. After construction of an appropriate panel, PCR analysis is performed using marker- or gene-specific primers on template DNA isolated from each cell line within the panel. Each cell line is scored for the presence or absence of an amplification product, indicating the subset of cell lines within the panel that have retained the fragment of chromosome on which the marker or gene lies. More recently, the radiation of the intact donor cells containing a full complement of donor chromosomes, and subsequent fusion with rodent recipient cells, has enabled the procedure to be applied to the mapping of entire genomes with a single panel of radiation hybrids. This whole genome radiation hybrid (WG-RH) method was first described by Walter *et al.* (1994).

In one study (Cox, 1992) it was shown that the retention frequency (number of cell lines analysed that contain the fragment of interest; typically expressed as a percentage) of individual chromosome fragments within an RH panel was relatively constant, regardless of their location within the chromosome. Other studies demonstrated an increased retention frequency for markers originating near the centromere of the donor chromosome. This effect has been proposed to indicate the increased stability of fragments that retain a centromere-like structure (Walter *et al.*, 1994). A similar effect at the telomere has also been demonstrated (Ceccherini *et al.*, 1992). Donor fragments may be maintained as separate structures in hybrid cells rather than being incorporated into recipient chromosomes. Therefore fragments containing telomeric or centromeric regions are likely to demonstrate a higher than average retention frequency in the recipient cells because these parts of chromosome are necessary for propagation of donor chromosome fragments that have not been incorporated into recipient chromosomes (Jones, 1996).

5.2 Results

5.2.1 *In silico* assignment (mapping) of the mouse *Spp2* gene

As discussed in Section 5.1.1, there is considerable conservation of synteny between humans and mice. Therefore, once the chromosomal location for a gene of interest is known in mouse or human, it is usually possible to predict the likely location of that gene in the other species. The human *SPP2* gene, encoding the spp24 protein has previously been assigned to the tip of the long arm of chromosome 2 (2q37-qter) using *in situ* hybridisation (Swallow *et al.*, 1997).

At the outset of this study, only the sequence of the *Spp2* cDNA was available (Bennett, 2002) and there was no available data about the location of the gene in mouse. Therefore it was decided to assign the location of the *Spp2* gene in mouse. There is a conservation of synteny between the distal region of the long arm of human chromosome 2 and mouse chromosome 1 (Strachan and Read, 1999). Therefore in the first attempt to assign the mouse *Spp2* gene, the genes on both sides of the human *SPP2* gene (spanning about 23 cM distance) and their orthologous genes on mouse chromosome 1 were determined (using the Entrez interface at NCBI). Six orthologous genes whose location had had previously been assigned in human were identified: *VILI, CRYBA2, DES, PTPRN, UGT1A1* and *HDLBP*. Based on these data, the mouse *Spp2* gene was provisionally assigned (mapped) to a 4.6 cM interval between 51.7 and 55.3 cM (approximately 8 Mb) of the mouse chromosome 1. Figure 5.1 schematically illustrates the assignment of the mouse Spp2 gene using this *in silico* method.

5.2.2 Interspecific backcross mapping of *Spp2*

This section begins by describing the work carried out prior to the start of work for this thesis. A mouse genomic PAC library RPCI21 (Osoegawa *et al.*, 2000) containing 254,217 clones with an insert of average size of 137 kb in the vector pPAC4 was screened for the mouse *Spp2* gene (Manship and Dalgleish, unpublished) and several positive clones identified carrying the gene. Subsequently, two of the PAC clones known to contain the *Spp2* gene were then used to make a small-insert sub-clone library (Swallow and Dalgleish, unpublished). This library, which provided a source of small fragments from the mouse *Spp2* gene, was screened by hybridisation using a $(CA)_n$ probe. Seven positive clones were identified and four of these were sequenced. Of these four clones, only three had a CA repeat of any significant length with clones E1/4E and H2/4F having the longest. A pair of PCR primers were designed for regions flanking the $(CA)_n$ repeat in H2/4F:

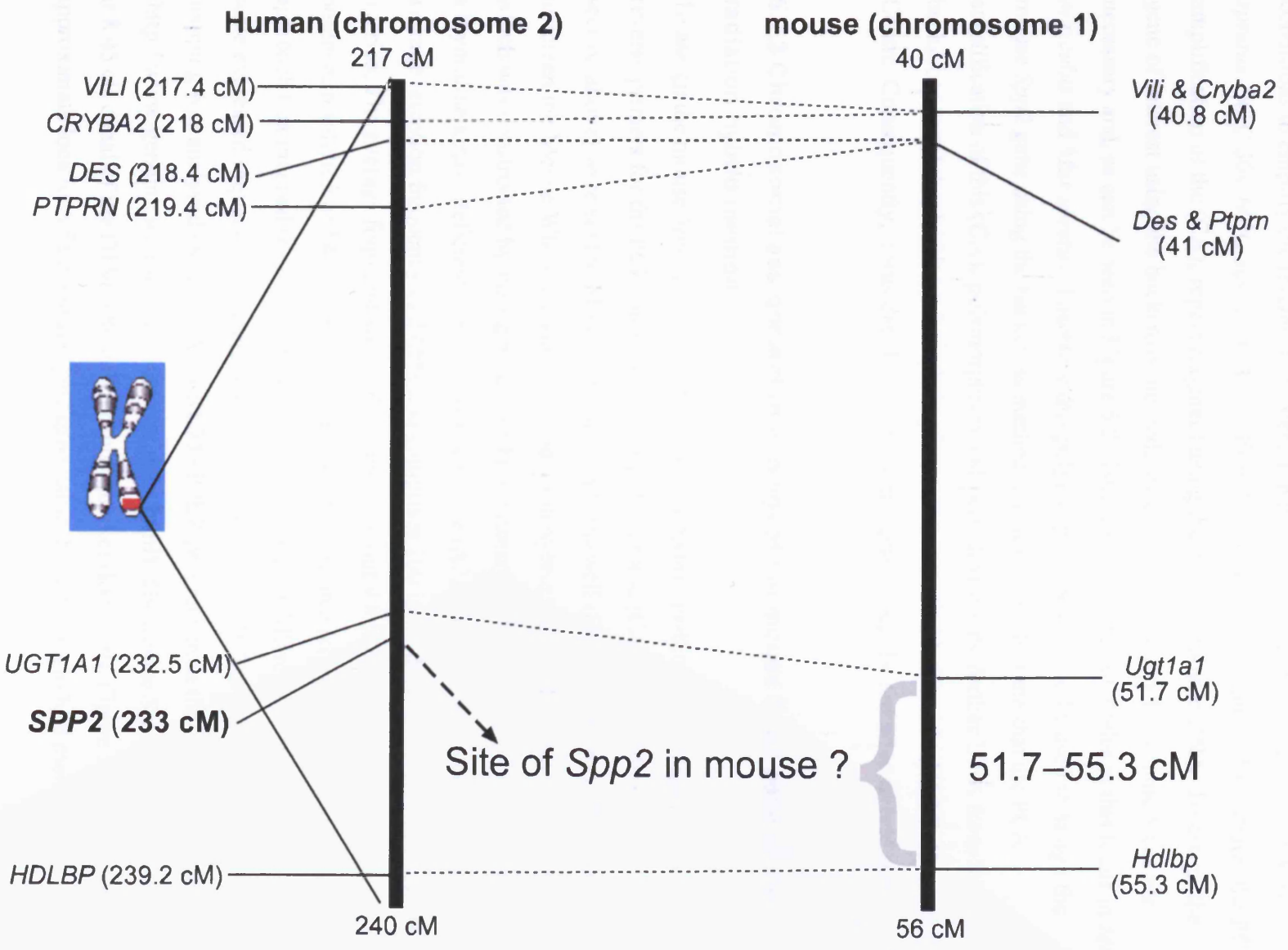


Figure 5.1. In silico assignment (mapping) of the mouse *Spp2* gene (not drawn to scale)

Forward 5′ AAACCCTTTCTGACTTGCATTTC 3′

Reverse 5′ TGACAAGCATGTTAAATAGTAGATGTG 3′

PCR amplification was carried out as described in Section 2.7.1 using the following PCR conditions to amplify the fragment of interest in mouse species *Mus musculus* and *Mus spretus*: 96°C 30s, 60°C 30s, 72°C 30s × 30 and 4.5 mM Mg^{2+}. Figure 5.2 illustrates the PCR amplification of the $(CA)_n$ repeat fragment using the mouse genomic DNA. To assign the gene of interest using the backcross method, different alleles at a polymorphic locus are necessary and, as can be seen in Figure 5.2, there are two different alleles at this locus in *Mus musculus* and *Mus spretus*. Therefore this polymorphic locus could be used to assign the mouse *Spp2* gene using the backcross method. However, by the time that the PCR amplification of this $(CA)_n$ polymorphism had been devised, no further DNA from the mouse backcross panel that had been intended to be used was available from the HGMP Resource Centre. Consequently, it was decided to assign the gene using the radiation hybrid method.

5.2.3 Chromosomal assignment (mapping) of the mouse *Spp2* gene using radiation hybrid method

To assign the mouse *Spp2* gene using the radiation hybrid method, the designed forward and reverse primers for the PCR amplification of polymorphic $(CA)_n$ repeat (described in the section above) were sent to Mouse Genome Unit, Harwell (Dr. Paul Denny). These were used to screen the Mouse Whole Genome T31 radiation hybrid panel (McCarthy *et al.*, 1997) which was constructed by fusing mouse embryo primary cells, irradiated with 3,000 rad, with a thymidine kinase deficient (TK^-) hamster cell line (A23). This mouse WG-RH panel has an average retention frequency of 27.6% and comprises 100 individual whole-genome-radiation hybrids. The average fragment size in this panel is about 9.8 Mb and has an estimated potential resolution of 145 kb, making it a powerful resource (comparing to *in silico* mapping approach that mapped the *Spp2* gene to an approximately 8 Mb region) for efficient large-scale expressed sequence tag mapping (McCarthy *et al.*, 1997). The result of radiation hybrid mapping was analysed using the Auto-RHMAPPER program from the Whitehead Institute (http://www-genome.wi.mit.edu/mouse_rh/index.html). The mouse *Spp2* gene was assigned at 8.45 cR distal of the D1Mit486 polymorphic marker (lod >3.0). Figure 5.3 illustrates the approximate location of the mouse *Spp2* gene using the radiation hybrid method.

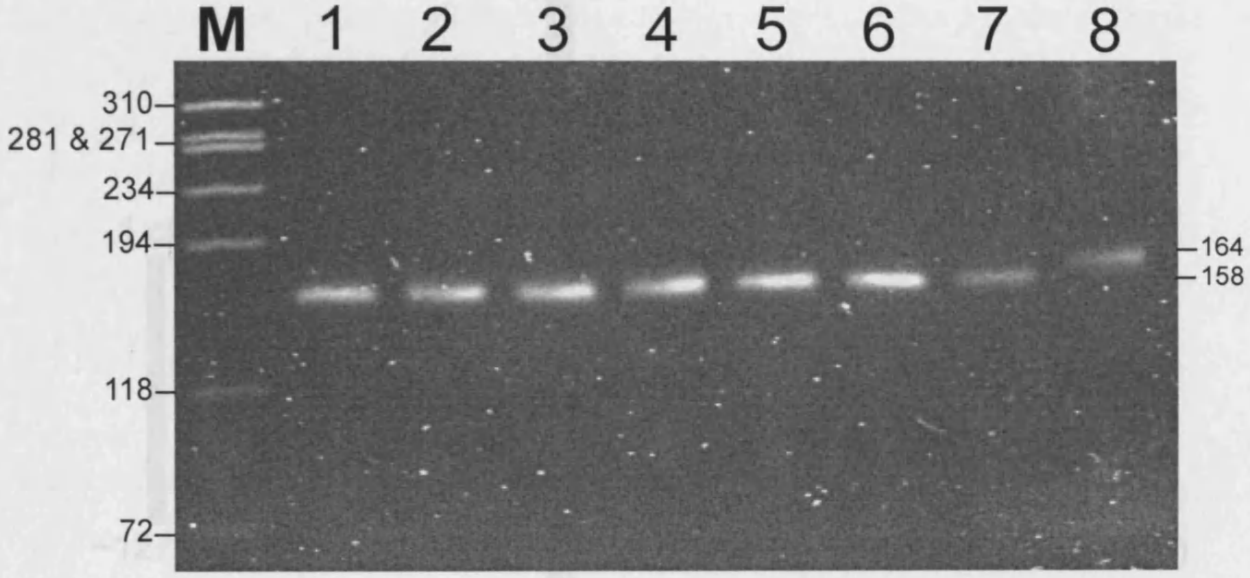


Figure 5.2. PCR amplification of the $(CA)_n$ repeat of clone H2/4F using mouse genomic DNA of *Mus musculus and Mus spretus*

Sample 1-7: different samples (tissues) of *Mus musculus*
Sample 8: *Mus spretus*

M = marker (ΦX174RF/*Hae*III)

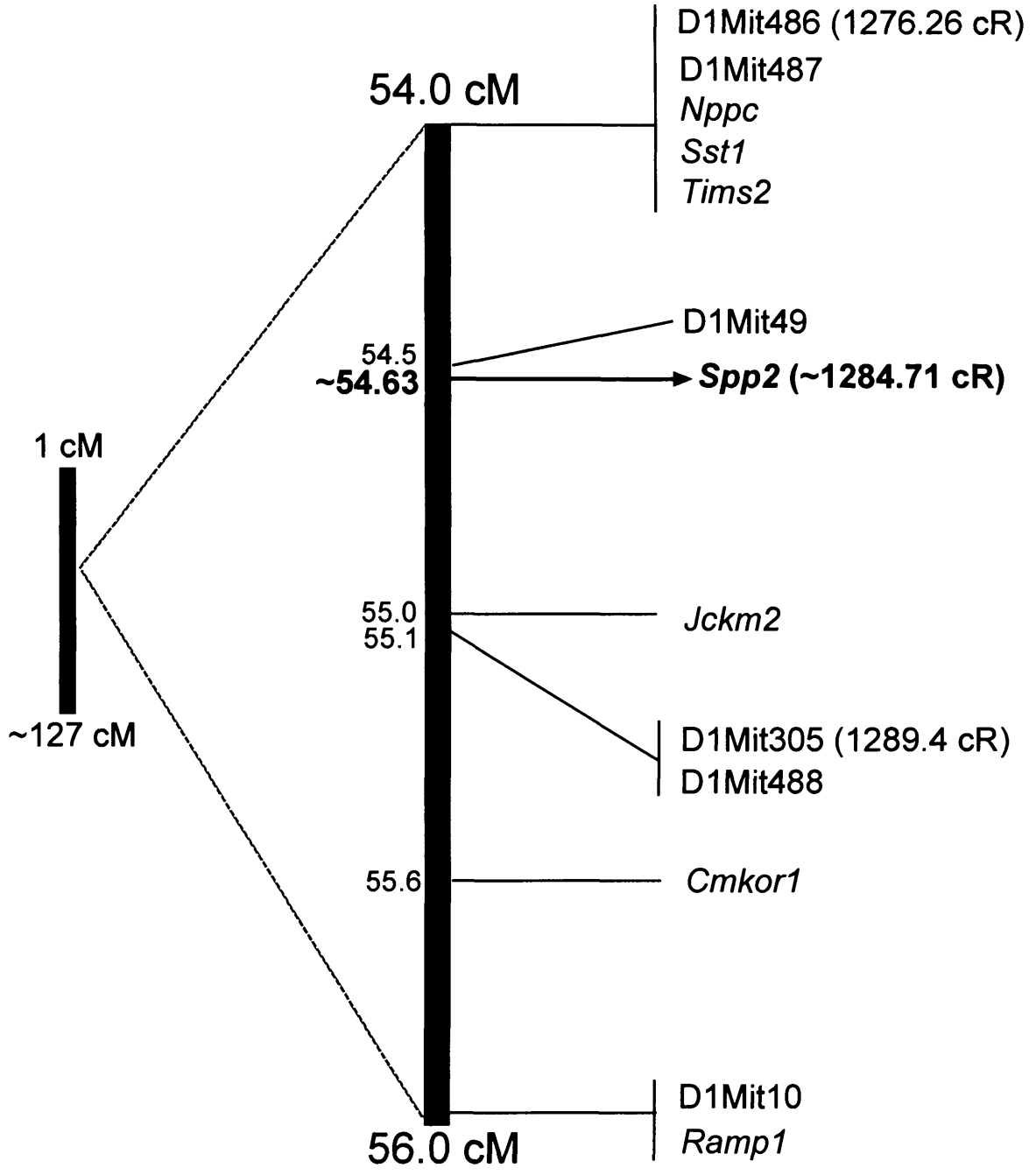


Figure 5.3. Chromosomal assignment (mapping) of the mouse *Spp2* gene to chromosome 1 using the radiation hybrid (RH) method.

5.2.4 BAC clones that cover the flanking and interval region of D1Mit486 and D1Mit305 markers

The RH mapping indicated that *Spp2* lies between the anchor markers D1Mit486 and D1Mit305, a 13.14 cR interval. To confirm this assignment, and to increase the resolution of the map, four overlapping BAC genomic DNA clones were identified covering the interval between the D1Mit486 and D1Mit305 markers along with their immediate flanking sequences (Figure 5.4A) using the mouse chromosome 1 contig map (http://mouse.ensembl.org). The distance between these two markers is about 365,000 bp and clones RP24-391C11, RP24-73K7, RP24-108I6 and RP23-334J22 have enough overlap to identify the approximate location of the mouse *Spp2* gene. These clones are from RPCI mouse genomic BAC libraries RCPI-23 and RPCI-24 (Osegawa and Taneto, 2000) and were purchased from BACPAC Resources, Children's Hospital Oakland Research Institute, Oakland, USA.

5.2.5 Southern analysis of BAC clones

All the work described in this section was carried out using the BAC and PAC preps, restriction enzyme digestion and Southern blotting protocols detailed in Sections 2.12.8.2, 2.3 and 2.26 of Chapter 2 respectively.

DNA mini-preps were prepared for the four BAC clones (RP24-391C11, RP24-73K7, RP24-108I6, RP23-334J22) and also from six of the PAC clones from the RPCI-21 mouse genomic library (RP21-397P17, RP21-475K16, RP21-555B9, RP21-614M2, RP21-643E5, RP21-644E5) to act as positive controls for the mouse *Spp2* gene and from the PAC clone RP21-399H13 which does not contain the mouse *Spp2* gene, as a negative control (the average insert sizes in RPCI23 and 24 are 197 and 155 kb respectively). The mini-preps were digested with *Eco*RI and, following agarose gel electrophoresis, the DNA fragments were Southern blotted and hybridised with a mouse *Spp2* cDNA probe (BAC clone RP23-334J22 was omitted from the analysis as it contained no identifiable genomic DNA insert).

BAC clones RP24-391C11, RP24-73K7 and all positive control PAC clones (RP21-397P17, RP21-475K16, RP-21555B9, RP21-614M2, RP21-643E5 and RP21-644E5) gave a positive signal for hybridisation to the mouse *Spp2* cDNA probe. BAC clone RP24-108I6 and PAC clone RP21-399H13 (negative control) did not give any positive signal. This indicates that both RP24-391C11 and RP24-73K7 contain the entire or part of the mouse *Spp2* gene sequence. Figure 5.4B shows the result of the Southern analysis. As can be seen, both the

Figure 5.4A. Schematic representation of the four overlapping BAC clones which encompass the D1Mit486 and D1Mit305 markers

The distance between the D1Mit486 and D1Mit305 markers is about 365,000 bp and clones RP24-391C11, RP24-73K7, RP24-108I6 and RP23-334J22 have enough overlap to identify the approximate site of the mouse *Spp2* gene. Clone RP23-334J22 extends to the distal part of the image (it is not shown). These clones are from RPCI mouse genomic BAC libraries 23 and 24 and were purchased from BACPAC Resources, Children's Hospital Oakland Research Institute, Oakland, USA. It should be mentioned that clone RP23-334J22 has no insert (presumably did originally had lost its site when we come to do this work). The scaling bar reffering to 100.82, 100.88 *etc* is 60 kb.

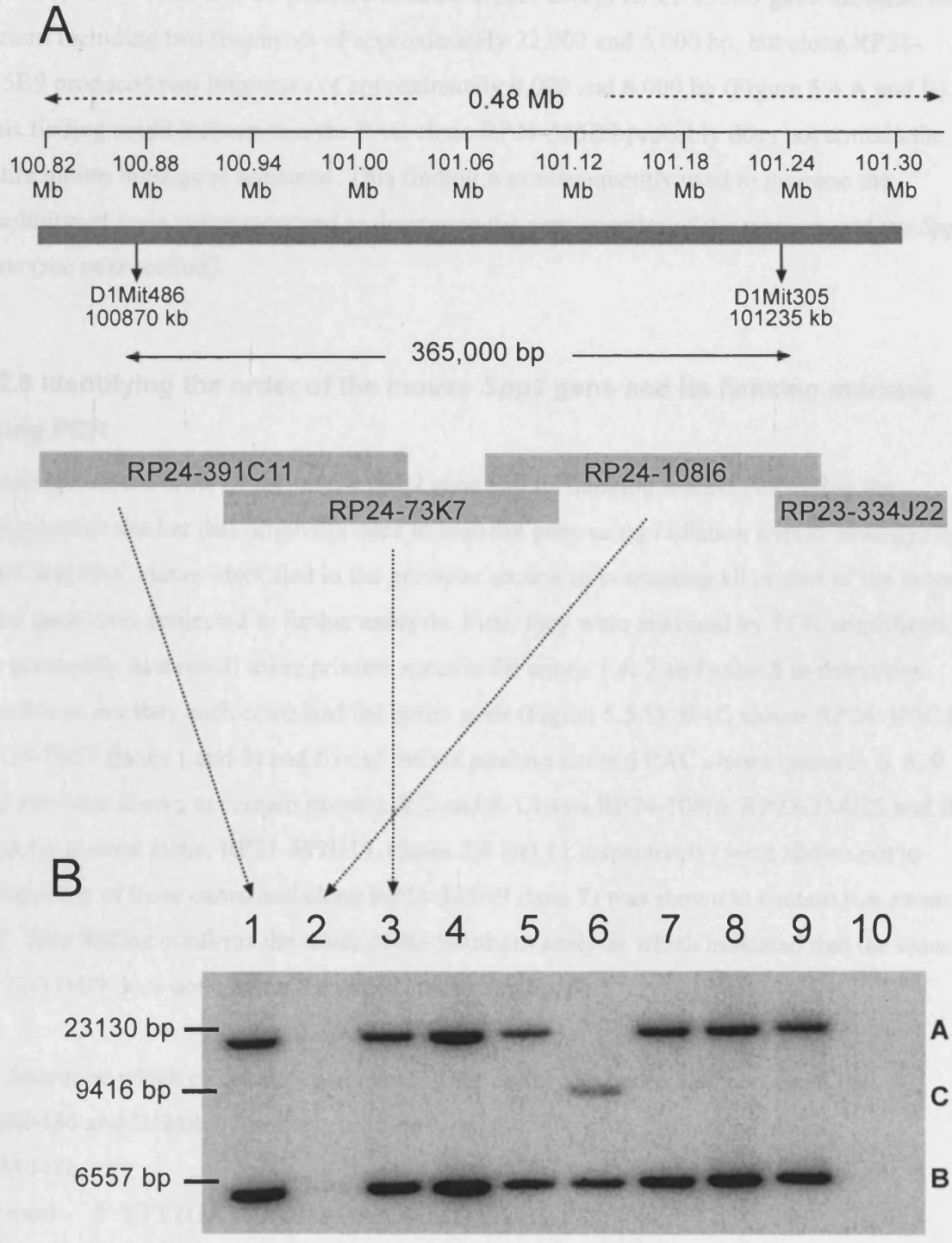
Figure 5.4B. Autoradiograph of the Southern blot of 10 mouse genomic clones digested with *Eco*RI and hybridised with a mouse *Spp2* cDNA probe.

The position of the three largest λ/*Hin*dIII marker fragments are indicated and only two visible bands for each sample are labelled (A and B or B and C). Arrows link the samples to their corresponding clones.

lane 1: RP24-391C11
lane 2: RP24-108I6
lane 3: RP24-73K7

Positive controls clones:
lane 4: RP21-379P17
lane 5: RP21-475K16
lane 6: RP21-555B9
lane 7: RP21-614M2
lane 8: RP21-643E5
lane 9: RP21-644E5

Negative control clone:
lane 10: RP21-399H13


A
0.48 Mb
100.82 Mb
100.88 Mb
100.94 Mb
101.00 Mb
101.06 Mb
101.12 Mb
101.18 Mb
101.24 Mb
101.30 Mb
D1Mit486
100870 kb
D1Mit305
101235 kb
365,000 bp
RP24-391C11
RP24-108I6
RP24-73K7
RP23-334J22
B
1 2 3 4 5 6 7 8 9 10
23130 bp
9416 bp
6557 bp
A
C
B

clones gave a positive hybridisation signal which indicates that the *Spp2* gene falls in the overlapping region of these two clones which is approximately 100,000 bp. Clones RP24-391C11, RP24-73K7 and all positive controls clones except RP21-555B9 gave the same band pattern including two fragments of approximately 22,000 and 6,000 bp, but clone RP21-555B9 produced two fragments of approximately 9,000 and 6,000 bp (Figure 5.4 A and B). This finding could indicate that the PAC clone RP21-555B9 probably does not contain the entire mouse *Spp2* gene sequence. This finding was subsequently used to increase the resolution of gene assignment and to determine the precise order of the markers and the *Spp2* gene (see next section).

5.2.6 Identifying the order of the mouse *Spp2* gene and its flanking markers using PCR

To determine the order of the mouse *Spp2* gene and its flanking marker (including the polymorphic marker that originally used to map the gene using radiation hybrid strategy) the BAC and PAC clones identified in the previous section as containing all or part of the mouse *Spp2* gene were subjected to further analysis. First, they were analysed by PCR amplification (as previously described) using primers specific for exons 1 & 2 and exon 8 to determine whether or not they each contained the entire gene (Figure 5.5A). BAC clones RP24-391C11, RP24-73K7 (lanes 1 and 3) and five of the six positive control PAC clones (lanes 5, 6, 8, 9 and 10) were shown to contain exons 1 & 2 and 8. Clones RP24-108I6, RP23-334J22 and the negative control clone, RP21-399H13, (lanes 2,4 and 11 respectively) were shown not to contain any of these exons and clone RP21-555B9 (lane 7) was shown to contain just exons 1 & 2. This finding confirms the result of the Southern analysis which indicated that the clone RP21-555B9 does not contain the entire mouse *Spp2* gene.

To determine which of the clones containing the entire *Spp2* gene also contained the D1Mit486 and D1Mit305 polymorphic markers, the following primers were used:

D1Mit486

Forward 5′ TTTTGATCAGGAGAAAGAGAGAGG 3′

Reverse 5′ TAAACCACATGGAAAGAATCACA 3′

D1Mit305

Forward 5′ GTGGGAACCTTCTACTATTTCTATGC 3′

Reverse 5′ GTGTACCTCCTTTCTGTTTATGGG 3′

Figure 5.5A. PCR amplification of exons 1 + 2 and 8 of the mouse *Spp2* gene

Exons 1 & 2 and 8 of the mouse *Spp2* gene amplification and agarose gel electrophoresis.

lane 1: RP24-73K7
lane 2: RP24-108I6
lane 3: RP24-391C11
lane 4: RP23-334J22 (the clone has no genomic insert)

positive control clones:
lane 5: RP21-379P17
lane 6: RP21-475K16
lane 7: RP21-555B9
lane 8: RP21-614M2
lane 9: RP21-643E5
lane 10: RP21-644E5

negative control clone:
lane 11: RP21-399H13

M = marker (ΦX174RF/*Hae*III)

Figure 5.5B. PCR amplification of D1Mit486 and D1Mit305 markers

PCR amplification of D1Mit486 and D1Mit305 markers in different clones and agarose gel
electrophoresis.

lanes 1 & 7: RP24-73K7
lanes 2 & 8: RP24-108I6
lanes 3 & 9: RP24-391C11
lanes 4 & 10: RP23-334J22 (the clone has no genomic insert)

negative control clone:
lanes 5 & 11: RP21-399H13

positive control:
lanes 6 & 12: mouse genomic DNA

M = marker (ΦX174RF/*Hae*III)

Figure 5.5C. PCR amplification of the $(CA)_n$ repeat marker

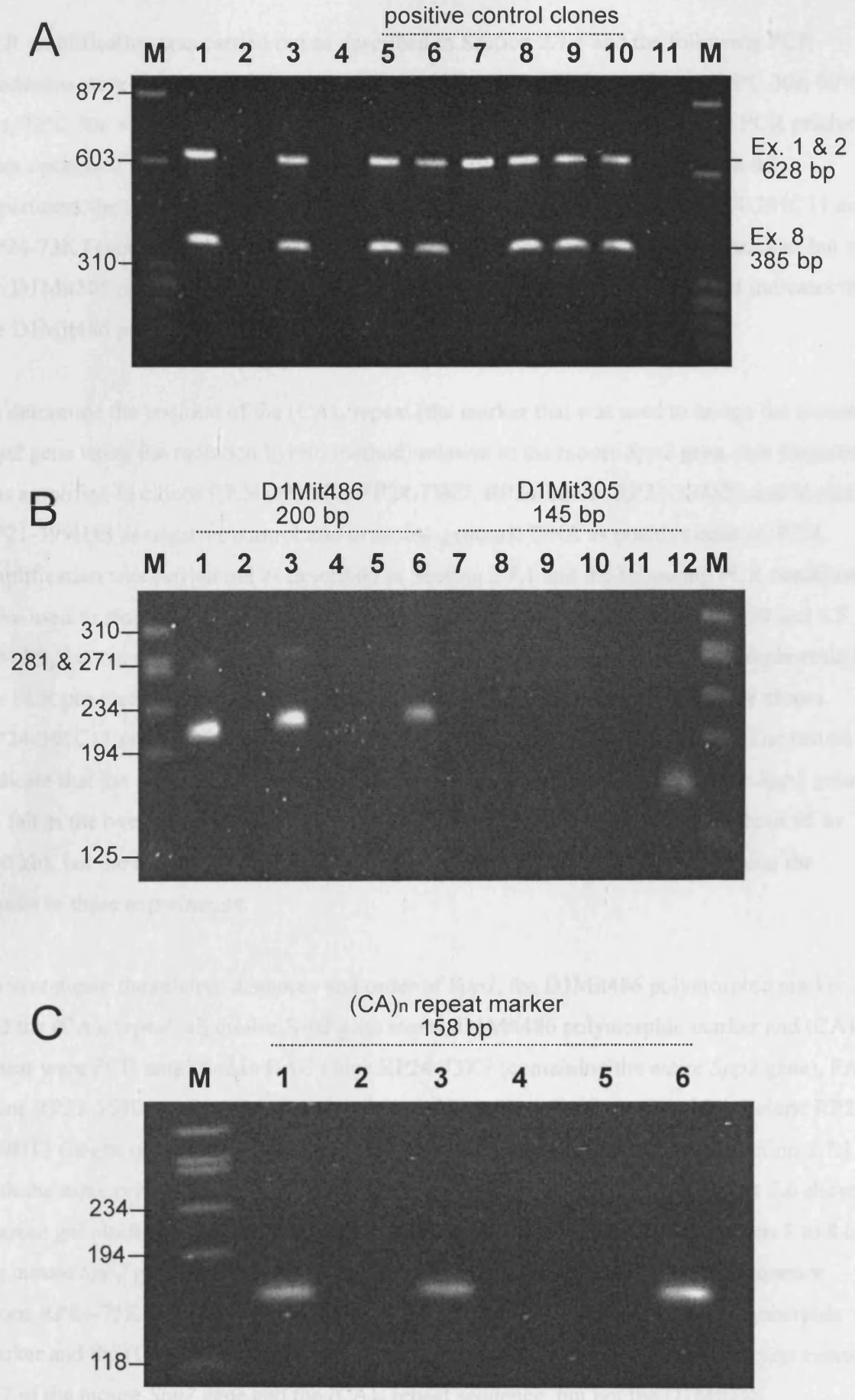
PCR amplification of the $(CA)_n$ repeat marker in different clones and agarose gel electrophoresis.

lane 1: RP24-73K7
lane 2: RP24-108I6
lane 3: RP24-391C11
lane 4: RP23-334J22 (the clone has no genomic insert)

negative control clone:
lane 5: RP21-399H13

positive control:
lane 6: mouse genomic DNA

M = marker (ΦX174RF/*Hae*III)


A
positive control clones
M 1 2 3 4 5 6 7 8 9 10 11 M
872
603
310
Ex. 1 & 2
628 bp
Ex. 8
385 bp
B
D1Mit486
200 bp
D1Mit305
145 bp
M 1 2 3 4 5 6 7 8 9 10 11 12 M
310
281 & 271
234
194
125
C
(CA)n repeat marker
158 bp
M 1 2 3 4 5 6
234
194
118

PCR amplification was carried out as described in Section 2.7.1 and the following PCR conditions were used to amplify the fragments of interest in different clones: 96°C 30s, 60°C 30s, 72°C 30s × 30 and 4.5 mM Mg^{2+}. Figure 5.5B shows an agarose gel of the PCR products from optimised reactions for D1Mit486 and D1Mit305 polymorphic markers. In this experiment the mouse genomic DNA was used as positive control. Clones RP24-391C11 and RP24-73K7 (lanes 3 and 1 respectively) were shown to contain the D1Mit486 marker, but not the D1Mit305 marker (lanes 9 and 7 respectively). The result of this experiment indicates that the D1Mit486 polymorphic marker is closer to the *Spp2* gene than D1Mit305.

To determine the position of the $(CA)_n$ repeat (the marker that was used to assign the mouse *Spp2* gene using the radiation hybrid method) relative to the mouse *Spp2* gene, this fragment was amplified in clones RP24-391C11, RP24-73K7, RP24-108I6, RP23-334J22 and in clone RP21-399H13 as negative control and in mouse genomic DNA as positive control. PCR amplification was carried out as described in Section 2.7.1 and the following PCR conditions were used to amplify the fragments of interest: 96°C 30s, 60°C 30s, 72°C 30s × 30 and 4.5 mM Mg^{2+} concentration (Section 5.2.1). Figure 5.5C shows an agarose gel electrophoresis of the PCR products from optimised reactions for the $(CA)_n$ repeat marker. Both the clones RP24-391C11 and RP24-73K7 were shown to contain the $(CA)_n$ repeat marker. The results indicate that the $(CA)_n$ repeat, the D1Mit486 polymorphic marker and the mouse *Spp2* gene all fall in the overlapping region of the clones RP24-391C11 and RP24-73K7 (about 95 to 100 kb), but the distances between them and their order cannot be determined using the results of these experiments.

To investigate the relative distances and order of *Spp2*, the D1Mit486 polymorphic marker and the $(CA)_n$ repeat, all mouse *Spp2* gene exons, D1Mit486 polymorphic marker and $(CA)_n$ repeat were PCR amplified in BAC clone RP24-73K7 (containing the entire *Spp2* gene), PAC clone RP21-555B9 (containing just some part of the mouse *Spp2* gene) and PAC clone RP21-399H13 (negative control). PCR amplification was carried out as described in Section 2.7.1., with the same primers as before (Table 4.2.3 and Sections 4.2.2 and 5.2.5). Figure 5.6 shows agarose gel electrophoresis of the PCR products from optimised reactions for exons 1 to 8 of the mouse *Spp2* gene, the D1Mit486 polymorphic marker and the $(CA)_n$ repeat sequence. Clone RP24-73K7 was shown to contain the entire *Spp2* gene, the D1Mit486 polymorphic marker and the $(CA)_n$ repeat sequence. Clone RP21-555B9 was shown to contain just exons 1 to 7 of the mouse *Spp2* gene and the $(CA)_n$ repeat sequence, but not the D1Mit486 polymorphic marker. The PCR amplification of exon 7 from clone RP21-555B9 was not as

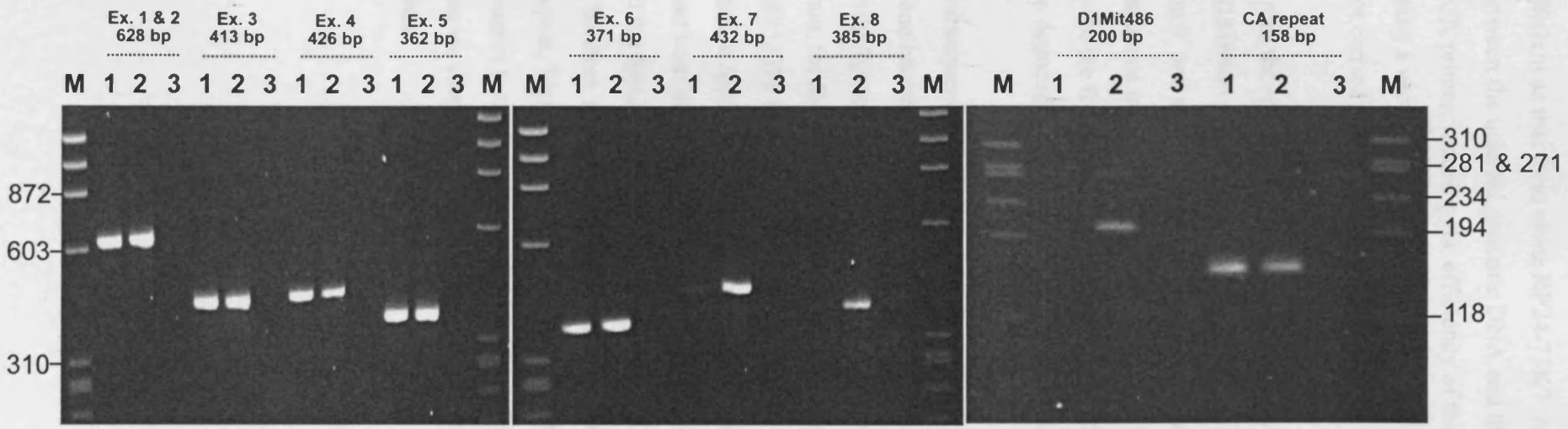


Figure 5.6. PCR amplification of exons 1 to 8 of the mouse *Spp*2 gene, D1Mit486 polymorphic marker and $(CA)_n$ repeat sequence.

Exons 1 to 8 of the mouse *Spp2* gene, D1Mit486 polymorphic marker and $(CA)_n$ repeat sequence amplification and agarose gel electerophoresis.

lane 1: RP21-555B9 (containing part of the mouse *Spp2* gene)
lane 2: RP24-73K7 (containing the entire *Spp2* gene)
lane 3: RP21-399H13 (negative control)

M = marker (ΦX174RF/*Hae*III)

efficient as that from clone RP24-73K7. A possible explanation for this is that the boundary between the inserted genomic DNA and the vector falls with the binding site for the reverse PCR primer, reducing the efficiency of the reaction. DNA sequence analysis of the boundary using a vector-specific sequencing primer could have resolved this issue, but the analysis was not carried out.

From the results of the PCR amplification of all the mouse *Spp2* gene exons and the D1Mit486 polymorphic marker, it can be deduced that D1Mit486 must lie within or beyond the 3′ end of *Spp2* as this marker is not amplified from clone RP21-555B9. The $(CA)_n$ repeat sequence must lie 5′ of exon 7 as it amplifies from both clones RP21-555B9 and RP24-73K7, since the former does not appear to extend beyond exon 7. The location of D1Mit305 cannot be deduced from these results as it failed to amplify from any of the genomic clones analysed.

Subsequent to these analyses, the DNA sequence of mouse chromosome 1 around the *Spp2* gene became available. The sequence was found at the UCSC Genome Web site (http://genome.ucsc.edu) by searching the data with their "BLAT" search tool. Using UCSC data, the *Spp2* gene was mapped (from the start of exon 1 to the end of exon 8) to 88832387–88853226 bp of the mouse chromosome 1. The sequence data reveal, unexpectedly, that the mouse *Spp2* gene does not fall between the D1Mit486 and D1Mit305 polymorphic markers as had been thought from the RH mapping data. Instead, D1Mit486 and D1Mit305 both lie distal to the *Spp2* gene with D1Mit486 closer. The $(CA)_n$ repeat sequence lies proximal to the gene. Therefore the correct order (from centromere to telomere of mouse chromosome 1) is $(CA)_n$ repeat, *Spp2* gene, D1Mit486 and D1Mit305 (Figure 5.7). In spite of the discrepancy with respect to the locations of D1Mit486 and D1Mit305, the initial regional assignment of *Spp2* to mouse chromosome 1 is entirely consistent with the final mouse genomic DNA sequence data.

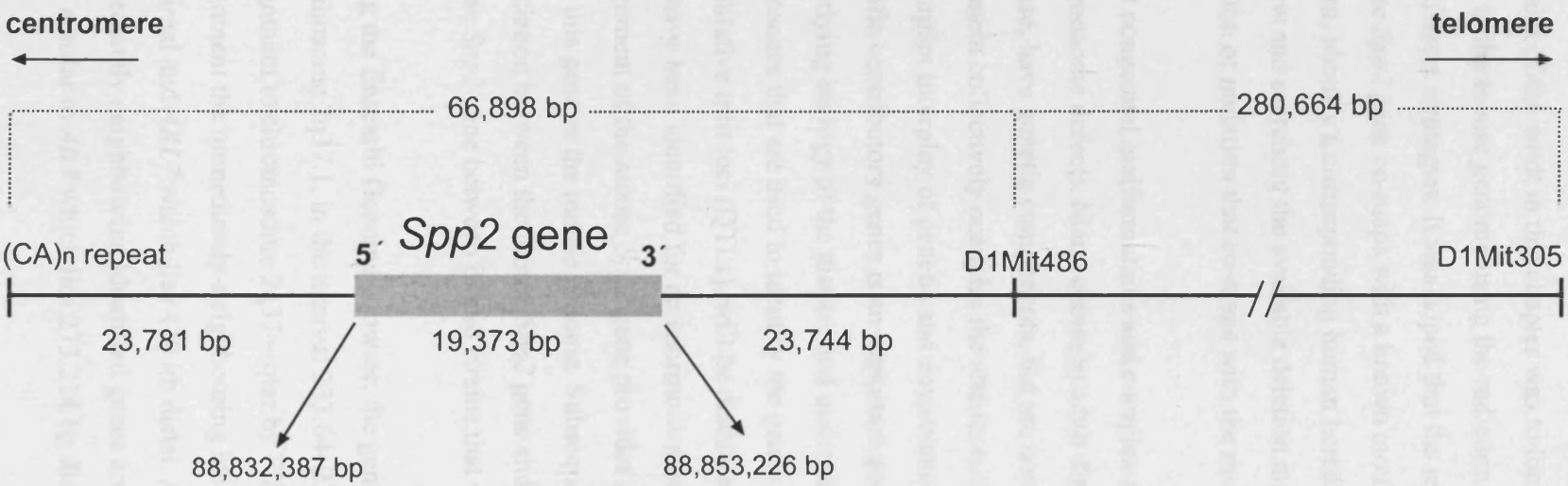


Figure 5.7. Mapping the locations of (CA)n repeat, *Spp2* gene, DiMit486 and D1Mit305 polymorphic markers

This figure shows the mapping sites of the (CA)n repeat, the *Spp2* gene, the DiMit486 and D1Mit305 polymorphic markers and their interval sizes (in bp) by analysis of the DNA sequence at the UCSC GenomeWeb site by searching their data with their ''BLAT'' search tool.

5.3 Discussion

The aim of the work in this chapter was to localise the gene orthologous to the human *SPP2* gene in the mouse genome using the radiation hybrid method and confirmation of the result by different strategies. It was hoped that the result of this study might have revealed that the mouse *Spp2* gene co-maps with a known mouse single-gene mutant phenotype which could, in turn, identify a corresponding human hereditary disorder. After an extensive literature review and searching the available deletion mapping databases of the mouse genome, no deletion or mutation that co-maps with the mouse *Spp2* gene could be found.

Most congenital malformations and complex disorders are not caused by single genes or chromosome defects. Many common adult diseases, such as osteoporosis, cancer and heart disease, have genetic components, but are not caused by single genes. These diseases, whose treatment collectively occupies the attention of most health care practitioners, are the result of a complex interplay of genetic and environmental factors. Nevertheless, the identification of specific contributory genes is an important goal, since only then can we start to understand the underlying biology of the disease and undertake correcting the defect. One of the several approaches that are used to identify the genes underlying multifactorial traits, known as quantitative trait loci (QTLs), will be discussed in more detail in Chapter 6. Quantitative trait loci have been identified for an enormous number of different mouse phenotypes and assignment of the mouse *Spp2* gene provides an opportunity to identify any QTLs that co-map with this gene in the mouse genome. Subsequently it might be possible to investigate the association between the mouse *Spp2* gene and any QTL using sequence alteration of the mouse *Spp2* gene between inbred strains that vary for the trait of interest.

Using the Ensembl Genome Browser, the gene encoding human spp24 was located to chromosome 2q37.1, in the interval 233.64–233.67 Mb which is consistent with the initial assignment to chromosome 2q37→qter by FISH (Swallow *et al.*, 1997). According to this assignment the immediately-neighbouring identified genes are *TRPM8* which lies 60 kb proximal and *ARL7* which lies 410 kb distal. As expected, in the mouse genome the immediately neighbouring identified genes are *Trpm8* which lies 17,921 bp proximal and a gene similar to *Alr7* which lies 273,214 bp distal.

