STUDIES ON THE MOLECULAR GENETICS

OF THE HUMAN FIBRILLAR COLLAGENS

A thesis submitted for the degree of Ph.D

University of Leicester

by

Baljinder Singh Mankoo

1988
To my family and friends,

for everything.
Some of this work has been published:


INDEX

CHAPTER 1. INTRODUCTION.

1.1 The extracellular matrix. 1
1.2 Collagen. 2
1.3 Collagen biosynthesis. 12
1.4 Collagen genes. 15
1.5 Control of gene expression. 23
1.6 Inherited collagen disorders. 29
1.7 The molecular basis of inherited collagen disorders. 32
1.8 Aims. 39

CHAPTER 2. MATERIALS AND METHODS.

2.1 Enzymes, antibiotics, chemicals and reagents. 41
2.2 Media and general DNA handling techniques. 42
2.3 DNA. 45
2.4 Plasmids, bacteriophage and bacterial strains. 46
2.5 Agarose gel electrophoresis. 47
2.6 Recovery of DNA from preparative gels. 48
2.7 Transformation of E. coli. 49
2.8 Plasmid DNA. 50
2.9 Preparation of radio-labelled DNA fragments. 53
2.10 Southern blotting. 55
2.11 Hybridisations. 56
2.12 Packaging extracts for cosmid cloning. 58
2.13 Cosmid cloning. 59
2.14 Screening human cosmid library H: A. 64
2.15 Preparation of λL47.1 arms for lambda cloning. 65
CHAPTER 3. DNA SEQUENCE ANALYSIS OF cDNA OF THE HUMAN TYPE III PROCOLLAGEN GENE.

3.1 Introduction. 82
3.2 Cloning of the human type III procollagen cDNAs. 83
3.3 Sequencing. 84
3.4 Sequence of type III procollagen cDNAs. 84
3.5 Analysis of the α-chain sequence. 86
3.6 Analysis of telopeptide and carboxyl-propeptide. 87
3.7 Evolution of the carboxyl-propeptide. 93
3.8 Analysis of 3' untranslated sequence. 94
3.9 A sequence anomaly. 95
3.10 Construction of a mini-library following fragment enrichment. 96
3.11 Discussion. 98
Appendix.

CHAPTER 4. ATTEMPTED COSMID CLONING OF THE HUMAN TYPE III PROCOLLAGEN GENE.

4.1 Introduction. 111
4.2 Preparation of cosmid DNA. 114
4.3 Preparation of insert DNA. 115
4.4 Screening of cosmid library. 120
4.5 Does the cosmid library contain type III collagen sequences? 124
4.6 Discussion. 125
CHAPTER 5. A VARIABLE LENGTH REPEAT SEQUENCE IN THE
COL1A1 GENE?

5.1 Introduction. 131
5.2 Isolation of the repeat sequence. 132
5.3 KP600 detects a polymorphic locus. 134
5.4 Cloning of the highly variable locus. 136
5.5 Characterisation of the clone \(\LambdaKP20.1\). 137
5.6 Mendelian inheritance. 138
5.7 Chromosomal assignment and localisation. 138
5.8 Discussion. 139

REFERENCES 144
ATP  adenosine triphosphate
RNA  ribonucleic acid
DNA  deoxyribonucleic acid
bp   base-pairs
kb   kilo base-pairs
OI   osteogenesis imperfecta
EDS  Ehlers-Danlos syndrome
MS   Marfan syndrome
ACKNOWLEDGEMENTS

I wish to thank Dr. Raymond Dalgleish for providing the facilities to enable the work contained in this thesis, and for his supervision of this work. I am deeply grateful to the many people in the Genetics Department who helped me both practically or with advice. In particular, I wish to thank Alain Blanchetot, John Brookfield, Terry Burke, Mick Pocklington, Melanie Price and Nicola Royle. Most importantly, I would like to thank everyone for providing such a warm and friendly atmosphere, at work and socially, during my sojourn at Leicester. I express my gratitude to the Marquis of Wellington, Redfearn, Black Horse, New Road Inn, Phoenix Arts, Leicester Haymarket, Jokespace, and Gill Harding's pottery classes for preserving my sanity. I thank Jan Svankmajer, Stephen Roche and the Super Complex Boys for much needed inspiration. To a moist-nosed, small-eared friend for being my technician for a week, thanks.

I am deeply indebted to Lysiane Schneider for typing up the final version of the thesis at such short notice. She managed with great skill to decipher the almost illegible scrawl of my handwriting.

I must acknowledge receipt of a Medical Research Council Studentship; however, the generosity of the Housing Benefit Office at Leicester City Council was required to keep my finances solvent.

During the initial period of my Ph.D I was a most surprised, but very grateful, recipient of the Charles Edward Layton Research Award from King's College, University of London, which I acknowledge with pleasure.

Baljinder Singh Mankoo
Strasbourg,
April 1988.
CHAPTER 1

INTRODUCTION
Most cells in multicellular organisms are in contact with a meshwork of interacting, extracellular macromolecules that constitute the extracellular matrix. It is now clear that the matrix not only holds the cells together in tissues, and tissues together in organs, but it also influences the development, polarity and behaviour of the cells it connects.

The macromolecules that constitute the extracellular matrix are secreted by local cells, especially fibroblasts, which are widely distributed in the matrix. In specialised matrix structures, such as cartilage and bone, these molecules are secreted by more specialised cells; for example, chondrocytes form cartilage, and osteoblasts form bone.

The matrix contains three major fibre-forming proteins - collagen, elastin and fibronectin - which are embedded in a hydrated gel formed by a network of glycosaminoglycan chains.

The term connective tissue is often used to describe the extracellular matrix plus the cells found in it, such as fibroblasts, macrophages, and mast cells. The amount of connective tissue in the body varies greatly: skin and bone are composed mainly of connective tissue, whereas the brain and spinal cord contain very little. Moreover, the relative amounts of the different types of matrix macromolecules and the way that they are organised within the extracellular matrix vary enormously; giving rise to a diversity of forms, each adapted to the functional requirements of the particular tissue. Thus, the matrix can become calcified to form rock-hard structures of bone and teeth, or it can assume almost crystalline order to form the transparent matrix of cornea; or it can take on the rope-like organisation of the collagen fibres in tendons, which gives them their high tensile strength.
1.2 COLLAGEN

Collagen is the major protein in the extracellular matrix of connective tissue. In fact it is the most abundant protein in the animal kingdom. A distinctive characteristic of collagen is that it forms insoluble fibres that have a very high tensile strength. In addition to its main function in the provision and maintenance of physical support for extracellular matrices, collagen has been acknowledged to have several physiological roles. The collagens have an extremely important role during embryonic development (Löhler et al., 1984). In the chick embryo, the collagens are produced in specific combinations during the growth and development of major connective tissues such as bone, cartilage and skin (Merlino et al., 1983). Collagen gene expression is also altered in a variety of human pathologic states, such as fibrosis and muscular dystrophy (Duance et al. 1980). Collagens are involved in chemotaxis, platelet adhesion and aggregation, and cell attachment to substrata. These properties have been extensively reviewed elsewhere (Bornstein and Sage, 1980) and will not be discussed further.

1.2.1 Collagen structure

The fundamental structural unit of collagen is tropocollagen, a long (300 nm), thin (1.5 nm diameter) protein that consists of three coiled polypeptide subunits called α chains, each of about 1050 amino-acids. The most important identifying criterion of collagen is the folding of a region of polypeptide chain to participate in a right-handed triple helix. Each α chain contains lengthy sequences of repeating glycine-X-Y units, in which X and Y positions are often occupied by proline and hydroxyproline residues respectively. The presence of the repetitive triplet structure is an absolute requirement for the formation of a semi-rigid triple-helical molecule. The triple helix is formed by hydrogen bonding between the amino group of a glycine residue in one chain and the carboxyl group of a proline or other residue in the X position of an adjacent chain. Glycine, being the smallest amino acid residue, fits
into the space where the three chains meet. The amino acid side chains are on the outside of the molecules and result in interactions which cause spontaneous aggregation of mature collagen molecules into microfibrils, which then assemble into higher order structures - fibrils and fibres. The fibrous elements may vary in diameter from 16-1500 nm depending on the tissue and location in which they are deposited. The molecular architecture of the fibres is best considered with reference to a microfibrillar unit, several of which may aggregate to form fibres of different diameter. X-Ray diffraction studies revealed that there are five molecules per unit cell in a fibril cross-section (Fraser et al., 1983). These five molecules are arranged circularly, but staggered axially by integral multiples of the distance D, 67 nm. As each molecule is 4.4D in length, this arrangement produces alternating overlap and gap regions on the surface of the microfibril. The length of an overlap zone is 27 nm and that of a gap zone is 40 nm (Miller, 1982). However, there are tissue-specific differences in the lateral packing of individual molecules into fibrils; and even the characteristic axial D period shows some variation (Bonar et al., 1985). This diversity in fibril structure undoubtedly relates to the specific mechanical and structural properties required for different connective tissues.

The packing arrangement of molecules within collagen fibrils imposes several requirements on those molecules. The molecules must be capable of assuming and maintaining a rod-like conformation. They must also contain regularly spaced interaction sites that permit lateral aggregation and specify the axial displacement between molecules. The molecules must have a minimum length of 268 nm to allow each molecule to interact with four additional molecules at integral multiples of the displacement distance D. Furthermore, they should be no longer than 340 nm, the maximum permissible length that can accommodate a gap zone of 40 nm.

On the basis of size and whether the molecules are capable of forming fibrous aggregates, Miller (1985) has separated the collagen family of proteins into three broad
groups. Group 1 comprises collagen types I, II, III, V and IX. These molecules consist largely of a single uninterrupted helical domain approximating 300 nm in length. Types I, II and III are incorporated as the principal constituents of collagen fibres, and are known as the fibrillar collagens. To date it is not known whether type V and K form fibrous aggregates. Although peptic fragments of type V collagen can form fibres \textit{in vitro} (Chiang et al., 1980) the constituent chains of the molecules deposited extracellularly are 1.5-2.0 times as large as the chains derived from peptic fragments (Narayanan and Page, 1983). Analysis of genomic clones of the \( \alpha_2(V) \) gene (Weil et al., 1987) indicate a close evolutionary relationship to the major fibrillar collagens.