To further investigate the organisation, order and orientation of the functional genes in the vicinity of the genes encoding spp24 protein in human and mouse, a comparative analysis was

performed between the human map around the *SPP2* locus and the corresponding mouse syntenic region around the *Spp2* locus using the Entrez interface at NCBI. There is extensive conservation of synteny between human chromosome 2 and mouse chromosome 1 in the vicinity of *SPP2* and *Spp2*. The orders of genes proximal and distal to the human *SPP2* gene (spanning about 6.23 Mb) and the mouse *Spp2* gene (spanning about 4.65 Mb) were determined. Seventeen functional genes on both sides of the human *SPP2* and their orthologous genes on the mouse chromosome 1 were identified. These genes are *NPPC, ECEL1, CHRND, CHRNG, NGEF, NEU2, INPP5D, SAG, TRPM8, ALR7, GBX2, COLA6A3, MLPH, RAB17, LRRFIP1, RAMP1* and *SCLY*. Figure 5.8 shows human chromosome 2 from 232.75 Mb to 239.02 Mb and the syntenic region on mouse chromosome 1 from 86.84 Mb to 91.52 Mb. Mapped human genes and their orthologous gene in mouse and their separation in bp are shown. As can be seen for the entire region, the order and orientation of the seventeen genes is absolutely conserved. Analysis was not extended beyond these genes and the synteny may extend further than this study reveals. Since the human and mouse genomic sequences are now almost complete, whole human-mouse conserved synteny maps are now also available from several sources and can be accessed from gateways such as (http://www.ncbi.nlm.nih.gov/Homology).

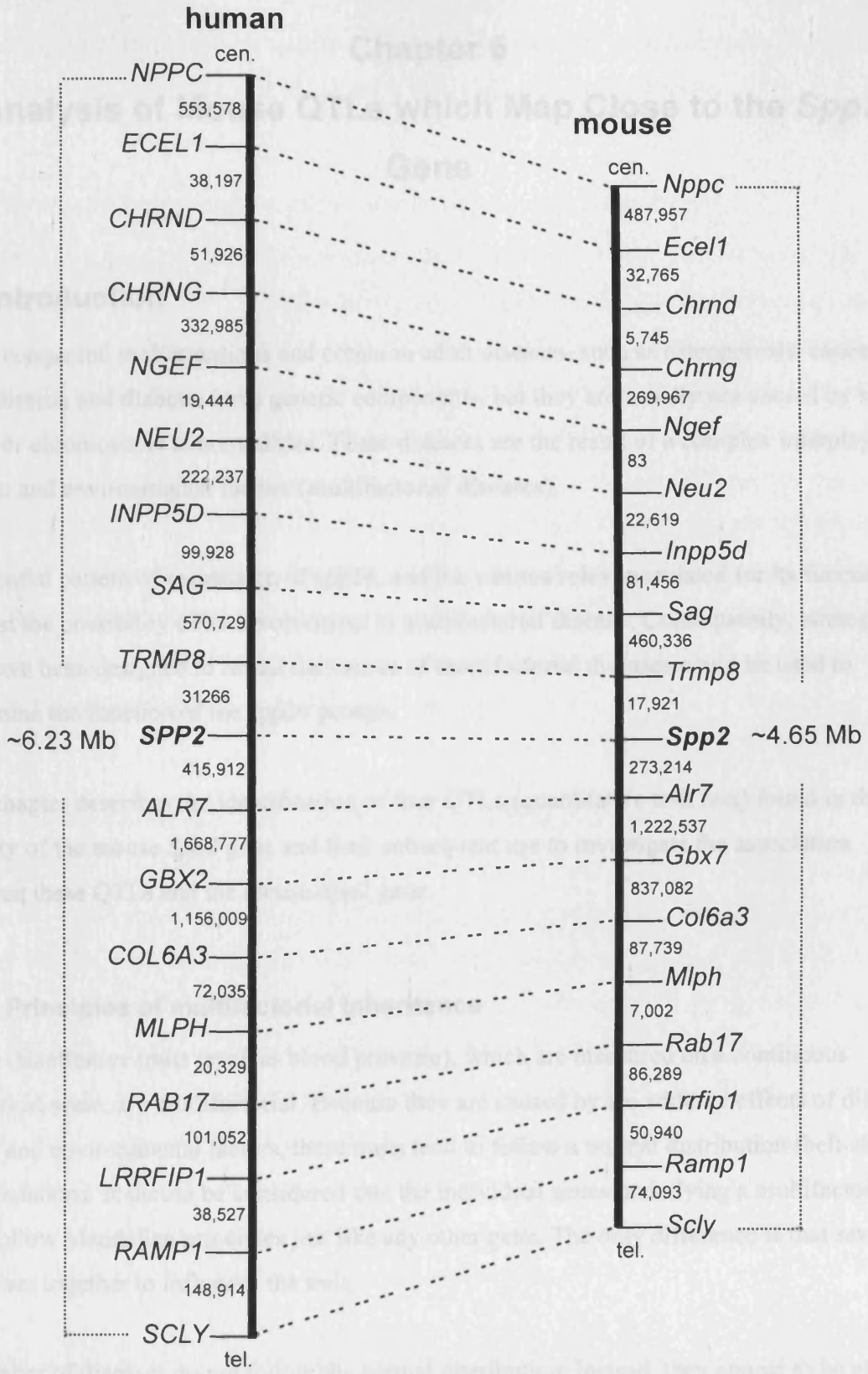


Figure 5.8. Physical map of the regions around the genes encoding the spp24 protein in human and mouse

This figure shows the part of human chromosome 2 from 232.75 Mb to 239.02 Mb and its mouse syntenic part on chromosome 1 from 86.84 Mb to 91.52 Mb. Mapped human genes and their orthologous gene in mouse and their separation are shown in base pairs (not to scale).

Chapter 6
Analysis of Mouse QTLs which Map Close to the *Spp2* Gene

6.1 Introduction

Many congenital malformations and common adult diseases, such as osteoporosis, cancers, heart disease and diabetes have genetic components, but they are usually not caused by single genes or chromosome abnormalities. These diseases are the result of a complex interplay of genetic and environmental factors (multifactorial diseases).

The spatial pattern of expression of spp24, and the various roles speculated for its function, suggest the possibility of its involvement in multifactorial disease. Consequently, strategies that have been designed to reveal the causes of multifactorial diseases could be used to determine the function of the spp24 protein.

This chapter describes the identification of four QTLs (quantitative trait loci) found in the vicinity of the mouse *Spp2* gene and their subsequent use to investigate the association between these QTLs and the mouse *Spp2* gene.

6.1.1 Principles of multifactorial inheritance

Many quantitative traits (such as blood pressure), which are measured on a continuous numerical scale, are multifactorial. Because they are caused by the additive effects of different genes and environmental factors, these traits tend to follow a normal distribution (bell-shaped) in populations. It should be considered that the individual genes underlying a multifactorial trait follow Mendelian principles just like any other gene. The only difference is that several genes act together to influence the trait.

A number of diseases do not follow the normal distribution. Instead, they appear to be either present or absent and they do not follow the patterns expected of single-gene disorders either. These types of disorders follow the liability distribution in a population. For multifactorial diseases (such as pyloric stenosis) that are either present or absent, it is thought that a

threshold of liability must be crossed before the disease is expressed. Below the threshold the individual appears normal but above it, he or she is affected by the disease.

6.1.2 Estimating the relative influence of genes and environment

Two research strategies, twin studies and adoption studies, are often used to estimate the relative influence of genes and environment. Twin studies usually consist of comparisons between monozygotic (MZ) and dizygotic (DZ) twins. If both members of a twin pair share a trait, they are said to be concordant and if they do not, they are discordant. Comparisons of correlations and concordance rates in monozygotic and dizygotic twins allow the estimation of heritability, a measure of the proportion of population variation in a disease that could be attributed to genes. Adoption studies provide a second means of estimating the influence of genes on multifactorial diseases. They consist of comparing disease rates among the adopted offspring of affected parents with the rates among adopted offspring of unaffected parents. Although both twin and adoption studies provide valuable information, they are also affected by certain biases that should be considered in any study.

6.1.3 Finding the underlying genes using quantitative trait loci (QTLs)

Although both twin and adoption studies provide valuable information, they are not designed to reveal specific genes that cause multifactorial diseases. Although the identification of specific causative genes in multifactorial diseases is an important goal, it is also a formidable task. Fortunately, recent advances in gene mapping and molecular biology make this goal more attainable. There are different approaches that can be used to identify the genes underlying multifactorial traits (or genes of small effect), often known as quantitative trait loci (QTLs). A QTL analysis can allow us to address specific questions concerning genetic architecture, such as the number of loci potentially affecting the trait, the distribution of gene effects, and the underlying patterns of gene action including dominance, sex specificity, epistasis and pleiotropy (reviewed in Falconer and Mackay, 1996).

One way to search for QTLs is to use conventional linkage analysis. In this method (often termed a genome scan) the disease families are collected, a single-gene mode of inheritance is assumed (inferred by segregation analysis) and linkage analysis is undertaken with a large series of polymorphic markers that span the genome. If a sufficiently large LOD score (logarithm of odds) is found with a polymorphic marker, it is assumed that the region in the vicinity of this marker may contain a QTL. This method is sometimes very helpful, especially

when there are large families in which a single-gene mode of inheritance is observed (*e.g.* familial breast cancer).

In most multifactorial disorders, to find large disease families with a single-gene mode of inheritance is difficult, therefore using linkage analysis is often impractical. An alternative to linkage analysis is the affected sib-pair method originally described by Penrose (1953). In this approach, following the collection of DNA samples from a large number of sib pairs (in which both members of the pair are affected by the disease) a genome scan is undertaken. If it is found that the siblings share the same allele for a polymorphic marker more frequently than half the time, this would be evidence that the marker probably is linked to a susceptibility locus or a QTL. The affected sib-pair method has the advantage that the investigator does not have to speculate a specific mode of inheritance and the method is not affected by reduced penetrance because both members of the sib pair should be affected to be included in the study. The weakness of this strategy is that it requires large sample sizes to yield a meaningful and significant result and it tends to have low resolution (Jorde *et al.*, 2000).

The final method combines genome scanning and the use of animal models. To use this method, selective animal breeding must be carried out. Crosses are carried out with experimental animals, such as mice, to select strains that have the extreme value of a trait (*e.g.* mice that have increased resistance to tuberculosis). These are then crossed with normal strains to produce offspring animals that each have one normal chromosome and one affected chromosome that presumably contains genes that confer resistance to tuberculosis. These offspring are in turn mated with the normal strain mice (backcross mating). This cross produces a third generation of animals in whom one chromosome has only normal genes, with the homologous chromosome being a recombinant between the normal and affected chromosomes of the previous generation. This series of matings produces progeny that are suitable for linkage analysis.

Once a linked marker or markers have been identified, it may possible to isolate the actual functional gene responsible for the trait using a candidate-gene strategy. Once the candidate genes for the phenotype of interest are identified the DNA sequences and expression profile of those genes can be compared and analysed between the normal strain and the strain that has the extreme values of the trait. When a functional gene is isolated and cloned in the experimental animal, it can be used as a "probe" (if the orthologous gene in human is not

known) to search the human genome for a gene with high DNA sequence homology that may have the same function.

This approach has the advantage that animals can easily be selected with extreme values of a trait and any desired breeding scheme can be used to generate useful recombinants. Animals, of course, do not necessarily model humans accurately and this technique detects only individual genes that cause disease in the animal model, but it cannot assess the pattern of interaction of these genes.

6.2 Results

6.2.1 Identification of QTLs in the vicinity of the *Spp2* gene in mouse

After an extensive literature review using a positional candidate gene strategy and by searching the mouse genome database (using the Entrez interface at NCBI), four QTLs were identified in the vicinity of the *Spp2* in mouse. Figure 6.1 illustrates these four QTLS and their approximate locations and distances to the mouse *Spp2* gene. These QTLs are as follows:

1- Susceptibility to tuberculosis (*Sst1*) at 54 cM, which controls the progression of tuberculosis (TB) in a lung-specific manner (Kramnik *et al.*, 2000).

2- Body weight QTL1 (*Bwtq1*) at 54.3 cM, which controls the growth and body weight at the age of 10 weeks in the mouse (Morris *et al.*, 1997; Vaughan *et al.*, 1999).

3- A major region in the vicinity of the mouse *Spp2* gene, which controls the susceptibility of the mouse to autoimmune sialadentitis or Sjögren syndrome (Boulard *et al.*, 2002).

4- Loss of righting due to ethanol 1(*Lore1*), which controls the sensitivity of the central nervous system to alcohol (Bennett *et al.*, 1994; Markel *et al.*, 1996; 1997).

6.2.2 Analysis of the relationship between the *Sst1* QTL (susceptibility to tuberculosis) and the *Spp2* gene in mouse

Tuberculosis (TB), historically the largest single cause of death in man, remains the seventh most important cause of premature morbidity and mortality. Currently there are 8 million new cases and 2–3 million deaths worldwide annually from TB, and it is estimated that a total of 225 million new cases and 79 million deaths will occur between 1998 and 2030 (Murray and Salomon, 1998). It is estimated that only 10% of those who become infected by *Mycobacterium tuberculosis* will develop clinical disease and only in a limited number of cases has an obvious risk factor (such as diabetes, advanced age, alcohol abuse, AIDS infection or corticosteroid usage) been identified (Murray *et al.*, 1990).

It is a common misunderstanding that mortality from multifactorial diseases such as cancer and cardiovascular problems are influenced by genetic factors, but that mortality and morbidity from infections are due solely to environmental factors. A study of 960 adoptees in Denmark indicated that the genetic component of susceptibility to premature death from infectious diseases is greater than for cancer and cardiovascular problems (Sorensen *et al.*,

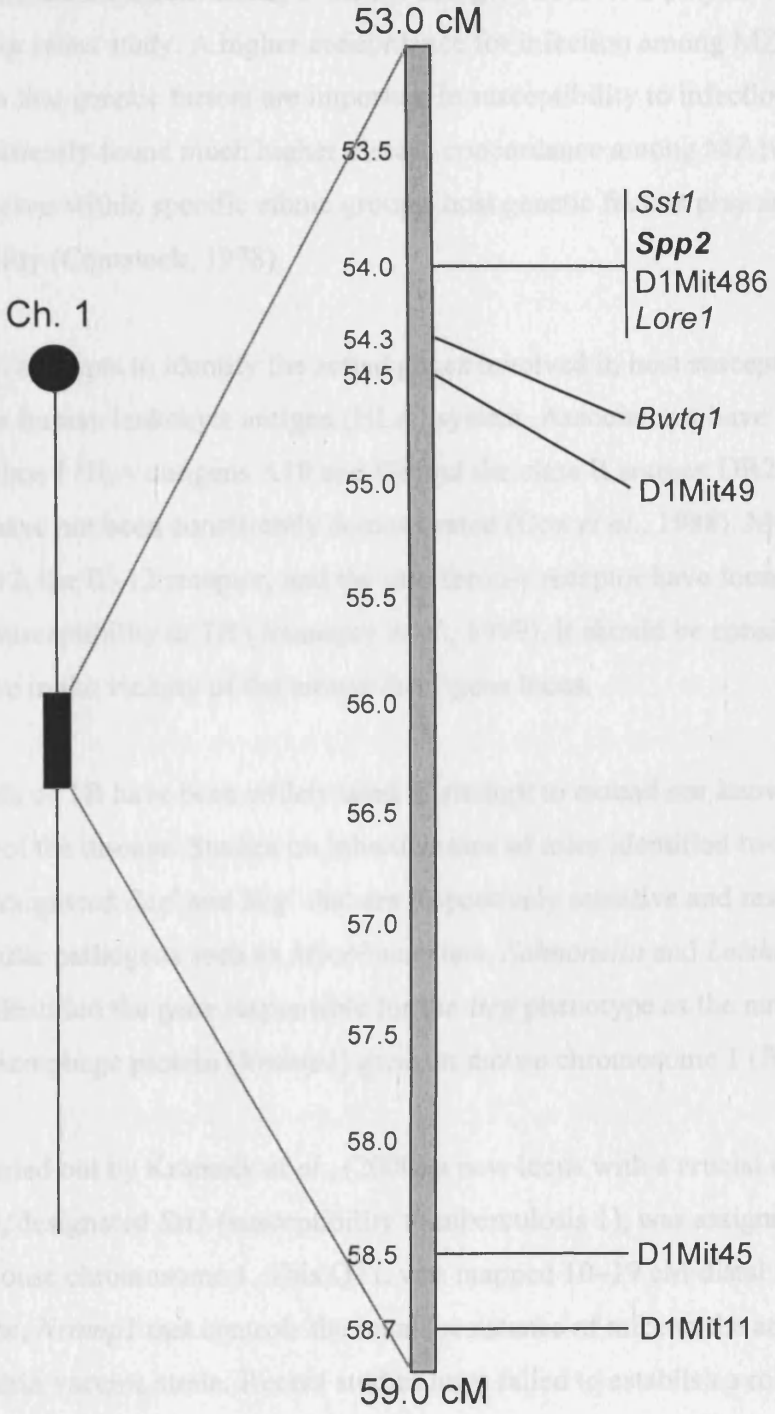


Figure 6.1. Approximate locations of the QTLs in the vicinty of the mouse *Spp2* gene

The major QTL which controls the susceptibility of the mouse to autoimmune sialadentitis or Sjögren syndrome extends over 50 cM, with several peaks having a LOD score of more than 3. The strongest association was with the D1Mit11 marker (MGD position 58.7 cM), but one of these peaks lies at about 54 cM, which is in the vicinity of the mouse *Spp2* gene.

Sst1: susceptibility to tuberculosis
Bwtq1: body weight QTL1
Lore1: loss of righting due to the ethanol 1

1988). Probably the most convincing evidence that genetic factors play an important role in TB comes from twins study. A higher concordance for infection among MZ than among DZ twins suggests that genetic factors are important in susceptibility to infection. Twin studies in TB have consistently found much higher disease concordance among MZ twins which could indicate that, even within specific ethnic groups, host genetic factors play an important role in TB susceptibility (Comstock, 1978).

Until recently, attempts to identify the actual genes involved in host susceptibility to TB focused on the human leukocyte antigen (HLA) system. Associations have been found between the class I HLA antigens A10 and B8 and the class II antigen DR2. However, these associations have not been consistently demonstrated (Cox *et al.*, 1988). Mutations in genes encoding IL-12, the IL-12 receptor, and the interferon-γ receptor have found to be responsible for selective susceptibility to TB (Jouanguy *et al.*, 1999). It should be considered that none of these genes are in the vicinity of the mouse *Spp2* gene locus.

Animal models of TB have been widely used to attempt to extend our knowledge of the pathogenesis of the disease. Studies on inbred strains of mice identified two distinct phenotypes designated Bcg^s and Bcg^r that are respectively sensitive and resistant to infection with intracellular pathogens such as *Mycobacterium*, *Salmonella* and *Leishmania*. Positional cloning has identified the gene responsible for the *Bcg* phenotype as the natural resistance associated macrophage protein (*Nramp1*) gene on mouse chromosome 1 (Bellamy, 1998).

In a study carried out by Kramnik *et al.*, (2000) a new locus with a crucial effect on TB susceptibility, designated *Sst1* (susceptibility to tuberculosis 1), was assigned to a 9 cM interval on mouse chromosome 1. This QTL was mapped 10–19 cM distal to a previously identified gene, *Nramp1* that controls the innate resistance of mice to the attenuated bacillus Calmette-Guérin vaccine strain. Recent studies have failed to establish a role for *Nramp1* in protection against virulent *M. tuberculosis*. The phenotypic expression of this new locus is distinct from *Nramp1* gene in that *Sst1* controls progression of tuberculosis infection in a lung-specific manner. Lung pathology in congenic *Sst1*-susceptible mice is characterised by extensive necrosis and unrestricted extracellular multiplication of virulent mycobacteria, but *Sst1*-resistant mice develop interstitial granulomas and controlled multiplication of bacilli.

According to this study and based on the median survival times, mouse inbred strains could be classified into highly susceptible (CBA/N, C3HeB/FeJ and DBA/2), intermediate (BALB/c

and 129/SvJ) and relatively resistant (C57BL/6J and C57BL/10J). Based on this study, the strongest evidence for linkage was observed with D1Mit49 ($\chi^2_{2DF} = 170.65$, $P = 2.6 \times 10^{-40}$). The results of a Kruskal-Wallis nonparameteric test of the data confirms linkage of susceptibility to tuberculosis to the region of mouse chromosome 1 around D1Mit49 (*K* value = 202.49, $P < 0.0001$). Figure 6.1 illustrates the location of this QTL and its approximate distance to the *Spp2* gene.

As discussed in Chapter 1, the spp24 protein may have an antimicrobial function like cathelins. The non-cystatin-like domain may be responsible for this activity, as it is the most divergent domain between species (Chapter 4) therefore, based on a positional candidate gene strategy, it was speculated that the *Spp2* gene in mouse could be a good candidate gene for the *Sst1* QTL. To determine any association between the *Sst1* QTL and the *Spp2* gene in mouse, three strains of mice, one highly susceptible (C57BL/6H), one intermediate (BALB/c) and one relatively resistant (DBA/2J) were selected. Using DNA samples of these three strains (donated by Dr. Mark Plumb, Department of Genetics, and Dr. Catrin Prichard, Department of Biochemistry, University of Leicester) all exons and the promoter region of the mouse *Spp2* gene were PCR amplified, gel purified and sequenced in both the forward and reverse directions. PCR amplification, recovery of DNA fragments from agarose gels by electro- elution onto dialysis membrane and sequencing were carried out as described in Sections, 2.7.1, 2.5.3 and 2.10.1 respectively. To PCR amplify the 8 exons and the promoter region (about 1000 bp upstream of the transcription initiation site) of the mouse *Spp2* gene, the previously designed primers and conditions were used (Section 4.2.2 and Table 4.3).

The sequences of all exons and the promoter region of the gene in the three strains were aligned using the program pileup and manually edited in the program SeqLab, both programs being part of the GCG Molecular Biology package. No sequence difference was found among these three strains, therefore it was concluded that the *Spp2* gene in mouse is not a good candidate gene for the *Sst1* QTL.

6.2.3 Analysis of the relationship between the *Bwtq1* QTL (body weight QTL 1) and the *Spp2* gene in mouse

Growth rate in mice has long been studied as a model for quantitative traits. This trait does not show classic Mendelian inheritance attributable to a single genetic locus, but is strongly heritable. Body weight variation in mice is normally distributed and seems to be controlled by

many genes, each having a relatively small and additive effect on the phenotype, therefore making it as an ideal trait for a QTL analysis (Falconer, 1953).

Falconer *et al.* (1978) showed that body size and growth rate in rodents appear to be determined through independent general physiological mechanisms that act at different life stages. Other studies confirmed that separate growth rate and body size QTLs indicate separate genetic and physiological systems for early and late growth in mice, as Falconer suggested (Vaughn *et al.*, 1999).

Other studies have been carried out to identify the QTLs responsible for growth rate and body size in mouse. In one of these, carried out by Morris *et al.* (1999), a murine genome scan was performed on animals in an F_2 population of 927 C57BL/6J × DBA/2J mice intercrossed to identify QTLs associated with tail length and age-related body weight. DBA/2J mice are slightly heavier than C57BL/6J mice. In this study the QTLs for body weight at 3 weeks mapped to mouse chromosome 4 from 40–60 cM (*Bwtq2*, LOD=4.4 at 55 cM), mouse chromosome 9 from 0–50 cM (*Bwtq4*, LOD=8.3 at 8cM) and mouse chromosome 11 from 20–55 cM (*Bwtq5*, LOD=9.5 at 27.8 cM). QTLs for body weight at 10 weeks of age were assigned to mouse chromosome 1 from 40–70 cM (*Bwtq1*, LOD=8.8 at 54.3 cM), mouse chromosome 6 from 0–29 cM (*Bwtq3*, LOD=4.4 at 4 cM) and mouse chromosome 15 from 25–50 cM (*Bwtq6*, LOD=7.6 at 41.2 cM). A major QTL for tail length was mapped to mouse chromosome 1 from 40–70 cM (*Tlln*, LOD=33 at 50 cM).

As can be seen from this study, one of the QTLs with a high LOD score (8.8) has been mapped in the vicinity of the mouse *Spp2* gene. If spp24 is a thiol protease inhibitor, as has been suggested, and is found in bone, then it might inhibit the thiol protease activity in bone and therefore could play an important role in bone metabolism and turnover (Hu *et al.*, 1995). Therefore, it was speculated that the mouse *Spp2* gene could be a good candidate for the *Bwtq1* QTL. Because the sequences of all exons and the promoter region of the mouse *Spp2* gene had already been determined and were shown to be identical in these two strains (C57Bl/6H and DBA/2J), it was concluded that *Spp2* gene is not a good candidate gene for the *Bwtq1* QTL.

6.2.4 Analysis of the relationship between Sjögren syndrome (SS) and the *Spp2* gene in mouse

Sjögren syndrome (SS) is a chronic autoimmune disorder characterised by lymphocyte and plasma cell infiltration and destruction of exocrine glands such as salivary and lacrimal glands causing oral and ocular dryness respectively. The disease may occur alone (primary SS) or in association with other connective tissue diseases (secondary SS), most notably rheumatoid arthritis, systemic lupus erythematosus and other autoimmune diseases. The aetiology of SS is presently unknown. It is likely to involve multiple factors, including immunologic, genetic, hormonal and perhaps viral infections (Fox, 1996).

Several animal models for studying immune-mediated sialadenitis including autoimmune-prone mice which develop the disease spontaneously, and non-autoimmune-prone mice in which the disease can be induced by different experimental manipulations. Autoimmune disease-prone mice such as NZB, (NZB × NZW)F_1, lupus-prone MRL/1pr, SL/Ni and NOD (non-obese diabetic) strains have been used as animal models for the investigation of Sjögren syndrome in humans (Hayashi, 1995). The NOD strain is known primarily as a reference for insulin-dependent diabetes mellitus (*Iddm*). However, NOD mice also develop inflammatory infiltration of exocrine glands, notably including salivary and lacrimal glands. This model of sialadentitis has now been well characterised, and two features make it useful for the study of human SS. First, the incidence of sialitis is gender-biased and is higher in females than in males (similar to human SS patients) and, secondarily, the inflammatory infiltration of the salivary glands is associated with a loss of secretory function (Hu *et al.*, 1992).

The molecular mechanisms that underlie sialadenitis in the NOD strain mice are not known, but must be determined in part by genetic factors. In a study, a susceptibility locus was mapped to chromosome 1 which was validated by the analysis of congenic mice (Garchon *et al.*, 1991; Brayer *et al.*, 2000). The genetic control of NOD sialitis could be polygenic, as is the case for most autoimmune diseases.

In a recent study (Boulard *et al.*, 2002), the whole mouse genome was scanned in search of sialitis loci in two F_2 inter crosses, including (NOD × C57Bl/6J)F_2 cross and (NOD × NZW)F_2 cross. Information from the (NOD × C57Bl/6J) × NOD backcross was also used and for the purposes of data analysis, sialitis was considered as a quantitative trait. Mice with at least one inflammatory site in their salivary glands were considered affected. Analysis of sialadenitis in NOD parents showed complete penetrance at 10 months of age in both males

and females. A genome-wide scan was carried out on the (NOD × C57Bl/6J)F_2 progeny with a total 156 microsatellite markers. Their average spacing was 10 cM, the largest gap being 34 cM. Parametric data analysis identified six chromosomal regions associated with sialitis. Of these, only chromosome 1 was associated at a significant LOD score (4.8 at D1Mit11). This major region extended over 50 cM, from D1Mit300 (MGD position 32.8 cM) to D1Mit104 (MGD position 79 cM), with several peaks having a LOD score greater than 3. The strongest association was with the D1Mit11 marker (MGD position 58.7 cM).

In summary, this study indicated a complex picture of the genetic control of lymphocytic infiltration associated with SS in mice. Multiple loci were detected depending on sex and on the combination of strains used in the matings. Despite the complexity, the middle region of chromosome 1 plays a crucial role in this syndrome as it was observed in all crosses studied so far. Its effect was most significant in the (NOD × C57Bl/6J)F_2 intercross where a QTL-LOD score of 4.8 was reached. This major region extended over 50 cM, with several peaks having a LOD score greater than 3. One of these peaks lies at about 54 cM, which is in the vicinity of the mouse *Spp2* gene. If spp24 circulates and acts like fetuins, it could have a role in processes such as inflammation and immune response (Chapter 1). Therefore, it was speculated that the mouse *Spp2* gene could be a good candidate gene for this major susceptibility region on chromosome 1.

To determine any association between this region and the *Spp2* gene in mouse, the NOD strain of mice was analysed. Using DNA samples of this strain (the DNA sample was purchased from the Jackson Laboratory), all exons and the promoter region of the mouse *Spp2* gene were PCR amplified and sequenced as described in Section 6.2.2. No sequence difference was found between the NOD and C57BL/6H strains. Therefore, it was concluded that the *Spp2* gene in mouse is not a good candidate gene for this major susceptibility region on chromosome 1.

6.2.5 Analysis of the relationship between *Lore1* (loss of righting due to ethanol 1) QTL and *Spp2* gene in mouse

It is estimated that approximately 14 million Americans suffer from alcoholism and alcohol abuse, at annual cost to society of $184.6 billion in 1998. Although environmental factors influence the likelihood of development of alcoholism, it has been established through adoption and twin studies that genetic factors play a role in alcohol dependence (Ehringer *et*

al., 2001). In order to better understand, treat and prevent the problems associated with alcoholism, research has focused on identifying the specific genetic and environmental elements that contribute to liability for, or protect against, alcohol abuse. As with other complex traits, alcoholism is a quantitative or polygenic trait in which the influences of multiple genes combine to contribute to the pathological state (Grisel, 2000). Several human studies have indicated that a person's initial level of response (LR) to alcohol is highly predictive of future risk for alcoholism. These studies demonstrated that individuals who exhibit a lower initial LR to alcohol are more prone to alcohol abuse (Schuckit, 1998). Over the past several years, different linkage studies of alcoholism have not identified any predisposing gene.

Quantitative trait loci (QTL) analysis allows the mapping of the underlying genes in this kind of multigenic trait. As an adjunct to the study of the genetics of alcoholism, different strains of mice have been used to identify QTLs for a variety of alcohol-related phenotypes. Many recent reports have described the identification and localisation of QTLs for the action of alcohol in mice, but most of these studies have exploited the C57BL/6J by DBA/2J series and their intercrosses. The most extensively studied mouse models for this purpose have been the inbred long-sleep (ILS) and inbred short-sleep (ISS) mice (McClearn and Kakihana, 1981). These strains were selected for differential sensitivity to the hypnotic effects of ethanol as measured by the loss of righting reflex. This mouse phenotype appears to be strikingly similar to the LR phenotype found to be important in human alcoholism. Because of the apparent relevance of this phenotype to alcoholism these two strains (ILS/ISS) have been studied extensively. One of the results of these extensive studies has been the identification of seven QTLs, four of which (*Lore1* on chromosome 1, *Lore2* on chromosome 2, *Lore4* on chromosome 11 and *Lore5* on chromosome 15) were confirmed independently. It is believed that these four QTLs contain genes contributing to the alcohol sensitivity differences in ILS and ISS strains (Bennett *et al.*, 1994; Markel *et al.*, 1996; 1997).

The *Lore1* QTL was mapped to mouse chromosome 1 from 43–59 cM between the D1Mit180 and D1Mit45 markers with a maximum LOD of 5.4 at 54 cM. The syntenic region of this QTL on the human genome falls on human chromosome 2 at 2q31 to 2q37 between the D2S156 and D2S163 markers.

In another study, Ehringer *et al.*, (2001) used high-throughput comparative gene sequencing, in the search for genes underlying these four QTLs, sequencing over 1.7 million bases of

DNA from the ILS and ISS strains. This comprised 68 candidates genes (from 478 initial genes mapping in the regions of the four QTLs), corresponding to a survey of over 36,000 amino acids. Eight central nervous system genes, located within these QTLs, were identified that harbour a total of 36 protein coding changes. In *Lore1*, 11 genes were sequenced from an initial 101 genes that mapped to that interval. Three genes were identified in *Lore1* that contained DNA differences between ILS and ISS, with two of these genes having differences predicted to result in a total of seven amino acid changes. All of these genes lie distal to the mouse *Spp2* gene which was not itself included in this published study.

Due to the fact that the mouse *Spp2* gene falls in the vicinity of the *Lore1* QTL and it is expressed in mouse brain and liver, it was speculated that the gene could be a good candidate gene for the *Lore1* QTL. DNA samples of the ILS and ISS strains of mice (DNA samples were kindly donated by Dr. Beth Bennett, Institute for Behavioural Genetics, University of Colorado) were PCR amplified and sequenced, as before, for all exons and the promoter region of the mouse *Spp2* gene. Once again, no sequence difference was found between these two strains and it was concluded that the *Spp2* gene in mouse is not a good candidate gene for the *Lore1* QTL.

6.3 Discussion

Using animal models to find QTLs for multifactorial diseases has the advantage that animals can easily be selected with extreme values of a trait, and any desired breeding scheme can be used to generate useful recombinants. However, it should be also acknowledged that animals do not necessarily model humans accurately. Furthermore, this approach detects only individual genes that cause disease in the animal model and it cannot assess the pattern of interaction of these genes with others. The nature of these interactions may be critically important and may differ in humans and in experimental animals. In spite of these reservations, this approach shows the way in which progress in molecular genetics and gene mapping may increase our knowledge of the genes responsible for multifactorial diseases.

To carry out this study, 4,211 bases of DNA sequence was compared in six different mouse strains (C57BL/6H, DBA/2J, BALB/C, NOD, ILS and ISS) corresponding to the whole coding sequence of the mouse *Spp2* gene and its promoter region. Data were derived from both strands of the amplified DNA. No base change was identified (even in the flanking introns) between these six strains, therefore it was concluded that the mouse *Spp2* gene is not a good candidate gene for *Sst1*, *Bwtq1*, *Lore1* and major region for susceptibility to Sjögren syndrome. Although this study did not identify any association between the *Spp2* gene and these QTLs, the approach was justified on the basis of the speculated functions of the spp24 protein (Chapter 1) and due to this fact that all of these QTLs fall in the vicinity of the *Spp2* gene. Furthermore, in this study the *Spp2* sequence was determined in six different mouse strains and these data can now be used to test any association between this gene and any new QTL mapped in the vicinity of gene, identified using any of these six strains.

Although no difference was identified in the sequences of the coding region of the *Spp2* gene, acceptor sites, donor sites and the immediate sequences of the flanking introns among these 6 different strains, other sequence difference in the introns of the *Spp2* gene can potentially alter important binding site sequences (such as enhancers and silencers) and consequently change the expression level of the gene. Therefore identification of any SNP or other sequence differences in the introns of the gene can be helpful in understanding differences in the expression of the gene in the various mice strains.

To find any variation, various Mouse SNP databases can be used. The Celera Mouse Reference SNP database (http://www.celera.com) includes approximately 3.4 million SNPs

covering five different mouse strains. About 40% of Celera's mouse SNPs are found within gene-coding regions, with an average density of 40 SNPs per gene. More than 90% of the SNPs have been validated as true polymorphisms. Celera is not alone in this effort and the mouse SNP database at NCBI contains about 6,000 unique mouse SNPs (available for public use) that come from 15 strains of mice. The region that covers the *Spp2* gene in chromosome 1 was searched using the Mouse SNP database (http://mousesnp.roche.com) to find any SNP in 15 different strains (including 129/Sv, A/HeJ, A/J, AKR/J, B10.D2-H2/oSNJ, BALB/cByJ, BALB/cJ, C3H/HeJ, C57BL/6J, CAST/Ei, DBA/2J, MRL/MpJ, NZB/BlnJ, NZW/LaC and SPRET/Ei), but no SNP was found. Because of the limited number of SNPs it seems that this database is not reliable and in the case of any new QTL, sequencing of the coding and promoter regions of the gene is necessary in the relevant strains (except the six strains in which the gene have been sequenced).

Following this study, case-control association studies could still be carried out to examine the role of the spp24 protein in humans in relation to the mouse QTLs discussed in this study. This is worthwhile even though the present study failed to find evidence that *Spp2* is involved in disease susceptibility, because of the greater power of association studies (Bellamy, 1998). If it is speculated from spp24 functional studies that the protein could play a role in a multifactorial disease, then an association study could be carried out. There are three previously characterised RFLPs (probably resulting from SNPs), three characterised short tandem repeats (Bennett and Dalgleish, unpublished) and two new variants (Chapter 3) in the human *SPP2* gene that could be used in any future association study.

Chapter 7

The Pattern of Expression of the Gene Encoding the spp24 Protein in Mouse, Chicken and Human

7.1 Introduction

The expression profile of a gene provides information about the sites and times at which it is expressed. This information can indicate possible new functions and also provide evidence to confirm any previously speculated functions. In the case of the spp24 protein, because its function is unknown, it is crucial to build up an expression profile, preferably in different organisms, to provide some clue to what the function might be. The expression profile will also help to design valid approaches for further functional studies.

The spp24 protein was originally isolated from a demineralised extract of bovine cortical bone (Hu *et al.*, 1995). This indicates the presence of the protein in this tissue, but it cannot show that the gene is also expressed here. However, Hu *et al*, (1995) also reported the results of a northern blot analysis on bovine bone periosteum, liver, heart, kidney, lung and spleen. A single transcript corresponding to the size of the deduced cDNA sequence was seen in bovine bone periosteum and liver, but not in heart, lung, spleen or kidney. The highest level of expression was seen in liver.

The pattern of expression of spp24 in human and mouse tissues was also assessed in this laboratory. In human, using hybridisation of a full-length spp24 cDNA probe to a human multiple tissue expression (MTE) array, the strongest hybridisation was identified in adult liver and kidney and weaker hybridisation to foetal liver (Bennett *et al.*, manuscript submitted). In mouse, using an RT-PCR method, a single transcript corresponding to the size of the deduced cDNA sequence was seen in mouse liver, kidney, brain and diaphragm tissues, but not in heart, muscle, testis, eye and stomach (Bennett and Dalgleish, unpublished). The highest level of expression was again seen in liver and kidney.

Because in all three species mentioned the gene encoding spp24 protein is expressed at the highest level in liver, it was therefore decided to identify the spatial expression of this gene in mouse liver tissue. This first part of this chapter describes the spatial pattern of expression of

the *Spp2* gene in mouse liver cells using a non-isotopic *in situ* hybridisation method and the expression profile of the gene encoding spp24 protein for mouse, chicken and human using data from various sources and experiments. Additional information on the expression profile of the gene encoding spp24 has been gained from ESTs, RT-PCR and microarrays, and is also presented in this chapter.

7.1.1 Study of the spatial expression of a gene

In multicellular organisms the majority of expressed genes are common to all cells and only 2–3% of these genes are differentially expressed, resulting in cell differentiation and specialisation. Most tissues are composed of different cell types in close proximity and therefore analysis of the differentially expressed gene can only be reliably achieved by a method that can directly show these differences (*in situ* method). Once the mRNA or total RNA is extracted from the tissue for RT-PCR or northern blotting, information about the site of RNA synthesis is then lost. Therefore, to investigate gene expression and cell differentiation, knowledge of the spatial distribution and quantification of RNAs in the tissue is needed (Levy and Herrington, 1995).

7.1.2 Tissue *in situ* hybridisation to detect specific mRNAs

In tissue *in situ* hybridisation, a labelled probe is hybridised to RNA in tissue sections. This identifies which cells are expressing the gene of interest as well as providing qualitative information about the level of gene expression.

The use of a probe to detect mRNA rather than detecting the gene product, and for determining the site of synthesis, can also be useful in a number of circumstances such as:

1. When the protein (gene product) is rapidly degraded in the cell.
2. When the protein is secreted from the cells as in the case of spp24. In this situation immuno-histochemical methods will fail to localise the cells expressing the gene product.
3. Antibodies are not available for the gene product.
4. When the gene product is mutated, such that the epitope in the mutated protein may not be recognised by antibodies.
5. Immuno-histochemical methods fail because antigens are not preserved in formalin-fixed paraffin-embedded tissue.

6. Frequently the half-life of mRNA differs from its protein product, and RNA detection may give more information about the steady-state synthesis of protein, whereas detection of protein reflects both stored and newly produced protein.

7.1.3 Rationale for using Digoxigenin-labelled riboprobes with *in situ* hybridisation

In situ hybridisation was first used in the 1960s and is based on the specific hybridisation of a labelled nucleic acid probe to a target RNA or DNA. An appropriate detection system then allows visualisation of the specific signal in the tissue section (Gall and Pardue, 1969; Farquharson *et al.*, 1990). *In situ* hybridisation can be carried out with DNA or RNA probes (Komminoth, 1992; Komminoth *et al.*, 1992). Although double-stranded cDNAs have been used as probes, single-stranded RNA probes (riboprobes) are preferred. In spite of the fragility of RNA in an environment with RNase activity, an RNA probe will bind to a mRNA target molecule with the strongest affinity of any nucleic acid-nucleic acid interaction. Due to the relatively high T_m (the mean thermal denaturation temperature at which 50% of the hybrids dissociate) of RNA-RNA hybrids, high stringency conditions and post-hybridisation washes are possible, ensuring a low background and maximal signal strength (Komminoth, 1992; Komminoth *et al.*, 1992). The sensitivity of initially single-stranded probes is generally higher than that of double-stranded probes, probably because a proportion of the denatured double-stranded probes renatures to form probe homoduplexes. cRNA riboprobes that are complementary to the mRNA of a gene are known as antisense riboprobes and can be obtained by cloning a gene's cDNA in the reverse orientation in a suitable expression vector.

In situ hybridisation originally utilised radiolabelled probes (Gall and Pardue, 1969), but recently the digoxigenin steroid hapten, derived from the plant cardiac glycoside digoxin, has been used as labelling molecule (Farquharson *et al.*, 1990). The decision to use digoxigenin-labelling over radiolabelling of probes was based on various factors:

1. The specificity of the digoxigenin is equal to that of a radiolabelled probe of the same length but the former has the advantage of higher stability.
2. The rapid development speed of the enzymatic or fluorescent detection system.
3. The use of radiolabelled probes often results in blooming of the signal on the autoradiogram, yielding poor resolution of signal detection.

4. The use of digoxigenin removes the risks associated with the use of radioactive materials, both to the researcher and to the environment (Komminoth, 1992; Komminoth *et al.*, 1992).

Riboprobes offer several advantages over synthetic oligonucleotide probes, as discussed above. One of the best methods for production of high specific activity digoxigenin-labelled riboprobes is *in vitro* transcription (the method that was used in this study) from a cDNA template cloned in a suitable expression vector with an RNA polymerase (T7, T3 or SP6 RNA polymerase), digoxigenin-UTP and other unlabelled ribonucleoside triphosphates. The size of the cloned cDNA fragment can be up to 500 bp. This is considered the maximum length of polynucleotide that can effectively penetrate the surface of the *in situ* hybridisation section to bind to the target mRNA molecule (Komminoth, 1992; Komminoth *et al.*, 1992).

7.1.4 A primary expression profile derived from expressed sequence tag (EST) data

Expressed sequence tags (ESTs) are short cDNA sequences that have been obtained from cDNA clone libraries. ESTs of many species are available from several EST databases. These ESTs can be used to build up a DNA sequence contig of a full-length (or nearly full-length) cDNA, but they can also provide valuable information regarding expression. Although quality of some ESTs is not perfect, the data are usually sufficiently good to allow unequivocal identification of the gene from which the EST is derived. Although ESTs cannot provide any quantitative information about the level of expression of a gene in a given tissue, numerous ESTs for a gene derived that tissue (perhaps from several independent cDNA libraries) are a good indication that the gene is genuinely expressed in that tissue.

In this chapter, different sources are used to build up the expression profile of the gene coding the spp24 protein in mouse, chicken and human. The ESTs were from three main sources, the TIGR human and mouse indices (http://www.tigr.org/tdb/), University of Delaware Chicken EST database (http://www.chickest.udel.edu) and the UniGene human and mouse databases (http://www.ncbi.nlm.nih.gov/UniGene).

7.1.5 The use of northern blot, ribonuclease protection assay and RT-PCR to obtain information about gene expression

To obtain direct information about the expression of a particular gene there are various common techniques including northern blot, ribonuclease protection assay and RT-PCR.

Northern blot hybridisation is a variant of Southern blotting in which the target nucleic acid is RNA instead of DNA. A principle use of this method is to obtain information on the expression patterns of specific genes. Once a gene has been cloned, it can be used as a probe and hybridised against samples of RNA isolated from variety of different tissues. It provides information about the size of transcript and the relative abundance of transcripts. Additionally, by revealing transcripts of different sizes, it may provide evidence for the use of alternative promoters, splice sites or polyadenylation sites. However, the technique is not without disadvantages. Northern blot analysis is intolerant of degradation and the isolated RNA should be of a high quality and of the three techniques, northern blotting is the least sensitive.

In ribonuclease protection assay an antisense probe is hybridised to an RNA sample in solution. Un-hybridised probe and RNA are then degraded by ribonuclease and the hybridised fragments are separated on a polyacrylamide gel. This method is more sensitive than northern blot analysis (approximately 10–100-fold) and is more tolerant of RNA degradation. Hybridisation in solution is also more efficient than filter hybridisation. Ribonuclease protection assays are quantitative and it is possible to carry out a multi-probe analysis. The disadvantages of the method are the lack of information regarding size, as the protected fragment is determined by the size of the probe, and the probe must be antisense RNA. RT-PCR (reverse transcriptase-PCR) is the most sensitive technique for mRNA detection and quantitation currently available. This method is moderately tolerant of partially degraded RNA, but is intolerant of RNA contaminated with DNA and the isolated RNA must be pure and DNA-free. This technique can be used to quantitate the RNA, but the methods need to be optimised for this purpose.

In this chapter, the results of RT-PCR on various mouse and human RNA samples are presented. This method was chosen due to its high sensitivity, simplicity and its tolerance to partially-degraded RNA.

7.1.6 The use of microarrays to obtain expression data

Recently-developed DNA microarrays have provided an increase in scale in hybridisation assay technology because of their huge capacity for miniaturisation and automation. With the advance of robotics it is now possible to spot thousands of samples onto filters, plates or

chips. The arrays are then probed with a fluorescent probe, and each individual spot (sample) is then scanned to obtain their relative hybridisation intensities.

To conduct a multiplex gene expression screen, a microarray needs to be designed to contain cDNA clones or gene-specific oligonucleotides to represent every gene of interest. To investigate expression of the chosen genes, RNA is prepared from the selected target cells, converted into cDNA using standard methods, labelled with a fluorescent tag and hybridised to the microarray. In this way, the expression pattern of tissues can be compared to identify the level of expression of the genes of interest. Microarrays can also be used to investigate the effect on gene expression of various chemical treatments.

This chapter presents data obtained on the gene encoding spp24 in mouse from the RIKEN cDNA Expression Array Database (READ) (Miki *et al.*, 2001). In this study Miki *et al.* (2001) arrayed 18,816 cDNAs and characterised the gene expression profiles for 49 adult and developing mouse tissues. It was estimated there were approximately 13,600 non-redundant genes in the array and all tissues were compared to pooled male and female 17.5-day embryos. In this study, a cluster analysis was performed which allowed the definition of sets of genes that were expressed ubiquitously and sets of genes that were expressed in similar groups of tissues. This group also clustered the genes coding for known enzymes into 78 metabolic pathways and this revealed coordination of expression within each pathway among different tissues, demonstrating how expression profiles can be helpful in revealing possible functions for a protein.

The RIKEN data with respect to the expression of the mouse *Spp2* gene have previously been analysed in detail (Bennett, 2002).

7.2 Results

7.2.1 Cloning the mouse *Spp2* cDNA fragment

An spp24 cDNA encoding the whole mouse protein (excluding the signal peptide) had previously been cloned into the vector pDNR-2 (Clontech) using restriction enzyme sites for *Eco*RI and *Bam*HI that had been incorporated into the forward and reverse PCR primers respectively (Figure 7.1).

To make positive and negative-control expression riboprobe constructs for the spp24 cDNA, the pBluescript SK+ phagemid vector was chosen. This vector is a 2958-bp phagemid derived from plasmid pUC19. This plasmid contains T7 and T3 RNA promoters on either side of the polylinker which can be used for *in vitro* transcription of the cloned insert using T7 or T3 RNA polymerases in both sense and anti-sense directions.

To subclone the whole spp24 cDNA into pBluescript SK+, a sufficient amount of pBluescript SK and pDNR-2 vector containing the spp24 cDNA were double digested by *Eco*RI and *Bam*HI restriction enzymes. The products of restriction digestion (cut spp24 cDNA and pBluescript SK+) were purified from an agarose gel. The restriction digestion and the recovery of DNA from agarose gel were carried out using the protocols described in Sections 2.3 and 2.5.2 respectively. The 556-bp spp24 cDNA was ligated into the expression vector pBluescript SK+ using the restriction enzyme sites *Eco*RI and *Bam*HI (Figure 7.2) using the protocols described in section 2.22. The ligation products were transformed into *E.coli* DH5α using the chemical transformation protocol described in section 2.12.5. A control transformation with pUC19 DNA was used to assess the efficiency of transformation. Transformants were plated onto Luria agar/X-gal/IPTG/ampicillin plates and grown for 16 hours at 37°C. A transformation efficiency of 7×10^5 was achieved for the construct. Isolation of plasmid DNA for the new expression construct was carried out using the method described in Section 2.12.8.3.

7.2.2 Labelled and unlabelled *in vitro* transcription, using linearised cloned spp24 cDNA template

Linearisation of the expression construct (usually within the cloned insert, or immediately distal to it) is a prerequisite for transcript 'run-off', allowing production of either antisense or

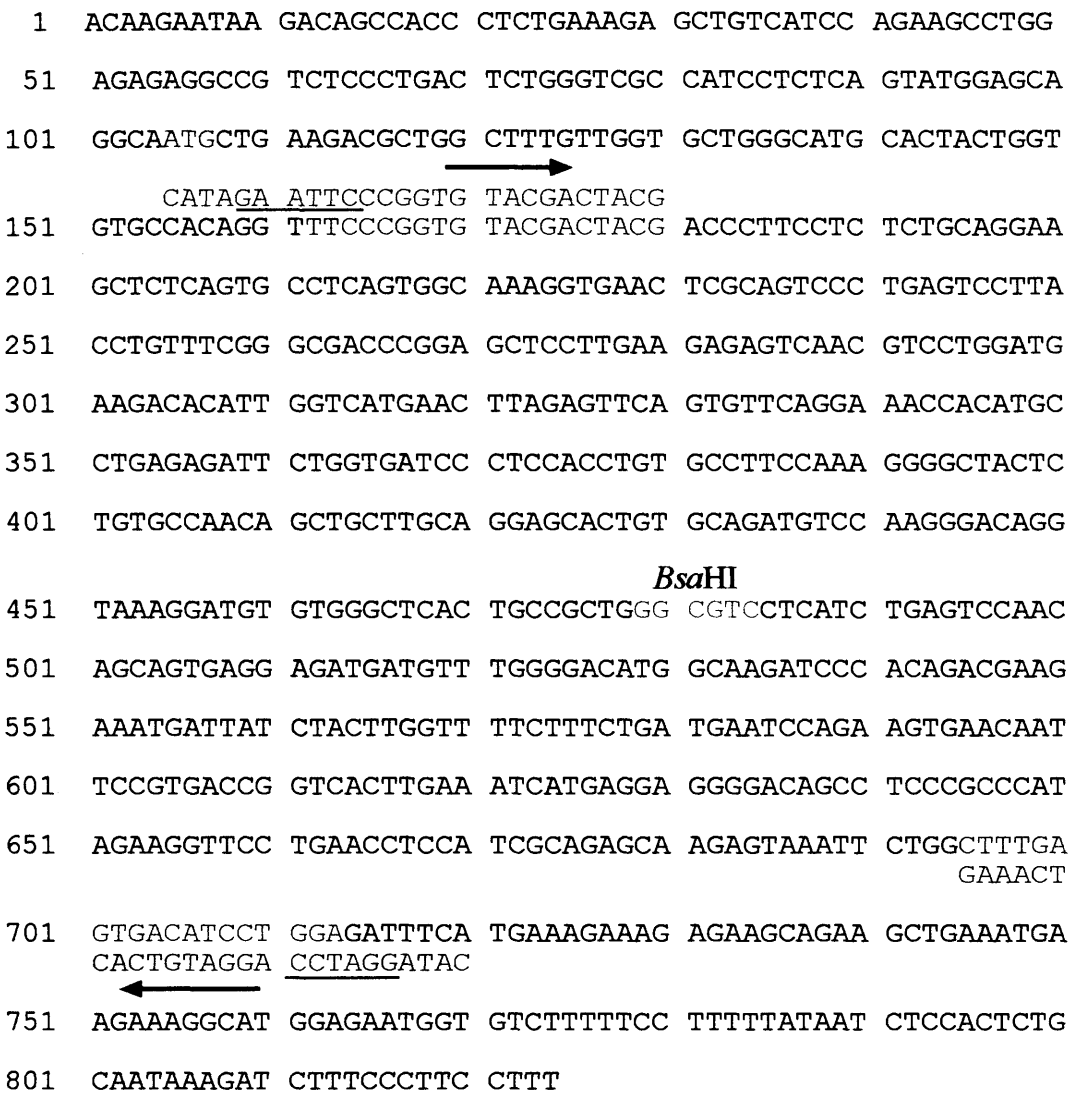


Figure 7.1. The position of the primers used to generate and clone the coding region of the mouse spp24 cDNA in the pDNR-2 vector

The 'ATG' and 'TGA' start and stop codons are shown in red. The primer sequence and annealing sequence in the cDNA are shown in blue. The restriction sites *Eco*RI and *Bam*HI are underlined in the forward and reverse primers respectively. The restriction site *Bsa*HI used to linearised the construct is shown in green.

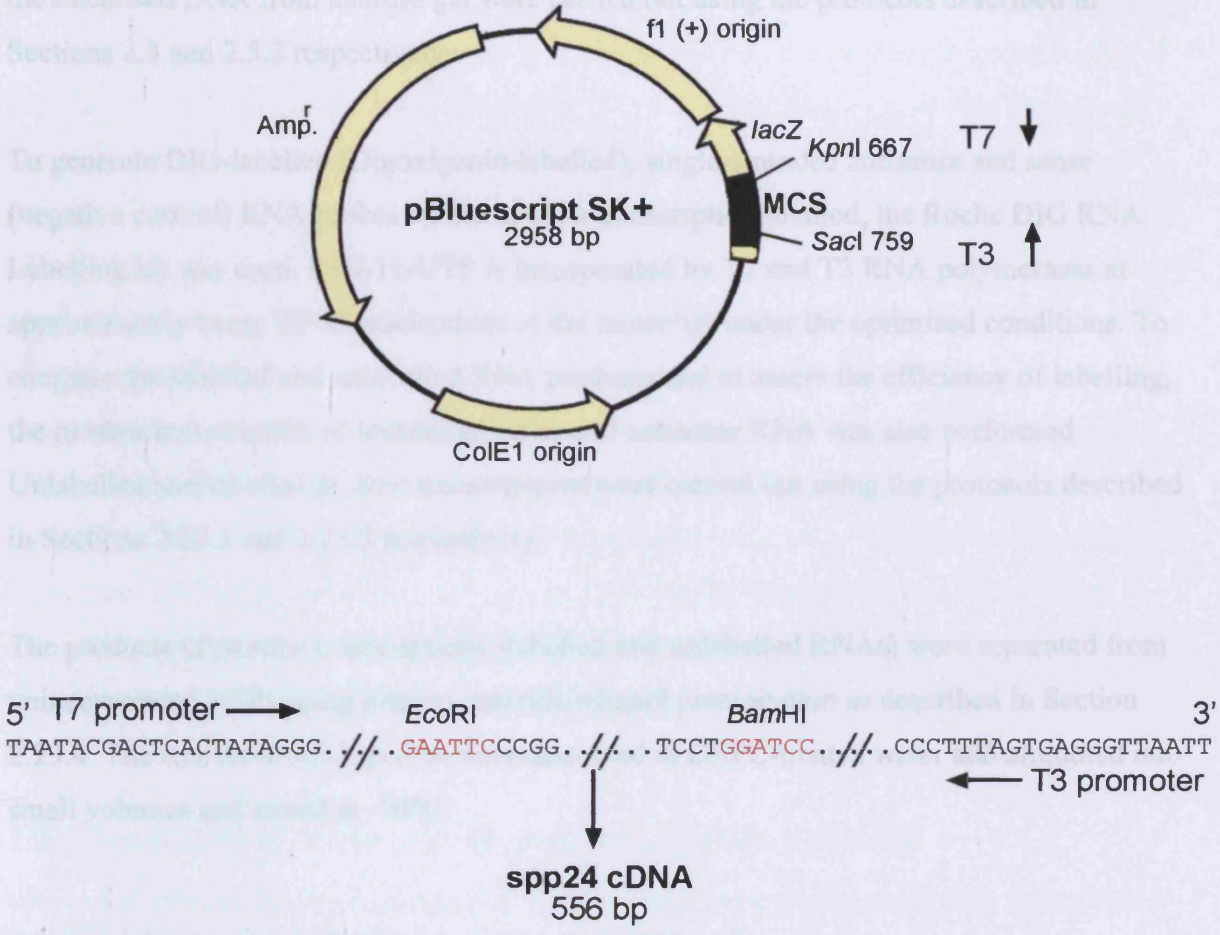


Figure 7.2. The pBluescript SK+ phagemid expression vector

The pBluescript SK+ phagemid is a 2,958-bp phagemid derived from pUC19. The SK designation indicates the polylinker is oriented such that *lacZ* transcription proceeds from *Sac*I to *Kpn*I. This phagemid contains an ampicillin resistance gene and the truncated *lacZ* gene for blue/white selection. This phagemid allows transcription from the T7 and T3 RNA polymerase promoter sites located on each side of the cloned insert.

sense RNA. Digestion of the construct by the *Bsa*HI restriction enzyme, cuts the spp24 insert at the site shown in Figure 7.1, allowing the *in vitro* transcription of a 303-bp antisense RNA (using T3 RNA polymerase) and a 396-bp sense RNA (using T7 RNA polymerase). The enzyme also cuts the vector at a single location. The restriction digestion and the recovery of the linearised DNA from agarose gel were carried out using the protocols described in Sections 2.3 and 2.5.2 respectively.

To generate DIG-labelled (Digoxigenin-labelled), single-stranded antisense and sense (negative control) RNA probes by the *in-vitro* transcription method, the Roche DIG RNA Labelling kit was used. DIG-11-UTP is incorporated by T7 and T3 RNA polymerases at approximately every 20–25 nucleotides of the transcript under the optimised conditions. To compare the labelled and unlabelled RNA products and to assess the efficiency of labelling, the *in vitro* transcription of unlabelled sense and antisense RNA was also performed. Unlabelled and labelled *in vitro* transcriptions were carried out using the protocols described in Sections 2.23.1 and 2.23.2 respectively.

The products of *in vitro* transcriptions (labelled and unlabelled RNAs) were separated from unincorporated NTPs using lithium chloride/ethanol precipitation as described in Section 2.23.4. The recovered RNA pellets were dissolved in DEPC-treated water and aliquoted into small volumes and stored at -70°C.

7.2.3 Analysis of labelled and unlabelled RNA transcripts

Quality and quantity of the transcript can be analysed by non-denaturing agarose gel electrophoresis and ethidium bromide staining. If the transcription reaction has worked well, the ethidium bromide-stained transcript RNA band should appear approximately 10-fold stronger than the stained template DNA. The size and amount of transcript can be estimated by comparison with known amounts of RNA. To determine the efficiency of the DIG-labelling reaction, the size of the labelled and unlabelled RNAs can be compared. In a successful labelling reaction, the labelled transcript should be heavier as judged by agarose gel electrophoresis. The efficiency of the labelling reaction can also be determined by detection of DIG-labelled RNA and comparisons of signals between labelled and unlabelled RNAs.

The labelled and unlabelled antisense and sense RNAs were analysed using 1% (w/v) non-denaturing and denaturing agarose gel electrophoresis. The signal intensity of transcript bands (in both labelled and unlabelled antisense and sense RNAs) were about 9 fold stronger than that from the template DNA, with a predominant high intensity RNA band indicating good quantity and quality of the synthesised RNAs. The sizes of labelled riboprobes were bigger than the sizes of unlabelled RNA transcripts which indicates the successful incorporation of the DIG-11-UTP into the labelled riboprobes. The result of denaturing gel electrophoresis confirmed the expected sizes of the T3 antisense and T7 sense RNA transcripts (data not shown).

Next, northern blot analysis was used to confirm the efficiency of the DIG-labelling reaction and to determine the efficiency of bonding of Anti-DIG-Ab to the labelled riboprobes. The RNAs separated by non-denaturing gel electrophoresis were northern blotted and detected using the protocol described in Section 2.27. The DIG-labelled riboprobes RNAs were detected by immunoassay with anti-DIG-AP conjugate and the chemiluminescent substrate CDP-Star. This experiment indicated that the DIG-labelled riboprobes can be detected effectively using the anti-DIG-AP conjugate and chemiluminescent substrate CDP-Star and the efficiency of the labelling reaction was good (data not shown).

7.2.4 Confirmation of riboprobes sensitivity in RNA dot blots

The sensitivities of the T3 antisense and T7 sense riboprobes were assessed by hybridisation to membrane-bound serial dilutions of liver total RNA in dot blots, as described in Section 2.28. In one experiment the mouse spp24 antisense (T3 antisense) and sense (T7 sense) riboprobes were separately hybridised to titrated total RNA (6 μg–0.06 μg) from mouse liver. After 3 hours exposure, the results indicated that the antisense riboprobes were able to detect a signal from 0.06 μg of total RNA.

7.2.5 Localising the different structures in the mouse liver

Mouse tissues were obtained by dissection and were immediately fixed using two alternative methods described in Section 2.18. Individual tissue structures and hepatic cells were localised and identified in 4 μm paraffin-embedded mouse liver sections using Haematoxylin-Eosin staining (H&E staining, Section 2.20.). This method of staining indicates the condition of morphology and preservation within the tissue. The five-month mouse (strain B10.5, male) was supplied by Carole Yauk, Department of Genetics, University of Leicester. Figure 7.3

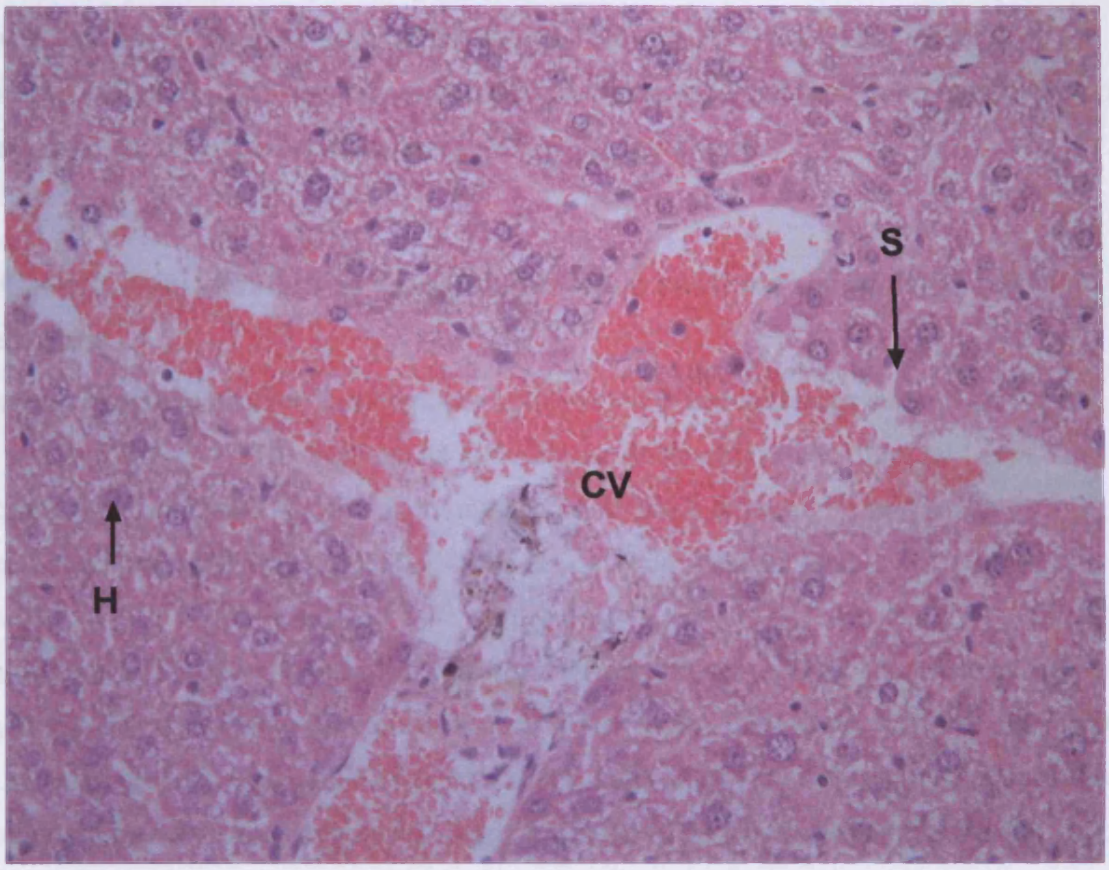


Figure 7.3. Haematoxylin-Eosin (H&E) stained mouse liver tissue section through a central vein

The image shows the centre of a lobule in mouse liver. Several sinusoids terminate into the central vein. The space between sinusoidal endothelium and hepatocytes is called the space of Disse. The tissue section provides evidence of good fixation and morphology within the paraffin-embedded liver block. Structures indicated include the central vein (CV), sinusoids (S) and hepatocytes (H).

shows a typical H&E-stained mouse liver section indicating the centre of a lobule in mouse liver and several sinusoids which are joined to the central vein. The space between sinusoidal endothelium and hepatocytes is called the space of Disse (not visible in this figure). Sinusoidal endothelial cells are highly fenestrated, which allows virtually unimpeded flow of plasma from the sinusoidal endothelium into the space of Disse.

7.2.6 *In situ* hybridisation to mouse liver tissue sections using *Spp2* mRNA riboprobes

Numerous experiments were carried out to optimise the individual steps and to achieve the best results (Table 7.1). The best results and staining were achieved at 10 $\mu g.ml^{-1}$ pre-treatment with proteinase K, 50% formamide, 0.15 M NaCl in pre-hybridisation and hybridisation and 50% formamide, 0.5× SSC in post-hybridisation wash.

Figures 7.4A and B show the results of *in situ* hybridisation using *Spp2* antisense and sense (negative control) riboprobes respectively in different mouse liver tissue sections. *Spp2* probe-free negative controls were almost free of staining, indicating very little endogenous alkaline phosphatase activity (data not shown). Using the mouse T3 antisense riboprobes *Spp2* expression in mouse liver tissue was found in:

1. Hepatocytes across the liver, but the expression is most prominent adjacent to vessels especially around the portal vein
2. Endothelial cells
3. Smooth muscle cells of the arteries (suggestive)
4. Connective tissues around the vessels especially around the portal vein (suggestive)
5. It is not expressed in polymorphonuclear cells (neutrophils) (suggestive)

In situ hybridisation using the T7 sense riboprobes (as negative control) gave no specific signals (Figure 7.4).

7.2.7 Expression data obtained from EST databases for mouse, chicken and human

In total, 57 mouse spp24 ESTs were identified from the UniGene EST database (http://www.ncbi.nlm.nih.gov/UniGene) and TIGR Mouse Gene Index (MGI) (http://www.tigr.org/tdb/mgi/). These EST sequences were deposited from Washington University School of Medicine, the National Institute of Dental and Craniofacial Research and

Experimental stage	Procedure	Condition/Parameter
Tissue sample	Five different mouse tissues were immediately fixed using two methods	Liver, heart, kidney, bowel and brain
	a- Fixation in 10% formal saline	24–48 hours
	b- Frozen section	Samples were kept in liquid nitrogen
Section pre-treatment	Proteinase K concentration	Optimised for each tissue sample and each probe (2–20 $\mu g.ml^{-1}$)
Hybridisation	P47 antisense probes T3 antisense and T7 sense probes	150–500 $ng.ml^{-1}$
	Formamide concentration	30–60%
	SSC	2–0.2×
	NaCl	0.6–0.05 M
Probe detection	NBT/BCIP incubation	Up to 9 hours

Table 7.1. Optimisation of *in situ* hybridisation experiments

Figure 7.4. *In situ* hybridisation of *Spp2* gene antisense and sense riboprobes to mouse liver tissue sections

A: *In situ* hybridisation using T3 antisense riboprobes shows the expression of the *Spp2* gene in following cells:

Hepatocytes (H) adjacent to vessels

Endothelial cells (EN)

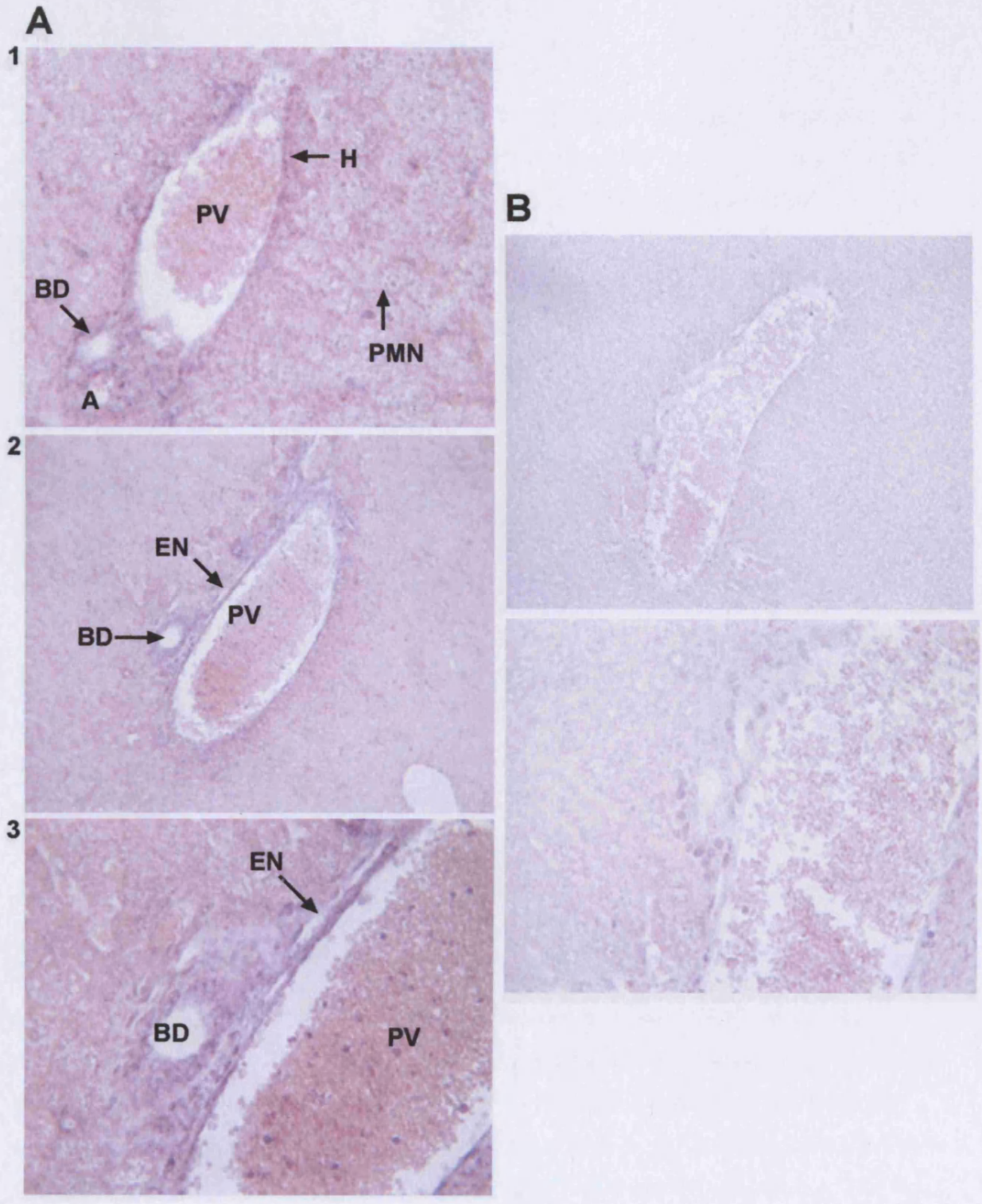
Smooth muscles (SM) of the arteries (suggestive)

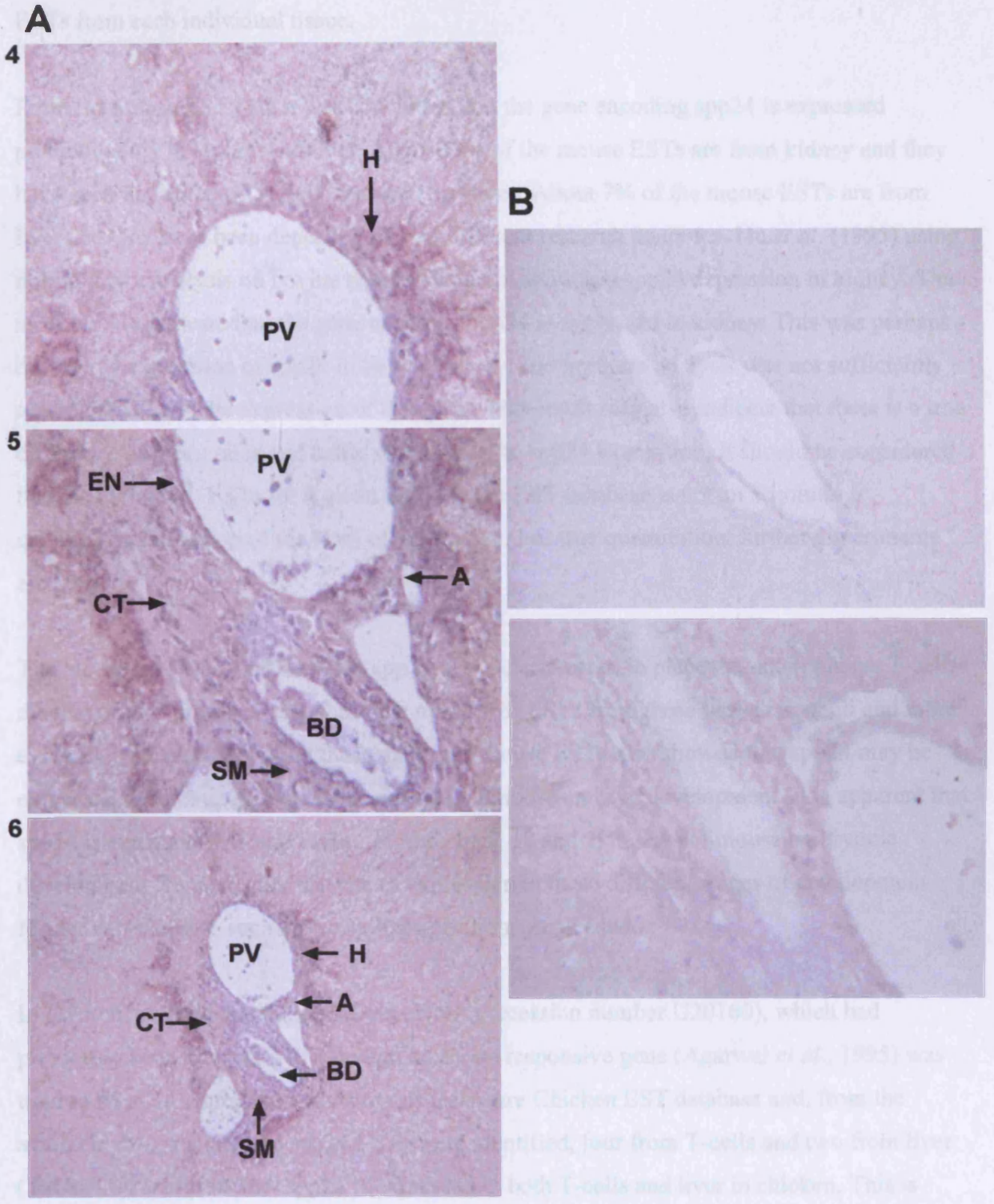
Connective tissues (CT) around the vessels (suggestive)

It is not expressed in polymorphonuclear (neutrophils, PMN) cells (suggestive)

Other structure are shown including portal vein (PV) and bile duct (BD).

B: *In situ* hybridisation using T7 sense riboprobes (negative control) gave no specific signals. The negative control in the absence of probe showed no staining (data not shown), indicating low background levels of endogenous alkaline phosphatase activity.


A
1
H
PV
BD
PMN
A
2
EN
PV
BD
3
EN
BD
PV
B


A
4
H
PV
5
EN
PV
A
CT
BD
SM
6
PV
H
A
CT
BD
SM
B

RIKEN (The Institute of Physical and Chemical Research). Table 7.2 illustrates the number of ESTs from each individual tissue.

From the mouse EST data it was concluded that the gene encoding spp24 is expressed predominantly in kidney and liver. About 33% of the mouse ESTs are from kidney and they have been deposited by various research institutes. About 7% of the mouse ESTs are from liver and they have been deposited by two different research institutes. Hu *et al.* (1995) using northern blot analysis on bovine tissues could not show any spp24 expression in kidney. The mouse ESTs indicate that the gene encoding spp24 is expressed in kidney. This was perhaps due to low expression of spp24 in bovine kidney, and northern analysis was not sufficiently sensitive to detect the expression of this gene. This result may also indicate that there is a true difference between mice and cattle with respect to spp24 expression. It should be considered that the number of ESTs for a given tissue in the EST database is not an absolutely quantitative indication of the level of expression. For true quantitation, further experiments are needed.

The mouse ESTs also indicate that spp24 may be expressed in placenta, macrophage, T-cell, diaphragm and proximal colon, but the number of ESTs from these tissues is small and more evidence is needed to support these data. The mouse ESTs also showed that spp24 may be expressed differentially in individual stages of mouse embryo development. It is apparent that spp24 is expressed in some tissues at 13.5, 14.5, 18 and 19.5 days of mouse embryonic development. To determine the site of expression in these different stages of development, further experiments, such as *in situ* hybridisation, are needed.

In chicken, a single matching chicken cDNA (accession number U20160), which had previously been identified as a growth hormone responsive gene (Agarwal *et al.*, 1995) was used to BLAST search the University of Delaware Chicken EST database and, from the available data, six chicken spp24 ESTs were identified, four from T-cells and two from liver (Table 7.3). It is clear that spp24 is expressed in both T-cells and liver in chicken. This is consistent with the liver result reported by Hu *et al.* (1995) the liver and T-cell results from the mouse ESTs discussed above.

In total, 23 human spp24 ESTs were identified from the UniGene EST database and the TIGR Human Gene Index (HGI). These ESTs were deposited from different sources including Washington University School of Medicine, the National Cancer Institute, the Beijing

Table 7.2. The number of mouse ESTs from various tissues

ESTs were identified from the UniGene EST database cluster Mm.28247 (http://www.ncbi.nlm.nih.gov/UniGene) and the TIGR Mouse Gene Index (MGI) (http://www.tigr.org/tdb/bgi/). These EST sequences were deposited to these databases from the National Institute of Dental and Craniofacial Research, RIKEN (The Institute of Physical and chemical Research), and Washington University School of Medicine.

Table 7.3. The number of chicken ESTs from various tissues

ESTs were identified from the University of Delaware Chicken EST database (http://www.chickest.udel.edu), using blast search of the single cDNA (accession number U20160), which had previously been identified as a growth hormone responsive gene (Agarawal *et al.*, 1995).

Table 7.4. The number of human ESTs from various tissues

ESTs were identified from the UniGene EST database cluster Hs.12230 (http://www.ncbi.nlm.nih.gov/UniGene) and the TIGR Human Gene Index (HGI) (http://www.tigr.org/tdb/hgi/). These EST sequences were deposited to these databases from Washington University School of Medicine, the National Cancer Institute, the Beijing Institute of Radiation Medicine, Pohang Institute of Science and Technology.

Source (Tissue/tissues)	Number of ESTs
Kidney	19
Whole mouse	6
13.5-14.5 day whole foetus	5
Kidney day 7	4
Liver	4
18 day embryo	3
Uterus	3
Kidney day 0	2
Macrophage	2
Placenta	2
19.5 day whole foetus	2
Proximal colon	1
8.5 embryo craniofacial subtraction library	1
Embryonic carcinoma	1
T-cell	1
Diaphragm	1
Total	57

Table 7.2. The number of mouse ESTs from various tissues

Source (Tissue)	Number of ESTs
T-cell	4
Liver	2
Total	6

Table 7.3. The number of chicken ESTs from various tissues

Source (Tissue/Tissues)	Number of ESTs
Foetal liver and spleen	10
Liver	5
Foetal lung, testis, B-cell	4
22 week foetal liver	2
Muscle (skeletal)	1
Foetal liver	1
Total	23

Table 7.4. The number of human ESTs from various tissues

Institute of Radiation Medicine, Pohang Institute of Science and Technology. Table 7.4 shows the number of ESTs from each tissue in human.

Approximately 22% of the human ESTs are from adult liver, 13% from foetal liver and 43% from a foetal liver and spleen library. Therefore these results enable us to conclude reliably that spp24 is expressed in liver. This is consistent with the result reported by Hu *et al.* (1995) and the data from the mouse and chicken ESTs discussed above.

Four human ESTs were identified in a foetal lung, testis and B-cells library. This suggests that spp24 is expressed in at least one of these tissues at some level, but it is impossible to determine the exact tissue of origin of these ESTs. There is also a single EST from skeletal muscle, but a reliable conclusion cannot be drawn just on this single EST.

Expression of the gene encoding the spp24 protein in mouse liver and kidney had also previously been demonstrated by RT-PCR (Bennett and Dalgleish, unpublished), consistent with the result reported by Hu *et al.* (1995) and the human and mouse EST expression profiles. That study also demonstrated expression of spp24 in brain and diaphragm, consistent with the RIKEN data and the identification of the single mouse diaphragm EST.

7.2.8 Determining the expression of mouse spp24 in bone and kidney tissue using RT-PCR

To further investigate the expression pattern of mouse spp24, RNA was extracted from kidney and bone of a 16-week adult female mouse and a 12-day immature female mouse (C57BL/6J, supplied by Biomedical Services, University of Leicester) using the RNAzol method described in Section 2.16.1. The quality of the extracted RNA was assessed using non-denaturing agarose gel electrophoresis and the RT-PCRs were then carried out on 4 μg of total RNA using the method described in Section 2.7.3. The PCR conditions used were: (96°C 30s, 65°C 30s, 72°C 40s) × 15–50 cycles in 5-cycle increments (15, 20, 25, 25, 30, 35, 40, 45 and 50).

The primers used for the RT-PCR and their position in the mouse cDNA are shown in Figure 7.1. The forward and reverse primers were tagged with *Eco*RI and *Bam*HI restriction enzyme sites respectively to enable cloning of any products if required. The primers span the region coding the mature spp24 protein and the size of the expected RT-PCR product is 566 bp.

To indicate the efficiency of the RT-PCR reactions and that equal amounts of RNA have been used in different RT-PCR reactions, the *GAPDH* gene (glyceraldehydes 3-phosphate dehydrogenase) was chosen as a control. To allow the use of common primers to amplify the *GAPDH* gene in human and mouse, a region of high similarity between human and mouse was identified and the primers were designed based on the human *GAPDH* sequence as follows:

Forward 5′ AGAACATCATCCCTGCCTC 3′

Reverse 5′ GCCAAATTCGTTGTCCATACC 3′

To carry out PCR, the following conditions were used: 96°C 30s, 55°C 30s, 72°C 30s × 30. The primers span the middle region of the *GAPDH*-encoded protein and the size of the expected RT-PCR product is 348 bp.

To estimate the approximate amount of *Spp2* and *Gapdh* mRNA in the adult and immature bone and kidney tissues, the starting amount of reverse transcription product was diluted between 10 to 10^{11} times for each individual tissue, and subsequently PCR was carried out for 40 cycles. The RT-PCR products were electrophoresed on 1.4% agarose gels along with the products of a control reaction containing no RNA in the initial reverse transcription. These results are shown in Figure 7.5.

RT-PCR products of the expected size (566 bp) were seen in kidney tissues of both adult and immature mice from 25 cycles to 50 cycles, but not in the bone tissue of adult mice, even at 50 cycles (Figure 7.5). In immature mouse, RT-PCR products of the expected size were seen in bone tissue at 40, 45 and 50 cycles (Figure 7.5). These PCR products must be from template cDNA that has been reverse transcribed from RNA as there are no products seen in the no-RT controls (40 cycles). This confirms that the total RNA preparations are free from contaminating DNA that would otherwise have amplified in the PCR to give a product. The size of the products also confirms that the template is cDNA that has been reverse transcribed as the primers span several exon/intron boundaries. The size of the product is consistent with there being no introns present, therefore the template must have been reverse transcribed from mRNA.

RT-PCR products of the expected size for the *Gapdh* gene (348 bp) were seen in kidney and bone tissues of both adult and immature mice in different dilutions (10, 100, 1,000 and 10,000 fold dilutions) (Figure 7.5C). This result indicates the presence of approximately equal amounts of total RNA in kidney and bone tissues of the adult and immature mice.

Figure 7.5. RT-PCR performed on RNA from adult and immature mice tissues (kidney and bone)

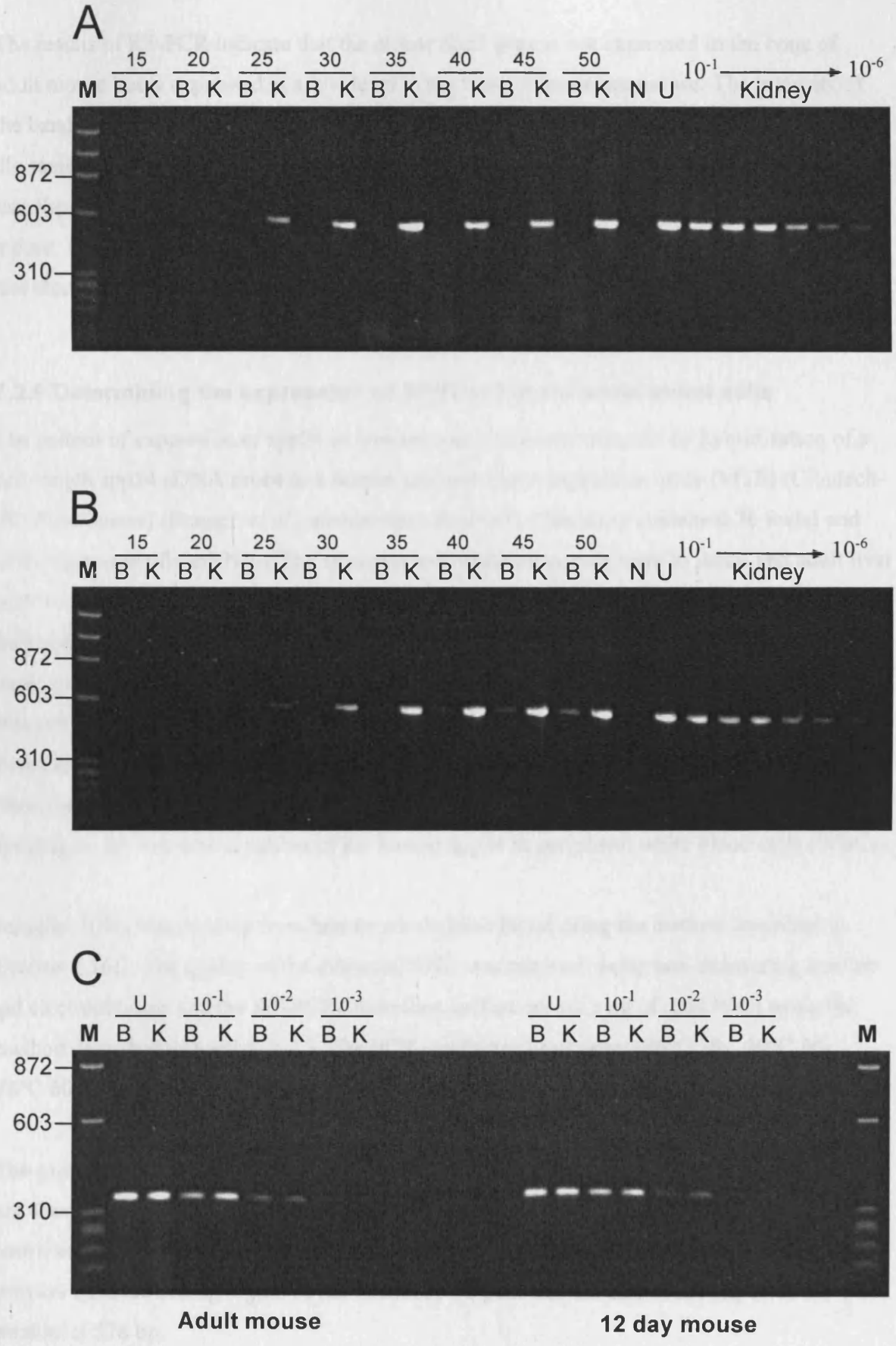
RT-PCR was performed on 4 μg total RNA as described in section 2.7.3. The total RNA was extracted from kidney and bone of a 16-week adult female mouse and a 12-day immature female mouse (C57BL/6J, supplied by Biomedical Services, University of Leicester) using the RNAzol method described in Section 2.16.1. The RT-PCR products were electrophoresed on a 1.4% agarose gel.

A: The RT-PCR products (*Spp2* gene) obtained in kidney (K) and bone (B) of the adult mouse in different PCR cycles. Lane 'N' shows no-RNA control (RT-PCR with no RNA in the initial reverse transcription). The following lanes show the PCR products of different dilutions of the original reverse transcription product in kidney.

B: The RT-PCR products (*Spp2* gene) obtained in kidney (K) and bone (B) of the immature mouse in different PCR cycles. Lane 'N' shows no-RNA control (RT-PCR withno RNA in the initial reverse transcription). The following lanes show the PCR products of different dilutions of the original reverse transcription product in kidney.

C: The RT-PCR products (*Gapdh* gene) obtained in kidney (K) and bone (B) of the adult and immature mice in different dilutions of the original reverse transcription products in kidney and bone.