Group 2 collagens are unlikely to form fibres. Type IV molecules have multiple non-helical regions. They have a length greater than 400 nm and tend to aggregate through end-to-end association of monomers. Type VI molecules consist of a 140 nm helical domain plus an extensive globular domain. Electron-microscopic studies suggest that these molecules form microfibrillar arrays by end-to-end association of tetrameric units (Furthmayer et al., 1983). Type VII molecules are 450 nm in length, all but a small proportion of this length is in triple-helical conformation. It has been suggested that type VII molecules form a collagen fibre by unstaggered lateral aggregation. Type VIII molecules seem to contain one or more nonhelical segments (Sage et al., 1983).

The nature of the aggregates formed by type VIII is unknown. The group 3 molecules are types IX and X. It is not known whether these short-chain collagens (less than 95 000 molecular weight) are capable of forming fibres.

The present study is confined to the molecular genetics of the fibrillar collagens, types I, II and III. The following discussion is directed principally to the structure and function of these collagens.

1.2.2 Structural features of fibrillar collagens

Collagen molecules are derived by extracellular processing of precursor (procollagen) molecules, and since procollagen molecules are formed through lateral alignment
Figure 1.1

A diagrammatic representation of the structure of the fibrillar procollagen molecule.

A representation of type I procollagen is shown. There are two pro-α1(I) and one proα2(I) chains in the molecule. For the other fibrillar procollagens the molecule is a homotrimer of only one type of α-chain. The amino and carboxyl propeptides are cleaved off when the molecule is secreted out of the cell. The mature molecule, consisting mainly of a triple helical domain, assembles into fibrils. Glc denotes glucose, Gal galactose, Man mannos, and GlcNac is N-acetylglucosamine.

This figure is adapted from that of Prockop et al. (1979).
and aggregation of three pro-\(\alpha\) chains, the structural features of the molecules are best understood with reference to the primary structure of individual pro-\(\alpha\) chains. It must be noted that whilst the term collagen refers correctly to the final product released into the extracellular matrix, it is often used in a more general sense to encompass the precursor, procollagen, molecules also. Hence, whilst the genes for collagen are formally referred to as procollagen genes, the term collagen is often used in place.

Practically all of the polypeptide of a fibrillar collagen is in a triple-helical configuration, except for a stretch of about a dozen residues at each end (telopeptides). The triple-helical region contains about 1000 amino acid residues, every third amino acid being glycine. The molecular formula of an \(\alpha\) chain can thus be approximated as (GLY-X-Y)\(\_3\).

The procollagen precursor molecules have more extensive non-helical propeptides at the amino and carboxyl ends of about 100 amino acid residues. The basic structure of a procollagen molecule is depicted in Figure 1.1. The amino-terminal propeptide has a molecular weight of about 20,000 and contains a globular amino-terminal domain, a central triple-helical domain, and another short globular domain. The carboxyl-terminal propeptide has a molecular weight of 30,000 to 35,000, and a globular conformation without any triple-helical domains.

The amino-terminal propeptides contain cysteine residues which form only intrachain disulphide bonds; whereas the carboxyl-terminal propeptide cysteines are involved in both intrachain and interchain bonds.

The carboxyl terminal telopeptide shows a variability in length between different types, and also for the same \(\alpha\)-chain in different species. Furthermore, type III collagen is unique in having two adjacent cysteine residues, at the carboxyl terminus of the helical domain, which can form interchain disulphide bonds (Seyer and Kang, 1981; Loidl et al., 1984).
1.2.3 Collagen types

Collagen types are, by definition, products of different genetic loci and are, therefore, nonallelic.

At least twelve different collagen types have been identified, and are encoded by at least twenty different genes. An outline of the biochemical and functional properties of each collagen type is presented below.

Type I collagen

Type I collagen is a heterotrimer of two \( \alpha_1(I) \) and one \( \alpha_2(I) \) chains. It generally forms well-structured, cross-striated fibres of substantial size, greater than 40 nm in diameter and several dm in length (Gay and Miller, 1983). The properties of the fibres are influenced by the degree of intra and intermolecular cross-linking, as well as orientation and density of the fibrils. Their disposition may be in parallel (tendons and bones) or cross-alignment (skin and organ capsules) depending on the location and functional demands. Its distribution indicates that type I collagen plays a key role as a supporting element of high tensile strength and limited elasticity. Type I constitutes 80-85\% of dermal collagen and is present to varying extents in virtually all major connective tissues. It is the only collagen normally present in bone. Interestingly, type I collagen mRNAs are transcribed in differentiated chondrocytes, albeit low levels, though these cells do not synthesise type I collagen (Kosher et al., 1986). That type I collagen is essential to embryonic development has been demonstrated by the insertion of a Mov-13 retrovirus into the \( \alpha_1(I) \) collagen gene of chick embryos (Löhler et al., 1984).

Type II collagen

Type II collagen is unique among the abundant collagens in showing a high degree of tissue specificity. It is found essentially only in cartilage, and certain portions of the eye (von der Mark et al., 1977). It is initially expressed in specialised embryonic cells that have chondrogenic potential. It is important not only for those structures that remain as
cartilage but also for bone formation because hypertrophic cartilage forms the framework onto which the bone matrix is deposited during ossification, which occurs in the foetus.

Type II collagen makes up 40% of the protein in the cartilage matrix. Nonexpression of type II collagen leads to lethal newborn achondrogenesis, an indication that this collagen is essential to normal development (Eyre et al., 1986).

A cartilage matrix protein, chondrocalcin, has been identified as the carboxyl-propeptide of type II collagen. It has been suggested that it may play a role in endochondral ossification once cleaved from the type II procollagen molecule (van der Rest et al., 1986).

Type III collagen

Type III collagen is found in most tissues that also synthesise type I except bone; and makes 10-15% of total collagen in the human body. Skin fibroblasts synthesise types I and III in a 5 to 1 ratio (Breul et al., 1980). Type III is prevalent in the more distensible tissues, such as the uterine wall, blood vessels, gastrointestinal tract, lung, liver and lymphoid organs. It forms a loose reticular network which provides support without inhibiting rapid diffusion and metabolic exchange.

The essential functional role of type III collagen is illustrated by patients suffering from Ehlers-Danlos syndrome type IV, who have defective type III collagen production (Pope et al., 1975; Pyeritz et al., 1984; Stolle et al., 1985).

Plasma von Willebrand factor, which mediates platelet adhesion, is activated by binding to type III collagen of the subendothelium (Roth et al., 1986).

Type IV collagen

Type IV collagen is the main collagenous component of basement membrane. The molecule is 400 nm in length and has a noncollagenous globular, NCI, domain at its carboxyl terminal. The molecule is comprised of $\alpha_1$(IV) and $\alpha_2$(IV) chains each of approximately 1700 amino-acids residues (Dixit
Type IV molecules form a network by end-to-end aggregation. Four molecules are bound to one another through their amino-terminal 30 nm-long triple-helical segments, known as the 7S domain. At their opposite ends the carboxyl-terminal globular NCI domains of two molecules are joined (Duncan et al., 1983). Stabilisation of this alignment is achieved by intermolecular cross-links, including reducible disulphide bridges and non-reducible, presumably lysine-derived, bonds located in the 30 nm overlap zone (Bailey et al., 1984). In contrast to the fibrillar collagens, the helical regions of the α1(IV) and α2(IV) chains are frequently interrupted by non-helical segments (Oberbaumer et al., 1985; Killen et al., 1987). The non-helical areas appear to be important structural features of the triple helix, necessary for the regulation of not only the flexibility of the molecules, but also the elastic properties of the macromolecular network of collagen IV.

The NCI globular domain has autoantigenic properties. Injection of rabbit antiserum against mouse NCI into mice produced pathologic changes similar to those involved in the human autoimmune disorder, Goodpasture's syndrome (Wieslander et al., 1984). Patients with this syndrome suffer from lung bleeding and renal complications due to antibodies that react preferentially with glomerular and alveolar basement membranes.

Type V collagen

Surveys of the location of type V collagen were first made by chemical extraction (Bornstein and Sage, 1980) and then by immunolocalisation at the light and electron microscope levels. It can be concluded that it occurs widely as a minor component, forming a fine filamentous mesh over the plasmalemma of smooth muscle cells and is synthesised in small quantities by other connective tissues. Human placental type V collagen has been resolved into two fractions, one containing α1(V) and α2(V) chains in a 2:1 ratio, and the other containing α1(V), α2(V) and α3(V) chains in equimolar ratios (Niyibizi et al., 1984). In the human amnion, type V collagen is a 12 nm unbanded fibril which
extends from the lamina densa of the basal lamina into the adjacent interstitial matrix. It has been proposed (Modesti et al., 1984) that type V functions as a network of anchoring fibrils between the cell basal lamina and the extracellular matrix, particularly with type I collagen fibres. In contrast to the fibrillar collagens in which only very small nontriple helical domains (telopeptides) are retained in the fully processed molecules, type V collagen retains a globular domain at one end of the triple helix (Broek et al., 1985).

A fourth chain for type V collagen, α4(V) has been reported to exist in tendons only (Fessler et al., 1985). This finding has yet to be confirmed. If it does exist, this raises the possibility that different type V molecules might be adapted to serve a particular ultrastructural function which is used in a particular tissue.

Type VI collagen

Type VI collagen is widely distributed throughout connective tissues. It represents a highly disulphide-linked protein containing relatively short triple-helical domains, contributing to one-third or less of the total mass of the protein (Furthmayer et al., 1983). It is composed of three different polypeptide chains α1(VI), α2(VI) and α3(VI), which are disulphide-linked to form large aggregates (Trüeb and Winterhalter, 1986). A model has been proposed for the assembly of type VI collagen (Engvall et al., 1986), in which monomers assemble intracellularly into tetramers, and these associate extracellularly into fibrils by end-to-end contact.

Little is known about the biological function of type VI collagen.