M = marker ΦX174RF/*Hae*III


A
15 20 25 30 35 40 45 50
M B K B K B K B K B K B K B K B K N U
10-1 Kidney 10-6
872
603
310
B
15 20 25 30 35 40 45 50
M B K B K B K B K B K B K B K B K N U
10-1 Kidney 10-6
872
603
310
C
U 10-1 10-2 10-3
M B K B K B K B K
U 10-1 10-2 10-3
B K B K B K B K M
872
603
310
Adult mouse
12 day mouse

The results of RT-PCR indicate that the mouse *Spp2* gene is not expressed in the bone of adult mouse but is expressed at a low level in the bone of immature mouse. The intensity of the band representing the bone RT-PCR product at 40 cycles in Figure 7.5 is approximately the same as that seen in the same figure for the 10^6 dilution of the kidney cDNA. This implies that *Spp2* is expressed in bone at a considerably lower level than in kidney in the immature mouse. If it is expressed at all in adult mouse, it is below the level of detection of this system; less than 10^{-6} relative to kidney.

7.2.9 Determining the expression of *SPP2* in human white blood cells

The pattern of expression of spp24 in humans was previously assessed by hybridisation of a full-length spp24 cDNA probe to a human multiple tissue expression array (MTE) (Clontech-BD Biosciences) (Bennett *et al.*, manuscript submitted). This array contained 76 foetal and adult tissue-specific mRNAs. The strongest hybridisation signals were to foetal and adult liver with and a weaker hybridisation to foetal kidney. The array did not contain mRNA from bone, therefore comparisons with the expression of spp24 in bovine bone was not possible. The array contains mRNA from peripheral blood leukocytes, but no hybridisation was observed to this mRNA. However, since there is evidence that the gene encoding spp24 is expressed in both chicken and mouse T-cells (Section 7.2.7) and is also expressed in the mouse thymus (Section 7.2.10) which is an important site for maturation of T lymphocytes, it was decided to investigate the expression pattern of the human spp24 in peripheral white blood cells (WBCs).

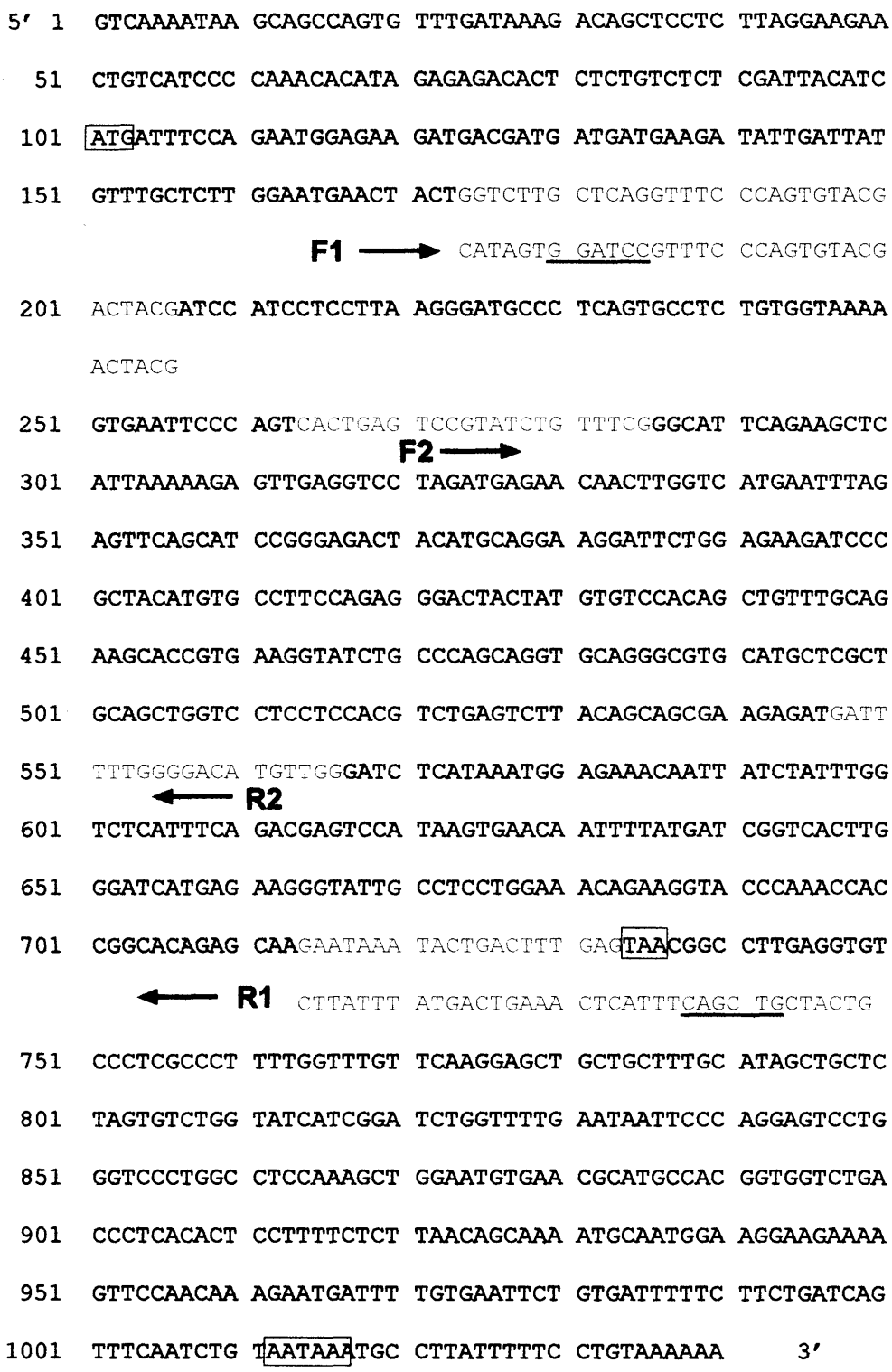
Initially, RNA was isolated from human whole fresh blood using the method described in Section 2.16.2. The quality of the extracted RNA was assessed using non-denaturing agarose gel electrophoresis and the RT-PCRs were then carried out on 2 μg of total RNA using the method described in Section 2.7.3. The PCR conditions used were: (96°C 30s, 50°C 30s, 68°C 60s) × 40.

The primers used for the RT-PCR (F1 and R1) and their position in the human spp24 cDNA are shown in Figure 7.6 The forward and reverse primers were tagged with *Bam*HI and *Sal*I restriction enzyme sites respectively to allow cloning of the PCR product if necessary. The primers span the coding region of the human *SPP2* gene and the size of the expected RT-PCR product is 576 bp.

Figure 7.6. The position of human RT-PCR primers with respect to the human cDNA sequence

This figure illustrates the human *SPP2* cDNA sequence. The start codon (ATG), the stop codon (TAA) and polyadenylation signal are boxed. The primer sequences (F1, R1, F2 and R2) and the annealing sequences (F1 and R1) in the cDNA sequence are shown in blue.

The restriction enzyme sites *Bam*HI and *Sal*I are underlined in the F1 and R1 primers respectively (which were designed originaly to generate an spp24 cDNA clone).



Forward1 primer (F1) 5′ CATAGTGGATCCGTTTCCCAGTGTACGACTACG 3′

Reverse1 primer (R1) 5′ GTCATCGTCGACTTACTCAAAGTCAGTATTTATTC 3′

Forward2 primer (F2) 5′ CACTGAGTCCGTATCTGTTTCG 3′

Reverse2 primer (R2) 5′ CCAACATGTCCCCAAAAATC 3′

To assess the efficiency of the RT-PCR reactions, the human *GAPDH* gene was used as template for positive-control RT-PCR reactions. The designed primers (Section 7.2.8) span the middle region of the human *GAPDH* and the size of expected RT-PCR product is 348 bp. RT-PCR was then carried out on 2 μg of total RNA using the method described in Section 2.7.3. The PCR conditions used were: (96°C 30s, 55°C 30s, 72°C 30s) × 30.

The RT-PCR products of the human *SPP2* and *GAPDH* genes from human WBCs were electrophoresed on 1% agarose gels along with a negative control containing no RNA in the initial reverse transcription. An RT-PCR product of the expected size (576 bp) was observed in human WBCs (Figure 7.7).Again, the size of the amplified product confirms that it is derived from template DNA that has been reverse transcribed as the PCR product from genomic DNA would span several exon/intron boundaries. The size of the amplified product is consistent with there being no introns in the fragment, therefore the template DNA has been transcribed from *SPP2* mRNA.

The result of this study indicates that the gene encoding the spp24 protein is expressed in human WBCs though the type of WBC cell/cells cannot be identified. There is clear evidence from EST data that spp24 is expressed in chicken T cells and in mouse T cells and macrophages. Further experiments are required to identify whether the human WBC amplification products derive from T cells or macrophages, or from one of the other white cell types such as B cells, monocytes, neutrophils, eosinophils or basophils.

7.2.10 Determining the expression of the human *SPP2* gene in lymphocytes, endothelial cells and monocytes using RT-PCR

Like most other tissues, the WBCs are composed of a variety of cell types including lymphocyte, monocyte, neutrophil, eosinophil and basophil. Once the RNAs have been extracted from the WBCs for northern blot or RT-PCR, information about the site of synthesis is lost. Therefore, to determine the spatial expression of a specific gene in WBCs, the different cells should be separated and RNA isolated from each individual cell type.

The mononuclear lymphocytes and monocytes were isolated from whole fresh human blood (from an Asian individual) using the method described in Section 2.14. Monocyte cells adhere to the surface of polystyrene dishes and grow as monolayers, so this biological property was used to isolate them from mononuclear lymphocytes as described in Section 2.14. To study

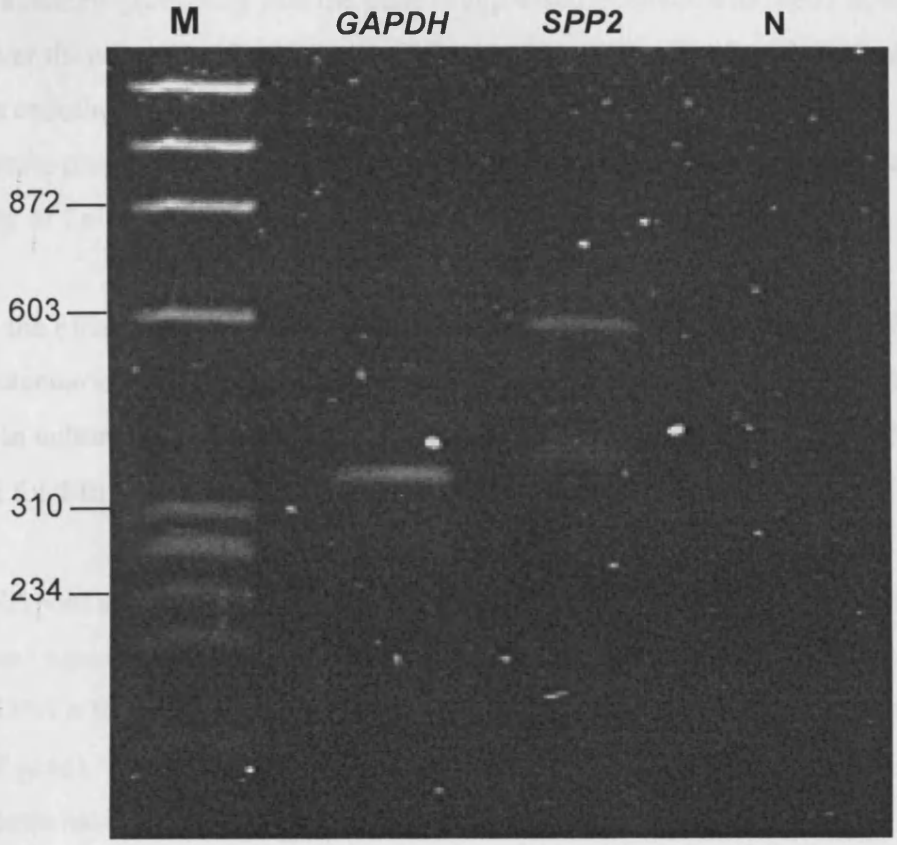


Figure 7.7. RT-PCR performed on RNA isolated from human WBCs

RT-PCR was performed as described in section 2.7.3 on 2 μg total RNA isolated from from human whole fresh blood using the method described in section 2.16.2. The RT-PCR products were electrophoresed on a 1% agarose gel.

This gel shows the RT-PCR products (*GAPDH* and *SPP2*) obtained in human WBCs. Lane N shows the no-RNA control (RT-PCR with no RNA in the initial reverse transcription).

M = marker ΦX174RF/*Hae*III

the expression of the human *SPP2* gene in endothelial cells, and to confirm the result of the *in situ* hybridisation (indicating that the gene is expressed in endothelial cells of vessels in mouse liver tissue) and to show whether this gene is expressed only in liver endothelial cells or also in endothelial cells of other tissues, endothelial cells from the umbilical cord of a neonate were prepared (endothelial cells were provided kindly by the Department of Surgery, University of Leicester).

To study the effects of different substances on the expression of the human *SPP2* gene, lipopolysaccharide (LPS) and tumour necrosis factor alpha (TNFα - Cachectin) were added to the cells in culture. Depending on the cell types, they were incubated in the presence of LPS or TNFα for 1 to 24 hours in conditions described in Section 2.14.

Total RNA was extracted from the mononuclear cells (lymphocytes and monocytes) before incubation (time 0), and 16 and 24 hours after incubation in both the presence and absence of LPS and TNFα (if spp24 has a role in immunity, LPS and TNFα may affect the expression of the *SPP2* gene). Total RNA was also extracted from monocytes 16 hours after incubation and from endothelial cells 24 hours after incubation in the presence and absence of TNFα. Total RNA extraction was carried out using the method described in Section 2.16.3. RT-PCRs were then performed on 2 μg total RNA using the method described in Section 2.7.3. The PCR conditions used were: (96°C 30s, 58°C 30s, 72°C 60s) × 40.
The primers used for the RT-PCR (F2 and R2) and their position in the human spp24 cDNA are shown in Figure 7.6. The primers span the middle of the human *SPP2* cDNA and the size of the expected RT-PCR product is 303 bp.

To assess the efficiency of the RT-PCR reactions and to show that equal amounts of RNA have been used in individual reactions, amplification of human *GAPDH* cDNA was used as a positive control. The designed primers (Section 7.2.8) span the middle region of the human *GAPDH* cDNA and the size of the expected RT-PCR product is 348 bp. RT-PCR was carried out on 2 μg of total RNA using the method described in Section 2.7.3. The PCR conditions used were: (96°C 30s, 55°C 30s, 72°C 30s) × 30.

The RT-PCR products of the human *SPP2* and *GAPDH* cDNA from each cell type under different conditions and incubation times were electrophoresed on 2% agarose gels (Figure 7.8). At time 0, there are no observable *SPP2* RT-PCR products of the expected size (303 bp) in mononuclear cells (Figure 7.8A, lanes MC0). However, at 16 hours (lane MC16-LPS) and

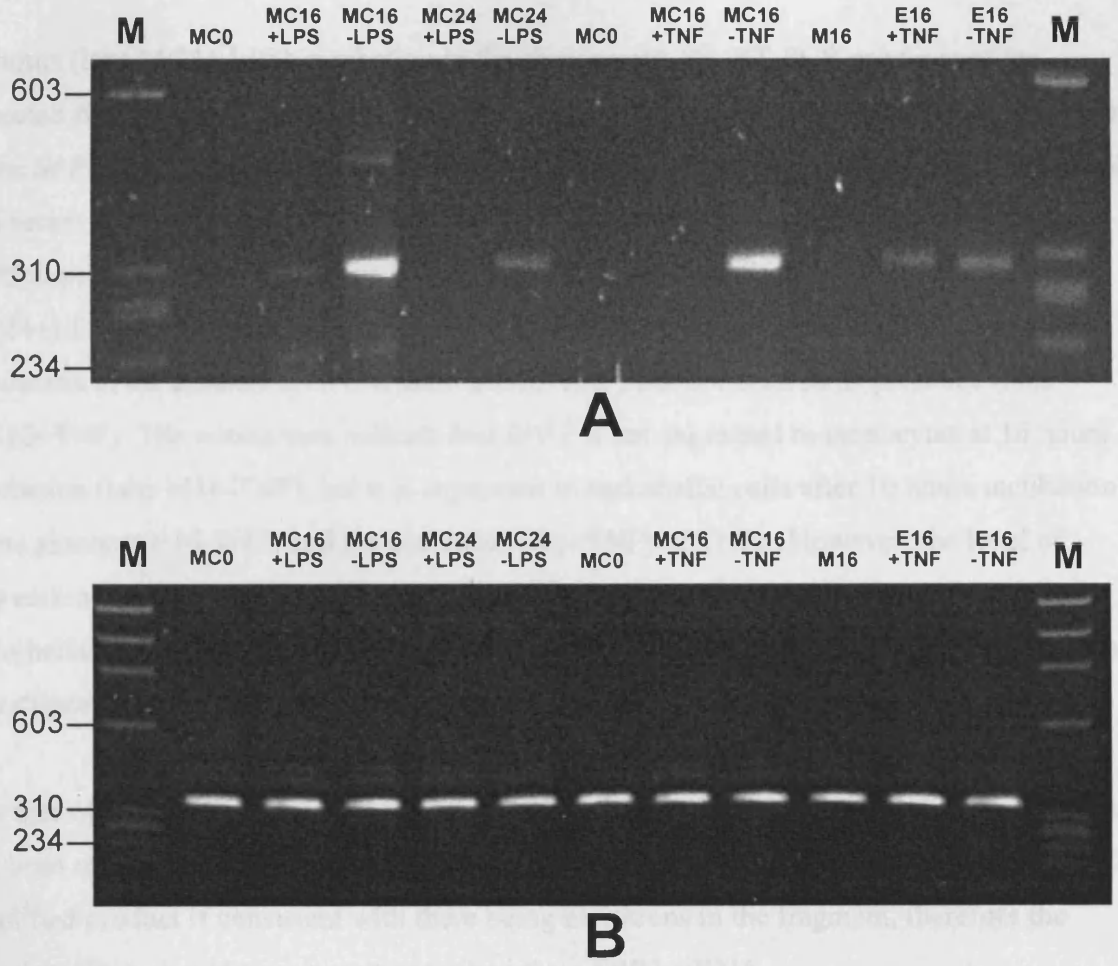


Figure 7.8. RT-PCR performed on RNA extracted from mononuclear cells (MC), monocytes (M) and endothelial cells (E) under different conditions

RT-PCR was performed on 2 μg total RNA as described in section 2.7.3. The total RNA was extracted from mononuclear cells (MC), monocytes (M) and endothelial cells (E) incubated for different times in the absence or presence of LPS and TNFα as described in section 2.16.3. The RT-PCR products were electrophoresed on 2% agarose gel.

A: The RT-PCR products of the human *SPP2* gene obtained:

In mononuclear cells before incubation (MC0), after 16 and 24 hours incubation in the presence of LPS (+LPS) and in the absence of LPS (-LPS)
In mononuclear cells before incubation (MC0) and after 16 hours incubation in the presence of TNFα (+TNF) and in the absence of TNFα (-TNF)
In monocytes after 16 hours incubation (M16)
In endothelial cells (E) after 16 hours incubation in the presence and absence of TNFα respectively (+TNF and -TNF)

B: The RT-PCR products of the human *GAPDH* gene obtained:

The lanes correspond with those in A.

M = marker ΦX174RF/*Hae*III

24 hours (lane MC24-LPS) incubation in the absence of LPS, RT-PCR products of the expected size are seen. Incubation of the mononuclear cells in culture has induced expression of the *SPP2* gene, possibly as a consequence of the action of some factor present in the foetal calf serum of the culture medium. Addition of LPS to the culture medium markedly reduces *SPP2* expression at 16 hours (lane MC16+LPS) and totally inhibits it at 24 hours (lane MC24+LPS). The *SPP2* RT-PCR product was seen in mononuclear cells at 16 hours incubation in the absence of TNFα (lane MC16-TNF) but not at all in its presence (lane MC16+TNF). The results also indicate that *SPP2* is not expressed in monocytes at 16 hours incubation (lane M16-TNF), but it is expressed in endothelial cells after 16 hours incubation in the absence (E16-TNF) and the presence (E16+TNF) of TNFα. However, the level of expression in the presence TNFα is lower than in its absence. Due to the unavailability of endothelial cells the expression of the *SPP2* gene at time 0 and in the presence of LPS was not investigated.

The size of the amplified product (303 bp) confirms that it is derived from template DNA that has been reverse transcribed as the primers span several exon/intron boundaries. The size of amplified product is consistent with there being no introns in the fragment, therefore the template DNA should have been transcribed from *SPP2* mRNA.

Mononuclear cells are composed of lymphocytes and monocytes and, since the human *SPP2* gene is not expressed in monocytes even after 16 hours of incubation (lane M16-TNF), it suggests that the gene is expressed in lymphocytes. However, the results cannot rule out the possibility that *SPP2* is also expressed in polymorphonuclear cells such as neutrophils, basophils and eosinophils. This will be discussed further in Section 7.3.

The results of this experiment indicate that the expression of the human *SPP2* gene can be up-regulated by some component of the culture medium, an effect also observed for other genes (Grove and Plumb, 1993). The expression of *SPP2* increases in lymphocytes after 16 and 24 hours in culture and in endothelial cells at 16 hours, and both LPS and TNFα inhibit or down-regulate this expression in both cell types. It should be mentioned that when the experiment was repeated using WBCs of a second (oriental Asian) individual, no expression of *SPP2* was detected in any samples even though the *GAPDH* controls were all positive. The significance of this finding is not clear.

In Figure 7.8, in addition to the expected 303 bp RT-PCR product for the human *SPP2* gene, smaller and larger fainter bands are seen in the RT-PCR products of mononuclear cells after 16 hours incubation (Lanes MC16-LPS and MC16+LPS). These fainter bands may show the possibility of the occurrence of low levels of alternatively spliced or incorrectly spliced transcripts. To investigate this possibility, the products seen in the RT-PCR products of mononuclear cells ought to be sequenced and analysed to identify the mRNAs from which they are derived. During a previous similar expression study in mouse, using the RT-PCR method, two smaller but fainter extra bands with sizes different from the expected *SPP2* RT-PCR product were observed in liver and brain. These two fragments were cloned and sequenced, but neither of them was derived from spp24 mRNA and it was concluded that in mouse there is a single transcript of spp24 and alternative splicing does not occur (Bennett, 2002). Although there is the possibility of alternatively spliced or incorrectly spliced transcripts for the *SPP2* gene in human, it is assumed because of the mouse result that the extra faint bands are not *SPP2*, though they were not sequenced and analysed to confirm this.

7.2.11 Expression data for the gene encoding the spp24 protein in mouse obtained from the RIKEN READ database

Miki *et al.* (2001) developed an expression database using microarray technology. In this study the RIKEN mouse cDNA libraries, which originally were constructed by Carninci and Hayashizaki, (1999), were enriched for full-length cDNA and used to collect target cDNAs. In total 18,816 unique cDNAs were arrayed and the gene expression profiles for 49 adult and developing mouse tissues were characterised. It was estimated there were approximately 13,600 non-redundant genes in the array and their expression in all tissues was compared to that in samples prepared form pooled male and female 17.5-day embryos. This was chosen as a reference as the 17.5-day embryo has a relatively complex pattern of expression and it was thought that it was reproducible.

Miki *et al.* (2001) developed a web-based database search engine that was named READ (RIKEN cDNA Expression Array Database) (http://genome.gsc.riken.go.jp/READ/). This database was searched for "secreted phosphoprotein 24" and a clone was obtained (RIKEN ID 1600023D11) that was "very similar to *R. norvegicus* spp24". The sequence of this EST was compared with the mouse spp24 cDNA sequence (Chapter 4) and it was determined that the RIKEN clone was mouse spp24. Table 7.5 illustrates the microarray expression data for

Table 7.5. Microarray expression results obtained for mouse spp24 from READ (RIKEN cDNA Expression Array Database)

The RIKEN cDNA Expression Array Database was searched (http://genome.gsc.riken.go.jp/READ/) for 'secreted phosphoprotein 24'. A clone was identified (ID 1600023D11) that was very 'similar to *R. norvegicus* spp24 protein'. The sequence of the EST was compared with the mouse spp24 cDNA sequence (accession number AJ315513, chapter 4) and it was concluded the clone was the mouse spp24 protein. All tissues values are expressed relative to pooled male and female 17.5-day embryos (reference value is 0).

The colour of each individual box indicates whether the gene is down regulated or up regulated relative to 17.5-day mouse embryos (reference value is 0). A red box indicates that the gene is up regulated compared to the reference and a green box indicates that the gene is down regulated. A black box indicates that no difference was found in the level of expression between the tissue and the reference and white box indicate that there is no result for this tissue.

RIKEN Clone ID	Kidney	Brain	Heart	Lung	Liver	Cerebellum	Placenta	Testis	Spleen	Pancreas	Small Intestine	Stomach
1600023D11	0.955	-2.236	-1.04	-2.155	2.35	-2.082	3.032	-0.427	-2.359	-0.612	-2.227	-1.84

Tongue	Embryo 13 liver	Embryo 10	Embryo 11	Embryo 12 head	Embryo 13 head	Embryo 17 head	Embryo 13	Embryo 15 head	Embryo 16 head	Thymus Preg. 1 day	Embryo 14 liver	10 day lactating mammary gland
-2.451	0.592	0.724	0.325	-0.556			0.346	-1.519		-1.332	-2141	2.813

Skin neonate 0 day	Skin neonate 10 day	Ovary, uterus preg. 11	Intestine neonate 10 day	Thymus	Embryo 12 head	Medulla oblongata	Olfactory brain	Cerebellum Neonate 10 day	Embryo 12 Wolffian duct	Eyeball	Cortex	Seminal vesicule
	-0.646			1.45	-1.32	-1.719		2.248	-1.206	-2.33	-1.26	-0.591

Uterus	Embryo 16 lung	Colon	Cecum	Bone	Sv40t	Lung neonate 0 day	Muscle	Neonate 0 day whole head	Neonate 6 day whole head	Neonate 10 day whole head	Description
-1.152	-1.286	-1.864			0.77		1.267			-0.975	ESTs, highly similar to Spp24 in *R. norvegicus*

Table 7.5. The microarray expression results obtained for mouse spp24 from READ (RIKEN cDNA Expression Array Database)

mouse spp24 in mouse from the RIKEN database. All values are expressed as ratios relative to 17.5-day mouse embryo.

The microarray expression data indicate that the gene encoding spp24 is expressed in placenta, 10-day lactating mammary gland, liver, 10-day neonate cerebellum, thymus, muscle, kidney, hepatocellular carcinoma (Sv40t), the liver of 13-day embryo and whole embryo days 10, 13 and 11. Based on the previous studies, the results observed in liver, kidney and thymus were expected, but the positive results observed in cerebellum, placenta, lactating mammary gland and muscle were new. The highest levels of expression were observed in placenta, 10-day lactating mammary gland, 10-day neonate cerebellum and liver. These values were approximately 2–8 times higher than the other positive ratios observed in other tissues. The Sv40t is a tumour of liver tissue (hepatocellular carcinoma) from a transgenic mouse which contains the SV40 virus under the control of the MUP (major urinary protein) promoter. The data indicate that the expression of the *Spp2* gene is higher in the Sv40t tumour than in the reference 17.5-day embryo, but is down-regulated relative to normal liver.

7.3 Discussion

In three species (human, mouse and cattle) the gene encoding the spp24 protein is expressed at its highest level in liver (Section 7.1.1), therefore it was decided to identify the spatial expression of this gene in mouse liver tissue. Liver, like most other tissues, is composed of different cell types in close proximity, therefore analysis of a differentially expressed gene can only be reliably achieved by a method that can show these differences. Due to the fact that spp24 is a secretory protein, detection of spp24 mRNA using *in situ* hybridisation is the method of choice. Non-isotopic detection of the spp24 mRNA in this study showed that this method could be very useful in tissues such as liver which has a high level of expression of the gene. However it might be less successful in the detection of low copy mRNAs which are dispersed throughout the cellular compartment. In such cases, RNA *in situ* hybridisation using radioactively labelled probes is generally more successful.

In liver the parenchymal cells, hepatocytes, are supported by a reticular fibre stroma and are arranged in cords around a central vein (all central veins terminate into the hepatic vein) to form the classic hepatic lobule. Between adjacent hepatic lobules, the portal lobules are formed and these contain connective tissues, hepatic arteries, portal venules and bile ductules. Tissues are supplied by terminal branches of the hepatic artery and portal vein. Cells nearest these vessels are first to receive oxygen and nutrients. In each lobule, blood flows from vessels in the portal canal through sinusoids that run between cords of hepatocytes into a central vein. The non-parenchymal cells, which constitute about 20% of the total liver volume, consist of Kupffer cells (resident liver macrophage), stellate (fat storing or Ito) cells, endothelial cells and pit cells (large granular lymphocytes).

An enormous number of functions are carried out by parenchymal and the four main types of non-parenchymal cells, either alone or in cooperation. Although the liver tissue is uniform at the histological level, it is heterogeneous at the histochemical level. Hepatocytes around the afferent (periportal, zone-1) vessels differ from those around the efferent (perivenous, zone-3) vessels in their contents of key enzymes and therefore have different metabolic capacities. This is the basis of the model of metabolic zonation, according to which glucose release from glycogen via gluconeogenesis, amino acid utilisation and ammonia detoxification, protective metabolism, bile formation and the synthesis of certain plasma proteins such as albumin and fibrinogen occur mainly in the periportal area, whereas glucose utilisation and the formation of other plasma protein such as alpha 1-antitrypsin or alpha-fetoprotein occur predominantly

in the perivenous zone (Jungermann and Kietzmann, 1996; 1997). Non-parenchymal cells also all exhibit moderate histochemical differences along the sinusoids, with the differences being generally more numerous in the periportal region.

There are different essential factors that are thought to control the zonation in liver, including oxygen, metabolites, signalling molecules, liver-enriched transcription factors and compartment-defining factors. During the passage of blood from the periportal region of the sinusoids to the perivenous region, the concentration of oxygen has been estimated to fall by 50%. It has been shown that the expression of a series of genes in liver is sensitive to oxygen (such as tyrosine aminotransferase) and they are expressed at higher levels with a higher concentration of oxygen. The sinusoidal concentration gradients of most carbon substrates, such as glucose, lactate, amino acids and fatty acids are estimated to be too shallow to play a major role in the regulation of metabolic zonation. However many amino acids and related components are taken up and released by active transport which by itself is zonated. Many hormones are taken up and degraded by the hepatocytes during the passage through the liver, making them plausible candidates as zonal regulators. Hormonal signals elicited under various pathological or nutritional states may affect the spatial expression pattern of a liver-enriched transcription factors and in this way indirectly modulate the zonal transcription of other genes. Glutamine synthetase (GS) is the classical example of a gene whose expression is modulated by compartment-defined signals. Diffusible cellular signals released from endothelial cells of the terminal hepatic venules probably allow the higher expression of this gene (Oinonen and Lindros, 1998).

As was explained briefly in Section 7.2.7, the mouse *Spp2* gene is expressed in hepatocytes all over the liver, but its expression is much higher in the hepatocytes adjacent to portal veins. This higher expression in the vicinity of blood vessels is compatible with the secretary nature of the spp24. The zonation pattern of the *Spp2* gene expression probably is partly due to the higher concentration of oxygen and metabolites in the periportal region, but it also could be due to a signalling molecule or hormone in this region. At present, this is just speculation and it needs further study and investigation to identify the major factors in zonation of the *Spp2* expression pattern. It is probably notable that the gene encoding spp24 in chicken is expressed widely in the developing chicken embryo in response to growth hormone, but its expression is limited to subsets of cells within individual organs (Harvey *et al.*, 2001).

In addition to hepatocytes, *Spp2* gene is expressed in other non-parenchymal cells including endothelial cells, smooth muscles of arteries (suggestive) and connective tissues in the interlobular septum as well. Although *Spp2* is expressed in the endothelial cells of central veins, it seems that its expression is higher still in the endothelial cells of portal veins.

Non-parenchymal liver cells (Kupffer cells, sinusoidal endothelial cells, Ito cells and liver-associated lymphocytes) interact with hepatocytes and with each other by soluble mediators and direct cell-cell contact. The acute phase response is a non-specific reaction of the organism to trauma, injury or infection and its main constituents (the acute phase proteins) are produced by hepatocytes (Knolle *et al.*, 1995). Therefore, as *Spp2* is expressed in a series of non-parenchymal cells and hepatocytes, it may play a role in the acute phase response in the reaction to infections or injury.

Table 7.6 lists a summary of the mouse spp24 protein expression data obtained from four different sources. A single EST was found from the EST profile for uterus and colon, and the READ microarray data indicates negative value results for both these tissues, which is consistent with low level expression of the gene in these two organs. One contradictory result was observed between the READ microarray result for muscle and the RT-PCR of muscle. The READ microarray data show a positive value for the spp24 expression in muscle, but the RT-PCR indicates a negative result. In spite of the high sensitivity of the RT-PCR method, without further experiments it is difficult to draw any conclusion about this result. It should be noted that spp24 expression was not previously detected in mouse skeletal muscle or heart by RT-PCR (Bennett, 2002) but the READ microarray analysis of these tissues indicates the presence of spp24 mRNA in both, with the levels in heart and skeletal muscle diminished and increased respectively with respect to the 17.5-day embryo reference. It is not clear why there is this discrepancy in the results.

The RT-PCR results also show that spp24 is expressed in the mouse adult brain and the READ microarray data show a negative value in this tissue, which may indicate a low level of expression of spp24 in brain. The READ microarray data show that spp24 is expressed in the cerebellum of a 10-day neonate mouse, but not in this tissue in the adult mouse. This finding may indicate that spp24 is expressed in this part of the central nervous system (CNS) at a particular stage of CNS development in mouse, possibly playing a role in the development of balance.

Tissue/Cell type	Mouse ESTs	RT-PCR	READ	*In situ*
Liver	Yes	Yes	2.35	Yes
Kidney	Yes	Yes	0.955	-
Brain	Suggested	Yes	-2.236	-
Placenta	-	-	3.032	-
Embryo 13 day liver	-	-	0.592	-
Embryo 10 day	-	-	0.724	-
Embryo 11 day	-	-	0.325	-
Embryo 13 day	Suggested	-	0.346	-
Mammary gland lactate day 10	-	-	2.813	-
Thymus	-	-	1.45	-
Cerebellum neonate day 10	-	-	2.248	-
Uterus	Suggested	-	-1.152	-
Colon	Suggested	-	-1.864	-
Sv40 liver tumour	-	-	0.77	-
Muscle (striate)	-	No	1.267	-
Diaphragm	Suggested	Yes	-	-
Macrophage	Suggested	-	-	-
T-cell	Suggested	-	-	-
19.5 day total foetus	Suggested	-	-	-
Embryonic carcinoma	Suggested	-	-	-
Endothelial cells of vessels in liver (especially portal vein)	-	-	-	Yes
Mononuclear cells around the vessels in liver	-	-	-	Suggested
Smooth muscles of the arteries in liver	-	-	-	Suggested
Connective tissues around the liver	-	-	-	Suggested
Adult mouse bone	-	No	-	-
Neonate mouse bone	-	Yes	-	-

Table 7.6. A summary of the mouse spp24 expression data obtained from different sources and experiments

This table summarises the expression data obtained for the mouse spp24 protein from two different external sources (ESTs profile and READ microarray results in 'Mouse ESTs' and 'READ' columns respectively), the study that was carried out by Bennett (2002) (an expression study using RT-PCR method in 'RT-PCR' column), study the expression of the mouse spp24 in bone tissue using RT-PCR method (in column 'RT-PCR', section 8.2.2) and detection of Spp2 expression using *in situ* hybridisation in mouse liver tissue (in '*in situ*' column, discussed in Chapter 7). The table shows whether the gene encoding spp24 protein in mouse is expressed for that specific tissue or cell type in the mouse ESTs, the RT-PCRs and *in situ* hybridisation results, with the words 'Yes' or 'No'. In the READ microarray column, the results indicate the expression ratio of the gene in various organs relative to the reference (17.5-day mouse embryo). A dash indicates that there are no available data for that particular tissue or cell type from that source or experiment. 'Suggested' indicates a positive result that cannot be reliably concluded and further study should be performed to confirm it.

The spp24 protein was originally isolated from bovine cortical bone (Hu *et al.*, 1995) and a northern blot showed a positive result for that tissue. Due to the fact that bone is a difficult tissue from which to isolate RNA, the expression of spp24 was not investigated in previous RT-PCR analyses. The RT-PCR results in this study show that spp24 is expressed in the bone tissue of a 12-day neonate mouse, but not in adult mouse bone. The main aim of this experiment was to confirm the expression of spp24 in bone, not to quantify or compare the expression between the neonate and adult bone tissues. Such a comparison would be best carried using kinetic amplification (real-time RT-PCR). During a PCR, the reaction goes through two different phases, the exponential phase and the plateau phase. In the exponential phase every cDNA is copied by the polymerase. This phase take place in the early to middle cycles of the PCR reaction and the number of cycles before a PCR reaction enters this phase depends on the amount of the starting template. Due to the differences in the amount of mRNA in individual reactions, the actual cycle numbers that comprise the exponential phase are difficult to determine. Ideally, it should be determined practically for each individual system. In the plateau phase, components of the reaction become limiting and the synthesis of new products becomes less predictable. Therefore, the quantification of the RT-PCR should be carried out based on the results of the exponential phase (Willard *et al.*, 1999). To carry out a simple and relative quantitation, different dilutions of the reverse transcription reaction products (cDNA) for bone and kidney tissues were prepared. The results of the PCR of different dilutions of *Spp2* cDNA in kidney (used as a positive control in both adult and immature mice) indicated that the method is able to detect *Spp2* mRNA in dilutions of the original cDNA up to 10^6. Therefore if the *Spp2* gene is expressed in the bone of adult mouse, its expression should be less than 10^{-6} relative to kidney. In summary, *Spp2* is not expressed in adult mouse bone, however it is in the bone of 12-day mice, though its level of expression there is significantly lower than that in kidney (the expression of *Spp2* gene in the bone of 12-day mouse is about 10^{-5}-10^{-6} relative to kidney).

Hu *et al.* (1995) speculated about a role for spp24 in the process of bone turnover because of the isolation of the protein from bovine cortical bone and a positive result for bovine northern blot. The result from human microarray data purchased from Incyte Corporation showed that spp24 is not expressed at any significant level by osteoblasts at different stages of development (Bennett, 2002). This result suggests that the spp24 in bone might instead be expressed in osteoclasts as this is where some possible thiol proteinases target proteins may be expressed. However, Kobori *et al.*, (1998) report the isolation of osteoclast-specific genes in the rabbit by the preparation of a subtracted cDNA library. According to his report spp24

does not appear in this library (Kobori, personal communication to Dalgleish). This result suggests that the expression of spp24 in bone reported by Hu *et al.* (1995) occurs in another cell type present in bone other than the osteoblast or osteoclast. However, these results may simply reflect species differences. Therefore in summary prevous syudies have shown that spp24 is expressed in the bone of cattle and rat and our study indicated that it is expressed in very low level in the bone of 12-day immature mouse.

Fetuin is a protein discussed as being similar in some respects to spp24 (Chapter 1) and is expressed mainly in the liver but also shows some expression in the brain. The human fetuin, α_2HS-glycoprotein, is a protein that is expressed in the neurons of the cortex in the developing embryonic brain (Dziegielewska *et al.*, 1987), but it cannot be detected in the human adult brain and it is thought that this is due to death of the particular cell population rather than loss of expression (Saunders *et al.*, 1992). The expression results for spp24 in the mouse brain and bone tissues may indicate that, in a similar way to α_2HS-glycoprotein, spp24 is expressed by a particular cell population that is formed at a specific stage of development in the infant mouse cerebellum. If this cell population were to then diminish or die, spp24 would not be found in the brain and bone tissues of the adult mouse. The detection of spp24 in the 13-week adult mouse brain using RT-PCR (Bennett, 2002) could perhaps be explained by the sensitivity of this method. RT-PCR is the most sensitive method to detect expression of genes and is able to detect small amounts of a transcript, whereas hybridisation techniques do not have this level of sensitivity.

As described earlier in Sections 7.2.9 and 7.2.10, the *SPP2* gene is expressed in total WBCs isolated from whole fresh blood but not in isolated mononuclear cells (lymphocytes and monocytes). This could be due to the *SPP2* being expressed by polymorphonuclear cells, or it might be an inhibitory effect of the reagents present in the Lymphoprep used to isolate the mononuclear cells. It could also be explained in terms of the deprivation of mononuclear cells of stimuli or components to which they are normally exposed when circulating in the body. To resolve this, the method of choice to investigate *SPP2* expression in different types of circulating WBCs would be *in situ* hybridisation of a peripheral blood film.

In contrast to the results in the previous paragraph, RT-PCR indicates that spp24 is expressed in human lymphocytes in culture, but not in monocytes. A small number of ESTs in mouse also suggest that spp24 may be expressed in T-cells and macrophages, but these cells types have not been analysed using any other technique. However, a major proportion of the

chicken spp24 ESTs also originated from T-cell-enriched splenocytes. All these findings in different species indicate that spp24 is expressed in lymphocytes, especially T lymphocytes and may indicate some involvement in the immune system.

Both T-lymphocytes and B-lymphocytes are key components of the adaptive immune responses. These cells are generated in the body's primary lymphoid tissues (thymus for T-cells and the bone marrow for B-cells). In the thymus, T-cells that can tolerate the body's own proteins are selected and those that cannot are destroyed and eliminated. Spp24 expression was also observed in mouse thymus at a high level (READ data), but a negative result was observed in human thymus (MTE data). This finding could be due to technical issues or may reflect true species differences, and further experiments would be required to resolve this.

It has been suggested that many tissue-specific genes are expressed at a very low level rate in many different cell types (Sarker and Sommer, 1989). This property of tissue-specific genes is called illegitimate or ectopic transcription and is the mechanism by which T-cells become tolerant to the body's own proteins. These genes are thought to be expressed transiently and the produced protein rapidly degraded to peptides (Linsk *et al.*, 1989). As explained in Section 7.2.10, no expression of human spp24 was found in the isolated lymphocytes from the second person (oriental Asian). Possible explanations include that one of the two individuals from whom the lymphocytes were isolated, had experienced some sort of acute response prior to the cells being cultured that resulted in the difference in *SPP2* expression that was observed.

The expression of spp24 was also observed in human endothelial cells from the lining of the umbilical cord using the RT-PCR technique. This result confirmed and supported the *in situ* hybridisation results which indicated spp24 expression in the endothelial cells lining the vessels in mouse liver tissue.

Table 7.7 summarises the human spp24 protein expression data obtained from three different experiments or sources, including one contradictory result. The RT-PCR result indicates that spp24 is expressed in the human peripheral blood leukocytes but the Clontech human MTE array results are negative in these cells. The detection of spp24 in the human peripheral blood leukocytes using RT-PCR could be explained by the sensitivity of this method.

Tissue/Cell type	Human ESTs	Clonetech human MTE array	RT-PCR
Liver	Yes	Yes	-
Foetal liver	Yes	Yes	-
Kidney	-	No	-
Foetal kidney	-	Yes	-
Thymus	-	No	-
Peripheral blood leukocytes	-	No	Yes
Monocytes	-	-	No
Lymphocyte	-	-	Yes
Endothelial cells	-	-	Yes

Table 7.7. A summary of the human spp24 expression data obtained from different sources and experiments

This table summarises the expression data obtained for the human spp24 protein from the ESTs profile, the study that was carried out by Bennett (2002) and the expression of spp24 in WBCs using the RT-PCR method (Sections 7.2.9 and 7.2.10). The table shows whether the gene encoding spp24 is expressed for that specific tissue or cell type in the human ESTs, the Clonetech human MTE array and the RT-PCRs with the words Yes or No. A dash indicates that there are no available data for that particular tissue or cell type from that source or experiment.

The results presented in this chapter also indicate that the expression of the human *SPP2* gene can be up-regulated by the foetal calf serum in the culture media (Grove and Plumb, 1993) and that both LPS and TNFα prevent, or down-regulate, the expression of the gene in lymphocytes and endothelial cells. The tumour necrosis factors (TNFs) are cytokines which are all small molecular weight peptides or glycopeptides. Many of the cytokines are produced by multiple cell types such as lymphocytes, monocytes/macrophages, mast cells, eosinophils and even endothelial cells lining blood vessels. The genes encoding TNFs and many similar proteins are members of a gene superfamily. The receptors for these proteins also constitute a TNF receptor (TNFR)-related gene superfamily. Most TNF/TNFR superfamilies are expressed in the cells of the immune system and their rapid and potent signalling properties are essential in coordinating the protective functions of these cells which are responsible in reaction against different pathogens. TNF secretion can be induced by different structural elements common to microbial pathogens, including components such as lipopolysaccharide (LPS) and peptidoglycan (Locksley *et al.*, 2001).

As described by Locksley *et al.* (2001), the most important functions of TNF/TNFR superfamilies can be summarised as:

- They communicate between cells and can organise permanent multicellular structures such as bone, lactating mammary gland, lymphoid organs and hair follicles.
- They coordinate the responses of the immune system against different antigens and also participate in acute immune response.
- They have a fundamental role in mediating cell death.
- They regulate the bone and mammary gland homeostasis.
- They play some role in hair follicle and sweat gland development.

In addition to lacking secondary lymph nodes, mice deficient in receptor activator of NF-kappaB ligand (RANKL) or its receptor (RANK) have severe osteopetrosis. Bone-resorbing agents including vitamin D_3, PTH and cytokines, such as IL-1 and TNF, up-regulate RANKL which mediates the differentiation and activation of osteoclasts from monocyte precursors and subsequently the induction of bone resorption. Thus, in both mice and humans TNF/TNFR superfamilies play central roles in the regulation of osteoclast differentiation and activation and therefore play an important role in calcium metabolism. In addition to their essential role in bone homeostasis, RANKL/RANK signals also induce and regulate the terminal differentiation of mammary gland alveolar buds to create the lobulo-alveolar structure component for lactation (Locksley *et al.*, 2001).

In another study it was also suggested that LPS is involved directly in inflammatory bone loss (by stimulating the survival and fusion of preosteoclasts, independent of RANKL, IL-1 and TNF) and also indirectly through the production of LPS-induced host factors such as TNFα and IL-1 (Suda *et al.*, 2002). In a recent study, Beckman *et al* (2003, abstract submitted) showed that spp24 is produced in several calciotropic organs such as kidney and bone in response to alterations in serum calcium (hypocalcaemia) and also in response to parathyroid hormone (PTH). According to these studies and the result of the experiment described in Section 7.2.10, it can be concluded that expression of spp24 may be down-regulated by TNF and LPS (directly or indirectly through TNF). It can be also speculated that the down regulation occurs following bone resorption by TNF and LPS which lead to hypercalcaemia and decrease in the secretion of PTH (in presence of LPS and TNF). Therefore down regulation of spp24 may be one of the mechanisms by which LPS and TNF can increase bone resorption, resulting subsequently in osteoporosis. It has been also indicated that TNFα has a key role in skeletal diseases in which it promotes reduced bone formation by mature osteoblasts. It was also shown that TNFα inhibits differentiation of osteoblast from precursor cells (Gilbert *et al.*, 2000). As previously mentioned, the gene encoding spp24 in chicken is expressed widely in the developing embryo in response to growth hormone, but its expression is limited to subsets of cells within individual organs (Harvey *et al.*, 2001). Therefore if spp24 has a role in growth and development of different cells or tissues (as in chicken), it is possible that one of the mechanisms by which TNFα inhibits the development of osteoblasts is through down regulation of the gene encoding the spp24 protein.

The READ microarray data give a strongly positive signal for the expression of spp24 in lactating mammary gland (Section 7.2.11), but this tissue has not been tested in other species. The Clontech human MTE array contained a sample from mammary gland but this was not from lactating mammary gland (Bennett, 2002). These results may suggest a role for spp24 in the development of lactating mammary glands from the non-lactating state. The expression of spp24 in lactating mammary glands is also interesting as it suggests the possibility that the protein may be secreted and found in milk. If so, it could be isolated from milk for further investigation. The expression of spp24 in mouse lactating mammary gland and placenta also suggests a possible antimicrobial function as well. Immunity can be passed from mother to foetus through both milk and placenta. This speculated role is supported by the fact that spp24 shows some similarity to the bovine neutrophil antibiotic peptide bactenecin precursor (Hu *et al.*, 1995).

Price *et al.*, (2002) described the fetuin-mineral complex (FMC), a high molecular weight complex of the proteins fetuins and matrix Gla protein (MGP) with calcium phosphate mineral that was originally discovered in the serum of rats treated with the bisphosphonate etidronate, and apparently playing an important role in inhibiting *in vivo* calcification (Price and Lim, 2003). Spp24 is now another protein that is recognised as a component of FMC (Price *et al.*, 2003). The fetuin-mineral complex reaches maximal levels 6 to 9 hours after etidronate treatment (the bone-active bisphosphonate etidronate inhibits bone mineralisation and increases the concentration of total serum calcium) and causes an approximately 4-fold increase in total serum calcium without causing any increase in the amount of ionic calcium. Acute inhibition of bone mineralisation by etidronate is accompanied by bone resorption by osteoclasts and a sharp increase in the concentrations of calcium and phosphate in the aqueous phase of the compartment. This increase spontaneously leads to formation of calcium phosphate crystal nuclei, but their further growth is inhibited by formation of FMC. At 6 to 24 hours following etidronate injection, almost the entire FMC is removed from the blood. It is not clear what targets the FMC for removal from blood, what receptor is involved or which cells are responsible for this removal. Both fetuin and spp24 are serum proteins synthesised in the liver, reaching the bone remodelling compartment by way of the continuous flow of blood through this compartment, supplying a reservoir of serum fetuin and spp24 for the continued formation of FMC within the compartment.

Maintenance of the body calcium stores and plasma calcium is tightly regulated and depends on dietary intake, its absorption from gastrointestinal tract and renal excretion. Renal excretion generally parallels sodium excretion and is affected by many of the same factors that regulate sodium transport in the proximal tubule, but PTH increases distal tubular reabsorption independently of sodium. The calcium in FMC cannot be reabsorbed by the kidneys as efficiently as free ionic calcium and therefore is excreted, which probably is one of the mechanisms that decreases the calcium in hypercalcaemic conditions. These data suggest that spp24 probably, like fetuin and MGP, plays a crucial role in inhibiting calcification.

As mentioned before, spp24 is expressed at a considerable level in mouse lactating mammary gland. Due to the fact that milk contains a variety of minerals especially calcium, FMC formation is possibly one of the mechanism that facilitates the excretion of milk and inhibits calcium deposition in mammary glands and tubules.

In summary, the results of this chapter and previous studies indicate the tissue-specific nature of spp24 expression, and the diversity of tissues and cells in which it is expressed support the notion that spp24 has multiple specific functions. These may include a role in liver function, a circulatory plasma protein function, a role in immune system, an antimicrobial function and finally a role in bone turnover and calcium metabolism.

Table 7.8 lists and compares a summary of the spp24 expression data obtained from various experimental sources for human, mouse, bovine and chicken.

Tissue/Cell type	Human	Mouse	Bovine	Chicken
Liver	Yes	Yes	Yes	Yes
Foetal liver	Yes	Yes	-	-
Kidney	No	Yes	No	-
Foetal kidney	Yes	-	-	-
Brain-cerebellum (infant)	-	Yes	-	-
Bone (adult)	-	No	Yes	-
Bone neonate day 12	-	Yes	-	-
Diaphragm	-	Yes	-	-
Placenta	No	Yes	-	-
Thymus	No	Yes	-	-
Mammary gland lactate day 10	-	Yes	-	-
Uterus	No	Suggested	-	-
Colon	No	Suggested	-	-
Muscle	No	Suggested	-	-
T-cell	-	Suggested	-	Yes
Macrophage	-	Suggested	-	-
Peripheral blood leukocytes	Yes	-	-	-
Monocytes	No	-		-
Lymphocyte	Yes	Suggested	-	-
Endothelial cells of vessels in liver	-	Yes	-	-
Endothelial cells of umbilical cord	Yes	-	-	-
Mononuclear cells around the vessels in liver	-	Suggested	-	-
Smooth muscles of the arteries in liver	-	Suggested	-	-
Connective tissues around the liver	-	Suggested	-	-

Table 7.8. A summary of the spp24 protein expression data from human, mouse, bovine and chicken

This table summarises the conclusions with respect to spp24 expression for human, mouse, bovine and chicken tissues (from results in table 8.2, table 8.5, table 8.6 and by Hu *et al.*, 1995). The words 'Yes' or 'No' simply indicates the expression or no expression respectively. A dash indicates that there are no available data for that particular tissue or cell type from that source or experiment. 'Suggested' indicates a positive result that cannot be reliably concluded and further study should be performed to confirm it.

Chapter 8
Construction of a Mouse Pooled-Tissues cDNA Library and Use of the Yeast Two-Hybrid System to Find Proteins Interacting with spp24

8.1 Introduction

The function of a gene of interest may not be obvious even after a broad study of its structure and expression patterns. In these circumstances it might be helpful to try to identify those proteins with which it may interact, especially if these proteins have previously been extensively studied and their functions are well known. In early years, to identify protein interactions different approaches such as co-immunoprecipitation and co-fractionation were used. These physical methods are often not so effective. However, recently, other powerful methods have been developed including the now widely-used yeast two-hybrid system.

Due to the fact that spp24 is a secrotory protein the best candidate interacting protein is an extracellular protein. In an earlier attempt to identify potential interacting proteins for spp24, the yeast two-hybrid system was used to test the interaction between spp24 and cathepsin K, and to screen for interactions between spp24 and a random peptide library and also a mouse whole embryo cDNA library. No interactions were identified (Bennett and Dalgleish, unpublished).

The lack of positive results from the two library screens and the specific interaction might have been due to any of the following reasons:
1- The spp24 protein may not be compatible with the yeast two-hybrid system, because there are known limitations related to protein folding in yeast.
2- It is possible that the target protein for spp24 is not expressed in any tissue of the mouse during embryonic life (or not at the embryonic stage used for the construction of the cDNA library) and is only expressed after birth or after exposure of mice to stimuli such as infectious or inflammatory agents.
3- If the target protein for spp24 acts as a trans-membrane receptor, it is possible that it does not adopt its native conformation before insertion into the membrane and so, using this method, we will not be able to obtain a positive interaction.

4- Spp24 is a secreted phosphoprotein, which contain two di-sulphide bands and undergoes some modification (phosphorylation) after translation, which might limit using this method to obtain a positive interaction.
5- It is possible that for interaction between spp24 and target proteins, longer amino-acid sequences are needed than found in the random peptide library, the average length of which was 16 amino acids.

Although limitations 1, 3 and 4 may still hold true, but in order to address points 2 and 5, it was decided to carry out this experiment for the second time. This chapter describes the construction of a mouse pooled tissues cDNA library and the use of this library in the two-hybrid system to identify proteins interacting with the mouse spp24 protein.

8.1.1 cDNA libraries

As gene expression varies in different cells and at different stages of development, individual cDNA libraries constructed from the RNA from specific tissues or specific stages of development will contain different complements of cDNA clones. From total RNA, poly A^+ mRNAs are selected by specific binding to complementary single-stranded oligonucleotides bound to a solid matrix. The isolated poly A^+ mRNAs can be converted to double-stranded cDNAs using reverse transcriptase. To assist cloning in a suitable vector, oligonucleotide linkers which contain appropriate restriction enzyme sites can be ligated to each end of the cDNA.

Every good cDNA library should have three key characteristics:
1- It should be large enough to contain representatives of all sequences of interest, some of which may be derived from low copy number mRNAs.
2- It includes a minimal number of clones that contain small (often defined as inserts smaller than 500 bp) cDNA inserts.
3- It is composed of cDNA inserts that are near-full length copies of the mRNA from which they were derived.

Traditional cDNA cloning methods often require the use of restriction enzymes to adapt the ends of the cDNA to fit into a cloning vector. These restriction enzymes may cut some of the cDNA inserts into multiple segments, reducing the chance of obtaining full-length cDNA clones. This limitation can be avoided using technologies such as "Gateway" (Invitrogen)

which utilise *in vitro* recombination to introduce the cDNA into the cloning vector (Nagy, 2000), hence maximising the number of full-length clones. However, this technology was not available as a commercial kit at the time that the cDNA was constructed, and so the traditional method (using restriction enzymes) was used.

8.1.2 Yeast two-hybrid system

The yeast two-hybrid system is an *in vivo* genetic assay for detecting protein-protein interactions by exploiting the ability of a pair of interacting proteins to bring into close proximity a DNA-binding domain (DB) and a transcription activation domain (AD) that regulate the expression of a reporter gene (Chien *et al.*, 1991).

Due to the several advantages of the yeast two-hybrid system, this method was chosen for this study. First, it provides a sensitive method to detect relatively weak and transient protein interactions. Such interactions might not be detectable biochemically, but might be essential for the proper functioning of a biological pathway. In addition, due to the fact that the two-hybrid system assay is carried out *in vivo*, the proteins are more likely to be in their correct conformation, which sometimes increases the sensitivity and accuracy of the detection (Guarente, 1993). Secondly, the transformation and manipulation of yeast is simple and all necessary reagents, vectors and yeast strains were available in the department. Thirdly, many of the protein-protein interactions may also happen naturally in yeast, and therefore the interactions may be facilitated by the host. Also, non-nuclear proteins are targeted to the nucleus. In spite of the fact that spp24 is naturally a secreted protein, it is targeted in this system to the nucleus where it can be tested for an interaction with a specific protein or proteins expressed from library constructs. Other reasons for using this system are that most protein-protein interactions will activate the transcription of reporter genes, because the fused protein domains are not required to be in a perfect conformation. Also, it is possible to quantitate the affinity of an interaction (Fields and Sternglanz, 1994).

The yeast two-hybrid system used in this study took advantage of the *S. cerevisiae GAL4* protein. *GAL4* is a transcriptional activator crucial for expression of the genes encoding the enzymes required for galactose metabolism and utilisation (Johnson, 1987). The *GAL4* protein has two distinct domains, both of which are necessary for the protein function. The N-terminal domain (DB) binds to a specific DNA sequences and the acidic C-terminal domain (AD) is essential for transcription activation (Ma and Ptashne, 1987).

The *GAL4*-based two-hybrid system requires the construction of a *GAL4* DNA binding domain fused to one protein of interest (bait, X) and a *GAL4* activation domain fused to the other protein/proteins of interest (prey, Y). Even after fusion of these two domains to another unrelated protein, they retain sufficient activity to activate and regulate the expression of the reporter genes. Figure 8.1 illustrates the strategy of the *GAL4* yeast two-hybrid system. Yeast two-hybrid proteins are constructed to contain one of the interacting proteins, either X (bait) or Y (prey). "Bait" normally refers to the protein fused to the *GAL4* DNA-binding domain and "prey" to the protein fused to the *GAL4* activation domain. Neither of these two fusion proteins can activate transcription of the reporter gene(s) on their own, but when an interaction occurs between bait and prey proteins, the *GAL4* DNA-binding domain and activation domains are brought into enough proximity to be able to activate transcription of the reporter gene(s).

The yeast two-hybrid system used in the study presented in this chapter used *S. cerevisiae* strain AH109 (Section 2.13.3) (James *et al.*, 1996). Strain AH109 (derivative of strain PJ69-2A) is *TRP* and *LEU* deficient enabling selection of *TRP1* and *LEU2* plasmids. The strain has the *HIS3*, *ADE2* selectable reporter and the *MEL1* and *lacZ* reporter genes. Consequently, when an interaction occurs, transcription of *HIS3*, *ADE2*, *lacZ* and *MEL1* are activated. The *ADE2* reporter alone provides strong nutritional selection. The option of using *HIS3* selection reduces the incidence of false positives and allows us to control the stringency of the selection. Using this strain also provides the opportunity to use either *MEL1* or *lacZ*, which encode α-galactosidase and β-galactosidase respectively.

8.1.3 Yeast two-hybrid vectors

The hybrid proteins are generated by expression of DNA from two expression vectors encoding either the DNA-binding domain or the activation domain. The vectors used in the yeast two-hybrid system were pEXP-AD502 (Invitrogen, cat. number 11376027) containing the *GAL4* activation domain and pDEST 32 (Invitrogen, cat. number10835031) containing the *GAL4* DNA-binding domain.

8.1.3.1 Donor vector pDONR 201

It was decided to construct the cDNA clone library and the DNA-binding domain expression construct using vectors compatible with Invitrogen Gateway Cloning Technology. Gateway Cloning Technology is a novel universal system for cloning and sub-cloning DNA sequences

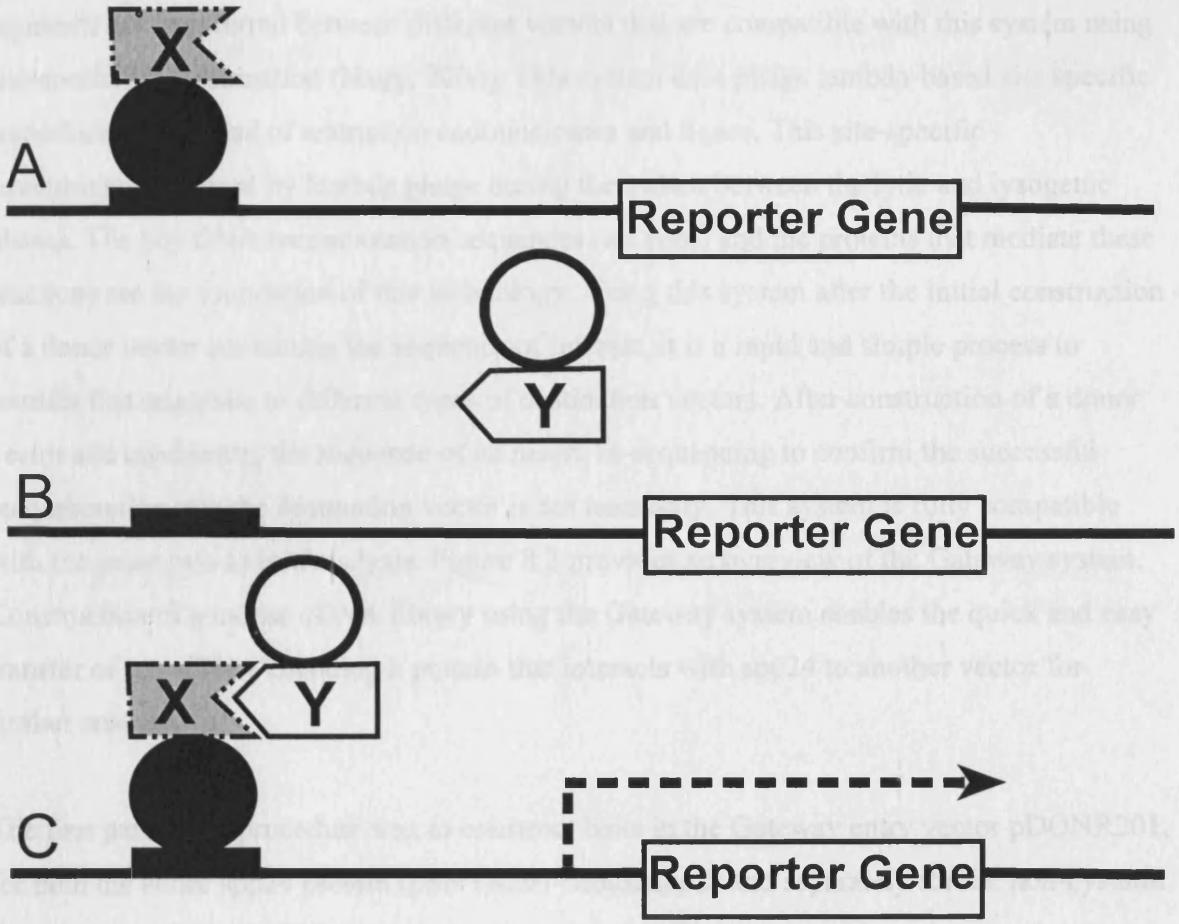


Figure 8.1. The yeast two-hybrid principle

The vectors are constructed containing one of the interacting proteins X and Y. Protein X (bait) is fused to the *GAL4* DNA-binding domain (illustrated by a black circle), and Y (prey) to the *GAL4* activation domain (illustrated by a white circle). Both proteins cannot activate transcription on their own (A and B). However, when bait and prey interact, the *GAL4* DNA-binding domain and the *GAL4* activation domain are brought into close enough proximity to be able to activate transcription of the reporter gene (C).
The sequence to which the *GAL4* DNA-binding domain binds is shown by a black rectangle. The reporter gene is boxed and transcription is indicated by an arrow.

that facilitates the analysis of gene function and protein expression. In this system, DNA segments are transferred between different vectors that are compatible with this system using site-specific recombination (Nagy, 2000). This system uses phage lambda-based site specific recombination instead of restriction endonucleases and ligase. This site-specific recombination is used by lambda phage during the switch between the lytic and lysogenic phases. The key DNA recombination sequences (*att* sites) and the proteins that mediate these reactions are the foundation of this technology. Using this system after the initial construction of a donor vector containing the sequence of interest, it is a rapid and simple process to transfer this sequence to different types of destination vectors. After construction of a donor vector and confirming the sequence of its insert, re-sequencing to confirm the successful recombination into the destination vector is not necessary. This system is fully compatible with the yeast two-hybrid analysis. Figure 8.2 provides an overview of the Gateway system. Construction of a mouse cDNA library using the Gateway system enables the quick and easy transfer of any cDNA encoding a protein that interacts with spp24 to another vector for further analysis.

The first part of the procedure was to construct baits in the Gateway entry vector pDONR201, for both the entire spp24 protein (pENTR201-mouse*Spp2*) and separately for the non-cystatin domain of spp24 (pENTR201-mouse non-cystatin*Spp2*).

The donor vector pDONR201, illustrated in Figure 8.3A, uses the BP reaction for production of kanamycin-resistant entry clones. DNA from the PCR amplification of spp24 cDNA replaces the region between the *att* sites. The vector has the pUC origin of replication and selectable markers that confer resistance to kanamycin and chloramphenicol thereby permitting standard propagation in *E. coli* cells. The vector contains T1 and T2 transcription terminators to minimise the possible toxic effects of cloned genes expressed from vector-encoded promoters. The vector with no insert must propagated in *E. coli* DB3.1 host cells because of the *ccd*B gene. The presence of this gene selects against the growth of clones in which the region between *att*P1 and *att*P2 sites has not been successfully replaced. After construction of entry clones and verification of sequence and reading frame, the DNA can then be transferred using the LR reaction into an appropriate expression vector with the correct sequence and reading frame being preserved.

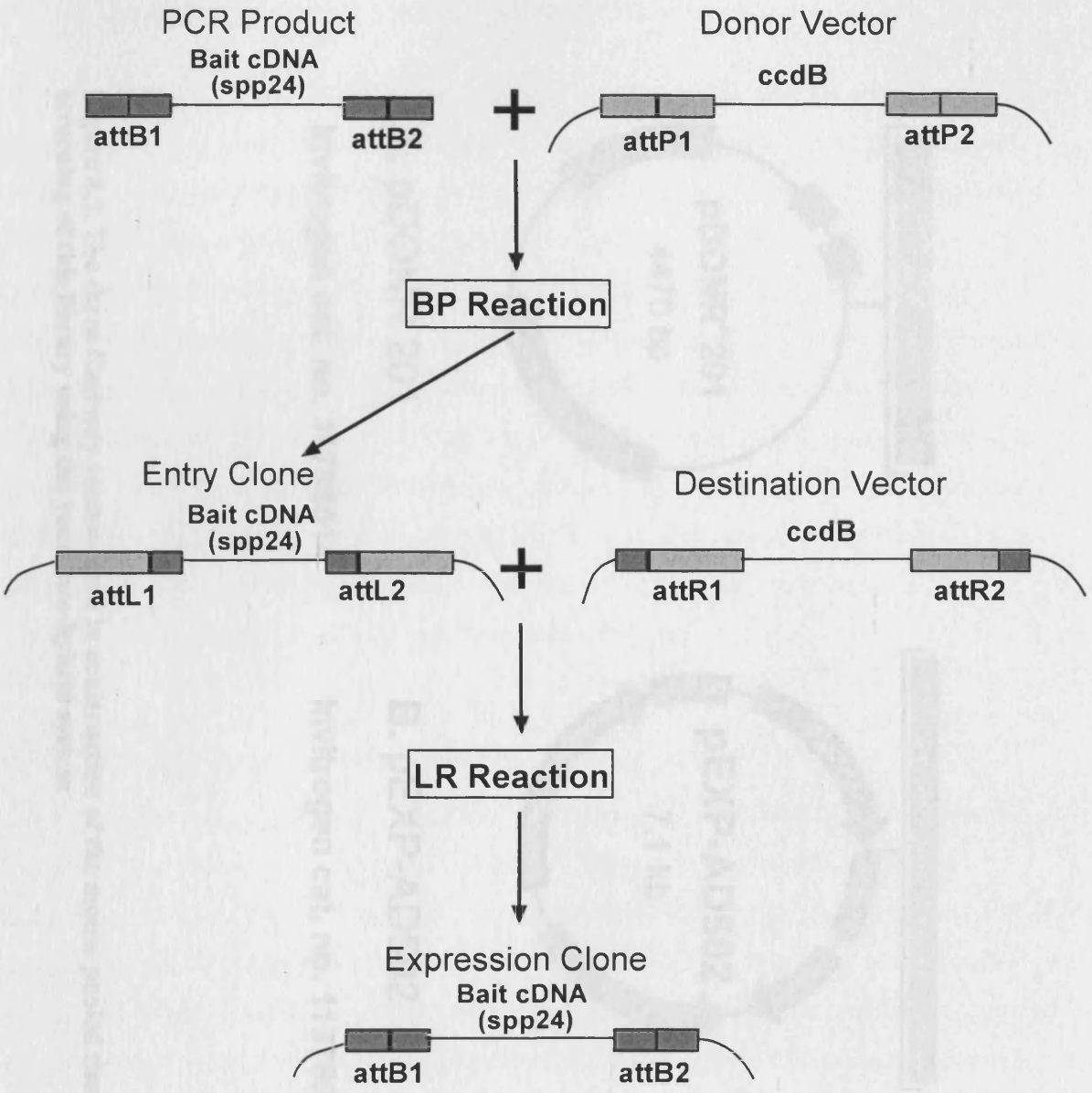
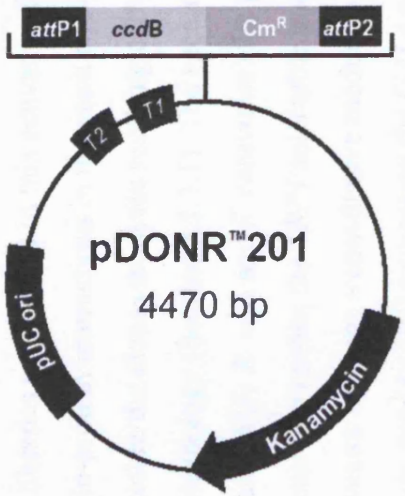
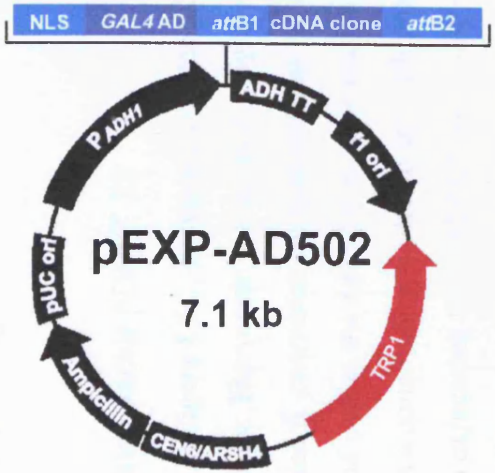
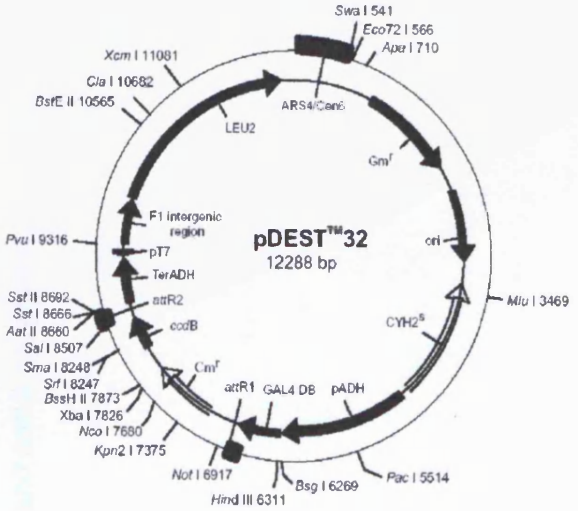


Figure 8.2 Invitrogen Gateway Cloning Technology as an operating system for cloning and subcloning DNA

This figure provides an overview of the Gateway cloning system and illustrates how an *att*B-tagged PCR product can be transfered into vectors by site-specific recombination.







A. pDONR 201

Invitrogen cat. no. 11798014

B. pEXP-AD502

Invitrogen cat. no. 11376027

C. pDEST 32

Invitrogen cat. no. 10835031

Figure 8.3. The three Gateway vectors used in construction of the mouse pooled tissues cDNA library and screening of this library using the yeast two-hybrid system

8.1.3.2 The activation domain expression vector pEXP-AD502

The expression vector pEXP-AD502 is illustrated in Figure 8.3B, and has been designed for directional cloning of DNA inserts. cDNA libraries constructed in this activation domain fusion vector are used for identifying protein-protein interactions in the yeast two-hybrid system. The vector contains *att*B sites, therefore the clones from the resulting cDNA library can be used in conjunction with Gateway technology (Section 8.1.3.1). The vector has the pUC origin of replication. As selectable markers (in *E. coli* and *S. cerevisiae*) the plasmid confers resistance to ampicillin for *E. coli* and also carries the *TRP1* selectable marker (nutritional marker) for *S. cerevisiae*. The vector contains a constitutive moderate-strength promoter for the yeast alcohol dehydrogenase gene (*ADH*) expressing the *GAL4* activation domain fused to the nuclear localisation signal (NLS) for SV40 upstream of the *att*B1 site followed by polylinker (containing six unique restriction enzyme sites) and *att*B2. Downstream of the *att*B2 site is the *ADH* termination signal.

8.1.3.3 The DNA-binding domain expression vector pDEST 32

The expression vector pDEST 32, illustrated in Figure 8.3C, contains *att*R sites allowing it to be used in conjunction with Gateway technology. The vector has the pUC origin of replication. As selectable markers in *E. coli* and *S. cerevisiae* the plasmid confers resistance to gentamicin and chloramphenicol in *E. coli* and carries the *LEU2* selectable marker (nutritional marker) for *S. cerevisiae*. The vector contains a constitutive moderate-strength promoter for the yeast alcohol dehydrogenase gene (*ADH*) expressing the *GAL4* DNA-binding domain fused to the nuclear localisation signal for SV40 upstream of the *att*R1 site followed by the chloramphenicol resistance and *ccd*B genes and *att*R2. Downstream of the *att*B2 site is the *ADH* termination signal.

8.2 Results

8.2.1 Construction of a mouse pooled tissues cDNA library

To perform a successful yeast two-hybrid analysis, a high quality cDNA library is needed containing clones for any protein with which spp24 might interact. The main site of spp24 synthesis (in different species) is liver (Chapter 7), but due to the fact that spp24 is a secreted phosphoprotein, it might interact with several proteins, synthesised in different and probably distant organs. Therefore construction of a cDNA library should not be limited to cells of organs in which *Spp2* is expressed. To achieve as comprehensive a search as possible for proteins which might interact with spp24, it was decided to use pooled tissues from all mouse organs to construct a cDNA library.

Total RNA was extracted from all tissues of two adult male mice (C57BL/6J, 12 and 13 weeks, supplied by Biomedical Services, University of Leicester), using the RNAzol B method described in Section 2.16.1. The quality of the extracted total RNA from all tissues was assessed using denaturing agarose gel electrophoresis as described in Section 2.17. To make mouse pooled-tissues total RNA, equal amounts of total RNA from each individual tissue were mixed together and polyA$^+$ mRNA was purified from the mixed total RNA using the method described in Section 2.16.2. To increase the concentration of polyA$^+$ mRNA, it was precipitated using the method described in Section 2.16.5.

The cDNA library was constructed from the polyA$^+$ mRNA using the SuperScript Plasmid System with Gateway technology method described in Section 2.24. This method employs the method of double-stranded cDNA synthesis (Okayama and Berg (1982) with modifications by Gubler and Hoffman (1983)).

A *Not*I primer-adaptor (consisting of an oligo-dT primer and 5′ *Not*I adaptor) was annealed to the polyA$^+$ mRNA and the first strand synthesised using SuperScript II reverse transcriptase and ^{32}P-labelled dCTP. Before synthesis of the second strand cDNA, the quantity and quality of the first strand cDNA was determined. The overall yield of the first strand reaction was calculated from the amount of acid-precipitable radioactivity determined as described in Section 2.24.2. The specific activity and the amount of the first-strand cDNA were determined as 1655 (cpm/pmol dCTP) and 0.18 μg respectively. The first-strand cDNA, was

also analysed by alkaline agarose gel electrophoresis (Section 2.24.3) to estimate the size range of product synthesised. Figure 8.4 shows that the distribution of the cDNA is centred in the 500 bp to 2 kb range, which indicates a reasonable quality (1-2 kb) for the cDNA.

The second strand was synthesised using *E. coli* RNase H, which introduces nicks into the mRNA strand of the cDNA, and *E. coli* DNA polymerase I, that repairs the gaps created by RNase H in a 5′ to 3′ direction. The fragments of DNA were then ligated together by *E. coli* DNA ligase. The double stranded DNA was then blunt-ended using T4 DNA polymerase, and a *Sal*I adapter ligated to the both ends of the cDNA. Following digestion of the cDNA with *Not*I restriction enzyme to cut the *Not*I adaptor, the cDNA was size fractionated by column chromatography and fractions of presumed different-sized cDNAs were collected. These were designated F1, F2 and F3. Each fraction was independently ligated into *Not*I/*Sal*I cut pEXP-AD502 vector using T4 DNA ligase and transformed into electro-competent *E. coli* DH10B by electroporation. Colonies were grown on media containing ampicillin since pEXP-AD502 has an ampicillin resistance gene (Section 8.1.3.2), however there was no means of identifying any background non-recombinant colonies as the plasmid vector used contains no indicator or selector for this purpose. A total of 1×10^5 clones were obtained with approximately the same number of clones resulting from each cDNA fraction. The transformation efficiency of the library was about 1.25×10^6 transformants per μg of cDNA. To obtain a comprehensive library with more independent primary clones, a second library was constructed using the same conditions. This time a total of 2.03×10^5 clones were achieved. The transformation efficiency of the library was about 2×10^6 transformants per μg of cDNA. Therefore, in total, 3.03×10^5 independent clones were obtained.

The primary independent clones were amplified on nine 140-mm plates; one plate each for each of the three cDNA fractions for the first library, and two plates each for each of the cDNA fractions of the second library. Transformed cells representing each of the cDNA fractions from both libraries were independently harvested into 15 ml of freezing medium as described in Section 2.25.2. (The number of independent clones in the library-2 aliquots is double that for the aliquots from library 1). These 15-ml aliquots were then sub-divided into 1-ml aliquots prior to freezing at -80°C.

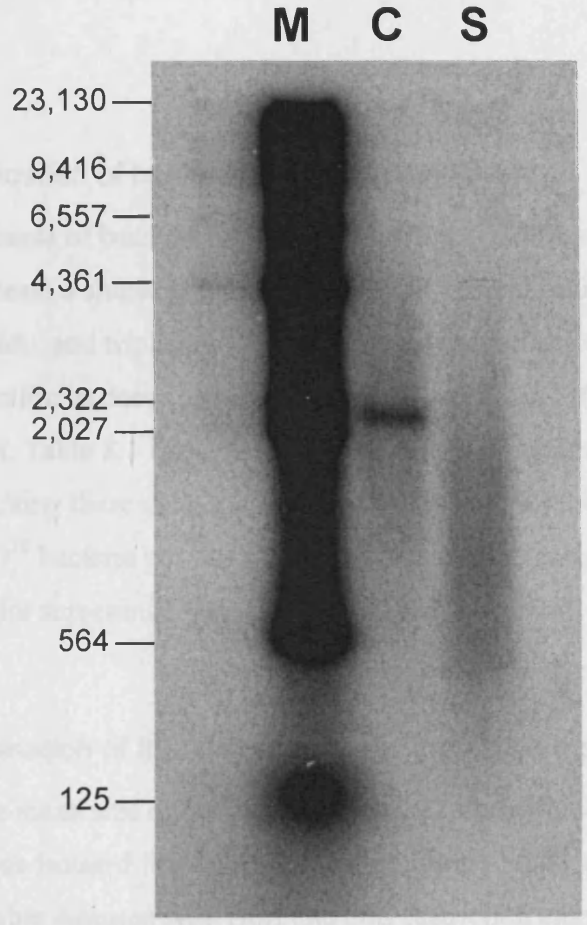


Figure 8.4. Alkaline agarose gel electrophoresis of the first strand cDNA synthesised with the SuperScript Plasmid System

Samples of ^{32}P-labelled first-strand cDNA made from control mRNA (C) and mRNA isolated from all different mouse tissues (S). The distribution of the cDNA is centred in the 500 bp to 2 kb range, which indicates a reasonable quality for the cDNA.

M = marker λ/*Hin*dIII

8.2.2 Characterisation of the mouse pooled-tissues cDNA library

Prior to screening the mouse pooled-tissues cDNA library in yeast two-hybrid analyses, the library was characterised. The bacterial cell density and the mean cDNA insert size were determined.

8.2.2.1 Quantification of bacterial density in the library

Appropriate amounts of bacteria representing the three cDNA size fractions for both libraries were mixed to create a single pooled library. Serial 10-fold dilutions of the pooled-library bacteria were made, and triplicate 50 µl aliquots of each dilution were plated on LUA plates containing ampicillin as described in section 2.12. Dilutions 10^{-3} to 10^{-5} produced too many colonies to count. Table 8.1 lists the number of colonies obtained on each of the 10^{-6} and 10^{-7} dilution plates. Using these data, the bacterial cell density of the pooled library was calculated as about 1.1×10^{10} bacteria per ml. This value was used to calculate the volume of the pooled library required for screening.

8.2.2.2 Determination of the average size of the cDNA inserts

To determine the mean size of the cDNA inserts, 72 library clones were selected at random. Plasmid DNA was isolated from each individual library clone, derived from well-separated colonies and double digested with *Not*I and *Sal*I restriction enzymes to release a linear plasmid fragment and the cDNA insert fragment. Isolation of plasmid DNA from *E. coli* and restriction digestion were carried out according to the methods described in Sections 2.12.8.1 and 2.3 respectively. Analysis by agarose gel electrophoresis allowed calculation of the size of each individual insert. Figure 8.5 shows electrophoresis data for samples 1–12, 37–48 and 61–72, and the sizes of inserts for all 72 samples are listed in Table 8.2.

Two of the colonies (numbers 32 and 65) harboured a cDNA that released more than one band following restriction digestion. This was probably due to digestion within the insert by *Sal*I restriction enzyme, because prior to ligation of cDNAs into the vector (pEXP-AD502), the cDNA fragment were digested by with *Not*I. This should mean that there are no *Not*I sites within any of the cloned cDNA fragments. Four clones (numbers 53, 58, 59 and 69) did not appear to have any insert. This may be because these inserts have a similar size to the plasmid fragment (7126 bp) and co-migrate with it, though this is unlikely given that so little of the original cDNA was as large as the vector (Figure 8.4). The alternative explanation is that no

Dilution factor	Count			Mean
	1	2	3	
10^{-6}	580	476	513	523
10^{-7}	50	55	48	51

Table 8.1. Number of colonies cultured from 10^{-6} and 10^{-7} dilutions of the mouse pooled tissues cDNA library

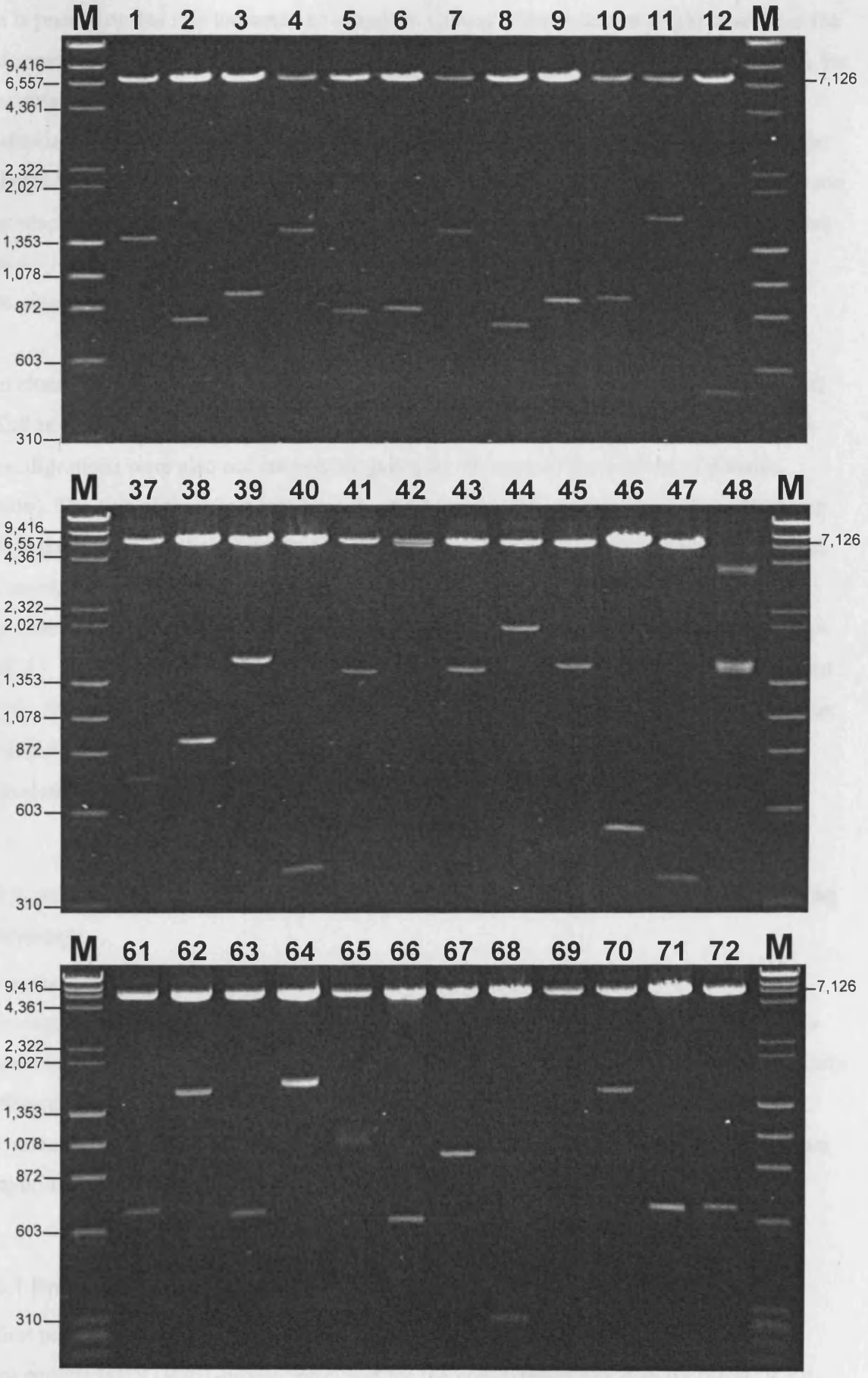
Clone Number	Insert Size	Clone Number	Insert Size	Clone Number	Insert Size	Clone Number	Insert Size
1	1401	19	887	37	730	55	796
2	793	20	844	38	943	56	1000
3	950	21	1886	39	1620	57	2385
4	1488	22	1018	40	404	58	unknown
5	843	23	781	41	1516	59	unknown
6	859	24	1761	42	6500	60	1538
7	1478	25	undigested	43	1527	61	688
8	769	26	1034	44	2017	62	1614
9	913	27	253	45	1561	63	682
10	926	28	941	46	533	64	1725
11	1613	29	840	47	383	65	1259 + 1159
12	477	30	1013	48	undigested	66	655
13	630	31	undigested	49	1118	67	1041
14	undigested	32	1253 + 684	50	1126	68	318
15	904	33	712	51	517	69	unknown
16	674	34	964	52	undigested	70	1655
17	571	35	undigested	53	unknown	71	711
18	undigested	36	244	54	undigested	72	717

Table 8.2. Clone number and insert size of 72 random cDNA clones

Figure 8.5. Determination of the cDNA insert fragment sizes of clone numbers 1-12, 37-48 and 61-72 using restriction digestion by *NotI* and *SalI*

Lanes 1-12, 37-48 and 61-72 represent the *NotI*/*SalI* digested plasmid DNA and insert fragments. The results for all 72 clones are summarised in Table 9.2. Marker DNA fragment sizes are shown on the left, and the calculated linear plasmid DNA fragment size on the right of the gel image.

M = marker λ/*Hin*dIII and ΦX174RF/*Hae*III


M 1 2 3 4 5 6 7 8 9 10 11 12 M
9,416
6,557
4,361
2,322
2,027
1,353
1,078
872
603
310
7,126
M 37 38 39 40 41 42 43 44 45 46 47 48 M
9,416
6,557
4,361
2,322
2,027
1,353
1,078
872
603
310
7,126
M 61 62 63 64 65 66 67 68 69 70 71 72 M
9,416
4,361
2,322
2,027
1,353
1,078
872
603
310
7,126

insert is present or that it is too small to visualise. Clones without inserts might arise from the transformation of bacteria with vector DNA that had escaped digestion with *Sal*I and *Not*I, but the manufacturer provided no data with respect to completeness of the digestion. Transformations of the vector DNA, ligated and un-ligated, into *E. coli* might have resolved this, but these were not carried out. The size fractionation of the cDNA prior to ligation to the vector would have adequately separated the digested adaptors from the cDNA but would not necessarily have excluded small cDNAs. It is most likely that the clones without visible inserts either really do lack inserts or that they are too small to visualise.

Seven clones (numbers 14, 18, 25, 31, 35, 48 and 52) were not digested properly using *Not*I and *Sal*I restriction enzymes. Although other restriction enzymes were used to digest these clones, digestions were also not successful (probably because of the method of plasmid isolation). The size of the insert fragment in sample number 42 appears to be about 6500 bp which is unlikely given the arguments presented above, however no other restriction digests were carried out on this clone to help clarify the issue. Therefore samples with no visible insert (numbers 53, 58, 59 and 69), samples with poor restriction digestion results (numbers 14, 18, 25, 31, 35, 48 and 52) and the sample with the large insert size (number 42) were not take into account for further calculations. For the 59 remaining colonies, the mean insert size was 1028 bp with a standard deviation of 543 bp (sample standard deviation) and 538 bp (population standard deviation).