Type VII collagen

Type VII collagen is a homopolymer of three identical chains which are genetically distinct from other collagen chains. The triple helical domain of 424 nm is 1.5 times the length of type I collagen (Bentz et al., 1983). Ultrastructural studies suggest that type VII collagen molecules associate in an overlapping, antiparallel fashion to form covalently stabilised dimers (Morris et al., 1986).
It has been postulated that type VII collagen is a major component of anchoring fibrils (Keene et al., 1987). Anchoring fibrils are accessory structures to the epithelial basement membranes of several tissues which are subjected to considerable frictional and torsional forces. These include skin, oesophagus, cornea, vaginal mucosa and chorioamniotic membranes.

**Type VIII collagen**

Originally designated endothelial collagen, due to its identification in aortic endothelial cells (Sage et al., 1980), it has subsequently been identified as a product of nonmuscular and tumour cells of diverse origin (Alitalo et al., 1983). Type VIII is the predominant collagen synthesised by bovine Descemet's membrane, a morphologically unique basement membrane in the cornea of the eye (Kapoor et al., 1986).

Benya and Padilla (1986) have proposed a structure in which type VIII collagen is composed of three α-chains of 61 000 molecular weight, arranged in a predominantly helical structure with non-helical domains at each end. These chains are stabilised by acid labile, SDS-labile, noncovalent cross-links in the larger nonhelical, NCI, domain. No interchain disulphide cross-links are present. This model is based on sedimentation analysis, and contradicts with an earlier model proposed by Sage et al. (1983) in which type VIII collagen consists of one continuous peptide of 550 000 molecular weight, with non-helical domains at one-third and two-third the length of the molecule.

**Type IX collagen**

While type II collagen forms the fibrillar network of cartilagenous tissues, accounting for over 90% of the total collagen in those tissues, type IX collagen accounts for 1-2% of the collagen in adult hyaline cartilage and 10% in foetal cartilage (Eyre et al., 1987).
Type IX is an unusual molecule of three different chains that consist of three triple helical domains joined by short non-triple-helical sequences (van der Rest et al., 1985; Ninomiya et al., 1985; Mayne et al., 1985). At the amino and carboxyl ends are noncollagenous domains which do not appear to be homologous to amino and carboxyl propeptides of fibrillar collagens.

Studies by immunofluorescence and immunoelectron microscopy revealed that type IX collagen is associated with type II fibrils, but essentially restricted to intersections of the fibrils (Muller-Glauser et al., 1986). Those workers suggest the three independent triple helical domains allow the protein to interact with two or more type II fibrils. Through such interactions type IX collagen would profoundly influence the material and properties of cartilage; the presence of type IX could prevent the lateral growth of type II fibrils.

Type X collagen

Type X is the first example of a transient, developmentally and topographically regulated collagen (Ninomiya et al., 1986). In the developing chick embryo type X collagen is synthesised by chondrocytes from regions of hypertrophy and not by chondrocytes from other regions (Castagnola et al., 1986; Reginato et al., 1986). This collagen may be involved in the extracellular matrix remodelling associated with the initiation and progression of tissue calcification.

The collagen is synthesised as a procollagen of three identical \( \alpha \) chains. Nucleotide sequence studies have identified three domains: a large non-collagenous carboxy-terminal propeptide of 120 amino acids, a short (132 nm) triple helical region of 460 amino acids, and a very short non-collagenous amino-terminal propeptide of 42 residues (Ninomiya et al., 1986). The same workers suggest that the 3' end of the molecule contains a transmembrane segment.

Disulphide-bonded type X collagen is present in bovine cartilage (Bashey et al., 1985; Ayad et al., 1987); the mature molecule in chicken does not contain disulphide links.
Type XI collagen

This collagen is also known as type K (1α, 2α, 3α). Chemical data on this collagen has been acquired largely from the peptic fragments isolated following limited proteolysis; more recent studies have employed chromatographic analysis of cross-linked chains (Morris and Bächinger, 1987).

The chains are approximately the same length as those of fibrillar collagens. The 3 chain is identical to the α1(II) chain (Furuto and Miller, 1983); while the 1α and 2α chains are similar to, but distinctly different from, type V chains (Burgeson et al., 1982). The functional role of type XI collagen is unknown.

Type XII collagen

A cDNA clone encoding a polypeptide similar to, but different from, type IX collagen has been designated as encoding type XII (Gordon et al., 1987). The same workers isolated genomic clones that have a similar intron/exon structure to the type IX gene. Whereas α2(IX) RNA is only detected in sternal cartilage, α1(XII) RNA is present in calvaria, tendon, cornea of chick embryos, and sternal cartilage. Pepsin-derived collagenous fragments of type XII collagen substantiate the similarity to type IX chains (Dublet and van der Rest, 1987).

1.3 Collagen Biosynthesis

The literature on collagen biosynthesis is extensive. Several reviews have examined this subject in detail (Fessler and Fessler, 1978; Prockop et al., 1979; Kivirikko and Myllyla, 1985) and so only an outline is presented.

The synthesis of collagen can be considered as occurring in two stages: intracellular steps are required to assemble
and secrete the procollagen molecules; and extracellular processing converts the procollagen molecules to collagen which is incorporated into stable, cross-linked fibrils. Type I collagen has been the most studied in this respect and is used as a model system. Almost all the specific processing events demonstrate an unusual relationship to the conformation of the protein being modified. To serve as substrates for the five intracellular enzymes - the three hydroxylases and the two transferases which add sugars to the hydroxylysine residues - the pro-α chains must be non-helical. Modification by these enzymes ceases when the protein folds into a triple helix. Once folded, the procollagen is secreted, and the time required for folding directly affects that required for secretion. Procollagen NH₂-terminal proteinases cleave the secreted proteins only if they are folded into the triple helix, and lysyl oxidase only acts on collagen after it has become incorporated into native-type fibrils. The nascent procollagen chains have signal sequences at the amino-terminal end. These signal peptides direct the movement of pro-α chains into the rough endoplasmic reticulum where all intracellular modifications take place. The signal peptides are cleaved off during or shortly after translocation across the membrane.

Most of the hydroxylation of proline and lysine residues, and glycosylation of hydroxylysine residues, occur as cotranslational processing, but continue as post-translational modifications until triple helix formation of the pro-α chains prevents further processing. Hydroxyproline plays an important role in the stabilisation not only of the triple helical collagen molecule, but also of this assembly into microfibrils. The 4-hydroxyprolyl residue often occurs in the Y position of GLY-X-Y triplets, but never in the X position. This allows better fitting of the triple-helices (Nemethy and Scheraga, 1986). Hydroxylation is catalysed by three enzymes: prolyl 4-hydroxylase, prolyl 3-hydroxylase and lysyl hydroxylase, all three hydroxylating proline or lysine in peptide linkage. All three enzymes require Fe²⁺ as a cofactor.
Glycosylation is catalysed by two specific enzymes; first, hydroxylysyl galactosyltransferase which transfers galactose to some of the hydroxylysine residues, and then galactosyl hydroxylysyl glucosyltransferase which transfers glucose to some of those galactosylhydroxylysine residues. The propeptides, unlike the triple helical domain, also contain asparagine-linked carbohydrate units. The asparagine residue being present in the sequence Asn-Ile-Thr in the four human fibrillar collagen carboxy-propeptides.

The propeptides of collagens contain both intrachain and interchain disulphide bonds. The enzyme responsible for this is protein disulphide isomerase (Myllyla et al., 1983). The function of the disulphide bonds is to provide covalent linkages that stabilise the molecules. A further important function of interchain disulphide bonds between carboxy-terminal propeptides is to facilitate triple helix formation. These propeptides direct the association of the three prochains into a procollagen molecule, and so the formation of interchain disulphide bonds is an absolute requirement for the formation of a procollagen molecule. The rate of triple helix formation is limited by the cis-trans isomerisation of peptide bonds (Bächinger, 1987). The procollagen molecule is then packaged in the Golgi apparatus and secreted.

Conversion of procollagen to collagen occurs by the enzymatic removal of the amino- and carboxyl-propeptides by specific proteases. There are several isozymes of the aminoprotease which cleave different procollagen types.

The collagen molecules spontaneously assemble into fibrils in the case of the fibrillar collagens, or other ordered structures for non-fibrillar collagens. By using enzymatically generated collagen, it has been demonstrated that fibril formation is an entropy-driven polymerisation process similar to that of actin and flagellae (Kadler et al., 1987).

The fibrils are cross-linked by a series of covalent bonds. This occurs in two or three steps. At first, oxidative deamination of ε-amino groups in certain lysyl and hydroxylysyl residues occurs to yield reactive aldehydes. These can subsequently form two major kinds of cross-links.
One kind is an intra-molecular cross-link that joins the chains of the same molecule and is formed by aldol condensation of two of the aldehydes. The second kind involves condensation between such an aldehyde and an -amino group of an, as yet, unreacted lysine, hydroxylysine or glycosylated hydroxylysine. It is this cross-linking of molecules which provides the characteristic tensile strength of collagen.

1.4 COLLAGEN GENES

As discussed above, the collagens provide the extracellular framework responsible for the structural integrity of nearly every organ and tissue in vertebrates. The correctly programmed spatial and temporal regulation of production of collagen molecules is critical for growth and development. Inhibition of pro-α 1(I) collagen expression, for example, by retroviral mutagenesis results in lethal development arrest in early mouse embryos (Jaenisch et al., 1983; Schnieke et al., 1983, Löhler et al., 1984). A number of genetic disorders in man seem to be related to the disfunction of collagen genes or their products (reviewed by Prockop and Kivirikko, 1984; Cheah, 1985), these will be discussed later. The analysis of the collagen gene family at the molecular level is consequently an important prerequisite for understanding their expression in diseased as well as normal, states.

In vertebrates there are at least twelve geneticaly distinct proteins which are coded for by at least twenty genes. The genes of those collagen types so far studied have been shown to be single copy by, for example, dot-blot hybridisation (Dalgleish et al., 1982) and pedigree analysis (Chu et al., 1985) for the human pro-α2(I) collagen gene; and segregation analysis of restriction fragment length polymorphism for all four fibrillar collagen genes. The genes encoding the human collagen are dispersed in the genome. The pro-α1(I) gene, COL1A1, is on chromosome 17 (Huerre et al., 1982), the pro-α2(I) gene, COL1A2, is on chromosome 7 (Junien
et al., 1982; Solomon et al., 1983). The pro-\(\alpha_1\)(II) gene, COL2A1, was at first incorrectly assigned to chromosome 7 (Weiss et al., 1982), but later studies locate it on chromosome 12 (Huerre-Jeanpierre et al., 1986; Solomon et al., 1985).