8.2.3 Construction of the mouse spp24 hybrid proteins using Gateway Cloning Technology

As described in Section 9.1.3.1, it was decided that the Invitrogen Gateway Cloning Technology would be used as, after the initial construction of a donor vector containing the sequence of interest, it is a simple and rapid process to transfer this sequence to a wide variety of different destination vectors. In this study, using two mouse spp24 hybrid proteins, (the entire mature peptide and the non-cystatin-like domain) a mouse pooled-tissues cDNA yeast two-hybrid library was screened.

8.2.3.1 Entry vectors construction

The first part of the procedure was to construct Gateway entry clones for both the entire mature peptide (pENTR201-mouse*Spp2*) and for the non-cystatin-like domain (pENTR201-mouse non-cystatin*Spp2*). To construct these two entry clones, primers modified with *att*B

sequences, were used to PCR amplify the mouse *Spp2* cDNA using I.M.A.G.E. clone 335916 as the template DNA. The modification of the primers with *att*B sites enables the resulting PCR products to be transferred into the donor vector by a site-specific recombination reaction (the BP reaction). The location of these primers (two forward and one reverse) with respect to the mouse *Spp2* cDNA is illustrated in Figure 8.6 and the structures of the primers are shown in Figure 8.7.

Two high-fidelity PCR amplifications using *Pfu* polymerase were carried out as described in Section 2.7.1, with the following conditions: (96°C 30s, 64°C 30s, 72°C 30s) × 30. *Pfu* DNA polymerase, derived from the hyperthermophilic archae *Pyrococcus furiosus*, has been shown to exhibit superior thermostability and proofreading properties compared to other thermostable polymerase. Unlike Taq DNA polymerase, highly thermostable *Pfu* DNA polymerase possesses 3' to 5' exonuclease proofreading activity that enables the polymerase to correct nucleotide-misincorporation errors. After PCR amplification, an aliquot of the product was electrophoresed on an agarose gel to assess the yield and purity of the PCR products. The remaining PCR products (from two pooled reactions) were then purified using the method described in Section 2.7.4.1.

To carry out BP reactions (for construction of entry clones, Figure 8.2 and 8.3), the purified PCR product was incubated with the donor vector pDONR201 to allow site-specific recombination to occur as described in section 2.22.3. Following the BP reactions, the products (pENTR201-mouse*Spp2* and pENTR201-mouse non-cystatin*Spp2*) were transformed into *E. coli* DH10B using the electroporation method (Section 2.12.6.2) and plated on selective LUA plates containing kanamycin. The resulting entry clone DNAs were each isolated from two independent bacterial colonies grown in LB broth using the method described in Section 2.12.8.3. Restriction digestion and agarose gel electrophoresis was performed to confirm that the recombination reactions had taken place successfully. For confirmation of the sequences of each of the two inserts, each construct was sequenced in both the forward and reverse directions using the following vector-specific primers:
Forward primer (proximal to *att*L1 site) 5′ TCGCGTTAACGCTAGCATGGATCTC 3′
Reverse primer (proximal to *att*L2 site) 5′ GTAACATCAGAGATTTTGAGACAC 3′
Sequencing of inserts was carried out as described in Section 2.10.1. A single sequence error was noted in one of the two DNAs sequenced for the entire spp24 construct. Although the sequence difference (an A→G transition) did not result in an amino acid change, this clone

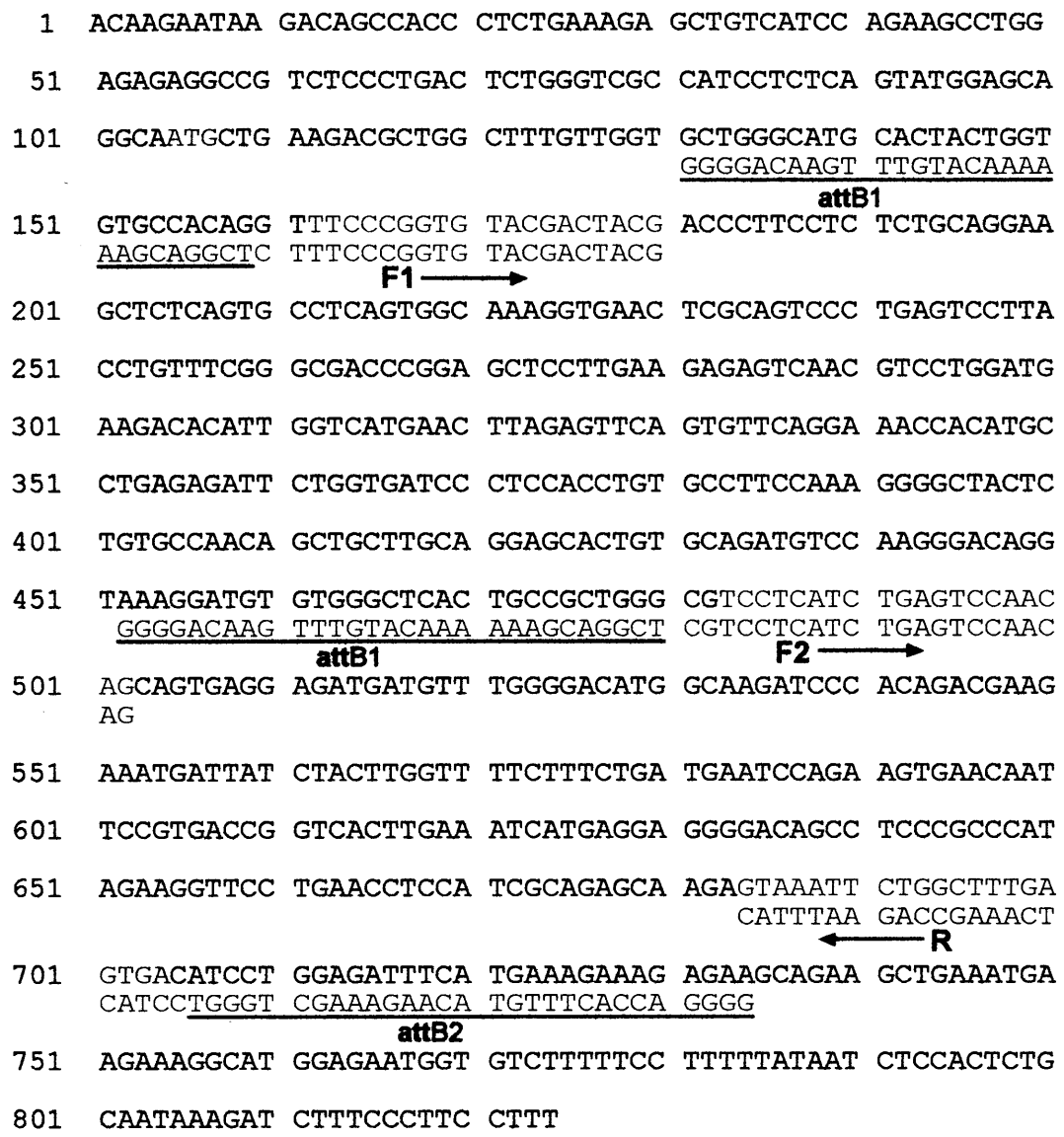


Figure 8.6. The position of the primers used to generate the spp24 yast two-hybrid bait construct in the mouse *Spp2* cDNA consensus sequence described in Chapter 4

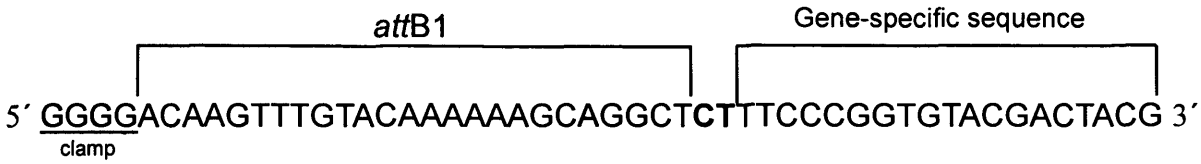
This figure illustrates the sequence of the mouse *Spp2* cDNA consensus sequence (Chapter 4). The 'ATG' and 'TGA' start and stop codons are shown in red. The primer sequences and the annealing sequences in the cDNA are illustrated in blue.
The *att*B1 and *att*B2 sites are underlined in the forward (F1 and F2) and reverse primers respectively. A four-base G clamp was added at the end of the *att*B1 and *att*B2 sites.
To establish the correct reading frame in the entire spp24 and the non-cystatin-like domain constructs of spp24, two extra bases were added between the *att*B1 sites and the annealed sequences in F1 (CT) and F2 (CG). For this purpose, one extra base (C) was also added between the *att*B2 sites and the annealed sequences of reverse primer.

Two forward primers are shown (F1 and F2) and one reverse primer is shown (R).
The following combinations of primers were used to generate the two spp24 constructs:

Entire spp24 (mature peptide): F1 and R

Non-cystatin-like domain of spp24 : F2 and R

Forward primer 1 (F1)



Forward primer 2 (F2)

Reverse primer (R)

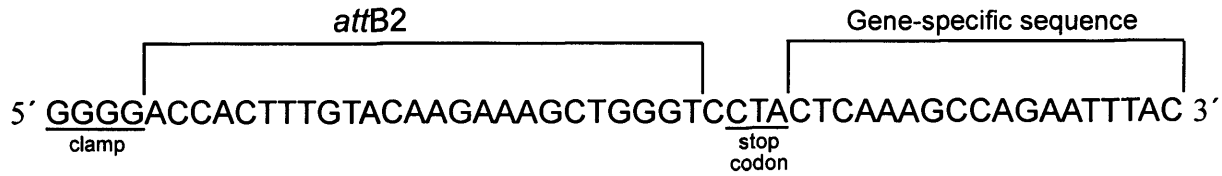


Figure 8.7. *att*B modified PCR primers

This figure shows the *att*B-modified PCR primers used to generate the mouse spp24 constructs (the entire spp24 and the non-cystatin-like domain of spp24 protein). The *att*B sequence enables DNA amplified with these primers to be incorporated into a suitable vector by site-specific recombination.

The *att*B1 and *att*B2 sites and gene-specific sequences of each individual primer are indicated for the forward (F1 and F2) and reverse primers respectively. A four-base G clamp were added at the end of *att*B1 and *att*B2 sites and is underlined.
To establish the correct reading frame in the entire spp24 and non-cystatin-like domain construct of spp24, two extra bases were added between the *att*B1 sites and annealed the sequences in F1 (CT) and F2 (CG). For this purpose one extra base (C) was also added between the *att*B2 sites and the annealed sequences of the reverse primer. The extra base pairs are shown in bold letters.

was not used in the two-hybrid screen. Both clones for the non-cystatin region were without sequence errors.

8.2.3.2 Yeast two-hybrid expression vector construction

After the construction of entry clones (pENTR201-mouse*Spp2* and pENTR201-mouse non-cystatin*Spp2*) the final expression vectors (pDEST32-mouse*Spp2* and pDEST32-mouse non-cystatin*Spp2*) for screening the cDNA library using the yeast two-hybrid method were generated. LR recombination reactions were carried out, incubating each of the entry clones with the destination vector (pDEST32, Figure 8.3), allowing the *Spp2* insert to be transferred into pDEST32 by site-specific recombination (Section 2.22.3 and Figure 8.2). Following the LR reactions, the products (pDEST32-mouse*Spp2* and pDEST32-mouse non-cystatin*Spp2*) were transformed into *E. coli* DH10B using the electroporation method (Section 2.12.6.2) and plated on selective LUA plates containing gentamicin. The resulting destination plasmid DNAs were each isolated from two independent bacterial colonies grown in LB broth using the method described in Section 2.12.8.3. Restriction digestion and agarose gel electrophoresis was performed to confirm that the recombination reactions had taken place successfully.

8.2.4 Yeast two-hybrid interactions

To screen the constructed cDNA library in the yeast two-hybrid system, the two bait plasmids (pDEST32-mouse*Spp2* or W*Spp2* and pDEST32-mouse non-cystatin*Spp2* or N*Spp2*) were first transformed independently into AH109 (a yeast cell strain) using small-scale transformations as described in Sections 2.13.14 and 2.13.15.

The transformed cells were first assessed to verify that the DNA-binding domain plasmids (the baits) did not autonomously activate the reporter genes and that there were no toxic effects introduced by the baits. Next, these cells were transformed with DNA prepared from the cDNA library (the prey constructs) in large-scale transformation (Section 2.13.15) to carry out the yeast two-hybrid screen.

8.2.4.1 Small-scale transformation of two baits (W*Spp2* and N*Spp2*) and positive and negative control vectors

Using small-scale transformation the two constructed baits (W*Spp2* and N*Spp2*) were transformed individually into the AH109 strain. The MATCHMAKER Two-Hybrid System 3 (Clontech) provides positive and negative control vectors. Vector information is provided in the MATCHMAKER *GAL4* Two-Hybrid System 3 & Libraries User Manual. pCL1 encodes the full-length, wild type *GAL4* protein and provides a positive control for all the reporter genes. pGBKT7-53 and pGADT7-T encode fusions between the *GAL4* DNA-BD and AD and murine p53 and SV40 large T-antigen, respectively. p53 and large T-antigen interact in a yeast two-hybrid assay and provide a positive interaction control for this system. pGBKT7-Lam encodes a fusion of the DNA-BD with human lamin C and provides a control for a fortuitous interaction between an unrelated protein and either the pGADT7-T control or the AD/cDNA library plasmids used in this study. Lamin C neither forms complexes nor interacts with most other proteins and provides a negative interaction control for this system. In all steps of the study, these control vectors were used to compare the transformation efficiency and the strength of interactions.

Following the small-scale transformation of the two bait constructs into the yeast cells, they were plated on SD/−Leu medium and transformant cells containing the DNA-binding domain bait (W*Spp2* and N*Spp2*) were selected according the method described in Section 2.13.6. For the next step of transformation, competent cells were prepared from these transformants using the method described in Section 2.13.4.

8.2.4.2 Screening the mouse pooled-tissues cDNA library

To isolate enough plasmid for the yeast-two hybrid analyses, further amplification of the library was carried as described in Section 8.2.1 using 2×10^6 bacterial colonies grown on ten plates. Maxi-scale plasmid DNA isolation was then carried out using a Qiagen kit as described in section 2.12.8.3.

To screen the library, the plasmids containing the activation domain and cDNA inserts were then transformed into the AH109 cells already containing the DNA-binding domain plasmid baits (W*Spp2* or N*Spp2*) using the large-scale transformation (Section 2.13.5). Following the large-scale transformation of the cDNA library into the yeast cells, the transformants were divided into three equal fractions and plated on SD/−Leu/−Trp (low stringency),

SD/−Leu/−Trp/−His (medium stringency) and SD/−Leu/−Trp/−/His/−Ade (high stringency) media as described in Section 2.13.6.

Table 8.3 shows the number of mouse cDNA library colonies screened with each spp24 bait. Also shown is the number of potential positive clones that were identified on the basis of vigorous growth (relative to presumed background colonies) on SD/−Leu/−Trp/−His (medium stringency) and SD/−Leu/−Trp/−/His/−Ade (high stringency) media.

To confirm the results, all 430 potential positive clones were first replica plated onto medium-stringency media, then re-plated on the high stringency media. All 25 positive colonies of the W*Spp2* interaction in the high stringency medium and 74 of the 111 positive colonies of the W*Spp2* interaction in the medium stringency grew on high stringency media. These 74 were the same 74 that had originally grown strongly under medium selection (Table 8.3). Twenty two of the 294 positive colonies of the N*Spp2* interaction under medium stringency selection also grew on high stringency media. All 22 grew after 4-5 days incubation, and although the colonies were smaller than 2 mm in diameter, they were still readily distinguishable from background colonies. In summary, 121 colonies were able to grow under high stringency selection, but only those from the W*Spp2* interaction yielded any strongly-growing colonies.

Thirty five of the original 430 potential positive clones were selected at random and their plasmids were isolated using the method described in Section 2.13.7. To determine the size of inserts in these positive clones, the plasmids were digested with *Not*I and *Sal*I restriction enzymes using the method described in Section 2.3. Among these 35 clones, only seven different insert sizes were identified (Figure 8.8), ranging from 1210 bp to 2050 bp. Representative clones for each of the seven insert sizes were sequenced in the forward and reverse directions as described in Section 2.10.1 using the following vector-specific primers:

pEXP-AD502 For 5′ TATAACGCGTTTGGAATCACT 3′
pEXP-AD502 Rev 5′ GTAAATTTCTGGCAAGGTAGAC 3′

Using the BLAST search program, the non-redundant nucleotide database of NCBI was searched and the genes encoding these potential interacting proteins were identified (Figure 8.8). These proteins include the granulin precursor also known as acrogranulin/epithelin (*Grn*), tissue specific transplantation antigen P35B (*Tsta3*), keratin complex 1, acidic, gene 18 (*Krt1-18*), keratin complex 1, acidic, gene 13 (*Krt1-13*), vimentin (*Vim*), protein phosphatase

Spp24 hybrid baits used in screen	Number of mouse cDNA library colonies screened	Number of potential positives identified	
		Medium stringency	High stringency
Whole protein (W*Spp2*)	2×10^6	111 colonies (74 strong, 37 weak)	25 colonies (17 strong, 8 weak)
Non-cystatin-like protein (N*Spp2*)	2×10^6	294 colonies (211 strong, 83 weak)	0

Table 8.3. The number of mouse cDNA library colonies screened with each spp24 hybrid bait and the number of potential positives identified

The mouse pooled-tissues cDNA library was screened with two different spp24 hybrid baits, whole and non-cystatin-like as described in Section 9.2.4.2. The number of colonies from the mouse pooled-tissues cDNA library that were screened with each hybrid bait are shown and the number of potential positive identified.

The strong positives are defined as colonies that grew on SD/−Leu/−Trp/−/His/−Ade (high stringency) or SD/−Leu/−Trp/−/His (medium stringency) media after 2 days incubation and whose diameter was ≥ 2 mm.

The weak positives yielded colonies of the same size when plated on the same media, but took 4-5 days to grow.

Positive controls under the same conditions yielded colonies after 2 days.

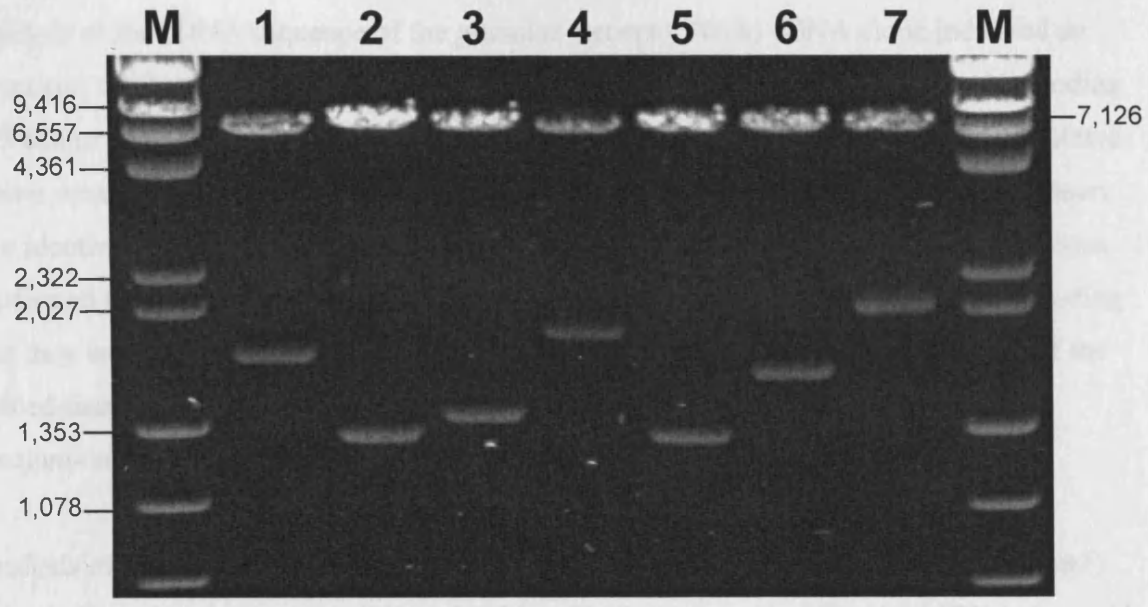


Figure 8.8. Restriction digestion and agarose gel electrophoresis of the seven different cDNA inserts encoding the potential positive interacting proteins with spp24 in mouse

Thirty five of the potential positives identified from the cDNA library screening were initially selected and their plasmids were isolated using the method described in Section 2.13.7. To determine the size of inserts in the positive clones, all isolated plasmids were digested with *Not*I and *Sal*I restriction enzymes using the method described in Section 2.3. Seven inserts with different sizes were identified as follows:

lane 1: acrogranulin/epithelin (*Grn*), 1680 bp
lane 2: tissue specific transplantation antigen P35B (*Tstap35b*), 1250 bp
lane 3: keratin complex 1, acidic, gene 18 (*Krt1-18*), 1360 bp
lane 4: vimentin (*Vim*), 1800 bp
lane 5: similar to protein phosphatase 1, regulatory (inhibitor) subunit 12C, 1210 bp
lane 6: keratin complex 1, acidic, gene 13 (*Krt1-13*), 1600 bp
lane 7: alpha actinin 4 (*Actn4*), 2050 bp

M = marker λ/*Hin*dIII and ΦX174RF/*Hae*III

1, regulatory (inhibitor) subunit 12C (no gene symbol yet assigned) and alpha-actinin-4 (*Actn4*).

Analysis of the cDNA sequence of the granulin precursor (*Grn*) cDNA clone indicated an insert size of about 1680 bp. The insert contains an incomplete open reading frame encoding 443 amino acids out of 589, and the entire 3′ untranslated region. Eighteen of the 35 isolated clones were judged to be granulin precursor cDNA on the basis that they each had an insert size identical to the original clone that was sequenced, and sequencing of a further 6 clones confirmed this. Interestingly, the 7 sequenced clones are identical to one another, suggesting that they were all amplified from a single initial clone subsequent to the construction of the pooled-tissues cDNA library. All 18 clones interacted with both W*Spp2* and N*Spp2* on medium- and high-stringency media.

Analysis of the cDNA sequence of the tissue specific transplantation antigen P35B (*Tsta3*) clone indicated an insert size of about 1250 bp. The insert contains the entire 321-amino acid open reading frame of the gene (cDNA), as well the entire 3′ untranslated region and 12 bp of the 5′ untranslated region. Two of the 35 isolated clones were tissue specific transplantation antigen P35B and it interacted with W*Spp2* on high-stringency media.

Analysis of the cDNA sequence of the keratin complex 1, acidic, gene 18 (*Krt1-18*) clone indicated an insert size of about 1360 bp. The insert contains the entire 426-amino acid open reading frame of the gene (cDNA), as well as the entire 3′ untranslated region and 29 bp of the 5′ untranslated region. Six of the 35 isolated clones were keratin complex 1, acidic, gene 18 (*Krt1-18*) and it interacted with WS*pp2* on high-stringency media and with N*Spp2* on medium-stringency media.

Analysis of the cDNA sequence of the keratin complex 1, acidic, gene 13 (*Krt1-13*) clone indicated an insert size of about 1600 bp. The insert contains an incomplete open reading frame of the gene (cDNA) encoding 430 amino acids out of 438, as well as and the entire 3′ untranslated region. Two of the 35 isolated clones were keratin complex 1, acidic, gene 13 (*Krt1-13*) and it interacted with N*Spp2* on medium-stringency media.

Analysis of the cDNA sequence of the vimentin (*Vim*) clone indicated an insert size of about 1800 bp. The insert contains the entire 467-amino acid open reading frame of the gene

(cDNA), as well as the entire 3′ and 5′ untranslated regions. Two of the 35 isolated clones were vimentin and it interacted with W*Spp2* on medium-stringency media.

Analysis of the cDNA sequence of the alpha-actinin-4 (*Actn4*) clone indicated an insert size of about 2150 bp. The insert contains an incomplete open reading frame of the gene (cDNA) encoding 573 amino acids out of 913, as well as the entire 3′ untranslated region. Only one of the 35 isolated clones was alpha-actinin-4 and it interacted with N*Spp2* on medium-stringency media.

Analysis of the cDNA sequence of the protein phosphatase 1, regulatory (inhibitor) subunit 12C isolated clone indicated an insert size of about 1210 bp. The insert contains an incomplete reading frame of the gene (cDNA) encoding 230 amino acids out of 782, as well as the the entire 3′ untranslated region. Four of the 35 isolated clones were protein phosphatase 1, regulatory (inhibitor) subunit 12C and interacted with W*Spp2* on medium-stringency media. Table 8.4 summarises the conditions under which these 35 positive clones were initially isolated.

Following analysis of the all cDNA sequences, an attempt was then made to reconstruct some of the positive interactions (including granulin precursor, vimentin, alpha-actinin-4 and the tissue specific transplantation antigen P35B) by sequentially transforming AH109 with the spp24 hybrid bait plasmids (W*Spp2* and N*Spp2*) followed individually by the four potential positive plasmids using the protocol described in Section 2.13.5. None of these four proteins had any toxicity effects or activated the reporter genes on their own. All four were demonstrated to interact, suggesting none of the initial interactions were false positives. Because of time limitations, this confirmation study was not carried out for the three other potential positive interactions. Table 8.5 summarises the results of analysis of reconstruction of these four positive interactions.

In summary, in the screening of the mouse pooled-tissues cDNA library using the yeast two-hybrid method, 430 potential positive clones were identified. About 8% (35) of these clones were analysed and seven proteins potentially interacting with the spp24 hybrid baits were identified. Four positive interactions were reconstructed and reconfirmed by sequentially transforming AH109 with the spp24 hybrid baits.

Interacting protein	Number of isolated clones	W*Spp2*		N*Spp2*	
		HS	MS	HS	MS
granulin precursor	18	12		5	1
tissue specific transplantation antigen P35B	2	2			
keratin complex 1, acidic, gene 18	6	2			4
keratin complex 1, acidic, gene 13	2				2
vimentin	2		2		
protein phosphatase 1, regulatory (inhibitor) subunit 12C	4		4		
alpha actinin 4	1				1
Total	35	16	6	5	8

Table 8.4. The record of the conditions in which the 35 randomly selected potential positive clones were isolated

Thirty five of the original 430 potential positive clones were selected at random and their plasmids were isolated using the method described in Section 2.13.7. Among these 35 clones, only seven different insert sizes were identified, ranging from 1210 bp to 2050 bp. Representative clones for each of the seven insert sizes were sequenced in the forward and reverse directions.
This table lists the conditions under which the 35 positive clones were first isolated.

HS = High stringency media (SD/−Leu−Trp−His−Ade)
MS = Medium stringency media (SD/−Leu−Trp−His)

Interacting prey protein	W*Spp2*		N*Spp2*	
	HS	MS	HS	MS
granulin precursor	+	+	+	+
tissue specific transplantation antigen P35B	+	+	−	−
vimentin	−	+	−	+
alpha actinin 4	+	+	−	+

Table 8.5. The results of analysis of four reconstructed positive interactions

Following analysis of all the cDNA sequences, an attempt was made to reconstruct some of the positive interactions (including granulin precursor, the tissue specific transplantation antigen P35B, vimentin and alpha actinin 4 and) by sequentially transforming AH109 with the spp24 hybrid bait plasmids (W*Spp2* and N*Spp2*) followed individually by the four potential positive plasmids using the protocol described in Section 2.13.5. All four were demonstrated to interact, suggesting that none of the initial interactions were false positives. Plus (+) indicates positive interaction and minus (−) indicates lack of an interaction.

HS = High stringency media (SD/−Leu−Trp−His−Ade)
MS = Medium stringency media (SD/−Leu−Trp−His)

8.3 Discussion

Screening of the mouse pooled-tissues cDNA library using the yeast two-hybrid method and the identification of several positive interactions suggests that the use of the spp24 protein as a bait is compatible with the yeast two-hybrid system. Using this system, 430 potential positive clones were identified of which about 8% (35) were analysed and seven proteins potentially interacting with the spp24 hybrid baits were identified. Four positive interactions including granulin precursor, vimentin, alpha-actinin-4 and the tissue specific transplantation antigen P35B were reconstructed and reconfirmed by sequentially transforming AH109 with the spp24 hybrid baits and these four preys. All four were demonstrated to interact, suggesting that none of the initial interactions were false positives.

Eighteen of the 35 isolated clones were judged to be granulin precursor cDNA on the basis that they each had an insert size identical to the original clone that was sequenced, and sequencing of a further six clones confirmed this. All 18 granulin clones interacted with both W*Spp2* and N*Spp2* on medium- and high-stringency media (Tables 8.4 and 8.5). This suggests that the interaction between spp24 and the granulin precursor involves regions in both the cystatin and noncystatin domains of spp24. Due to the fact that the granulin precursor comprises more than 50% of all the positive interactions, it will be discussed in more detail.

The granulins/epithelins are a family of cysteine-rich polypeptides that were first identified as peptides of approximately 6 kDa that were isolated from human neutrophils and rat bone marrow and some of which have growth modulatory activity (Bateman *et al.*, 1990). The granulins are exteracellular proteins and widespread presence of their mRNA in different cells including the haematopoietic system and in epithelia indicates the important function of this protein in these tissues. The granulins have no exact homology with any other regulatory proteins, but have several similarities with the epidermal growth factor/transforming growth factor alpha (EGF/TGFα) family, particularly with respect to their structural organisation. Several granulins (in human including granulins A, B, C, D and F) have been isolated. Predicted protein sequence of the precursor for the human, rat, mouse and guinea pig have been reported. All known mammalian granulins are generated from a common precursor, progranulin, which contains a secretory signal peptide and seven and a half repeats of the 12-cysteine granulin motif. The human gene is located on chromosome 17 and the mouse gene on chromosome 11 (Bhandari *et al.*, 1992; Bateman and Bennett, 1998). Two granulins, granulinA/epithelin1 and granulinB/epithelin2, are biologically active (the actions of the other

granulins, if any, are unknown) and have multiple and sometimes opposite effects on the growth of different cells. GranulinA promotes the growth of keratinocytes and, in combination with TGF-β, supports the growth of normal rat kidney fibroblasts. GranulinB inhibits keratinocyte growth and antagonises the proliferative actions of granulinA. Both peptides inhibit the growth of other epithelial cells (Zhiheng and Bateman, 1999).

Several studies have been reported of examples in which the granulin precursor may be important as a biologically active entity in its own right (Bateman and Bennett, 1998). Progranulin is constitutively expressed in a number of epithelia, particularly in the skin, GI tract and reproductive system. Other epithelia express the gene less strongly. Progranulin is expressed in immune cells and specific neurons in the brain. Little expression has been detected in muscle cells, connective tissues and endothelium. Different studies suggested that progranulin is a multifunctional gene with important roles in epithelial homeostasis, reproductive, immunological and neuronal function (Daniel *et al.*, 2000). The homologue of progranulin was also found in the sperm of mouse and guinea pig (acroganin) and shown to have growth modulation properties (Baba *et al.*, 1993). For epithelial cells such as SW-13 (adrenal carcinoma cells) and MDCK (non-transformed renal epithelia) the rate of proliferation, clonogenicity in semisolid agar and mitosis in monolayer culture is proportional to the level of expression of the progranulin gene, whereas a decrease in the gene expression impairs growth of these cells (Zhiheng and Bateman, 1999). Non-epithelial cells also respond mitogenically to progranulin. PC cells (highly tumourigenic murine mesenchymal teratoma cells) require progranulin (called PCDGF in this context) to sustain their growth and antisense targeting of PCDGF dramatically decreases their tumourigenicity. The PC cell line is a highly tumourigenic, insulin-independent, teratoma-derived cell line isolated from the non-tumourigenic, insulin-dependent 1246 cell line (Zhang and Serrero, 1998). Progranulin is the only growth factor able to overcome the cell cycle block that occurs in murine fibroblasts following deletion of a functional IGF-1 (insulin-like growth factor 1) receptor (Xu *et al.*, 1998).

Using northern and western blot analyses, it has been shown that PCDGF mRNA and protein expression is low in non-tumourigenic cells and is increased in human breast carcinoma cell lines exhibiting a positive correlation with their tumourigenicity. Inhibition of PCDGF expression by antisense PCDGF cDNA transfection inhibits tumourigenicity of the human breast carcinoma cell line MDA-MB-486. In another study it was also shown that the expression of PCDGF is stimulated by estradiol in human breast cancer cells and this

stimulation is completely inhibited by treatment with the anti-estrogen 4-hydroxytamoxifen (Lu and Serrero, 1999; 2000).

Li *et al.* (2000) identified the expression of progranulin mRNA in glial tumours of the brain, with lower levels in spleen, kidney and testis, whereas expression was not detected in non-tumour brain tissues. The differential pattern of expression, tissue distribution and the implication of this protein in growth regulation suggested a potentially important role in the pathogenesis and malignant progression of primary brain tumours.

In another study, comparison of cDNA libraries prepared from low malignant potential (LMP) and invasive ovarian cancers indicated that progranulin (PCDGF) was present only in the invasive ovarian cancer libraries and was absent in the LMP libraries (Jones *et al.*, 2003). Comparison of mRNA levels by semi-quantitative RT-PCR in another study, showed that progranulin is over expressed in gastric cancer tissues relative to normal tissues (Line *et al.*, 2002).

As discussed in Chapter 7, the TNFα and TNFα-receptor superfamily are involved in the regulation of cell proliferation, cellular activation and differentiation, including control of death or survival of cells by apoptosis or necrosis. Jurisic *et al.* (2000) showed that treatment of PC cells with TNFα significantly increases the apoptotic and necrotic forms of cell death in these cells.

It can be concluded from these studies that progranulin is a potent growth factor and, if over expressed, it can act as an oncogene in different epithelial and non-epithelial cell types. The results of the yeast two-hybrid screen indicate that spp24 interacts strongly with progranulin, but the effects of spp24 in this interaction and downstream events need further investigation (it will be discussed briefly in future work in Chapter 9). It is probably notable that the gene encoding spp24 in chicken is expressed widely in the developing chicken embryo in response to growth hormone (Harvey *et al.*, 2001). It has also been shown in this study that the expression of human spp24 is down regulated by TNFα. Therefore, if spp24 has a role in growth and development of different cells or tissues (as in chicken), it is possible that it induces its growth effect in combination and interaction with progranulin. It is also possible that the effect of TNFα in increasing the apoptotic and necrotic forms of cell death in PC cells is achieved through down regulating the expression of the gene encoding the spp24 protein in this cell line. As was discussed in Chapter 1, because of the similarity between spp24 and

other members of the cystatin superfamily, one of the speculated functions for the protein is the prevention of formation and metastasis of some tumours. Therefore it is possible that the interaction of spp24 with progranulin inhibits the growth activity of granulin precursor and consequently prevents the formation of different neoplasms.

He *et al.* (2003) showed that in murine transcutaneous puncture wounds, progranulin mRNA is expressed in the infiltrating inflammatory cells and is highly induced in dermal fibroblasts and endothelial cells following injury. When progranulin is applied to a cutaneous wound, it increases the accumulation of neutrophils, macrophages, blood vessels and fibroblasts at the site of injury. It acts directly on isolated dermal fibroblasts and endothelial cells to increase and promote migration, division and formation of the capillary-like tubule structures. Therefore it was concluded that progranulin is a wound-related growth factor.

Calcium has an established function in the normal homeostasis of mammalian skin and acts as a mediator in keratinocyte proliferation and differentiation. Different experimental studies suggest that control of calcium metabolism is essential in wound healing and it is an extracellular regulator and intracellular modulator of cell proliferation in the mammalian epidermis. Sequential events in wound healing including the homeostatic, inflammatory, proliferative, remodelling and normalisation phases, require a conductive environment within the wound bed and a balance of different ions, of which at least calcium, zinc, magnesium, copper, iron, sodium and potassium are significant. In experimental wound-skin there is a substantial overlap between the inflammatory and the onset of the proliferative activity phases. Proliferation of keratinocytes in normal regenerating epidermis depends on suitable calcium signalling. In the early stage of the proliferative phase, extracellular calcium concentration of <0.1 mM are held to be conductive for signal proliferation and migration along the pathway leading to re-epithelialisation of the injury site. Excessive calcium at this phase can inhibit migration and proliferation of keratinocytes (Lansdown, 2002).

Due to the fact that progranulin is a potent growth factor and is mediator of wound healing, it should have a significant role in the proliferation phase of wound healing. Therefore if spp24 has a role in growth and development of different cells or tissues (as in chicken), it is possible that it induces its growth effect in combination and interaction with progranulin in the proliferation phase of wound healing. As was discussed in Section 7.3, spp24 probably, like fetuin and MGP, plays a crucial role in calcium metabolism and in inhibiting calcification. Therefore it is possible that spp24 induces the migration and proliferation of keratinocytes and

also inhibits calcification (as a side effect) at the site of wound healing, by decreasing the calcium concentration at the proliferative phase.

In an attempt to identify any other proteins interacting with progranulin, it was found that protein dlk (delta-like) which is a member of the epidermal growth factor (EGF)-like homeotic family interacts with progranulin. It is known that dlk participates in several differentiation processes including adipogenesis, haematopoiesis and adrenal gland differentiation. In addition, lack of dlk expression correlates with increased malignancies of undifferentiated tumours (Baladrón *et al.*, 2002). Although no sequence similarity can be identified between dlk and spp24, the interaction between dlk and progranulin might suggest the role of other proteins that interact with progranulin (like spp24) in growth and differentiation.

Ten of the 35 isolated clones were shown to be keratin complex 1, acidic, gene 18 (6 clones), keratin complex 1, acidic, gene 13 (2 clones) and vimentin (2 clones) (Tables 8.4 and 8.5). All these proteins are components of intermediate filaments (IF) which are one of the three types of the cytoskeletal elements. The two other elements are thin filaments (actin) and microtubules. These three elements frequently work together to enhance structural integrity, cell shape, and cell and organelle motility. Intermediate filaments range from 8–10 nm (intermediate in size compared to thin filaments and microtubules). They are prominent in cells that withstand mechanical stress and are the most insoluble part of the cell. There are five different types of intermediate filaments:

- Types I and II including acidic keratin and basic keratin respectively. These types of intermediate filaments are produced by different types of epithelial cells (bladder, skin, *etc.*). Keratins also have different subtypes that are unique to different epithelial cells (bladder, skin, *etc.*) or even different subsets of one cell type (like basal epidermal cells). This is very useful in the identification and detection of the origin of cells in tumours, especially cells that have metastasised. The nomenclature of the different subtypes of keratins can be retrieved from: (http://www.gene.ucl.ac.uk/nomenclature/genefamily/krt.html)
- Type III intermediate filaments are distributed in a number of cell types, and comprise vimentin (in fibroblasts, endothelial cells and leukocytes), desmin, glial fibrillary acidic factor and peripherin.
- Type IV filaments include neurofilament H, M, L and internexin.
- Type V filaments include lamins.

Expression of keratins in an epithelial tissue is always as pairs of type I and type II, because two polypeptides from each subfamily can form a heterotetramer which acts as a building block for the keratin cytoskeleton (Stewart, 1993). In a screening of the draft sequence of the human genome 65 IF genes were detected, placing IF among the 100 largest gene families in humans. All functional keratin genes map to the two known keratin clusters on chromosomes 12 (type II plus keratin 18) and 17 (type I) whereas other IF genes are not clustered. Of the 208 keratin-like DNA sequences in the human genome, only 49 reflect true keratin genes, while the others are inactive gene fragments and processed pseudogenes. Nearly 90% of these inactivated genes are related specifically to the genes of keratins 8 and 18 (Hesse *et al.*, 2001). Hepatocyte IFs are made solely of keratin 8 (type II) and 18 (type I). In animal models, the absence of keratin 8 and 18 or the presence of mutant keratins causes or promotes liver disease like cirrhosis. It was also suggested that these two keratins provide resistance to Fas-mediated apoptosis (Sorom *et al.*, 2002). There is also convincing evidence, on the basis of clinical and experimental studies, that keratins 8 and 18 in liver exert a non-skeletal protective function against different toxins such as alcohol and oxidative stress. Disturbances of the keratin system may significantly contribute to cell damage (Denk *et al.*, 2001). It has also been shown that the keratin 18 gene is expressed at a normal level in cells of non-tumourigenic clones derived from the SW613-S human colon carcinoma cell line, but is over expressed in cells of tumourigenic clones. A high level of expression was also demonstrated in the cells from 10 of 15 other human colon carcinoma cell lines (Fossar *et al.*, 1999).

K13 (keratin 13) with K4 (its type II partner), forms the major keratin network of most internal stratified epithelia and is not generally expressed in epidermis except in penile foreskin, anal epidermis and regenerating epidermis in wound healing. Amongst the factors reported to influence K13 expression, the retinoids and calcium are the most investigated. High calcium concentrations induce stratification and K13 expression in immortalised and transformed keratinocytes (Waseem *et al.*, 1998). In DNBA/TPA (7, 12-dimethylbenz[α]anthracine/ 12-*O*-tetradecanoyl-phorbol-13acetate) -induced epidermal tumours, K13 expression is induced and is considered to be an early marker for carcinoma progression. Keratinocytes derived from these tumours have characteristics similar to those from internal stratified epithelia (Sutter *et al.*, 1991), but in squamous cell carcinoma of the oral cavity and the female genital tract, its expression is severely reduced (Malecha and Miettinen, 1991). In another study it was suggested that the K13 gene might play an important role in laryngeal carcinogenesis, acting as a new tumour suppressor gene, and may be relevant to laryngeal squamous cell carcinoma diagnosis and prognosis (He *et al*, 2002).

Vimentin purified from native sources usually appears in several isoforms that are thought to be the result of differential phosphorylation. Vimentin is not expressed in certain sub-clones of cells derived from the human SW-13 cell line and the vimentin gene has been knocked-out from the mouse genome. Both systems support the fact that vimentin is not vital for growth, division or development. However, the cells that lack vimentin have a defect in lipoprotein/cholesterol metabolism (Sarria *et al.*, 1992; Colucci-Guyon *et al.*, 1994). Over-expression of the oncogene HER2/*neu* occurs in up to 30% of breast cancers and is correlated with reduced survival. In a study, it was shown that vimentin is down-regulated and K18 is up-regulated in HER2/*neu*-positive breast cancer cell lines (Wilson *et al.*, 2002).

Spp24 is a secretory protein, and all intermediate filaments are intracellular, therefore at the moment it is impossible to comment about the importance of the interaction between spp24 and these three types of IF components. Although spp24 is a secretory protein, it may return to cells alone or in combination with other proteins such as fetuin-mineral complex and take part in a regulatory process (such as regulation of calcium concentration and metabolism). Therefore it is possible that it interacts with intracellular proteins after re-entering cells. A common aspect about all IFs that interact with spp24 is that in some way they have a role in tumourigenesis, especially in epithelial cells. Therefore, as discussed earlier in this section with respect to progranulin, spp24 may inhibit or promote the expression of these genes in tumourigenesis.

Alpha-actinin-4 is an actin-binding protein of about 100 kDa that is associated with cell motility, endocytosis, and cancer invasion (Honda *et al.*, 1998). The alpha actinin family comprises two non-muscle isoforms (alpha-actinin-1 and -4) and two skeletal muscle isoforms (alpha-actinin-2 and -3). While alpha-actinin-4 is almost ubiquitously expressed, particularly high concentrations are found in glomeruli. On the sub-cellular level it is associated with actin stress fibres, but in certain cells it also localises to the nucleus (Critchley and Flood, 1999; Honda *et al.*, 1998).

Mutations in the human alpha-actinin-4 gene cause an autosomal dominant form of familial focal segmental glomerulosclerosis. Regulation of the actin cytoskeleton of glomerular podocytes may be altered in this group of patients (Kaplan *et al.*, 2000). A point mutation in the alpha-actinin-4 gene was found to generate an antigen peptide that is recognised by cytotoxic T lymphocytes in a human lung large cell carcinoma. It is possible that these cytotoxic T lymphocytes (recognising a truly tumour-specific antigen) play a role in the

clinical progression of the lung cancer in the patient studied (Echchakir *et al.*, 2001). It was also shown that cytoplasmic protein regulates the actin cytoskeleton and increases cellular motility and its inactivation, by transfer to the nucleus, abolishes the metastatic potential of human cancers (Honda *et al.*, 1998). In a study, it was shown that elevated intracellular Ca^{2+} inhibits the Na^{+}/H^{+} exchanger activity (NHE3). This inhibition is carried out through oligomerisation and endocytosis of NHE3, which occurs via formation of an NHE3-E3KARP-alpha-actinin-4 complex (E3KARP is a PDZ domain-containing protein). In addition, this Ca^{2+}-dependent inhibition requires Ca^{2+}-dependent association between alpha-actinin-4 and E3KARP (Kim, *et al.*, 2002).

Due to the fact that spp24 is expressed at a high level in kidney, it might play a role (in combination with alpha-actinin-4) in regulation of the actin cytoskeleton of glomerular podocytes. As was discussed earlier, spp24 is a protein that is recognised as a component of FMC and is able to decrease the Ca^{2+} concentration (Price *et al.*, 2003). Therefore it is possible that it plays a role in the regulation of Na^{+}/H^{+} exchanger activity (NHE3) through altering the Ca^{2+} concentration or changing the property of NHE3-E3KARP-alpha-actinin-4 complex via interaction with the alpha-actinin-4 protein.

Protein phosphatases (PPs) are enzymes that remove phosphate groups that have been attached to amino acid residues of proteins by protein kinases. These phosphates are important in signal transduction because they regulate the proteins to which they are attached. To reverse the regulatory effect, the phosphates have to be removed and this occurs by hydrolysis or is mediated by protein phosphatases. These enzymes, like protein kinases, are classified into two groups, serine/threonine and tyrosine protein phosphatases, based on the amino acid residues that they de-phosphorylate. Using a variety of naturally occurring toxins, (some them are potent tumour-promoting agents such as okadic acid and microcystin) that inhibit cellular protein phosphatase activity it has been shown that serine/threonine protein phosphatases play a critical role in the control of cell function (Barford, 1995). Serine/threonine-specific protein phosphatases are classified as PP1, PP2A, PP2B, PP2C, PP4 and PP5. PP1 appears to be the most prevalent serine/threonine protein phosphatase in mammalian cells (Siegal *et al.*, 1999). In rat, cDNA cloning has shown the existence of at least four isoforms for PP1, termed PP1α, PP1δ, PP1γ1 and PP1γ2 (Sasaki *et al.*, 1990). PP1s are involved in different cell functions, including glycogen metabolism, muscle contraction, protein synthesis, and intracellular transport (Morimoto *et al.*, 2002).

Using substrate screening in various tissues, it was determined that there are two major substrates for MRCKα-kinase including MBS130 and MBS85 (myosin binding subunit 85). The human gene encoding myosin binding subunit 85 has been mapped to 19q13.3-q13.4.and is expressed in breast tumour, prostate tumour, thymus, uterus, amygdale, choroid, melanotic melanoma cell lines, normal prostatic epithelial cells, lymphocytes and leiomyosarcoma cell lines.

Myosin binding subunit 85 (also known as protein phosphatase 1, regulatory (inhibitor) subunit 12C (*PPP1R12C*) and as DKFZP434D0412) contains three distinctive domains and is a component specifically of PP1δ. These domains include an N-terminal ankyrin repeat, an α-helical C terminus which contains a leucine-zipper motif, and a centrally located motif containing a threonine residue which is a substrate for the MRCKα-kinase. When this threonine is not phosphorylated, PP1δ adopts a conformation that allows it to interact with its natural substrate, myosin light chain 2 (MLC2), which results in actin-myosin disassembly. When the threonine is phosphorylated, the protein adopts a conformation that has a higher affinity for PP1δ and prevents PP1δ binding to MLC2 thus suppressing its dephosphorylation, resulting in actin-myosin assembly. Therefore, MBS85 has an important role in the regulation of actin-myosin assembly and the cytoskeleton (Tan *et al.*, 2001).

Red blood cells contain a specific NADP(H)-binding protein (red cell NADP(H)-binding protein FX, tissue specific transplantation antigen 3). Lenzerini *et al.* (1981) concluded that there is a common genetic polymorphism at the locus or loci that control the level of the FX protein. The conclusion was based on the finding of large variation in FX levels in unrelated persons and a very strong family effect. Tonetti *et al.* (1996) used PCR to obtain the complete sequence of the human FX cDNA and it was found that the FX gene encodes a 320-amino acid polypeptide with a predicted molecular mass of 35.7 kDa. Database analysis demonstrated that human FX protein is 92.6% identical to the mouse tumour rejection antigen P35B and has a lower level of homology to three bacterial proteins, one of which may be involved in sugar metabolism. They also determined the role of FX protein in GDP-D-mannose metabolism. GDP-fucose is synthesized from GDP-mannose in a three-step pathway. The first step is catalyzed by GDP-mannose 4,6-dehydratase, or GMD. The second and third steps are an epimerisation and a reduction reaction, respectively. It was found that FX could catalyse both the epimerase and the reductase reactions, converting GDP-4-keto-6-D-deoxymannose to GDP-L-fucose. GDP-L-fucose is the substrate of several fucosyltransferases involved in the expression of many glycoconjugates, including blood

group ABH antigens and developmental adhesion antigens. Using purified GMD and FX, Sullivan *et al.* (1998) showed that the two proteins alone are sufficient to convert GDP-mannose to GDP-fucose in vitro. They suggested that mutations in one of these 2 enzymes may cause leukocyte adhesion deficiency, type II (LAD2). Using fluorescence *in situ* hybridisation, this gene was localised to 8q24.3 in humans.

Smith *et al.* (2002) induced a null mutation in the *Fx* locus in mice. Mice with this mutation exhibited complete deficiency of cellular fucosylation and variable intrauterine fatality. Liveborn Fx-null mice exhibited postnatal failure to thrive that was suppressed with a fucose-supplemented diet. Homozygous adults suffered from an extreme neutrophilia, myeloproliferation, and absence of leukocyte selectin ligand expression reminiscent of LAD2.

At present, using the available information, no speculation can be made about the possible role of spp24 with respect to its interaction with protein phosphatase 1, regulatory (inhibitor) 12C subunit and tissue transplantation antigen P35B.

Finally, it should be emphasised that this is just the beginning of a much larger study and all 430 potential positive clones should be analysed and the interacting proteins identified. All positive protein interactions should be retested using different strategies such as using the alternative reporter genes including *MEL1* (for α-galactosidase activity) and *LacZ* (for β-galactosidase activity), cotransformation of baits and preys, yeast mating, confirmation of protein interactions via an *in vitro* co-immunoprecipitation, *in vivo* analysis and transferring the library insert from the AD vector to the DNA-BD vector and *vice versa*, and then repeat the two-hybrid assay. Following confirmation of the positive interactions, the reconstruction of interactions must be carried out using full-length cDNA clones in cases where the interacting proteins, such as progranulin, in the initial screening were not full-length.

Chapter 9
Concluding Remarks and Future Work

The primary aim of this study was to identify the functions of the spp24 protein but unfortunately it seems to have raised more question than it has answered. However, the results of the different experiments have suggested appropriate directions for future functional studies.

9.1 Concluding remarks

There does not seem to be any doubt now that the spp24 protein is a new member of the cystatin superfamily. This was originally suggested by Hu *et al.*, (1995) and part of the work presented in this study (Chapter 4) support this. The exact place of the spp24 protein within the cystatin superfamily is still not clear. Cornwall and Hsia (2003) grouped spp24 with kininogen, cathelins and fetuins in cystatin family 3 based on its structure which includes a C-terminal non-cystatin domain. In this study it was shown that the exon/intron structure of the entire gene encoding the spp24 protein is identical between human and mouse and the size and location of intron 1 is conserved between nine species (Chapter 4). The structure of the spp24 gene in all these nine species is unlike any other member of the cystatin superfamily in that it has an additional, very small, intron splitting what would otherwise be the first exon. This difference makes it a unique, and probably new, branch of this superfamily.

Although spp24 has been classified as a member of the cystatin superfamily, it is unlikely to show a typical cystatin function by inhibiting thiol proteases from the papain superfamily. Spp24 does not contain the residues identified as being crucial to the cystatin-papain interaction (Bennett, 2002) although its ability to inhibit papain should be tested practically before it is finally ruled out.

Expression studies of spp24 indicated that there are some differences in expression patterns between species, particularly with respect to kidney. It is unlikely that there is a difference in spatial expression of the protein, but more likely a difference in the temporal expression. It is clear that expression of spp24 is highly tissue specific and likely to be tightly regulated and this suggests that the protein has a specific function.

In all studied species, spp24 is predominantly expressed in liver and in this respect is similar to that of fetuin which is expressed mainly in liver. This suggests that spp24 is a plasma protein or that it has a role in processes that take place in the liver. Structural and expression similarities (Chapters 1 and 7) between spp24 and fetuin may suggest functional similarities as well. Many functions have been suggested for the human fetuin though some of them have not been proved. However, it is accepted that fetuin can inhibit calcium salt precipitation in serum and modulate apatite formation in bone and therefore has an important role in regulation of osteogenesis (Schinke *et al.*, 1996). A recent study has also shown that spp24 plays a role in the inhibition of calcium precipitation through its interaction with fetuin-mineral complex (FMC) (Price *et al.*, 2003). It also seems that spp24, like fetuin, has a role in the development of a particular populations of cells in the cerebellum at a specific stage of development (Chapter 7).

Hu *et al.* (1995) reported isolation of the bovine mature spp24 from cortical bone. In another study, the distribution of spp24 was analysed by real-time RT-PCR and found to have strong expression in rat liver, kidney and bone tissues (Beckman, personal communication to Dalgleish). In this study it was shown that spp24 is not expressed in adult mouse bone, but it is expressed in the neonate mouse bone tissue (Chapter 7). These results may suggest another species difference in expression of the spp24 protein in bone tissues, which requires more investigation. As microarray technology progresses, and more precise expression databases emerge, it may be possible to clarify the expression pattern of spp24 in bone.

Screening of the mouse pooled-tissues cDNA library using the yeast two-hybrid method and the identification of several positive interactions suggests that the use of the spp24 protein as a bait is compatible with the yeast two-hybrid system and post-translational modifications are not likely to be essential to the interaction of spp24 with other proteins. The yeast two-hybrid studies carried out in Chapter 8 suggest different functions for the spp24 protein. After confirmation of these interactions, any speculated functions must then be tested using different strategies.

The work presented in this project has led to the following speculation about the functions of the spp24 proteins:

- The protein may have a cystatin-like function although this is thought unlikely due to the lack of essential residues thought to be important for an interaction with papain (Chapter 1; Bennett, 2002).
- The protein may circulate as a plasma protein like fetuin and therefore plays a role in different processes such as inflammation, coagulation, the immune response, mineralisation, calcium metabolism and inhibition of calcification in different parts of the body such as bone, lactating mammary glands and wound-healing areas (Chapters 1, 7 and 8).
- The protein may have an antimicrobial function like cathelins. The non-cystatin-like domain may have this activity, as it is the most divergent domain of protein between species (Chapters 1, 6 and 7).
- The spp24 protein may play a role in cancer and progression of metastasis. Therefore it is possible that the interaction of spp24 with progranulin inhibits the growth activity of granulin precursor and consequently prevents the formation of different neoplasia (Chapter 8).
- If spp24 has a role in growth and development of different cells or tissues (as in chicken), it is possible that it induces its growth effect on its own or in combination and interaction with progranulin in different normal and pathological processes such as wound healing, tumour formation and metastasis (Chapters 1, 7 and 8).
- Due to the fact that spp24 interacts with different components of the cytoskeletons (intermediate filaments and myosin binding subunit 85), it might play a role in the formation and/or regulation of the cytoskeleton (Chapter 8).
- The serine-rich region is thought likely to be a regulatory domain for the function/functions of the spp24 protein with the extent of phosphorylation determining the functional state of the protein (Chapters 1 and 3).

9.2 Future work

It seems that for more functional studies of spp24 to be carried out, a purified protein is mandatory. Expression of the protein in insect or mammalian cells is likely to result in the correct structure and acquisition of the necessary post-translational modifications. Probably the signal peptide should be included to help achieve these aims. To investigate the functions of the individual domains (the cystatin-like domain and the non-cystatin-like domain) and the protein as a whole, three different constructs should be made. Extensive optimisation of expression and purification may be required. Spp24 expression and purification in insect cells and *in vitro* translation and purification of the protein are currently being performed by Tim Francis (Department of Genetics, University of Leicester). Once a purified protein has been obtained, different experiments can be designed and performed to assess the possible functions of the protein. The effect of the protein on other proteins such as papain and cathepsin can be tested. It would be relatively simple to examine the protein for antimicrobial effects by adding preparations of spp24 to agar plates containing various bacterial cultures. It is obvious that all three spp24 constructs should be used, so if a function is found, the responsible domain of the protein can be identified. Following expression and purification of protein, it can be used to raise and synthesise antibodies in experimental animals such as rabbits. Such antibodies can be used to identify the presence of protein in any tissue of interest or in biological fluids such as milk.

As an alternative to the expression of spp24 in insect or mammalian cells, it may be possible to purify it from milk. Spp24 is expressed in the lactating mammary gland of mouse, raising this possibility that it might be present in cow's milk and could be purified using methods such as HPLC. Any attempt to purify the protein probably needs collaboration with other research groups with expertise in this field.

The generation of a *Spp2* knock-out mouse may be another way of highlighting the potential roles of spp24. This strategy may provide a visible phenotype that would help in understanding the possible function of spp24. The complete structure of the mouse *Spp2* gene is now known (Chapter 4), enabling us to create a knock-out or transgenic mouse. It should be considered that there is always a possibility that the knock-out mouse will be perfectly normal and have no obvious defects. However, this knock-out mouse model would still be useful to test for other possible abnormalities. For example the mouse model can be used to investigate

susceptibility of the knock-out mouse to TB or other infections compared to normal mice. Generation of a knock-out mouse is currently is being performed by Tim Francis (Department of Genetics, University of Leicester) in collaboration with GlaxoSmithKline.

If it is speculated from spp24 functional studies that the protein could play a role in a multifactorial disease, then an association study could be carried out. There are three previously characterised RFLPs (probably resulting from SNPs), three characterised short tandem repeats (Bennett and Dalgleish, unpublished) and two new variants (Chapter 3) in the human *SPP2* gene that could be used in any future association study.

Screening of the mouse pooled-tissues cDNA library using the yeast two-hybrid method and the identification of several positive interactions suggests that the use of the spp24 protein as a bait is compatible with the yeast two-hybrid system. Following confirmation of the positive interactions, the reconstruction of each interaction must be carried out using full-length cDNA clones in cases where the interacting proteins, such as granulin, in the initial screening were not full-length. This would afford an opportunity to examine the effects of specific variations in the sequence of spp24 on these true positive interactions. This experiment might also provide information on whether or not the serine to phenylalanine substitution (Chapter 3) alters the interaction of spp24 with any other protein. The results of the yeast two-hybrid screen indicate that spp24 interacts strongly with progranulin, but the effects of spp24 in this interaction and any downstream events needs further investigation. As discussed in Section 9.3 the PC cell line is a highly tumourigenic, insulin-independent, teratoma-derived cell line isolated from the non-tumourigenic, insulin-dependent 1246 cell line. It was also shown that progranulin (granulin precursor) is responsible for this tumourigenicity (Zhang and Serrero, 1998). To clarify the role of spp24 in its interaction with progranulin, the following experiments seem to be worthwhile and helpful:

- Comparing the expression of spp24 between PC and 1246 cell lines.
- If spp24 protein is available, test for the effects of spp24 by adding preparations of the protein to the culture media of PC and 1246 cell lines.
- If spp24 protein is not available, transfection of both cell line with *Spp2* antisense RNA, to investigate the effect of spp24 on tumourigenicity.

Once the function of spp24 has been identified it will be helpful to investigate the regulation of its expression. Study of the promoter region should be performed to determine the precise region required for transcription and to determine whether or not the gene is regulated by

growth hormone or other growth factors. Due to the fact that spp24 is a phosphoprotein, it is highly probable that the action of the protein is regulated by the extent of phosphorylation in the serine-rich region of the protein. Therefore, in collaboration with other groups this subject can be further investigated.

Once a function has been identified for spp24, mutation analysis studies could be carried out to identify the exact regions of the protein that are essential for that function. If the three dimensional structure of spp24 is available by that time (using X-ray crystallography), it could be a great help to target residues for these studies.

Bibliography

Abrahamson M, Ritonja A, Brown MA, Grubb A, Machleidt W, Barrett AJ (1987) Identification of the probable inhibitory reactive sites of the cysteine proteinase inhibitors human cystatin C and chicken cystatin. J. Biol. Chem. 262: 9688-9694

Abrahamson M, Salvesen G, Barrett AJ, Grubb A (1986) Isolation of six proteinase inhibitors from human urine. Their physicochemical and enzyme kinetic properties and concentrations in biological fluids. J. Biol. Chem. 261: 11282-11289

Agarwal SK, Cogburn LA, Burnside J (1995) Comparison of gene expression in normal and growth hormone receptor-deficient dwarf chickens reveals a novel growth hormone regulated gene. Biochem. Biophys. Res. Commun. 206: 153-160

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

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25: 3389-3402

Alvarez-Fernandez M, Barrett AJ, Gerhartz B, Dando PM, Ni J, Abrahamson M (1999) Inhibition of mammalian legumain by some cystatins is due to a novel second reactive site. J. Biol. Chem. 274: 19195-19203

Auerswald EA, Genenger G, Assfalg-Machleidt W, Engh RA, Fritz H (1992) Recombinant chicken egg white cystatins variants of the QLVSG region. Eur. J. Biochem. 209: 837-845

Baba T, Hoff HB, Nemoto H, Lee H, Orth J, Arai Y, Gerton GL (1993) Acroganin, an acrosomal cysteine-rich glycoprotein, is the precursor of the growth-modulating peptides, granulins, and epithelins, and is expressed in somatic as well as germ cells. Mol. Reprod. Dev. 34: 233-243

Baggio R, Shi Y, Wu Y, Abeles RH (1996) From good substrates to good inhibitors: design of inhibitors for serine and thiol proteases. Biochemistry 35: 3351-3353

Baladrón V, Ruiz-Hidalgo MJ, Bonvini E, Gubina E, Notario V, Laborda J (2002) The EGF-like homeotic protein dlk affects cell growth and interacts with growth-modulating molecules in the yeast two-hybrid system. Biochem. Biophys. Res. Commun. 291: 193-204

Barford D (1995) Protein phosphatases. Curr. Opin. Struct. Biol. 5: 728-734

Barrett AJ (1981) Cystatin, the egg white inhibitor of cysteine proteinases. Meth. Enzymol. 80: 771-778

Barrett AJ (1987) The cystatins: a new class of peptidase inhibitors. Trends Biochem. Sci. 12: 193-196

Barrett AJ, Davies ME, Grubb A (1984) The place of human γ-trace (cystatin C) amongst the cysteine proteinase inhibitors. Biochem. Biophys. Res. Commun. 120: 631-636

Barrett AJ, Fritz H, Grubb A, Isemura S, Jarvinen M, Katunuma N, Machleidt W, Müller-Esteryl W, Sasaki M, Turk V (1986a) Nomenclature and classification of proteins homologous with the cystein-proteinase inhibitor chicken cystatin. Biochem. J. 236: 312-312

Barrett AJ, Rawlings N, Davies M, Machleidt W, Salvesen G, Turk V (1986b) Cysteine proteinase inhibitors of the cystatin superfamily. In: Barrett AJ, Salvesen G (eds) Proteinase Inhibitors. Elsevier, Amesterdam, pp 515-569

Bateman A, Belcourt D, Bennett HPJ, Lazure C, Solomon S (1990) Granulins, a novel class of peptide from leukocytes. Biochem. Biophys. Res. Commun. 173: 1161-1168

Bateman A, Bennett HPJ (1998) Granulins: the structure and function of an emerging family of growth factors. J. Endocrinol. 158: 145-151

Bellamy R (1998) Genetic susceptibility to tuberculosis in human populations. Thorax 53: 588-593

Bennett B, Markel PD, Beeson MA, Gordon LG, Johnson TE (1994) Mapping quantitative trait loci for ethanol-induced anesthesia in LS × SS recombinant inbred and F2 mice: methodology and results. Alcohol Alcohol Suppl 2: 79-86

Bennett CS (2002) Characterisation of secreted phosphoprotein 24, Ph.D. Thesis, University of Leicester

Benson G (1999) Tandem repeats finder: A program to analyze DNA sequences. Nucl. Acids Res. 27: 573-580

Bhandari V, Palfree RGE, Bateman A (1992) Isolation and sequence of the granulin precursor cDNA from human bone marrow reveals tandem cysteine-rich granulin domains. Proc. Natl. Acad. Sci. USA 89: 1715-1719

Björk I, Ylinnenjärvi K (1992) Different roles of the two disulphide bonds of the cysteine protease inhibitor chicken cystatin, for the conformation of the active protein. Biochemistry 31: 8597-8602

Björk L, Akesson P, Bohus M, Trojnar J, Abrahamson M, Olafsson I, Grubb A (1989) Bacterial growth blocked by a synthetic peptide based on the structure of a human proteinase inhibitor. Nature 337: 387-388

Blom N, Gammeltoft S, Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J. Mol. Biol. 294: 1351-1362

Bode W, Engh R, Musil D, Thiele U, Huber R, Karshikov A, Brzin J, Kos J, Turk V (1998) The 2.0 Å X-ray crystal structure of chicken egg white cyststin and its possible mode of interaction with cysteine proteinases. EMBO J. 7: 2593-2599

Bollengier F (1987) Cystatin C, alias post-gama-globulin: a marker for multiple sclerosis. J. Clin. Chem. Biochem. 25: 589-593

Boonacker E, Van Noorden CJ (2001) Enzyme cytochemical techniques for metabolic mapping in living cells, with special reference to proteolysis. J. Histochem. Cytochem. 49: 1473-1486

Borodovsky M, McIninch JD (1993) GeneMark: Parallel gene recognition for both DNA strands. Comp. Chem. 17: 123-133

Bossard MJ, Tomaszek TA, Thompson SK, Amegadzie BY, Hanning CR, Jones C, Kurdyla JT, McNulty DE, Drake FH, Gowen M, Levy MA (1996) Proteolytic activity of human osteoclast cathepsin K. Expression, purification, activation, and substrate identification. J. Biol. Chem. 271: 12517-12524

Boulard O, Fluteau G, Eloy L, Damotte D, Bedossa P, Garchon HJ (2002) Genetic analysis of autoimmune sialadenitis in nonobese diabetic mice: A mojor susceptibilty region on chromosome 1. J. Immunol. 168: 4192-4201

Brayer J, Lowry J, Cha C, Robinson CP, Yamachica S, Peck AB, Humphreys-Beher MG (2000) Alleles from chromosome 1 and 3 of NOD mice combine to influence Sjögren syndrome-like autoimmune exocrinopathy. J. Rheumatol. 27: 1896-1904

Brömme D, Okamoto K (1995) Human cathepsin O2, a novel cysteine protease highly expressed in osteoclastomas and ovary, molecular cloning sequencing and tissue distribution. Biol. Chem. Hoppe-Seyler 376: 379-384

Brown T, Mackay K (1999) Northern hybridisation of RNA fractionated by agarose-formaldehyde gel electerophoresis, In: Fredrick M, Brent R, Kingston R, Moore D, Seidman J, Smith J, Struhl K (eds) Current Protocols In Molecular Biology. Wiley & Sons Inc., Boston, pp 4.9.2-4.9.4

Brown W, Saunders N, Møllgård K, Dziegielewska K (1992) Fetuin an old friend revisited. Bioessays 14: 749-755

Brown WM, Dziegielewska KM (1997) Friends and relations of the cystatin superfamily–new members and their evolution. Protein Sci. 6: 5-12

Burge C (1997) Identification of genes in human genomic DNA. PhD Thesis, Stanford University

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

Burset M, Guigo R (1996) Evaluation of gene structure prediction programs. Genomics 34: 353-367

Cardon LR, Bell JI (2001) Association study design for complex diseases. Nat. Rev. Genet. 2: 91-99

Carninci P, Hayashizaki Y (1999) High-efficiency full-length cDNA cloning. Meth. Enzymol. 303: 19-44

Ceccherini I, Romeo G, Lawrence S, Breuning MH, Harris PC, Himmelbauer H, Frischauf AM, Sutherland GR, Germino GG (1992) Construction of a map of chromosome 16 by using radiation hybrids. Proc. Natl. Acad. Sci. USA 89: 104-108

Chang HN, Lee SY (1994) Generation of bacteriophage lambda lysogens by electroporation. Biotechniques 16: 206-208

Chauhan SS, Golstein LJ, M.M G (1991) Expression of cathepsin L in human tumours. Cancer Res. 51: 1478-1481

Chen JM, Dando PM, Rawlings ND, Brown MA, Young NE, Stevens RA, Hewitt E, Watts C, Barrett AJ (1997) Cloning, isolation, and characterization of mammalian legumain, an asparaginyl endopeptidase. J. Biol. Chem. 272: 8090-8098

Chien C-T, Barten PL, Sternglanz R, Fields S (1991) The two-hybrid system: A method to identify and clone genes for proteins that interact with a protein of interest. Proc. Natl. Acad. Sci. USA 88: 9578-9582

Choi SJ, Reddy SV, Delvin RD, Menna C, Chung H, Boyce BF, Roodman GD (1999) Identification of human asparginyl endopeptidase (legumain) an inhibitor of osteoclast formation and bone resorption. J. Biol. Chem. 274: 27747-27753

Clausen J (1961) Proteins in normal cerebrospinal fluid not found in serum. Proc. Soc. Exp. Biol. Med. 107: 170-172

Collella R, Chambers AF, Denhardt DT (1993) Anticarcinogenic activities of naturally occuring cysteine proteinase inhibitors. In: Troll W, Kennedy AR (eds) Protease Inhibitors as Cancer Chemopreventive Agents. Plenum Press, New York, pp 199-216

Colucci-Guyon E, Portier M-M, Dunia I, Paulin D, Pournin S, Babinet C (1994) Mice lacking vimentin develop and produce without an obvious phenotype. Cell 79: 679-694

Comstock GW (1978) Tuberculosis in twins: a re-analysis of the Prophit study. Am. Rev. Respir. Dis. 117: 621-624

Cornwall GA, Hsia N (2003) A new subgroup of the family 2 cystatins. Mol. Cell Endocrinol. 200: 1-8

Cotton RGH (1997) Slowly but surely towards better scanning for mutations. Trends Genet. 13: 43-46

Cox DR (1992) Radiation hybrid mapping. Cytogenet. Cell Genet. 59: 80-81

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

Cox RA, Downs M, Neimes RE (1988) Immunogenetic analysis of human tuberculosis and association with HLA-DR types. J. Infect. Dis. 158: 1302-1308

Critchley DR, Flood G (1999) Alpha actinins. In: Kreis T, Vale R (eds) Guidebook to the Cytoskeletal and Motor Proteins. Oxford University Press, Oxford, pp 24-27

Dalgleish R (1987) Southern blotting. In: Boulnois GJ (ed) Gene Cloning and Analysis. Blackwell Scientific Publications, Oxford, pp 45-60

Daniel R, He Z, Carmichael KP, Halper J, Bateman A (2000) Cellular localization of gene expression for progranulin. J. Histochem. Cytochem. 48: 999-1009

Davies ME, Barrett AJ (1984) Immunolocalization of human cystatins in neutrophils and lymphocytes. Histochemistry 80: 373-377

Delaissé J-M, Eeckhout Y, Vaes G (1980) Inhibition of bone resorption in culture by inhibitors of thiol proteinases. Biochem. J. 192: 365-368

Delaissé J-M, Eeckhout Y, Vaes G (1984) *In vivo* and *in vitro* evidence for the involvement of cysteine proteinases in bone resorption. Biochem. Biophys. Res. Commun. 125: 441-447

Delwek R, Funabiki T, Kreider BL, morishita K, Ihle JN (1993) Four of the seven zinc fingers of the Evi-1 myeloid-transforming gene are required for sequence-specific binding to GA(C/T)AAGA(T/C)AAGATAA. Mol. Cell Biol. 13: 4291-4300

Denk H, Stumptner C, Fuchsbichler A, Zatloukal K (2001) Alcoholic and nonalcoholic steatohepatitis. Histopathologic and pathogenetic consideration. Pathologie 22: 388-398

Devoto M, Shimoya k, Caminis J, Ott J, Tenenhouse A, Whyte MP, Sereda L, Hall S, Considine E, William CJ, Tromp G, Kuivaniemi H, Ala-Kokko L, Prockop DJ, Spotila LD (1998) First-stage autosomal genome screen in extended pedigrees suggests genes predisposing to low bone mineral density on chromosomes 1p, 2p and 4q. Eur. J. Hum. Genet. 6: 151-157

Drake FH, Dodds RA, James IE, Connor JR, Debouck C, Richardson S, Lee-Rykaczewski E, Coleman L, Rieman D, Barthlow R, Hastings G, Gowen M (1996) Cathepsin K, but not cathepsins B, L, or S, is abundantly expressed in human osteoclasts. J. Biol. Chem. 271: 12511-12516

Dziegielewska KM, Møllgård K, Reynolds ML, Saunders NR (1987) A fetuin-related glycoprotein (α_2HS) in human embryonic and fetal development. Cell Tissue Res. 248: 33-41

Echchakir H, Mami-Chouaib F, Vergnon I, Baurain JF, Karanikas V, Chouaib S, Coulie PG (2001) A point mutation in the alpha-actinin-4 gene generates an antigenic peptide recognised by autologous cytotoxic T lymphocytes on a human lung carcinoma. Cancer Res. 61: 4078-4083

Eddy SR, Durbin R (1994) RNA sequence analysis using covariance models. Nucl. Acids Res. 22: 2079-2088

Ehringer MA, Thompson J, Conroy O, Xu Y, Yang F, Canniff J, Beeson M, Gordon L, Bennett B, Johnson TE, Sikela JM (2001) High-throughput identification of gene coding variants within alcohol-related QTLs. Mamm. Genome 12: 657-663

Eichbaum QG, Iyer R, Raveh DP, Mathieu C, Ezekowitz RA (1994) Restriction of interferon gamma responsiveness and basal expression of the myeloid human $F_c$gamma R1B gene is mediated by a functional PU.1 site and a transcription initiator consensus. J. Exp. Med. 179: 1985-1996

Einstein J, Uberbacher E, Guan X, Mural R, Mann R (1991) GAP- A computer program for gene assembly. ORNL/TM-11924

Einstein JR, Mural RJ, Guan X, Uberbacher EC (1992) Computer-based construction of gene models using the GRAIL gene assembly program. ORNL/TM-12174

Elzanowski A, Barxer WC, Hunt LT, Seibel-Rose E (1988) Cystatin domains in alpha-2-HS-glycoprotein and fetuin. FEBS Lett. 227: 167-170

Everts V, Aronson DC, Beertsen W (1985) Phagocytosis of bone collagen by osteoclasts in two cases of pycnodysostosis. Calcif. Tissue Int. 37: 25-31

Falconer D, Gauld I, Roberts R (1978) Cell numbers and cell sizes in organs of mice selected for large and small body size. Genet. Res. 31: 387-401

Falconer D, Mackay T (1996) Introduction to Quantitative Genetics, Fourth edn. Longman, Harlow, Essex

Falconer DS (1953) Selection for large and small size in mice. J. Genet. 51: 470-501

Farquharson M, Harvie R, McNicol AM (1990) Detection of messenger RNA using a digoxigenin end labelled oligodeoxynucleotide probe. Clin. Pathol. 43: 424-428

Fichant GA, Burks C (1991) Identifying potential tRNA genes in genomic DNA sequences. J. Mol. Biol. 220: 659-671

Fields S, Sternglanz R (1994) The two-hybrid system: an assay for protein-protein interactions. Trends Genet. 10: 286-292

Fisher JE, Caulfield MP, Sato M, Quartuccio HA, Gould RJ, Garsky VM, Rodan GA, Rosenblatt M (1993) Inhibition of osteoclastic bone resorption *in vivo* by echistatin, an "arginyl-glycil-aspartyl" (RGD)-containing protein. Endocrinology 132: 1411-1413

Fong D, Man-Yin Chan M, Hsieh WT (1991) Gene mapping of human cathepsin and cystatins. Biomed. Biochem. Acta 50: 595-598

Fossar N, Chaouche M, Prochasson P, Rousset M, Brison O (1999) Deregulated expression of the keratin 18 gene in human colon carcinoma cells. Somat. Cell Mol. Genet. 25: 223-235

Fossum K, Whitaker JR (1968) Ficin and papain inhibitor from chicken egg white. Arch. Biochem. Biophys. 125: 367-375

Fox RI (1996) Clinical features, pathogenesis, and treatment of Sjögren syndrome. Curr. Opin. Rheumatol. 8: 438-445

Freeman WM, Walker SJ, Vrana KE (1999) Quantitative RT-PCR: pitfalls and potential. Biotechniques 26: 112-125

Freije J, Pendas A, Velasco G, Roca A, Abrahamson M, Lopez-Otin C (1993) Localization of the human cystatin D gene (CST5) to human chromosome 20p11.21 by *in situ* hybridization. Cytogenet. Cell Genet. 62: 29-31

Freije JP, Abrahamson M, Olafsson I, Velasco G, Grubb A, Lopez-Otin C (1991) Structure and expression of the gene encoding cystatin , a novel human cysteine proteinase inhibitor. J. Biol. Chem. 266: 20538-20543

Futuyma DJ (1986) Evolutionary Biology, Second edn. Sinauer Associates, Inc., Sunderland, Massachusetts, MA

Gall JG, Pardue ML (1969) Formation and detection of RNA-DNA hybrid molecules in cytological preparations. Proc. Natl. Acad. Sci. USA 63: 378-383

Ganguly A (2002) An update on conformation sensitive gel electrophoresis. Hum. Mutat. 19: 334-342

Garchon HJ, Bedossa P, Eloy L, Bach JF (1991) Identification and mapping to chromosome 1 of a susceptibility locus for preinsulinitis in non-obese diabetic mice. Nature 353: 260-262

Garnero P, Borel O, Byrjalsen I, Ferreras M, Drake FH, McQueney MS, Foged NT, Delmas PD, Delaisse JM (1998) The collagenolytic activity of cathepsin K is unique among mammalian proteinases. J. Biol. Chem. 273: 32347-32352

Gelb B, Spencer E, Obad S, Edelson G, Faure S, Weissenbach J, Desnick R (1996a) Pycnodysostosis: refined linkage and radiation hybrid analyses reduce the critical region to 2 cM at 1q21 and map two candidate genes. Hum. Genet. 98: 141-144

Gelb BD, Shi GP, Chapman HA, Desnick RJ (1996b) Pycnodysostosis, a lysosomal disease caused by cathepsin K deficiency. Science 273: 1236-1238

Gelb B, Willner J, Dunn T, Kardon N, Verloes A, Poncin J, Desnick R (1998) Paternal uniparental disomy for chromosome 1 revealed by molecular analysis of a patient with pycnodysostosis. Am. J. Hum. Genet. 62: 848-854

Gennari L, Becherini L, Falchetti A, Masi L, Massart F, Brandi ML (2002) Genetics of osteoporosis: role of steroid hormone receptor gene polymorphisms. J. Steroid Biochem. Mol. Biol. 81: 1-24.

Gietz D, St. Jean A, Woods RA, Schiestl RH (1992) Improved method for high efficiency transformation of intact yeast cells. Nucl. Acids Res. 20: 1425

Giguere Y, Rousseau F (2000) The genetics of osteoporosis: complexities and difficulties. Clin. Genet. 57: 161-169

Gilbert L, He X, Farmer P, Boden S, Kozlowski M, Rubin J, Nanes MS (2000) Inhibition of osteoblast differentiation by tumor necrosis factor-α. Endocrinology 141: 3956-3964

Gingrich JC, Boehrer DM, Garnes JA, Johnson W, Wong BS, Bergmann A, Eveleth CG, Langlois RG, Carrano AV (1996) Construction and characterization of human chromosome 2-specific cosmid, fosmid and PAC clone libraries. Genomics 32: 65-74

Goss SJ, Harris H (1975) New method for mapping genes in human chromosomes. Nature 255: 245-250

Grant SFA, Reid DM, Blake G, Herd R, Fugelman I, Ralston SH (1996) Reduced bone density and osteoporosis associated with a polymorphic Sp1 binding site in the collagen type Iα1 gene. Nat. Genet. 14: 203-205

Grisel JE (2000) Quantitative trait locus analysis. Alcohol Res. Health 24: 169-174

Grove M, Plumb M (1993) C/EBP, NF-kappa B, and c-Ets family members and transcriptional regulation of the cell-specific and inducible macrophage inflammatory protein 1 alpha immediate-early gene. Mol. Cell Biol. 13: 5276-5289

Grubb A, Weiber H, Löfberg H (1983) The γ-trace concentration of normal human seminal plasma is thirty-six times of normal human blood plasma. Scand. J. Clin. Lab. Invest. 43: 421-425

Guan X, Mann RC, Mural RJ, Uberbacher EC (1991a) On parallel search of DNA sequence databases. Proceedings of the Fifth SIAM Conference on Parallel Processing for Scientific Computing, pp 332-337

Guan X, Mann R, Mural R, Uberbacher EC (1991b) Searching consensus patterns on hypercube. Sixth Distributed Memory Computing Conference, Portland, OR , pp 470-472

Guan X, Mural RJ, Einstein JR, Mann RC, Uberbacher EC (1992) GRAIL: An integrated artificial intelligence system for gene recognition and interpretation. Eight IEEE Conference on AI applications, March 2-6, Monterey, CA. IEEE Computer Society Press, pp 9-13

Guan X, Uberbacher EC (1996) A fast look-up algorithm for detecting repetetive DNA sequences. Abstract in Proceedings of the First Pacific Symposium on Biocomputing, January 3-6, pp 718-719

Guarente L (1993) Strategies for the identification of interacting proteins. Proc. Natl. Acad. Sci. USA 90: 1639-1641

Gubler U, Hoffman BJ (1983) A simple and very efficient method for generating cDNA libraries. Gene 25: 263-269

Hall A, Dalbøge H, Grubb A, Abrahamson M (1993) Importance of the evolutionary conserved glycine residue in the N-terminal region of human cystatin C (Gly-11) for cysteine endopeptidase inhibition. Biochem. J. 291: 123-129

Hanahan D (1983) Studies on transformation of *Escheriachia coli* with plasmid. J. Mol. Biol. 166: 557-588

Hansen Sk, Tjian R (1995) TAFs and TFILA mediate differential utilization of the tandem ADH promoters. Cell 82: 565-575

Harvey S, Lavelin I, Pines M (2001) Growth hormone (GH) action in early embryogenesis: expression of a GH-response gene in sites of GH production and action. Anat. Embryol. 204: 503-510

Hayashi Y (1995) Investigation of various animal models for Sjögren syndrome. Nippon Rinsho 53: 2383-2388

He G, Fu EN, Qiu GB, Zhao Z, Xu ZM, Sun XH, Sun KL (2002) Studies of the deletion and expression of cytokeratin 13 gene in laryngeal squamous cell carcinoma. Yi Chuan Xue Bao 29: 390-395

He Z, Bateman A (1999) Progranulin gene expression regulates epithelial cell growth and promotes tumor growth *in vivo*. Cancer. Res. 59: 3222-3229

He Z, Ong CHP, Halper J, Bateman A (2003) Progranulin is a mediator of the wound response. Nat. Med. 9: 225-229

Henskens YMC, Van der velden U, Veerman ECI, Nieuw Amerongen AV (1993) Protein, albumin and cystatin concentrations in saliva of healthy subjects and of patients with gingivitis or peridontitis. J. Peridont. Res. 28: 43-48

Henskens YMC, Veerman ECI, Nieuw Amerongen AV (1996) Cystatins in health and disease. Biol. Chem. Hoppe-Seyler 377: 71-86.

Hesse M, Magin TM, Weber K (2001) Genes for intermediate filament proteins and the draft sequence of the human genome: novel keratin genes and surprisingly high number of pseudogenes related to keratin genes 8 and 18. J. Cell Sci. 114: 2569-2575

Honda K, Yamada T, Endo R, Ino Y, Gotoh M, Tsuda H, Yamada Y, Chiba H, Hirohashi S (1998) Actinin-4, a novel actin-bounding protein associated with cell motility and cancer invasion. J. Cell. Biol. 140: 1383-1393

Horwitz JP, Cluna J, Curby RJ, Tomson AJ, DaRooge MA, Fisher BE, Mauricio J, Klundt I (1964) Substrates for cytochemical demonstration of enzyme activity I. Some substituted 3-indolyl-β-D-glycopyranosides. J. Med. Chem. 7: 574-574

Hsia N, Cornwall GA (2003) *Cres2* and *Cres3*: New members of the cystatin-related epididymal spermatogenic subgroup of family 2 cystatins. Endocrinology 144: 909-915

Hu B, Coulson L, Moyer B, Price PA (1995) Isolation and molecular cloning of a novel bone phosphoprotein related in sequence to the cystatin family of thiol protease inhibitors. J. Biol. Chem. 270: 431-436

Hu Y, Nakagawa Y, Purushotham KR, Humphreys-Beher MG (1992) Functional changes in salivary glands of autoimmune disease-prone NOD mice. Am. J. Physiol. 263: E607-614

Hutchinson KW, Halvorson HO (1980) Cloning of randomly sheared DNA fragments from a ϕ 10s lysogen of *Bacillus subtilis*: Identification of prophage-containing clones. Gene 8: 267-278

Inaoka T, Bilbe G, Ishibashi O, Tezuka K, Kumegawa M, Kokubo T (1995) Molecular cloning of human cDNA for cathepsin K: novel cysteine proteinase predominantly expressed in bone. Biochem. Biophys. Res. Commun. 206: 89-96.

Inui T, Ishibashi O, Inaoka T, Origane Y, Kumegawa M, Kokubo T, Yamamura T (1997) Cathepsin K antisense oligodeoxynucleotide inhibits osteoclastic bone resorption. J. Biol. Chem. 272: 8109-8112.

Isemura S, Saito E, Sanada K (1987) Characterization and amino acid sequence of a new acidic cysteine proteinase inhibitor (Cystatin SA) structurally closely related to cystatin S, from human saliva. J. Biochem. 102: 693-704

Isemura S, Saito E, Sanada K, Minakata K (1991) Identification of full-sized forms of salivary (S-type) cystatins (cystatin SN, cystatin SA, cystatin S and two phosphorylated forms of cystatin S) in human whole saliva and determination of phosphorylation sites of cysatin S. J. Biochem. 110: 559-564

Ish-Horowicz D, Burke JF (1981) Rapid and efficient cosmid cloning. Nucl. Acids Res. 9: 2989-2998

Itoh R, Kawamato S, Adachi W, Kinoshita S, Okubo K (1999) Genomic organization and chromosomal localization of the human cathepsin L2 gene. DNA Res. 6: 137-140

Jackson T, Mould AP, Sheppard D, King AM (2002) Integrin alphavbeta1 is a receptor for foot-and-mouth disease virus. J. Virol. 76: 935-941

James P, Haliday J, Craig EA (1996) Genomic libraries and a host strain designed for highly efficient two-hybrid selection in yeast. Genetics 144: 1425-1436

Järvinen M, Rinne A, Hopsu-Havu VK (1987) Human cystatins in normal and diseases tissues– a review. Acta Histochem. 82: 5-18

Javahery R, Kachi A, Zenzie-Gregory B, Smale ST (1994) DNA sequence requirments for transcriptional initiator activity in mammalian cells. Mol. Cell Biol. 14: 116-127

Jeffreys AJ, Neumann R, Wilson V (1990) Repeat unit sequence variation in minisatellites: A novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. Cell 60: 473-485

Johnson MR, Polymeropoulos MH, Vos HL, Ortiz de Luna RI, Francomano CA (1996) A nonsense mutation in the cathepsin K gene observed in a family with pycnodysostosis. Genome Res. 6: 1050-1055

Johnston C (1987) A model fungal gene regulatory mechanism: the *GAL* genes of *Saccharomyces cerevisiae*. Microbiol. Rev. 51: 458-476

Jones HB (1996) Pairwise analysis of radiation hybrid mapping data. Ann. Hum. Genet. 60: 351-357

Jones MB, Michener CM, Blanchette JO, Kuznetsov VA, Raffeld M, Serrero G, Emmert-Buck MR, Petricoin EF, Krizman DB, Liotta LA, Kohn EC (2003) The granulin-epithelin precursor/PC-cell-derived growth factor is a growth factor for epithelial ovarian cancer. Clin. Cancer Res. 9: 44-51

Jones P, Qiu J, Rickwood D (1994) RNA Isolation and Analysis, Chapter 3, Characterisation of RNA size. Bios Scientific Publishers, Oxford, pp 47-93

Jorde LB, Carey JC, Bamshad MJ, White RL (2000) Medical genetics, Mosby, St. Louis

Jouanguy E, Lamhamedi-Cherradi S, Lammas D, Dorman SE, Fondaneche MC, Dupuis S, Diffinger R, Altar F, girdlestone J, Emile JF, Ducoulombier H, Edgar D, Clark J, Oxelius V, Brai M, Noveli V, Heyne K, Fischer A, Holland SM, Kumararatne DS, Schreiber RD, Casanova JL (1999) A human IFNGRI small deletion hotspot associated with dominant susceptibility to mycobacterial infection. Nat. Genet. 21: 370-378

Jungermann K, Kietzmann T (1996) Zonation of parenchymal and nonparenchymal metabolism in liver. Ann. Rev. Nutr. 16: 179-203

Jungermann K, Kietzmann T (1997) Role of oxygen in the zonation of carbohydrate metabolism and gene expression in liver. Kidney Int. 51: 402-412

Junqueria LC, Carneiro J, Kelly RO (1992) Basic Histology, Seventh edn. Appleton and Lange, Norwalk, Conn.