The pro-\(\alpha_1\)(III) gene, COL3A1, and the pro-\(\alpha_2\)(V) gene, COL5A2, have been assigned to 2q 24-31 (Solomon et al., 1985; Emanuel et al., 1985). This is interesting as types III and V are usually coexpressed (Bornstein and Sage, 1980). Recently the pro-\(\alpha_3\)(VI) gene has been localised to 2q 37 (Weil et al., 1988). Thus five extracellular matrix genes, including elastin and fibronectin, have been mapped to the distal region of the long arm of chromosome 2. Weil et al. (1988) have also mapped the genes COL6A1 and COL6A2 to 21q 223. The genes for the pro-\(\alpha_1\)(IV) and pro-\(\alpha_2\)(IV) chains have been localised to within 400 kb of each other on chromosome 13 (Emanuel et al., 1986; Killen et al., 1987; Boyd et al., 1988). It is interesting that, although physically unlinked, expression of the fibrillar collagens is coordinated in embryonic development (Merlino et al., 1983). This is even more surprising as different embryonic cells must be responsible for synthesising specific dominant collagen types (I, II, or III in osteoblasts, chondroblasts, and embryonic dermal fibroblast cells, respectively). There is also coordinate expression during the in vitro differentiation of myoblasts. During chondrocyte differentiation there is a strict temporal regulation of cartilage collagen gene expression. There is an initial stage where type I collagen synthesis is replaced by type II and type IX expression; and a later stage characterised by expression of high levels of type X collagen (Castagnola, 1988).

The collagen genes have the most highly interrupted structure so far known. This was shown dramatically by electron microscopy of genomic clones hybridised to procollagen mRNA (Ohkubo et al., 1980). Portions of different collagen genes from a variety of species have been isolated and the intron/exon structure determined. That of the chick COL1A2 gene was the first to be determined and is 39 kb long, divided into 52 exons (Vogeli et al., 1980; Wozney et al.,
1981). The corresponding human gene is of similar size and also has 52 exons (de Wet et al., 1987). The human COL2A1 and COL3A1 genes are 30 kb and 38 kb in size respectively (Cheah et al., 1985). The size of the human COL3A1 gene is by estimation from the corresponding chicken gene. The human COL1A1 gene has 51 exons, but is only 18 kb in length, the smaller size of the gene is due to presence of shorter introns (Chu et al., 1984).

A distinctive feature of these genes is the presence of 45, 54, 99 and 108 bp exons, which encode multiples of the GLY-X-Y triplet, in the regions which code for the triple helical domains of all fibrillar collagens (Yamada et al. 1980, 1984; Solomon and Cheah, 1981; Wozney et al., 1981; Monson et al., 1982a). This observation led to the suggestion that the ancestor for the collagen genes had arisen by amplification of a genetic unit containing an exon of 54 bp embedded within intron sequences (Yamada et al., 1980). They suggested that exons which are shorter or longer than 54 bp presumably arose by recombination within exons. The only permissible additions or deletions being of 9 bp, or multiples thereof, this constraint being imposed by the absolute necessity to maintain the repeating GLY-X-Y structure in the protein.

A model of a primordial collagen gene which duplicates to form one chain would predict that internal homologies be found within that chain; provided, of course, that the sequence has been sufficiently conserved for these to be detected. A comparison matrix analysis on the amino acid sequence of the pro-α1(I) chain revealed significant regularities, occurring at a distance related to the stagger between adjacent triple helical molecules in the collagen fibre (McLachlen, 1976). These occur at a distance, D, of 234 amino acids, 78 triplets. McLachlen suggested that, if gene duplication has occurred, the repeated domain is likely to be length D, or a simple fraction of D. Analysis of the α1(I), α2(I) and α1(III) chains by others has revealed, not only the D repeat, but also smaller repeats of D/3, D/6, D/11 and D/13 (Hofman et al., 1980). This latter result is more significant as the repeats are defined mainly by interactive residues and

17
may better reflect the original genetic unit, such residues have not been exposed to selection pressure directed towards improved aggregation of the molecules into fibrils. Emerging from these data also is the fact that the homologies within one repeat to another are more conserved across species than from one repeat to another within a chain. The most likely explanation for this is early intrachain duplication with subsequent duplication of the whole chain.

The proposal by Yamada et al. (1981), that an ancestral 51 bp gene was duplicated to produce the present structure was based on partial sequencing data, only 54 bp exons were observed. Later studies revealed that only a proportion of exons are of 54 bp length. One half of the exons of the α2(I) chain triple helical region code for 15, 33 or 36 residues (Wozney et al., 1981). They suggested that the primordial gene arose by a series of homologous recombinational events within coding sequences and that introns were inserted at a later date. Support for this idea comes from the fact that most of the introns are unique, only a few contain moderately repeated sequences (Tate et al., 1983). Recombination between introns, as required by Yamada's model, would be infrequent. However, a recent study of patient suffering from osteogenesis imperfecta indicated a deletion of three α1(I) exons, the endpoints of the mutation being within two introns and bounded by two short, inverted repeats (Chu et al., 1985a). Further evidence that recombination events within intron have occurred during the evolution of collagen genes comes from the presence of a helical coding sequence in intron 2 of the chicken COL1A2 gene (numbering from the 5' end of the gene), nine GLY-X-Y triplets are interrupted by apparent insertions of 10 and 2 nucleotides. Intron 1 of this gene contains an inverted coding sequence of six GLY-X-Y triplets interrupted by a single nucleotide, this vestigial sequence also specifies both ochre and stop codons. Neither vestiges is flanked by splice junction sequences (Manson et al., 1982a). A sequence coding for three GLY-X-Y repeats and ending with a split codon is also found in an intron of the human pro α1(II) gene (Sangiorgi et al., 1985b). These repeats are flanked by acceptable variations of the intron
splice sequences. Its amino acid sequence, however, does not correspond to that in the rat $\alpha 1$(II) amino terminal propeptide (Kohno et al., 1984), and translation of this sequence would put the reading frames of both neighbouring exons out of phase.

Exons of the triple helical regions of the chick pro $\alpha 2$(I) gene shows an alternating pattern of large and small exons (Wozney et al., 1981; Tate et al., 1982). This pattern is absent in the 5' end of the helical coding region of this gene. A similar organisation is emerging for the other fibrillar collagen genes: pro $\alpha 1$(I) (Chu et al., 1984), pro $\alpha 1$(II) (Sandell et al., 1984) and pro $\alpha 1$(III) (Yamada et al., 1984; Chu et al., 1985b). It is also apparent that equivalent amino acid segments in the helical domain of each polypeptide are encoded by exons of equal size in the different genes. As yet only one major difference has been identified, the presence of a 108 bp exon in the pro $\alpha 1$(I) gene encoding residues 568-603, which in the genes of the other fibrillar collagens are encoded by two 54 bp exons (Chu et al., 1984). The occurrence of a fusion exon in the pro $\alpha 1$(I) gene must have occurred after duplication of the fibrillar collagen loci about 200 to 500 million years ago.

The fixation in size distribution of exons is found despite the presence of elements that should favour recombinational rearrangements, and contrasts with the pattern of sizes of exons encoding the propeptides of these exons (Ramirez et al., 1985). That uniformity in exon size pattern was achieved more recently by gene conversion is unlikely, because the sequences of a given exon are, in general, more analogous to the equivalent exon of homologous genes than paralogous genes. The complete divergence of intron sequences and the presence of repetitive sequences at different locations (Myers et al., 1983; Chu et al., 1985c) in these genes are consistent with this proposition. The fixation of the size of exons may have corresponded to the acquisition of an optimal length for the collagen molecules. A change in size of the helical region of the collagen polypeptide could not be tolerated, although changes in the sizes of the propeptides took place. This probably occurred due to the
highly structured fibrils of the interstitial collagens, which result from lateral associations of molecules in a regular and repeating pattern. Shorter or larger polypeptide chains which might be generated by recombination would disrupt the formation of these fibrils. In this context, it is interesting to note that the invertebrate collagens (Monson et al., 1982b; Cox et al., 1984) and vertebrate types IV, IX and X (Sakurai et al., 1986; Lozano et al., 1985; Ninomiya et al., 1986) have no 54 bp coding unit. In fact the type X gene contains a single open reading frame of 2000 bp without introns. The molecules produced from these genes do not form fibrils by lateral association of triple helical domains. There is, therefore, no selection against interruptions of the triple helix region and of any size changes. If the presence of introns in fibrillar collagen genes is to prevent recombinational rearrangements, and so maintain the precise length and distribution of charged and hydrophobic residues along the chains, it might be predicted that such events occur more frequently in the genes of the non-fibrillar collagens. If this were true, one would expect greater diversity among each non-fibrillar collagen type of different animal species than among fibrillar collagens.

Gilbert (1979) proposed that introns separate gene sequences encoding different conformational or functional domains, thereby allowing their independent evolution. As relics of the recombination process that brought the exons together, the introns would be long, random sequences that would drift rapidly in sequence and size since the last act that assembled the genes. In line with this proposition, different functions have been assigned to various exons of the triple helical region (Chu et al., 1984). Exon 12, for example, of the COL1A2 gene, encoding amino acid residues 760 to 780, has a GLY-ILE collagenase cleavage site and a putative fibronectin binding site. Exons 7 and 41 both have a core sequence of GLY-X-LYS-GLY-HIS-ARG which may be involved in interchain cross-linking with amino and carboxy terminal lysines. It is unlikely that every exon encodes a functional domain. This depends though on the criteria for
defining a functional domain; for example, charge
distribution along the length of the molecule, which could be
involved in intermolecular interactions, would be a function
that might not be obvious.