Juriaanase AC, Booij M (1979) Isolation and partial characterization of three proteins from human submandibular saliva. Archs. Oral. Biol. 24: 621-625

Jurisic V, Bogdanovic G, Srdic T, Kerenji A, Baltic M, Baltic V (2000) The kinetic of changes on PC cell line after TNF-alpha treatment *in vitro*. Ann. Oncol. 11, Suppl. 4: 15

Kaplan JM, Kim SH, North KN, Correia LA, Tong HQ, Mathis BJ, Rodriguez-Perez JC, Allen PG, Beggs AH, Pollak MR (2000) Mutations in *ACTN4*, encoding alpha-actinin-4, cause familial focal segmental glumerolosclerosis. Nat. Genet. 24: 251-256

Keilová H, Tomášek V (1974) Effect of papain inhibitor from chicken egg white on cathepsin B. Biochem. Biophys. Acta 334: 179-186

Kim JH, Lee-Kwon W, Park JB, Ryo SH, Yun CH, Donowitz M (2002) Ca(2+)-dependent inhibition of Na+/H+ exchanger 3 (*NHE3*) requires an NHE3-E3KARP-alpha-actinin-4 complex for oligomerization and endocytosis. J. Biol. Chem. 277: 23714-23724

Kitchen JA (1999) Isolation and function analysis of a human spp24 cDNA. A dissertation submitted for the degree of Batchelor of Medical Science, University of Leicester

Klein-Nulend J, Semeins CM, Ajubi NE, Nijweide PJ, Burger EH (1995) Pulsating fluid flow increases nitric oxide (NO) synthesis by osteocytes, but not periosteal fibroblast-correlation with prostoglandin upregulation. Biochem. Biophys. Res. Commun. 217: 640-648

Knolle P, Lohr H, Treichel U, Dienes HP, Lohse A, Schlaack J, Gerken G. (1995) Parenchymal and non-parenchymal liver cells and their interaction in the local immune response. Z Gastroenterol. 33: 613-620.

Kobori M, Ikeda Y, Nara H, Kato M, Kumegawa M, Nojima H, Kawashima H (2001) Large scale isolation of osteoclast-specific genes by an improved method involving the preparation of a subtracted cDNA library. Genes Cells 3: 459-475

Komminoth P (1992) Digoxigenin as an alternative probe labeling for *in situ* hybridisation. Diagn. Mol. Pathol. 1: 142-150

Komminoth P, Merk FB, Leav I, Wolf HJ, Roth J (1992) Comparison of ^{35}S- and digoxigenin-labeled RNA and oligonucleotide probes for *in situ* hybridization. Expression of mRNA of the seminal vesicle secretion protein II and androgen receptor genes in the rat prostate. Histochemistry 98: 217-228

Korant BD, Brzin J, Turk V (1985) Cystatin, a protein inhibitor of cysteine proteinases alters viral protein cleavages in infected human cells. Biochem. Biophys. Res. Commun. 127: 1072-1076

Körkkö J, Annuen S, Pihlajama T, Prockop DJ, Ala-Kokko L (1998) Conformation sensitive gel electrophoresis for simple and accurate detection of mutations. Comparison with denaturing gradient gel electrophoresis and nucleotide sequencing. Proc. Natl. Acad. Sci. USA 95: 1681-1685

Kos J, Lah TT (1998) Cysteine proteinases and their endogenous inhibitors: target proteins for prognosis, diagnosis and therapy in cancer (review). Oncol. Rep. 5: 1349-1361

Kovárová M, Dráber P (2000) New specificity and yield enhancer of polymerase chain reactions. Nucl. Acids Res. 28: E70

Kozak M (1989) A scanning model for translation: an update. J. Cell. Biol. 108: 229-241

Kramnik I, Dietrich WF, Demant P, Bloom BR (2000) Genetic control of resistance to experimental infection with virulent *Mycobacterium tuberculosis*. Proc. Natl. Acad. Sci. USA 97: 8560-8565

Krogh A (1997) Two methods for improving performanceof an HMM and their application for gene finding. Proc. Fifth Int. Conf. Intell. Syst. Mol. Biol. pp 179-186

Lah TT, Kos J (1998) Cysteine proteinases in cancer progression and their clinical relevance for prognosis. Biol. Chem. 379: 125-130

Lansdown ABG (2002) Calcium: a potential central regulator in wound healing in the skin. Wound Rep. Reg. 10: 271-285

Lee C-C, Bowman BH, Yang F (1987) Human α2-HS-glycoprotein : the A and B chains with a connecting sequence are encoded by a single mRNA transcript. Proc. Natl. Acad. Sci. USA 84: 4403-4407

Lenarcic B, Gabrijelcic D. Rozman B, Drobnic-kosorok M, Turk V (1988) Human cathepsin B and cysteine proteinase inhibitors (CPIs) in inflamatory and metabolic joint diseases. Biol. Chem. 369: S257-S261

Lennon GG, Auffray C, Polymerpolous M, Soares MB (1996) The I.M.A.G.E. consortium: an integrated molecular analysis of genome and their expression. Genomics 33: 151-152

Lenzerini L, Benatti U, Morelli A, Pontremoli S, De Flora A, Piazza A, Rinaldi A, Filippi G, Siniscalco M (1981.) Genetic variation in the quantitative levels of an NADP(H)-binding protein (*FX*) in human erythrocytes. Blood 57: 209-217

Leppävuori J, Kujala U, Kinnunen J, Kaprio J, Nissilä M, Heliövaara M, Klinger N, Partanen J, Terwilliger JD, Peltonen L (1999) Genome scan for predisposing loci for distal interphalangeal joint osteoarthritis: evidence for locus on 2q. Am. J. Hum. Genet. 65: 1060-1067

Lerner UH, Grubb A (1992) Human cystatin C, a cysteine proteinase inhibitor, inhibits bone resorption *in vitro* stimulated by parathyroid hormone and parathyroid hormone-related peptide of malignancy. J. Bone Miner. Res. 7: 433-440

Levy ER, Herrington CS (1995) Non-isotopic Methods in Molecular Biology. Oxford University Press, Oxford

Li YP, Alexander M, Wucherpfennig AL, Yelick P, Chen W, Stashenko P (1995) Cloning and complete coding sequence of a novel human cathepsin expressed in giant cells of osteoclastomas. J. Bone Miner. Res. 10: 1197-1202

Liau LM, Lallone RL, Seitz R, Buznikov A, Gregg JP, Kornblum HI, Nelson SF, Bronstein JM (2000) Identification of a human glioma-associated growth factor gene, granulin, using differential immuno-absorption. Cancer Res. 60: 1353-1360

Lindhal P, Alrikson E, Jörnwall I, Björk I (1988) Interaction of the cysteine proteinase inhibitor chicken cystatin with papain. Biochemistry 27: 5074-5082

Line A, Stengrevics A, Slucka Z, Li G, Jankevics E, Rees RC (2002) Serological identification and expression analysis of gastric cancer-associated genes. Br. J. Cancer 86: 1824-1830

Ling M, Merante F, Robinson BH (1995) A rapid and reliable DNA preparation method for screening a large number of yeast clones by polymerase chain reaction. Nucl. Acids Res. 23: 4924-4925

Linsk R, Gottesman M, Pernis B (1989) Are tissues a patch quilt of ectopic gene expression? Science 246: 261-261

Lo K, Smale ST (1996) Generality of a functional initiator consensus sequence. Gene 182: 13-22

Locksley RM, Killeen N, Lenardo MJ (2001) The TNF and TNF reseptor superfamilies: Integrating mammalian biology. Cell 104: 478-501

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

Lu R, Serrero G (1999) Stimulation of PC cell-derived growth factor (epithelin/granulin precursor) expression by estradiol in human breast cancer cells. Biochem. Biophys. Res. Commun. 256: 204-207

Lu R, Serrero G (2000) Inhibition of PC cell-derived growth factor (PCDGF, epithelin/granulin precursor) expression by antisense PCDGF cDNA transfection inhibits tumorigenicity of the human breast carcinoma cell line MDA-MB-486. Proc. Natl. Acad. Sci. USA 97: 3993-3998

Ma J, Ptashne M (1987) Deletion analysis of *GAL4* defines two transcriptional activating segments. Cell 48: 847-853

Machleidt W, Borchart U, Fritz H, Brzin J, Ritonja A, Turk V (1983) Protein inhibitors of cysteine proteinases. Primary structure of stefin, a cytosolic protein inhibitor of cysteine proteinases from human polymorphonuclear granulocytes. Biol. Chem. Hoppe-Seyler 364: S1481-S1486

Makalowski W, Mitchell GA, Labuda D (1994) Alu sequences in the coding region of mRNA: A source of protein variability. Trends Genet. 10: 188-193

Malecha MJ, Miettinen M (1991) Expression of keratin 13 in human epithelial neoplasms. Virchows Arch. A Patol. Anat. Histopathol. 418: 249-254

Markel P, Fulker D, Bennett B, Corley R, DeFries J, Erwin V, Johnson T (1996) Quantitative trait loci for ethanol sensitivity in the LS × SS recombinant inbred strains: interval mapping. Behav. Genet. 26: 447-458

Markel PD, Bennett B, Beeson M, Gordon L, Johnson TE (1997) Confirmation of quantitative trait loci for ethanol sensitivity in long-sleep and short-sleep mice. Genome Res. 7: 92-99

Maroteaux P, Lamy M (1964) Achondroplasia in man and animals. Clin. Orthop. 33: 91-103.

Matis S, Xu Y, Shah M, Guan X, Einstein JR, Mural RJ, Uberbacher EC (1996) Detection of RNA polymerase II promoters and polyadenylation sites in human DNA sequence. Comput. Chem. 20: 135-140

McCarthy LC, Terrett J, Davis ME, Knights C, Smith AL, Critcher R, Schmitt K, Hudson J, Spurr NK, Goodfellow PN (1997) A first-generation whole genome-radiation hybrid map spanning the mouse genome. Genome Res. 7: 1153-1161

McClearn G, Kakihana R (1981) Selective breeding for ethanol sensitivity: short-sleep and long-sleep mice. In McClearn G, Deitrich R, Erwin V (eds) Development of Animal Models as Pharmacogenetic Tools. U.S. Government Printing Office, Washington, DC, pp 147-159

McGuigan FE, Reid DM, Ralston SH (2000) Susceptibility to osteoporotic fracture is determined by allelic variation at the Sp1 site, rather than other polymorphic sites at the *COL1A1* locus. Osteoporos. Int. 11: 338-343

Melton L, Atkinson E, O'Fallon W, Wahner H, Riggs B (1995) Long-term fracture prediction by bone mineral assessed at different skeletal sites. J. Bone Miner. Res. 8: 1227-1233

Miki R, Kadota K, Bono H, Tomaru Y, Carninci P, Itoh M, shibata K, Kawai J, konno H, Watanabe S, Sato K, Tokusumi Y, Kikuchi N, Ishii Y, Hamaguchi Y, Nishizuka I, Goto H, Nitanda H, Satomi S, Yoshiki A, kusakabe M, DeRisi JL, Eisen MB, Iyer VR, Brown PO, Muramatsu M, Shimada H, Okazaki Y, Hayashizaki Y (2001) Delineating development and metabolic pathway *in vivo by* expression profiling using RIKEN set of 18,816 full-length enriched mouse cDNA arrays. Proc. Natl. Acad. Sci. USA 98: 2199-2204

Morimoto H, Okamura H, Haneji T (2002) Interaction of protein phosphatase 1 delta with nucleolin in human osteoblastic cells. J. Histochem. Cytochem. 50: 1187-1193

Morita M, Youshiuchi N, Arakawa H, Nishimura S (1999) *CMAP*: A novel cystatin-like gene involved in liver metastasis. Cancer Res. 59: 151-158

Morris KH, Ishikawa A, Keightley PD (1999) Quantitative trait loci for growth traits in C57BL/6J × DBA/2J mice. Mamm. Genome 10: 225-228

Morrison NA, Qi JC, Tokita A, Kelly PJ, Crofts L, Nguyen TV, Sambrook PN, Eisman JA (1994) Prediction of bone density from vitamin D receptor alleles. Nature 367: 284-287

Mullis KB, Faloona FA (1987) Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. Meth. Enzymol. 155: 335-50

Mural R, Guan X, Uberbacher E (1993) Computational methods for locating biological features in DNA sequences. Current Protocols in Human Genetics, Unit 6.5, Supplement 6

Mural RJ, Einstein JR, Guan X, Mann RC, Uberbacher EC (1991) An artificial intelligence approach to DNA sequence feature recognition. Trends Biotech. 10: 66-69

Murray CJL, Salomon J (1998) Modeling the impact of global tuberculosis control strategies. Proc. Natl. Acad. Sci. USA 95: 13881-13886

Murray CJL, Styblo K, Rouillon A (1990) Tuberculosis in developing countries: Burden, intervention and cost. Bull IUATLD 65: 6-24

Nagy A (2000) Cre recombinase: The universal reagent for genome tailoring. Genesis 26: 99-109

Nagy E, Maquat LE (1998) A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends Biochem. Sci. 23:198-199

Naito Y, Sasaki M, Umemoto T, Namikawa I, Sakae K, Ishihara Y, Isomura S, Suzuki I (1995) Bactericidal effect of rat cystatin on an oral bacterium *Porophyromonas gingivalis*. Comp. Biochem. Physiol. PartC, Pharmacol. Toxicol. Endocrinol. 110: 71-75

Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res. 11: 863-974

Ni J, Abrahamson M, Zhang M, Fernandez MA, Grubb A, Su J, Yu G-L, Li Y, Parmelee D, Xing L, Coleman TA, Gentz S, Thotakura R, Neguyen N, Hesselberg M, Gentz R (1997) Cystatin E is a novel human cysteine proteinase inhibitor with structural resemblance to family 2 cystatins. J. Biol. Chem. 272: 10853-10654

Ni J, Fernandez MA, Danielsson L, Chillakuru RA, Zhang J, Grubb A, Su J, Gentz R, Abrahamson M (1998) Cystatin F is a glycosylated human low molecular weight cysteine proteinase inhibitor. J. Biol. Chem. 273: 24797-24804

Nishida Y, Sunmi H, Mihara H (1984) A thiol protease inhibitor and its identity with low molecular weight kininogen. Biochemistry 23: 5691-5697

Niu T, Chen C, Cordell H, Yang J, Wang B, Wang Z, Fang Z, Schork NJ, Rosen CJ, Xu X (1999) A genome-wide scan for loci linked to forearm bone mineral density. Hum. Genet. 104: 226-233

Novina CD, Roy AL (1996) Core promoters and transcriptional control. Trends Genet. 12: 351-355

Ohba Y, Ohba T, Sumitani K, Tagami-Kondoh K, Hiura K, Miki Y, Kakegawa H, Takano-Yamamoto T, Katunuma N (1996) Inhibitory mechanisms of H^+-ATPase inhibitor bafilomycin A1 and carbonic anhydrase II inhibitor acetazolamide on experimental bone resorption. FEBS Lett. 387: 175-178

Ohnishi T, Nakamura O, Ozawa M, Arakaki N, Muramatsu T, Daikuhara Y (1993) Molecular cloning and sequence analysis of cDNA for a 59 kDa bone sialoprotein of the rat: Demonstration that it is a counterpart of human α2-HS glycoprotein and bovine fetuin. J. Bone Miner. Res. 8: 367-377

Oinonen T, Lindros KO (1998) Zonation of hepatic cytochrome P-450 expression and regulation. Biochem. J. 329: 17-35

Osoegawa K, Tateno M, Woon PY, Frengen E, Mammoser AG, Catanese JJ, Hayashizaki Y, de Jong PJ (2000) Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. Genome Res. 10: 116-128

Othani O, Fukuyama K, Epstein WL (1982) Biochemical properties of thiol proteinase inhibitors purified from psoriatic scales. J. Invest. Dermatol. 82: 280-284

Pavesi A, Conterio F, Bolchi A, Dieci G, Ottonello S (1994) Identification of new eukaryotic tRNA genes in genomic DNA databases by multistep weight matrix analysis of transcriptional control regions. Nucl. Acids Res. 22: 1247-1256

Penrose LS (1953) The general purpose sib-pair linkage test. Ann. Eugenics 18: 120-124

Price PA, Lim JE (2003) The inhibition of calcium phosphate precipitation by fetuin is accompanied by the formation of a fetuin-mineral complex. J. Biol. Chem. 278: 22144-22152

Price PA, Nguyen TM, Williamson MK (2003) Biochemical characterization of the serum fetuin-mineral complex. J. Biol. Chem. 13: 22153-22160

Price PA, Rise JS, Williamson MK (1994) Conserved phosphorylation of serines in the Ser-X-Glu/Ser(P) sequences of the vitamin K-dependent matrix GLA protein from shark, lamb, rat, cow and human. Protein Sci. 3: 822-830

Price PA, Caputo JM, Williamson MK (2002) Bone origin of the serum complex of calcium, phosphate, fetuin, and matrix GLA protein: biochemical evidence for cancellous bone-remodelling compartment. J. Bone Miner. Res. 17: 1171-1179

Quandt K, Frech K, Karas H, Wingender E, Werner T (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucl. Acids Res. 23: 4878-4884

Ralston S (2002) Genetic control of susceptibility to osteoporosis. J. Clin. Endocrinol. Metab. 87: 2460-2464

Ralston SH (1997) Osteoporosis. Brit. Med. J. 315: 469-472

Ritonja A, Machleidt W, Barrett AJ (1985) Amino acid sequence of the intracellular cysteine proteinase inhibitor cystatin B from human rat liver. Biochem. Biophys. Res. Commun. 131: 1187-1192

Roeder RG (1991) The complexities of eukaryotic transcription initiation: Regulation of preinitiation complex assembly. Trends Biochem. Sci. 16: 402-408

Roest Crollius H, Jaillon O, Bernot A, Dasilva C, Bouneau L, Fischer C, Fizames C, Wincker P, Brottier P, Quetier F, Saurin W, Weissenbach J (2000) Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. Nat Genet. 25: 235-238

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers of melanomas. Methods Mol. Biol. 132: 365-386

Rozhin J, Wade R, Honn KV, Sloane BF (1989) Membrane-associated cathepsin L: a role in metastasis. Biochem. Biophys. Res. Commun. 164: 556-561

Rubin JE, Rubin CT (1997) The osteoblast, osteocyte and osteoclast. Curr. Opin. Orthop. 8: 34-42

Salvesen G, Parkes C, Abrahamson M, Grubb A, Barrett AJ (1986) Human low-M_r kininogen contains three copies of cystatin sequence that are divergent in structure and in inhibitory activity for cysteine proteases. Biochem. J. 234: 429-434

Sambrook J, Fritsch EF, Maniatis T (1982) Molecular Cloning: A Laboratory Manual, Second edn. Cold Spring harbor Laboratory Press, Cold Spring Harbor, NY

Sambrook J, Fritsch EF, Maniatis T (1989) Molcular Cloning: A Laboratory Manual, Third edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Samejima T, Kaji H, Takeda A (1986) The interaction of papain molecule with thiol proteinase inhibitors from newborn rat epidermis. In: Turk V (ed) Cysteine Proteinases and Their Inhibitors. Walter de Ggruyter, Berlin, pp 561-568

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74: 5463-5467

Sarker G, Sommer SS (1989) Access to messenger RNA sequence or its protein product is not limited by tissue or species specificity. Science 244: 331-334

Sarria AJ, Panini SR, Evans RM (1992) A functional role for vimentin intermediate filaments in metabolism of lipoprotein-derived cholesterol in human SW-13 cells. J. Biol. Chem. 267: 19455-19463

Sasaki K, Shima H, Kitagawa Y, Irino S, Sugimura T, Nago M (1990) Identification of members of the protein phosphatase 1 gene family in the rat and enhanced expression of protein phosphatase 1 alpha gene in rat hepatocellular carcinoma. Jpn. J. Cancer Res. 81: 1272-1280

Saunders NR, Reynolds ML, Habgood MD, Ward RA (1992) Origin and fate of fetuin-containing neurons in the developing neocortex of the fetal sheep. Anat. Embryol. 186: 477-486

Sawyer JR, Hozier JC (1986) High resolution of mouse chromosomes: banding conservation between man and mouse. Science 232: 1632-1635

Schinke T, Amendt C, Trindl A, Poschke O, Muller-Esteryl W, Jahnen-Dechen W (1996) The serum protein α_2-HS glycoprotein/fetuin inhibits apatite formation *in vitro* and in mineralizing calvaria cells. J. Biol. Chem. 271: 20789-20796

Schuckit MA (1998) Biological, psycological and environmental factors of the alcoholism risk: a longitudinal study. J. Stud. Alcohol 59: 485-494

Schwyter DH, Huang J, Dubinicoff T, Courey AJ (1995) The decapentaplegic core promoter region plays an integral role in the spatial control of transcription. Mol. Cell Biol. 15: 3960-3968

Senapathy P, Shapiro MB, Harris NL (1990) Splice junctions, branch point sites and exons: sequence statistics, identification and applications to genome project. Meth. Enzymol. 183: 252-278

Shah M, Xu X, Einstein J, Guan X, hauser L, Matis S, Lee R, Mural R, Uberbacher E (1995) Gene discovery and sequence annotation in GRAIL 1.3. The Hilton Head DNA Sequence Conference, Hilton Head, S.C., September 16-20

Shah MB, Guan X, Einstein JR, Matis S, Xu Y, Mural RJ, Uberbacher EC (1996) User's guide to GRAIL and GENQUEST (Sequence analysis, gene assembly and sequence comparison system) E-mail servers an XGRAIL (Version 1.3c), GRAILCLNT (Version 1.3) command line interface and XGENQUEST (Version 1.1) client-server systems. Available by anonymous ftp from arthur.epm.ornl.gov (128.219.9.76) from directory pub/xgrail or pub/xgenQuest or pub/grailclnt as file Manual.grail1.3-genquest.

Shaw JP, Utz PJ, Durand DB, Toole JJ, Emmel EA, Crabtree GR (1988) Identification of a putative regulator of early T cell activation genes. Science 241: 202-205

Shi GP, Chapman HA, Bhairi SM, DeLeeuw C, Reddy VY, Weiss SJ (1995) Molecular cloning of human cathepsin O, a novel endoproteinase and homologue of rabbit OC2. FEBS Lett. 357: 129-34.

Siegal GJ, Agranoff BW, Albers RW, Fisher SK, Uhler MD (1999) Basic Neurochemistry, Molecular, Cellular, and Medical Aspects, Sixth edn. Lippincott, Williams & Wilkins, Philadelphia, PA

Siguret V, Ribba AS, Cherel G, Meyer D, Pietu G (1994) Effect of plasmid size on transformation efficiency by electroporation of *Escherichia coli* DH5α. Biotechniques 16: 422-426

Sloane B, Rozhin J, Hatfield J, Crissman J, Honn K (1987) Plasma membrane-associated cysteine proteinases in human and animal tumours. Exp. Cell Biol. 55: 209-224

Sloane BF (1990) Cathepsin B and cystatins: evidence for a role in cancer progression. Semin. Cancer Biol. 1: 137-152

Smale ST, Baltimore D (1989) The "initiator" as a transcription control element. Cell 57: 103-113

Smigilski EM, Sirotkin K, Ward M, Sherry ST (2000) dbSNP: a database of single nucleotide polymorphisms. Nucl. Acids Res. 28: 352-355

Smith D, Nance W, Kang K, Christian J, Johnson C (1973) Genetic factors in determining bone mass. J. Clin. Invest. 52: 2800-2808

Smith PL, Myers JT, Rogers CE, Zhou L, Petryniak B, Becker DJ, Homeister JW, Lowe JB (2002) Conditional control of selectin ligand expression and global fucosylation events in mice with a targeted mutation at the *FX* locus. J. Cell. Biol. 158: 801-815

Solovyev VV, Lawrence CB (1993) Prediction of human gene structure using dynamic programming and oligonucleotide composition. In: Abstracts of the Fourth annual keck symposium, Pittsburgh, pp 47-47

Solovyev VV, Salamov AA (1997) The Gene-Finder computer tools for analysis of human and model organisms genome sequences. In: Proc. Fifth Int. conf. Intell. Sys Mol. Bio. AAA1 Press, Menlo Park, CA

Solovyev VV, Salamov AA, Lawrence CB (1994a) The prediction of human exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. In: Altman R, Brutlag D, Karp R, Latrop R, Searls D (eds) The Second International Conference on Intelligent Systems for Molecular Bilogy, AAA1 Press, Menlo Park, CA, pp 354-362

Solovyev VV, Salamov AA, Lawrence CB (1994b) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. Nucl. Acids Res. 22: 5156-5163

Solovyev VV, Salamov AA, Lawrence CB (1995) Identification of human gene structure using linear discriminant functions and dynamic programming. Proc. Third Int. Conf. Intell. Sys. Mol. Biol. 3: 367-375

Sorensen TIA, Nielsen GG, Andersen PK (1988) Genetic and environmental influences on premature death in adult adoptees. N. Engl. J. Med. 318: 727-732

Sorom AJ, Nyberg SL, Gores GL (2002) Keratin, fas and cryptogenic liver failure. Liver Transpl. 8: 1195-1197

Sotiropoulou G, Anisowicz A, Sager R (1997) Identification, cloning and characterization of cystatin M, a novel cysteine proteinase inhibitor, down-regulated in breast cancer. J. Biol. Chem. 272: 903-910

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98: 503-517

Stewart M (1993) Intermediate filament structure and assembly. Curr. Opin. Cell Biol. 5: 3-11

Stewart TL, Ralston SH (2000) Role of genetic factors in the pathogenesis of osteoporosis. J . Endocinol. 166: 235-245

Strachan T, Read AP (1999) Human Molecular Genetics, second edn. BIOS Scientific Publishers, Oxford

Suda K, Woo JT, Takanmi M, Sexton PM, Nagai K (2002) Lipopolysaccharide supports survival and fusion of preosteoclasts independent of TNF-α, IL-1 and RANKL. J. Cell. Physiol. 190: 101-108

Sullivan FX, Kumar R, Kriz R, Stahl M, Xu G-Y, Rouse J, Chang X, Boodhoo A, Potvin B, Cumming DA (1998) Molecular cloning of human GDP-mannose 4,6-dehydratase and reconstitution of GDP-fucose biosynthesis *in vitro*. J. Biol. Chem. 273: 8193-8202

Sun H, Li N, Wang X, Liu S, Chen T, Zhang L, Wan T, Cao X (2003) Molecular cloning and characterization of a novel cystatin-like molecule, CLM, from human bone marrow stromal cells. Biochem. Biophys. Res. Commun. 301: 176-182

Sundquist K, Lakkakorpi P, Wallmark B, Vaananen K (1990) Inhibition of osteoclast proton transport by bafilomycin A1 abolishes bone resorption. Biochem. Biophys. Res. Commun. 168: 309-313

Sutter C, Nischt R, Winter H, Schweizer J (1991) Aberrant *in vitro* expression of keratin 13 induced by Ca^{2+} and vitamin A acid in mouse epidermal cell lines. Exp. Cell Res. 195: 183-193

Swallow JE, Merrison WK, Gill PK, Harris S, Dalgleish R (1997) Assignment of secreted phosphoprotein 24 (*SPP2*) to human chromosome band 2q37→qter by *in situ* hybridization. Cytogenet. Cell. Genet. 79: 142

Takagaki Y, Kitamura N, Nakanishi S (1985) Cloning and sequence analysis of cDNAs for human high molecular weight and low molecular weight prekininogens. Primary structures of two human prekininogens. J. Biol. Chem. 260: 8601-8609.

Takahashi M, Tezuka T, Katunuma N (1994) Inhibition of growth and cysteine proteinase activity of *Staphylococcus aureus* V8 by phosphorylated cystatin alpha in skin cornified envelope. FEBS Lett. 355: 275-278

Tan I, Ng CH, Lim L, Leung T (2001) Phosphorylation of a novel myosin binding subunit of protein phosphatase 1 reveals a conserved mechanism in the regulation of actin cytoskeleton. J. Biol. Chem. 276: 21209-21216

Taugner R, Buhrle C, Nobiling R, Kirschke H (1985) Coexistence of renin and cathepsin B in endotheloid cell secretory granules. Histochemistry 88: 102-108

Temin HM, Mizutani S (1970) RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature 226: 1211-1213

Tezuka K, Tezuka Y, Maejima A, Sato T, Nemoto K, Kamioka H, Hakeda Y, Kumegawa M (1994) Molecular cloning of a possible cysteine proteinase predominantly expressed in osteoclasts. J. Biol. Chem. 269: 1106-1109

Tjian R, Maniatis T (1994) Transcriptional activation: A complex puzzle with few easy pieces. Cell 77: 5-8

Tonetti M, Sturla L, Bisso A, Benatti U, De Flora A (1996) Synthesis of GDP-L-fucose by the human FX protein. J. Biol. Chem. 271: 27274-27279

Tsukamoto K, Yoshida H, Watanabe S, Suzuki T, Miyao M, Hosoi T, Orimo H, Emi M (1999) Association of radial bone mineral density with CA repeat polymorphism at the interleukin 6 locus in postmenopausal Japanese women. J. Hum. Genet. 44: 148-151

Turk B, Turk V, Turk D (1997) Structural and functional aspects of papain-like cysteine proteinases and their protein inhibitors. Biol Chem. 378: 141-50

Turk V, Bode W (1991) The cystatins: protein inhibitors of cysteine proteinases. FEBS Lett. 285: 213-219

Uberbacher EC (1994) ORNL Announce genQuest and X-GRAIL. Hum. Genome News 5: 8-9

Uberbacher EC (1995) Discovering the intelligence in molecular biology. Trends Biotech. 13: 497-500

Uberbacher EC, Einstein JR, Guan X, Mural RJ (1992) Gene recognition and assembly in the GRAIL system: Progress and challenges. The Second International Conference on Bioinformatics, Supercomputing, and complex Genome Analysis, pp 465-467

Uberbacher EC, Mann RC, Hand RC, Mural RJ (1991) A neural network-multiple sensor based method for recognition of gene coding segments in human DNA sequence data. ORNL/TM-11741

Uberbacher EC, Mural RJ (1991) Locating protein coding regions in human DNA sequences using a multiple sensor-neural network approach. Proc. Natl. Acad. Sci. USA 88: 11261-11265

Uberbacher EC, Xu Y, Mural RJ (1995a) Discovering and understanding genes in human DNA sequence using GRAIL. Computer Methods for Macromolecular Sequence Analysis, September

Uberbacher EC, Xu Y, Shah M, Matis S, Guan X, Mural RJ (1995b) DNA sequence pattern recognition methods in GRAIL. Presentation to be published as full article in DIMACS Workshop on Gene-Finding and Gene Structure Prediction., Philadelphia, PA, October 13-14

Ullmann A, Jacob F, Monod J (1967) Characterization by *in vitro* complementation of a peptide corresponding to an operator-proximal segment of the β-galactosidase structural gene of *Escherichia coli*. J. Mol. Biol 24: 339-343

Underhill PA, Jin L, Zemans R, Oefner PJ, Cavalli-Sforza LL (1996) A pre-Columbian Y chromosome-specific transition and its implications for human revolutionary history. Proc. Natl. Acad. Sci. USA 93: 196-200

Vaughn TT, Pletscher LS, Peripato A, King-Ellison K, Adams E, Erikson C, Cheverud JM (1999) Mapping quantitative trait loci for murine growth: a closer look at genetic architecture. Genet. Res. 74: 313-322

Walter MA, Spillet DJ, Thomas P, Weissenbach J, Goodfellow PN (1994) A method for constructing radiation hybrid maps of whole genome. Nat. Genet. 7: 22-28

Waseem A, Alam Y, Dogan B, White KN, Leigh IM, Waseem NH (1998) Isolation, sequence and expression of the gene encoding human keratin 13. Gene 215: 269-279

World Health Organisation Technical Report (1994) The WHO criteria for low bone mineral density and osteoporosis in women. WHO, series 834

Williams GW, Woollard PM, Hingamp P (1998) NIX: A nucleotide identification system at the HGMP-RC. URL: http://www.hgmp.mrc.ac.uk/NIX/

Wilson KS, Roberts H, Leek R, Harris AL, Geradts J (2002) Differential gene expression patterns in HER/*neu*-positive and negative breast cancer cell lines and tissues. Am. J. Pathol. 161: 1171-1185

Xia L, Kilb J, Wex H, Li Z, Lipyansky A, Breuil V, Stein L, Palmer JT, Dempster DW, Brömme D (1999) Localization of rat cathepsin K in osteoclasts and resorption pits: inhibition of bone resorption and cathepsin K-activity by peptidyl vinyl sulfones. Biol. Chem. 380: 679-687

Xu SQ, Tang D, Chamberlin S, Pronk G, Masiarz RF, Kaur S, Prisco M, Zanocco-Marani T, Baserga R (1998) The granulin/epithelin precursor abrogates the requirement for the insulin-like growth factor 1 receptor for growth in vitro. J. Biol. Chem. 273: 20078-20083

Xu Y, Mural RJ, Shah M, Uberbacher EC (1994a) Recognizing exons in genomic sequence using GRAIL II. Genetic Engineering, Principles and Methods, Plenum Press, Vol. 15

Xu Y, Einstein JR, Mural RJ, Shah M, Uberbacher EC (1994b) An improved system for exon recognition and gene modeling in human DNA sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. 2: 376-378

Xu Y, Mural RJ, Uberbacher EC (1994c) Constructing gene models from accurately-predicted exons: An application of dynamic programming. Appl. Biosci. 10: 613-623

Xu Y, Mural RJ, Uberbacher EC (1995a) An iterative algorithm for correcting sequencing errors in DNA coding regions. Presentation to be published as full article in DIMACS Workshop on Gene-Finding and Gene structure Prediction, Philadelphia, PA, October 13-14

Xu Y, Mural RJ, Uberbacher EC (1995b) Correcting sequencing errors in DNA coding regions using a dynamic programming approach. Comput. Appl. Biosci. 11: 117-124

Xu Y, Mural RJ, Uberbacher EC (1995c) An iterative algorithm for correcting sequencing errors in DNA coding regions. J. Comput. Biol. 3: 333-334

Xu Y, Uberbacher EC (1996a) Gene prediction by pattern recognition and homology search. The Forth International Conference on Intelligent System for Molecular Biology, St. louis, MO, June 13-15

Xu Y, Uberbacher EC (1996b) A polynomial-time algorithm for a class of protein threading problems. Proc. Int. conf. Intell. Sys. Mol. Biol. 12: 511-517

Yang Z, Wong GKS, Eberle MA, Kibokawa M, Passey DA, Hughes WR, Kruglyak L, Yu J (2000) Sampling SNPs. Nat. Genet. 26: 13-14

Zanetti M, Gennaro R, Romeo D (1995) Cathelicidins: a novel protein family with a common proregion and variable C-terminal antimicrobial domain. FEBS Lett. 374: 1-5

Zhang H, Serrero G (1998) Inhibition of tumorigenicty of the teratoma PC cell line by transfection with antisense cDNA for PC cell-derived growth factor (PCDGF, epithelin/granulin precursor). Proc. Natl. Acad. Sci. USA 93: 14202-14207

Zhang MQ (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc. Natl. Acad. Sci. USA 94: 565-568

Appendix A

```
ID   ENTRYNAME  standard; DNA; UNA; 20840 BP.
XX
HD * confidential 31-DEC-2004
XX
AC   ;
XX
DE   Mus musculus Spp2 gene for secreted phosphoprotein 24 precursor, exons
DE   1-8.
XX
KW   .
XX
OS   Mus musculus
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
XX
RN   [1]
RP   1-20840
RA   Dalgleish R.;
RT   ;
RL   Submitted (25-JUL-2002) to the EMBL/GenBank/DDBJ databases.
RL   Dalgleish R., Department of Genetics, University of Leicester, University
RL   Road, Leicester, LE1 7RH, UNITED KINGDOM.
XX
RN   [2]
RA   Bennett C.S, Khorram Khorshid H.R, Kitchen J.A.,.Arteta D., Denny P.,
RA   Smith T.P.L., Dalgleish R.;
RT   "Structure and expression analysis of the human and mouse genes encoding
RT   secreted phosphoprotein 24.";
RL   Unpublished.
XX
CC   TPA third party annotation created from mouse chromosome 1,
CC   88832387..88853226
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..20840
FT                   /organism="Mus musculus"
FT                   /db_xref="taxon:10090"
FT                   /strain="C57BL/6J"
FT   5'UTR           1022..1125
FT   LTR             3171..3449
FT                   /note="MaLR"
FT   LTR             complement(6536..6991)
FT                   /note="ERV1"
FT   LTR             complement(8259..8427)
FT                   /note="MaLR"
FT   LTR             complement(8412..8513)
FT                   /note="ERVK"
FT   LTR             complement(8544..8714)
FT                   /note="MaLR"
FT   LTR             complement(11559..11907)
FT                   /note="MaLR"
FT   LTR             complement(13568..13913)
FT                   /note="MaLR"
FT   polyA_signal    20373..20378
FT   repeat_region   4104..4576
FT                   /rpt_type=dispersed
FT                   /rpt_family="MER2_type"
FT   repeat_region   complement(4750..4853)
FT                   /rpt_type=dispersed
FT                   /rpt_family="Alu"
FT   repeat_region   complement(7027..7231)
FT                   /rpt_type=dispersed
FT                   /rpt_family="MER1_type"
FT   repeat_region   7408..7553
FT                   /rpt_type=dispersed
FT                   /rpt_family="MIR"
FT   repeat_region   complement(9660..9752)
FT                   /rpt_type=dispersed
FT                   /rpt_family="B4"
FT   repeat_region   complement(9940..10016)
FT                   /rpt_type=dispersed
FT                   /rpt_family="ID"
```

Appendix A

```
FT   repeat_region   complement(10068..10590)
FT                   /rpt_type=dispersed
FT                   /rpt_family="L1"
FT   repeat_region   complement(10720..10820)
FT                   /rpt_type=dispersed
FT                   /rpt_family="Alu"
FT   repeat_region   complement(10864..11058)
FT                   /rpt_type=dispersed
FT                   /rpt_family="B2"
FT   repeat_region   complement(13401..13512)
FT                   /rpt_type=dispersed
FT                   /rpt_family="Mariner"
FT   repeat_region   15682..15889
FT                   /rpt_type=dispersed
FT                   /rpt_family="B2"
FT   repeat_region   complement(16046..16207)
FT                   /rpt_type=dispersed
FT                   /rpt_family="B2"
FT   repeat_region   16344..16501
FT                   /rpt_type=dispersed
FT                   /rpt_family="B2"
FT   repeat_region   3666..3704
FT                   /rpt_type=tandem
FT                   /rpt_unit=3666..3685
FT                   /note="repeat unit = TAACTCAGCTACTCCCTCCC"
FT   repeat_region   4709..4738
FT                   /rpt_type=tandem
FT                   /rpt_unit=4709..4721
FT                   /note="repeat unit = CCTTCCTCTCTC"
FT   repeat_region   5826..5876
FT                   /rpt_type=tandem
FT                   /rpt_unit=5826..5827
FT                   /note="repeat unit = GT"
FT   repeat_region   8415..8497
FT                   /rpt_type=tandem
FT                   /rpt_unit=8415..8416
FT                   /note="repeat unit = AG  imperfect repeat containing some AC
repeat units"
FT   repeat_region   10634..10679
FT                   /rpt_type=tandem
FT                   /rpt_unit=10634..10637
FT                   /note="repeat unit = TTCC"
FT   repeat_region   15851..15902
FT                   /rpt_type=tandem
FT                   /rpt_unit=15851..15852
FT                   /note="repeat unit = AC"
FT   repeat_region   20423..20482
FT                   /rpt_type=tandem
FT                   /rpt_unit=20423..20452
FT                   /note="repeat unit = TAGAGCATCACTAAGTATCCAATTTCAGAT"
FT   mRNA            join(1022..1180,1281..1405,5073..5192,6215..6322
FT               11202..11256,12351..12401,14547..14642,20286..20395)
FT                   /gene="Spp2"
FT   CDS             join(1126..1180,1281..1405,5073..5192,6215..6322
FT                   ,11202..11256,12351..12401,14547..14632)
FT                   /codon_start=1
FT                   /gene="Spp2"
FT                   /product="secreted phospohoprotein 24 precursor"
/translation="MLKTLALLVLGMHYWCATGFPVYDYDPSSLQEALSASVAKVNSQS
LSPYLFRATRSSLKRVNVLDEDTLVMNLEFSVQETTCLRDSGDPSTCAFQRGYSVPTAA
CRSTVQMSKGQVKDVWAHCRWASSSESNSSEEMMFGDMARSHRRRNDYLLGFLSDESRS
EQFRDRSLEIMRRGQPPAHRRFLNLHRRARVNSGFE"
FT   sig_peptide     join(1126..1180,1281..1282)
FT                   /gene="Spp2"
FT   mat_peptide     join(1283..1405,5073..5192,6215..6322,11202..11256
FT                   ,12351..12401,14547..14629)
FT                   /gene="Spp2"
FT                   /product="spp24"
FT   exon            1022..1180
FT                   /number=1
FT                   /gene="Spp2"
FT   intron          1181..1280
FT                   /number=1
FT                   /gene="Spp2"
```

Appendix A

```
FT   exon            1281..1405
FT                   /number=2
FT                   /gene="Spp2"
FT   intron          1406..5072
FT                   /number=2
FT                   /gene="Spp2"
FT   exon            5073..5192
FT                   /number=3
FT                   /gene="Spp2"
FT   intron          5193..6214
FT                   /number=3
FT                   /gene="Spp2"
FT   exon            6215..6322
FT                   /number=4
FT                   /gene="Spp2"
FT   intron          6323..11201
FT                   /number=4
FT                   /gene="Spp2"
FT   exon            11202..11256
FT                   /number=5
FT                   /gene="Spp2"
FT   intron          11257..12350
FT                   /number=5
FT                   /gene="Spp2"
FT   exon            12351..12401
FT                   /number=6
FT                   /gene="Spp2"
FT   intron          12402..14546
FT                   /number=6
FT                   /gene="Spp2"
FT   exon            14547..14642
FT                   /number=7
FT                   /gene="Spp2"
FT   intron          14643..20285
FT                   /number=7
FT                   /gene="Spp2"
FT   exon            20286..20395
FT                   /number=8
FT                   /gene="Spp2"
FT   3'UTR           join(14633..14642,20286..20395)
```

Appendix B

```
ID   OAR544160  standard; RNA; MAM; 605 BP.
XX
HD * confidential 31-DEC-2005
XX
AC   AJ544160;
XX
SV   AJ544160.1
XX
DT   13-FEB-2003 (Rel. 74, Created)
DT   13-FEB-2003 (Rel. 74, Last updated, Version 0)
XX
DE   Ovis aries partial mRNA for secreted phosphoprotein 24 precursor (spp2
DE   gene)
XX
KW   secreted phosphoprotein 24 precursor; spp2 gene.
XX
OS   Ovis aries (sheep)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Cetartiodactyla; Ruminantia; Pecora; Bovoidea; Bovidae; Caprinae;
OC   Ovis.
XX
RN   [1]
RP   1-605
RA   Dalgleish R.;
RT   ;
RL   Submitted (12-FEB-2003) to the EMBL/GenBank/DDBJ databases.
RL   Dalgleish R., Department of Genetics, University of Leicester, University
RL   Road, Leicester, LE1 7RH, UNITED KINGDOM.
XX
RN   [2]
RA   Bennett C.S., Khorram Khorshid H.R., Kitchen J.A., Arteta D., Dalgleish R.;
RT   "Characterization of the human secreted phosphoprotein 24 gene (SPP2) and
RT   comparison of the protein with other species";
RL   Unpublished.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..605
FT                   /db_xref="taxon:9940"
FT                   /mol_type="mRNA"
FT                   /organism="Ovis aries"
FT                   /tissue_type="liver"
FT   CDS             <1..588
FT                   /codon_start=1
FT                   /gene="spp2"
FT                   /product="secreted phosphoprotein 24 precursor"
FT                   /protein_id="CAD66514.1"
/translation="LVIFVFGMNHWTCTGFPVYDYDPASLKEALSASVAKVNSQSLSPY
LFRAFRSSIKRVNALDEDSLTMDLEFRIQETTCRRESEADPATCDFQRGYHVPVAVCRS
TVRMSAERVQDVWVRCHWSSSSGSSSSEEMFFGDILGSSTSRNSHLLGLTPDRSRGEPL
YERSREMRRNFPLGNRRYSNPWPRARVNPGFE"
FT   sig_peptide     <1..45
FT                   /gene="spp2"
FT   mat_peptide     46..585
FT                   /gene="spp2"
FT                   /product="secreted phosphoprotein 24"
FT   3'UTR           589..>605
FT                   /gene="spp2"
XX
SQ   Sequence 605 BP; 145 A; 151 C; 169 G; 140 T; 0 other;
ttggtgatat ttgtctttgg aatgaaccac tggacctgta caggtttccc ggtgtatgac 60

tatgacccgg cttccctgaa ggaggctctc agcgcctctg tggcaaaagt gaattcccag 120

tcactgagcc cctatctgtt tcgagcattc agaagctcaa ttaaaagagt caacgccctg 180

gacgaggaca gcttgaccat ggacttggag ttcaggattc aagagacgac gtgcaggagg 240

gaatctgagg cagaccccgc cacctgtgac ttccagaggg gctaccacgt gcctgtggcc 300

gtttgcagaa gcaccgtgcg gatgtctgct gaacgcgtgc aggacgtgtg ggttcgctgc 360

cactggtcct ccagctctgg gtccagcagc agtgaagaga tgtttttttgg ggatatcttg 420
```

Appendix B

ggatcctcta catcaagaaa cagtcacctg cttggcctca ctcctgacag atccagaggt 480

gaaccgcttt atgaacgatc acgtgagatg agaagaaact ttcctcttgg aaataggagg 540

tactcgaacc cgtggcccag agcaagagta aaccctggct ttgagtgaca gccttgagca 600

aaatg 605