The presence of collagen genes, both vertebrate and
invertebrate, which contain so many introns as those of
fibrillar collagens, and the presence in fibrillar collagen
genes of "junction" exons, in both helical and propeptide
regions, has led to the suggestion that introns were inserted
in the fibrillar collagen genes in more recent evolutionary
times (Benviste-Schrode et al., 1985).

One problem this idea creates is, how were introns
inserted into the triple helical region without disrupting
the GLY-X-Y triplet? Indeed, all exons of these regions are
exact multiples of 9 bp and exon/intron junctions do not
disrupt any codons. In contrast, the mouse COL4A1 gene not
only has exons of sizes other than 54 bp, but every exon
examined contains a glycine codon split into two adjacent
exons at either one or both ends (Sakurai et al., 1986).
Since the glycine codon is GGX, the split between the first
and second bases could easily generate new splice signals.
The same phenomenon is seen in the human COL4A2 gene
(Soininen et al., 1986).

One recently reported observation was that expression of
gene constructs in transgenic mice was increased 10- to 100-
fold in transgenic mice, but not in cultured cells (Brinster
et al., 1988), when the constructs contained introns within
coding regions. The authors suggest the possibility that
introns contain sequences important during development,
either by regulating transcription directly or by affecting
the phasing of nucleosomes. At present, therefore, the reason
for the complex gene structure of the fibrillar collagens
remains unresolved.

Biochemical and molecular studies have indicated that
the amino-terminal propeptide region of the fibrillar
collagen genes exhibits a high degree of structural variation
among different chains as well as in the same molecule of
different species (Bornstein and Sage, 1980; Ramirez et al.,
1985).
The region can be subdivided into four distinct components: the signal peptide, an amino-terminal globular region, a triple helical segment and a short non-helical part. The latter is only five amino acid residues length in the chick COL1A2 gene (Tate et al., 1983). The signal peptides of all fibrillar pro-α chains are of the same size and have the sequence Met-X-Ser-Phe-Val-Asp. The second residue is different in various chains, but is identical in the same chain among different species (Chu et al., 1985c).

The amino-terminal propeptide remains attached to the triple-helical part of the molecule during triple helix formation, secretion out of the cell and during the initial stages of fibril formation (Fleischmajer et al., 1981). In dermatosporaxis the amino-terminal propeptide is not cleaved and the skin of affected individuals is hyperextensible and fragile. Both the pro-α 1(I) and pro-α 2(I) chains have a short globular domain of 86 amino acid residues (Chu et al., 1984). The pro-α 2(I) and pro-α 1(II) chains both lack the cluster of cysteine residues present in the globular domain which, in the other two fibrillar chains, are believed to participate in intra- and inter-chain disulphide bonds (Kohno et al., 1984). These authors also report that the chicken pro-α 1(II) chain collagenous domain, which is much longer than that of the other α chains, is interrupted once between residues 44 and 47.

Regardless of these structural differences, at the gene level all four amino-prepropeptides are encoded by six exons (Tate et al., 1982). The transitions between the four subdomains, and the amino-terminal propeptide and signal peptide, as well as the large triple helical domain, are encoded by junction exons (Chu et al., 1984). If as is proposed by Blake, Gilbert and others (see Gilbert, 1985 for review) that exons encode different functional domains of a protein, the junction exons may have resulted from the elimination of introns. There being no requirement for introns to stabilise the precise length of the coding
regions. Alternatively, these junction exons may encode less obvious functional domains. For example, the human pro-\(\alpha 1(I)\) and chickens pro-\(\alpha 2(I)\) signal peptidase cleavage sites are in the first exon (Chu et al., 1984; Aho et al., 1984). Similarly, the exons which overlap the 3' end of the amino-terminal propeptide and the telopeptide encode the propeptidase cleavage site (Dickson et al., 1985). In the rat COL2A1 gene, the sequence Gly-Pro-Pro is repeated four times in succession in exon 4, Kohno et al., (1984) proposed that triple helix formation may be initiated from this region, although this function is usually assigned to the carboxyl-propeptide.

As originally noted for the chicken COL1A2 gene (Wozney et al., 1981), the carboxyl-terminal propeptide of all fibrillar procollagenes is divided into four exons encoding distinct functional or conformational segments. The sizes of all four exons are highly conserved between all four genes, despite considerable variation in the size of introns and the 3' untranslated sequence (Ramirez et al., 1985). One paramount feature of the fibrillar collagen genes is the phylogenetic retention of the location of eight cysteines and their surrounding nucleotide sequences in the four exons of the carboxyl-propeptide. The first four of these cysteines are believed to be involved in interchain disulphide bonding, whereas the last four are required for intrachain linkage.

There are several lines of evidence that suggest the carboxyl-propeptides play a fundamental role in the selection and correct alignment of the pro-\(\alpha\) chains to form the procollagen molecule. The evidence and its relationship to the organisation of these genes is discussed later, in Chapter 3.

1.5 CONTROL OF GENE EXPRESSION

A variety of observations suggest that expression of the fibrillar collagen genes is subjected to intricate regulatory events at the transcriptional and post-transcriptional levels.
Modulation in the level of expression of the fibrillar collagen genes has been reported in a variety of developmental situations. The RNA levels for type I, II and III collagen show a 10-fold increase during development of chick embryos (Merlino et al., 1983). The expression of types I and III procollagen genes in myoblasts differentiating into myotubes demonstrate either a 15-fold or 10-fold increase in the levels of mRNAs. Once myoblast fusion is complete, both collagen types show a levelling off in mRNA synthesis (Gerstenfeld et al., 1984). Type I procollagen synthesis is turned off during chondrogenesis and replaced by synthesis of type II procollagen (Castagnola et al., 1988; Kosher et al., 1986). However, substantial amounts of type I collagen mRNAs are present in the cytoplasm of well-differentiated chondrocytes (Kosher et al., 1986; Focht and Adams, 1984), indicating that control of type I collagen gene expression is exerted at both transcriptional and post-transcriptional levels. Allebach et al. (1986) observed that virally transformed chick chondrocytes accumulate type I collagen mRNAs that are translated efficiently in vitro, but are not translated in the intact cells. Miskulin et al. (1986) reported that expression of the four fibrillar collagen genes is coordinately controlled during proliferating and nonproliferating growth of human fibroblasts. Liau et al., (1985) observed that the type I and III procollagen genes are coordinately expressed in mouse embryo fibroblasts transformed by Rous sarcoma virus. The same was found in various NIH3T3 murine cell lines transformed by members of the human ras oncogenes. However, in BALB3T3 Rous sarcoma virus transformed murine cell lines, there was a 4-fold increase in type III procollagen gene expression and a 8-fold decrease in type I procollagen mRNA levels. It is possible, that this discrepancy is due to amplification of the type III gene, translocation of gene sequences, or alterations to regulatory sequences. Studies on the expression of fibrillar collagen genes in cells after exposure to various compounds - such as ascorbic acid (Lyons et al., 1984), glucocorticoids (Walsh et al., 1987), growth factors (Ignoff et al., 1987), and cytokines (Goldring et al., 1987) - all reveal that
collagen synthesis is subject to control at the transcriptional level.

A number of approaches have been employed to study the regulation of transcription of the fibrillar procollagen genes. These include analysis of methylation patterns of the genes in cells with varying rates of collagen synthesis, comparison of nucleotide sequences of promoter and 5'-untranslated sequences of the genes, and localisation of nuclease hypersensitive sites. More recently, these studies have been extended to include functional deletion analysis of promoter sequences. The results of these analyses are discussed below.

In the nuclear DNA of higher eukaryotes, the sequence pCpG generally occurs at a much lower frequency than predicted by a random distribution of nucleotides in a sequence of a given base composition (Razin and Riggs, 1980). This is due to the deamination of cytosine, which can give rise to thymine which in turn escapes DNA repair mechanisms. The level of methylation of cytosine in these pCpG sequences within genes is generally inversely related to expression of the gene. However, in the case of the chicken COL1A2 gene, McKeon et al. (1982) found that a 5 kb DNA segment around the start site of transcription to be unmethylated in all tissues examined, whether or not these synthesised type I collagen. The central and 3' region of the gene was found to methylated to about the same extent in these same tissues. Fernandez et al. (1984) found no difference in the methylation state of the COL1A2 gene in chondrocytes or dedifferentiated chondrocytes. The 3' end of the COL2A1 gene has more methylation in tissues such as fibroblasts and erythrocytes, which do not synthesise this collagen chain, than in actively transcribing chondrocytes (Fernandez et al., 1984). These workers propose a hypothesis that hypermethylated collagen genes are not transcribed while demethylated collagen genes may or may not be expressed. These data were restricted to the 3' end of the COL2A1 gene; and it must be noted that differentiated chondrocytes, while not synthesising type I collagen, do transcribe type I procollagen genes at a high level (Focht and Adams, 1984;
Kosher et al., 1986). Parker et al. (1982) reported that a decrease in type I collagen synthesis in SV40-transformed human fibroblasts is associated with hypermethylation of the type I collagen genes. This contrasts with the finding that the methylation pattern of the COL1A2 gene is not altered in chick embryo fibroblasts after transformation by Rous sarcoma virus; despite the greater than 10-fold decrease in type I collagen mRNA synthesis in transformed cells (McKeon et al., 1982). It is possible that transformation of human and chick cells by DNA tumour viruses may result in different methylation pattern of collagen genes.

At present, therefore, the role of methylation in regulation of collagen gene expression is unclear. However, it is becoming clear that pCpG "islands", of 1-2 kb length, are often associated with the regions where transcription of genes begins (for review see Bird, 1987). One model proposed to explain the function of these islands suggests that they are rendered constitutively available to ubiquitous nuclear factors through an altered chromatin structure (Bird, 1986). Bound factors would deny access to the methylase. One consequence of this is that tissue-specific genes with islands would be continuously available to transcription factors. Bird suggests that these genes are silenced by efficient trans-acting repressors in non-expressing tissues. The gene for retinol-binding protein is associated with pCpG islands and is expressed only in liver. Upon transfection into other cells, it is expressed only in non-hepatic cells if a small region of the promoter is mutated (Colantuoni et al., 1987).

A conserved sequence of 50 bp has been identified around the translation start site of chicken types I and III collagen genes (Yamada et al., 1983b). This sequence contains an inverted repeat capable of forming a stem-loop structure. Rossi and de Crombrugghe (1987a) have demonstrated that this sequence can form intermolecular sense-antisense dimers. The equilibrium between monomers and dimers could be regulated to control the rate of translation.

Sensitivity to S1 nuclease is used to test for unusual
features in the secondary structure of a promoter. The S1
endonuclease enzyme cleaves supercoiled plasmids in which
perfect inverted repeats form cruciform structures under the
stress of supercoiling. The S1 nuclease will also cleave at
the boundaries between right-handed B-DNA and left-handed Z-
DNA. A strong correlation between S1 nuclease sensitivity and
the biological activity of an E. coli supercoiled promoter
has been demonstrated (Drew et al., 1985).

The promoters of both the mouse and chicken COL1A2 genes
contain a S1 sensitive region each, at -200 and -165
respectively (McKeon et al., 1984a; Finer et al., 1984). The
sequences at these sites are highly conserved between the two
species. They are rich in pyrimidines in one strand and with
an identical tandem repeat. The DNA of one tandem repeat
could base-pair with the opposite strand of the other tandem
repeat, creating a structure with two single-stranded loops
(Schmidt et al., 1984). This region might, therefore, not
bind nucleosomes as tightly as other segments, and would
allow dissociation of one or more nucleosomes under certain
conditions. Both the human COL1A1 and COL1A2 genes contain
analogous pyrimidine stretches, but it is not known whether
they are sensitive to S1 nuclease in vitro (Chu et al.,
1985c, Dickson et al., 1985). This pyrimidine stretch is
absent from the 5' flanking sequence of the chicken COL1A1
gene. Two S1 nuclease sensitive sites have been located at
250 and 350 nucleotides 3' to the first exon of this gene
(Finer et al., 1987). These sites correspond to the boundary
of unique DNA and a repeated purine-rich heptamer (fifteen
tandem repeats of GGGCAGA).

The chromatin structure of genes can be probed by
examining whether certain areas of the genes are more
susceptible to digestion by DNase I. The in vitro S1 nuclease
hypersensitive site also corresponds to the tissue-specific
in vitro DNase 1 hypersensitive site in each of the chicken
COL1A2 and mouse COL1A1 genes (McKeon et al., 1984b; Breindl
et al., 1984). Insertion of a single copy of the Moloney
murine leukaemia virus into the first intron of the mouse
COL1A2 gene results in the loss of the DNase I hypersensitive
site; and a 20- to 100-fold reduction in transcription of
the gene (Hartung et al., 1986). Although the human COL1A1 gene has a different pattern of DNase I hypersensitive sites in fibroblasts and placental tissue (Barsh et al., 1984); the only site that maps near the transcription start site is in the first intron, and is invariant in these tissues.

Functional deletion studies, using a transient expression assay in *Xenopus laevis* oocytes, demonstrated that in close proximity to the DNase I hypersensitive site in the first intron of the human COL1A1 gene is a 782 bp cis-acting region with enhancer activity (Roussouw et al., 1987). A variety of viral and cellular enhancers are found in nucleosome-free regions containing DNA elements for binding of constitutive and for regulatory factors (for review, see Kadonaga, 1986).

A 224 bp Ava I fragment from the first intron of the human COL1A1 gene was shown, in transient expression studies in chicken fibroblasts, to exhibit significant orientation specific inhibition of CAT activity (Bornstein et al., 1987). The normal regulation of COL1A1 gene transcription probably results from an interplay of positive and negative elements in the promoter region and the first intron.

An orientation dependent transcriptional enhancer has also been identified in the first intron of the mouse COL1A2 gene (Rossi and de Crombruggle, 1987b). The enhancer displays cell-specificity, since it functions in NIH3T3 murine fibroblasts but is completely inactive in a lymphoid cell line.

Most recently, transient expression studies have identified an orientation independent, chondrocyte specific, enhancer element. This has been located to within a 800 bp EcoR I-BamH I intron fragment, about 1.5 kb downstream of the first exon of the rat COL2A1 gene (Horton et al., 1987).

Whether transcriptional enhancers within the first intron are a general feature of collagen genes will require more functional analyses.

Expression of hybrid gene constructs in transgenic mice has demonstrated that all sequences responsible for correct developmental and differential expression of the human COL2A1 gene are within a region spanning 4-5 kb 5' and 2.2 kb 3'
flanking DNA (Lovell-Badge et al., 1987). Similarly, sequences within the region -2000 to +54 provide all the information required for developmental and tissue-specific expression of the mouse COL1A2 gene (Khillan et al., 1986).

1.6 INHERITED COLLAGEN DISORDERS

As a direct consequence of the complex gene structure and biosynthetic pathway of collagen, there are many steps at which errors can occur and so result in a disease state. These obviously affect primarily those tissues in which the proper development and integrity of connective tissue is of paramount importance. In osteogenesis imperfecta the bones are extremely weak and prone to fracture. Fragile skin, blood vessels, gut walls and ligaments are characteristic of Ehlers-Danlos syndrome. Marfan syndrome combines soft-tissue features such as aortic rupture, mitral valve prolapse and dislocation of the ligaments supporting the lens, with skeletal abnormalities such as scoliosis and arachnodactyly. Abnormal cartilage development is characteristic of the various chondrodysplasias.

Within each of these clinical groupings there is a variety of inheritance and a complete spectrum of expression and penetrance. Inheritance can be as an autosomal dominant or recessive trait as well as sex-linked. Disease states can vary from lethal to very mild forms.

Using the paradigm of globin gene mutations and thalassaemias, it has been proposed by several workers that connective tissue disorders might be due to mutations in the structural genes encoding the collagens.

There are, however, a number of technical problems in exploring this suggestion. One of these is the large size of the genes and their complex exon/intron structure. To date only the COL1A1, COL1A2 and COL2A1 genes have been covered completely by genomic clones (see Cheah, 1985, for a review of the clones isolated). Searching for small changes in such complex genes is difficult. The second technical problem lies in the complexity of procollagen biosynthesis. A defect in
one of the several post-translational enzymes could well mimic a mutation which alters a structural gene for one of the procollagens. In fact, the first true molecular defect in collagen metabolism to be identified was due to lysyl hydroxylase deficiency in Ehlers-Danlos (EDS) type VI (Pinnell et al., 1972). Furthermore, several different collagens may be present in the same tissue. Therefore, it is difficult at the outset to be certain which set of genes should be explored in examining any particular heritable disorder that might involve collagen.

The emphasis of the present discussion is on the use of molecular approaches to the analysis of connective tissue disorders associated with mutations of fibrillar procollagen genes, and the general conclusions that can be drawn from the published data. It is not intended to present a comprehensive catalogue of all the mutants analysed; extensive reviews of these have been published recently (Cheah, 1985; Prockop, 1985; Sykes, 1986; Tsipouras and Ramirez, 1987).

Osteogenesis Imperfecta

The term osteogenesis imperfecta (OI) includes a group of disorders which are highly heterogenous in their phenotypic expression. The disorders occur at a frequency of 1 in 25,000 births, about half the rate of phenylketonuria, and it is possible that many milder forms are undiagnosed.

Bone fragility is the principal characteristic feature of the OI syndromes. Short stature, joint laxity, easy bruising, blue sclerae, presenile hearing loss, and dentinogenesis imperfecta may also be present. The most widely accepted classification system is that of Sillence et al., (1979), whereby OI is divided into four broad groups.

OI type I is inherited as an autosomal dominant trait and is characterised by postnatal onset of fractures, blue sclerae, joint laxity, and dentinogenesis imperfecta.

OI type II is the lethal perinatal form characterised by extreme bone fragility resulting in death in utero or soon after birth. The disorder is thought to have both new dominant and recessive modes of inheritance.

OI type III is highly heterogenous and the principal
clinical feature is significant skeletal deformity. Autosomal recessive inheritance has been suggested; though in the majority of families studied the proband is the only affected individual, indicating spontaneous mutations are responsible.

OI type IV is inherited as an autosomal dominant trait, with mild to moderate bone fragility, short stature, white sclerae, and dentinogenesis imperfecta.

**Marfan Syndrome**

Marfan Syndrome (MS) is a group of dominantly inherited diseases characterised by arachnodactyly, pectus deformities, scoliosis, aortic root dilatation, mitral valve prolapses, myopia, and lens dislocation. Mean life span is reduced to about 45 years, with most premature deaths being the cardiovascular defects (Pyeritz and McKusick, 1981).

Fibrillar collagens are the major stress-bearing components of tissues most affected and have, therefore, been the favourite candidates for the basic defect in the gene. About 5% of cases are new mutations. The prevalence of Marfan syndrome is at least 1.5 per 100 000 live births.

**Ehlers-Danlos Syndrome**

EDS is a group of disorders more heterogenous than OI. The clinical manifestations of these disorders include laxity of joints, soft and extensible skin which may be fragile, and easy bruising. There are at least ten distinct types (for a comprehensive review see Byers and Holbrook, 1985).

EDS types I, II and III share many similar clinical features, a dominant mode of inheritance, and a common pattern of alteration in the morphology of collagen fibrils. As yet it is unclear whether the basic defect is in type I collagen or other genes.

EDS IV is the rarest but most severe form. It is defined on the basis of an abnormality of synthesis, secretion, or structure of type III collagen. In the majority of cases the disorder is inherited in an autosomal dominant form, but Pope et al. 1977) suggested that autosomal recessive forms also
exist.

EDS V and IX both show linkage to the X chromosome. As no collagen gene has yet been identified on this chromosome the phenotype is due to a mutation in a gene other than that of the major collagens. In both disorders the level of lysyl oxidase activity is affected.

EDS VI is the most common of the autosomal recessive forms of EDS. The disorder results from decreased lysyl hydroxylase activity (Pinnell et al., 1972). Prenatal diagnosis of EDS VI is fairly straightforward, depending on a measure of the lysyl hydroxylase activity of cultured cells with a substrate of type I collagen.

The principal feature of EDS type VII is gross joint instability and multiple dislocations. There is a correlation between the persistence of one or more N-propeptides and the EDS VII phenotype.

At present there is very little biochemical information on EDS types VIII and X.

1.7 THE MOLECULAR BASIS OF INHERITED COLLAGEN DISORDERS

1.7.1 Experimental strategies

Studies on the gross morphology of the affected tissue, usually skin, bone or cartilage, by light and electron microscopy is the simplest method used to detect alterations in collagen metabolism (for review see Hollister et al., 1982). Abnormalities in general collagen fibre structure are seen in a variety of connective tissue disorders. These studies also show that there are no unique disease-specific structural changes in collagen fibrils.

Because of the heterogeneity of inherited collagen diseases, it has been difficult to assign specific biochemical characteristics to each type. Comparisons of amino acid composition, levels of hydroxyproline and hydroxylysine, and the extent of glycosylation in collagen isolated from patient and control tissues may indicate the primary cause of the defect.

The approximate locations of structural defects can be
predicted from the analysis of the collagen molecule itself. Each \( \alpha \)-chain yields a characteristic set of cyanogen bromide peptides delineated by the position of methionine residues. Cysteine residues mark the characteristic position of intra- and inter-chain disulphide bonds. In the extracellular matrix the N- and C-proteinases cleave off the propeptides at specific sites.

Information obtained from the physical and chemical study of collagen and mRNA of known patients can be used to indicate the site of the primary defect. If it is a procollagen gene this can be cloned and examined in detail.

Direct analysis of collagen gene structure has been the most recent experimental approach utilised. Collagen gene probes can be used to identify restriction fragment length polymorphisms associated with a given gene; to test for coinheritance or cosegregation of a specific allele of a procollagen gene and a heritable disorder within a family. If coinheritance is found the specific allele can be cloned and examined in detail.

1.7.2 Structural mutations in collagen genes

Collagen fibrils are formed by the self-assembly of monomers of collagen molecules. For correct assembly, at best three features of the molecules are essential: first, the three \( \alpha \)-chains must fold into a unique triple helix conformation to produce a rigid molecule; secondly, the molecules must have the correct distribution of charge and hydrophobic amino acid residues so that they can assemble in a quarter-stagger array; and finally, the molecules must be of the correct length.

Mutations that alter the structure of the amino-terminal propeptides have their primary affect in the processing, by procollagen N-proteinase, of the procollagen molecule. Such cases are characterised by marked laxity of joints, and are generally classified as EDS type VII variants. The syndrome has been divided into two subtypes: EDS VIIA, inherited as an autosomal recessive trait, in which the biochemical defect is a partial deficiency in type I procollagen N-proteinase activity (Halila et al., 1986); and EDS VIIB which is
inherited as an autosomal dominant condition in which the type I procollagen molecules are altered due to a structural mutation.

N-proteinase requires a procollagen substrate that not only has the correct amino acid sequence but also the correct three-dimensional conformation. In one patient, Sippola et al. (1984) reported a deletion of between 10 and 30 amino acid residues in the amino-terminus of half of the \( \alpha_2(1) \) chains synthesised by fibroblasts. The deletion was about 100 residues removed from the N-proteinase cleavage site, but it disturbed the conformation of the site sufficiently to make the procollagen molecules resistant to proteinase action.

In another patient, removal of exon 47 of the \( \alpha_2(1) \) procollagen gene, which codes for the N-proteinase cleavage site and a cross-linking lysine residue, resulted in EDS type VIIB (Wirtz et al., 1987). The deletion of the entire amino terminal telopeptide of the pro-\( \alpha_1(1) \) chain was reported by Cole et al. (1986) to produce EDS type VII. Equivalent amounts of normal \( \alpha_1(1) \) chain and a larger pN-\( \alpha_1(1) \) chain were observed.

The removal of the N-terminal propeptide is believed to be a regulatory mechanism for fibril formation, while the persistence of pN-collagen molecules in collagen fibrils seems to be a physiological phenomenon. pN-collagens of types I and III can be demonstrated in foetal skin, and type III pN-collagen molecules are still present in adult skin collagen fibres (Fleischmajer et al., 1983). Interestingly, the patients studied so far all have a characteristic laxity of joints irrespective of the underlying mutated gene. The patients do not have symptoms related to bone even though pN-collagen probably persists in bone too. The results suggest that the removal of the amino-propeptide is not as critical a step for normal bone formation as it is for the formation of normal ligaments and tendons.

Several mutations in the central, triple helical domain of the procollagen molecule have been shown to decrease markedly the normal stability of the triple helix. One of the best studied examples is a patient with lethal OI in which collagen production was reduced in skin fibroblasts when
compared to normal cells (Barsh and Byers, 1981). The defect has been identified as the deletion of approximately 500 bp in the COL1A1 gene (Chu et al., 1983). Fibroblasts from the patient contained mRNAs for pro-\(\alpha\) 1(I) chains that were shortened by 250 nucleotides, whilst retaining the reading frame, hence shortened pro-\(\alpha\) chains were synthesised. The other allele for the COL1A1 gene was normal. The population of type I procollagen contained normal type I molecules, those with one short pro-\(\alpha\) (I) chain, and ones which had both mutant pro-\(\alpha\) (I) chains, in the ratio of 1:2:1. Three-quarters of the molecules were, therefore, abnormal and unable to form a stable triple helical molecule. Such molecules were rapidly degraded intracellularly or after secretion (Williams and Prockop, 1983). Recent work by Chu et al. (1985a) has demonstrated that the deletion is contained within two introns resulting in the elimination of three exons of the triple helical domain. Interestingly the termini of the deletion are located within two short inverted repeats. In this patient there was also an elevated type III collagen mRNA content. It has been suggested that this is a compensatory effect (Williams and Prockop, 1983), implying that type III collagen can substitute for type I. As this patient suffered from OI type II, and hence was fatally affected, this cannot be substantiated. Bateman et al. (1986), in a study of 17 OI type II cases, localised type III collagen to the uncalcified matrix of OI bone. It is likely, therefore, that type III collagen was produced as part of the fracture repair process. The proportion of type V collagen was also shown to increase in OI type II bone. This was probably because the number of osteoblasts, with which type V collagen is associated, is increased in OI bone.

The inactivation of normal pro-\(\alpha\) chains by shortened chains has been termed 'protein suicide' (Prockop, 1984). This phenomenon may explain how a heterozygous gene defect, which reduces the amount of a normal gene product by one half, can produce a dominantly inherited disease. A mutation which results in the complete nonfunctioning of one COL1A1 allele causes a disease but is less severe than one producing abnormal pro-\(\alpha\) chains. Skin fibroblasts of a patient with OI
type I synthesised pro-$\alpha$ 1(I) and pro-$\alpha$ 2(I) chains in a 1:1 ratio instead of the normal 2:1 ratio (Barsh et al., 1982). The lesser severity of this OI is because the type I molecules synthesised, albeit at a reduced rate, were normal.

There are several examples of OI where the disease state is caused by deletions or insertions in one of the type I procollagen genes, these have been reviewed by Cheah, Prockop, and Tsipouras and Ramirez, (1985; 1985; 1987). One conclusion that may be drawn is that the type I procollagen genes are prone to mutations which tend to produce shortened pro-$\alpha$ chains. However, a caveat must be issued. There are at least two reported cases where length variations in type I procollagen genes were identified as the defects responsible for the disease state. Byers et al. (1981) reported the insertion of 20 amino acids in the pro-2(I) chain of a Marfan patient. An insertion of 38 bp was found in an intron of the COL1A2 gene of that individual (Henke et al., 1985). Dalgleish et al. (1986) later showed that this 38 bp length variation is a common polymorphism. A 300 bp deletion has been described in the COL2A1 gene in families with OI II (Pope et al., 1984). Sykes et al. (1985) demonstrated that this region is highly polymorphic in normal individuals. In both cases the length polymorphism resides in non-coding DNA sequences.

A patient with perinatal lethal osteogenesis imperfecta was found to have $\alpha$1(I) dimers linked by cysteine located in the CB6 peptide. Incorporation of either one or two mutant chains into a molecule decreased the thermal stability of the triple helix, delayed secretion, and resulted in excessive lysyl hydroxylation of all chains, as a result of delayed triple-helix formation. It has been shown that the codon for glycine in position 988 of the helical protions has changed to cysteine by a single base mutation (Cohn et al., 1986). Similarly, Bateman et al. (1987) reported a case of OI II where there is a glycine to arginine substitution in the (I)CB8 peptide. In contrast, another patient with mild OI also has a cysteine substitution in the $\alpha$1(I)CB6 peptide. However, Steinmann et al. (1986) propose that the mutation
is in the X or Y position of the GLY-X-Y triplet; such a substitution, for steric reasons, would not disrupt the triple helix.

To date, only two mutations have been found that alter the structure of the carboxyl-terminal propeptides. In one case, classified as OI type III, the type I procollagen synthesised by the patient's fibroblasts contained an excess of mannose in the carboxyl-propeptide. The results suggest a change in the primary structure of the molecule which altered post-translational processing by glycosyl transferases (Peltonen et al., 1980). In another patient, a combination of S1 nuclease mapping, and the cloning and sequencing of the mutant allele showed a four base pair frame shift deletion, resulting in the complete alteration of the reading frame of the last 34 amino acids of the carboxyl-propeptide of the pro-α2 chain (Dickson et al., 1984; Pihlajaniemi et al., 1984).

The involvement of other collagen genes in inherited connective tissue disorders has not been well studied. An EDS IV patient with reduced levels of type III collagen content in the skin was reported to secrete abnormal and normal pro (III) chains. The defect is probably an insertion of 20 amino acids in the CB5 peptide (Pyeritz et al., 1984; Stolle et al., 1985).

Most information on the molecular basis of collagen disorders has been derived from the analysis of osteogenesis imperfecta variants, and therefore relates to the effect of mutations on type I collagen. However, some generalisations can be made that may be applicable in all fibrillar collagens.

The biological consequences of defects in the carboxyl-propeptide and the triple helical domain are significant. These two regions play a key role in the formation of the triple helix, which is essential for normal fibrillogenesis. Defects affecting the amino-propeptide, result in a distinct pathological phenotype, EDS VII, regardless of the nature of the chain involved. The synthesis of a structurally abnormal pro-α chain is usually more deleterious than the decreased synthesis of the same normal pro-α chain. Indeed, the
complete absence of pro-\(\alpha_2(I)\) chains was not lethal even in homozygotes (Pihlajaniemi et al., 1984). The unexpectedly mild phenotype is due to the ability of pro-\(\alpha_1(I)\) chains to form a homotrimer which can substitute for normal type I collagen.

Similar changes in different parts of the same molecule produce diseases that affect different tissues. Consequently, diseases such as OI and EDS that initially appeared to be discrete conditions are now seen to have a common unity at the molecular level. This explains why there is such heterogeneity and overlap in phenotype in the genetic diseases of connective tissues.

Mutations that change the length of pro-\(\alpha\) chains of type I procollagen, and probably other fibrillar collagen also, are relatively common. This may be due to the presence of repetitive triple helical coding sequences predisposing the genes to unequal crossover mutations during meiosis. The presence of a large number of introns would predispose the fibrillar procollagen genes to RNA splicing defects.

1.7.3 DNA polymorphisms and linkage studies

As mentioned before, one way of identifying the mutant allele in an individual suffering from an inherited collagen disease is by the use of restriction fragment length polymorphisms (RFLP). The detection of a collagen gene RFLP which segregates with an inherited connective tissue disorder is not only useful as identifying which gene is defective, but can also then be used for prenatal diagnosis.

Several RFLPs have been identified in close association with the fibrillar collagen genes. These have been listed by Tsipouras and Ramirez (1987). An Rsa I polymorphism in the COL1A2 gene was demonstrated to be coinherited with an OI phenotype in a family (Grobler-Rabie et al., 1985). This indicates that a mutation in that gene is responsible for the disease state in that family. Sykes et al. (1986) in a study of eleven OI pedigrees, using RFLPs, showed that the disease was inherited with either the COL1A1 or COL1A2 locus. However, in a pedigree with autosomal recessive inheritance of OI, the phenotype is not linked to either the COL1A1 or
COL2A1 loci (Aitchison et al., 1988). An EcoR I RFLP linked to the COL3A1 gene was used to show that there is a likelihood that EDS IV in two families cosegregates with this gene (Tsipouras et al., 1986a). An Ava II RFLP associated with the COL3A1 gene (Dalgleish et al., 1985) has been used to demonstrate very tight linkage of this locus and EDS IV in a large pedigree with eleven affected individuals (Nicholls et al., 1988). The use of RFLPs to exclude both the COL1A2 and COL3A1 loci in a Marfans pedigree has been reported (Tsipouras et al., 1986b; Dalgleish et al., 1987). More recently, a study of six pedigrees with Marfans syndrome demonstrated that in three pedigrees the disease segregates independently of both type I collagen loci (Ogilvie et al., 1987). In two of these families the disease segregated independently of both COL2A1 and COL3A1 loci also. This is an indication that Marfan syndrome is generally not due to mutations in the major fibrillar collagen genes. A report showing linkage of the COL2A1 locus to achondroplasia (Eng et al., 1985), has since been retracted.

1.8 Aims

Type III collagen accounts for 10 to 50% of total collagen in the mammalian body. Its tissue distribution is similar to type I but is absent from bone and tendon. The function of type III appears to be related to the maintenance of integrity of distensible tissues; and in its absence, larger vessels and viscera lack the strength to withstand the normal pressure. It is this feature which characterises the phenotypes of EDS IV. Type III collagen, as discussed earlier, plays a fundamental role in the genesis of this disorder. Collagen has also been implicated in acquired disorders such as scleroderma and a variety of fibrotic conditions (Prockop et al., 1979). Type III is especially important in terms of disease states as it occurs in rapidly proliferating tissues, not only embryonic but cancerous ones too.

At the time the project was begun, there were no reported genomic clones of the human type III procollagen
gene. Without cloning the defective gene the limit for detecting insertions or deletions in collagen genes within total DNA in about 100 bp. To study the molecular nature of inheritable disorders involving type III collagen, an attempt was made to clone the human COL3A1 gene using a cosmid vector.

A source of cloned DNA is a prime requirement for obtaining DNA probes specific for the COL3A1 gene. There are three main advantages to the use of large genomic DNA probes, such as those isolated from cosmid clones: the use of flanking sequences as a positive control for a deleted gene or part of a gene; greater amount of DNA can be screened for polymorphisms; and the use of larger sized fragments in hybridisation reactions results in increased sensitivity.

Another aim of the project was to study the structure and regulation of the gene. Any sequences which may be important in controlling gene expression can be identified by comparative sequencing studies. To this end, a sequencing project to characterise cDNA clones that overlap the 3' end of the COL3AI gene was initiated.

A thorough understanding of the nature of transcriptional control, however, will require the use of in vitro transcription systems; or of genes stably introduced into mammalian cells in culture.

The use of genomic clones will also provide the opportunity to analyse the role, if any, of conserved intron sequences in transcriptional control.
CHAPTER 2

MATERIALS AND METHODS
2.1 ENZYMES, ANTIHBiotics, CHEMICALS AND REAGENTS

The sources of the most important of these are listed below. All other chemicals were of analytical grade.

Restriction enzymes, unless otherwise stated, M13mp18 and M13mp19 RF DNAs, λ, λ/Hind III and φX174 RF/Hae III DNAs, DH5 and DH5α competent cells and caesium chloride were obtained from Gibco-BRL plc, Paisley, Scotland. Calf intestinal phosphatase was supplied by Boehringer Corporation (London) plc, Lewes, England.

Ampicillin, chloramphenicol, kanamycin, tetracycline, bovine serum albumin, dithiothreitol, ficoll 400, HEPEs (N-2-hydroxyethylpiperazine - N'-2-ethanesulphonic acid), IPTG (isopropyl- β-D-galacto-pyranoside), PEG6000 (polyethylene 6000), salmon sperm DNA (sodium salt), spermidine trichloride, TEMED (N,N,N',N'-tetramethyl-ethylene-diamine) and thymine were supplied by Sigma Chemical Company, Poole, England. Anglian Biotechnology, Colchester, England supplied BCIG (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside).

DNase I came from Worthington Biochemical Corporation, Freehold, New Jersey, U.S.A. New England Biolabs T4 DNA ligase was supplied by CP Laboratories, Bishop's Stortford, England.

All agarose used was supplied by FMC Corporation, Rockland, Maine, U.S.A.

Acrylamide, urea and Amberlite resin were obtained from Serva, Heidelberg, West Germany. N,N'-methylene-bisacrylamide was supplied by Uniscience, Cambridge, England, and ammonium persulphate by Bio-Rad Laboratories, Watford, England.

Fisons of Loughborough, England supplied phenol and polyvinylpyrrolidine (PVP). MOPS (3-(morpholino)propanesulphonic acid), TRIS (2-amino-2-(hydroxymethyl)propane-1,3-diol) and dimethyldichlorosilane solution (used for siliconising glass plates used for sequencing) were obtained from BDH, Poole, England.

Deoxyribonucleotides(dNTPs), dideoxyribonucleotides (ddNTPs), hexadeoxyribonucleotides (cat. no. 27-2166-01, for oligo-labelling), DNA polymerase I (Klenow fragment) and sephadex
G-50 and G-75 were obtained from Pharmacia, Milton Keynes, England.

Marvel dried milk, used in hybridisations, was obtained from Cadburys, Birmingham, England.

All radionucleotides were obtained from Amersham International plc, Little Chalfont, England.

Synthetic oligonucleotides including the M13 universal sequencing primer (dGTAAAACGACGGCCAGT) were synthesised by J. Keyte, Biochemistry Department, University of Leicester, England.

Media for growth of bacteria were obtained from Oxoid Ltd, Basingstoke, England; except for BBL trypticase (Becton Dickinson Ltd, Oxford, England), Difco Tryptone and Bacto Yeast Extract (Difco Ltd, East Molesley, England).

All other chemicals were of analytical grade.

2.2 MEDIA AND GENERAL DNA HANDLING TECHNIQUES

1) Media

Liquid and solid media were as follows:

Luria Broth (10g Difco Bacto Tryptone, 5g Difco Bacto Yeast Extract, 5g NaCl per litre of distilled water). Luria agar plates (LUA) were prepared by solidifying liquid media with 15g Difco bacto-agar per litre; for soft agar overlays 6g agar was used.

BLA contained 10g BBL trypticase, 5g NaCl and 15g Sterilin agar per litre. BTL soft agar overlay was as BLA but contained only 6g agar per litre.

E. coli strain JM101 was maintained on glucose supplemented minimal medium, this consisted of WA (15g Davis agar in one litre of water at pH 8.5) with 42 mM Na₂HPO₄, 22 mM KH₂PO₄, 18 mM NH₄Cl, 8 mM NaCl, 22 mM glucose, 0.1 mM CaCl₂, 1 mM MgSO₄, 3 M thiamine HCl and 0.17 mM proline.

S.O.C. contained 2% (w/v) bactotryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM MgCl₂, 10 mM MgSO₄, 20 mM glucose. Filter sterilised MgCl₂, MgSO₄ and glucose were added to the other components which had been dissolved in water and autoclaved. The pH should be 7.0.
ii) Phenol extractions

DNA solutions were mixed with 0.5 volume of phenol: chloroform: isoamyl alcohol: 8-hydroxyquinoline (100:100:4:0.1, w:v:v:w) saturated with 10 mM Tris-HCl, pH 7.5 and briefly centrifuged to separate the phases. The upper aqueous phase containing the DNA was removed and the phenol layer re-extracted with an equal volume of 10 mM Tris-HCl, pH 7.5. The phenol was AR grade and not redistilled.

iii) Ethanol precipitations

A standard procedure was as follows: DNA was precipitated from solution by the addition of 0.1 volume of 2M sodium acetate, pH 5.6, and 2 volumes of 100% ethanol. After mixing, the solution was chilled for 10 minutes in an IMS/dry-ice bath. DNA precipitates were pelleted by centrifugation at 13 000g for 20 minutes at 0°C (Sorvall HB4 rotor), or in a MSE Micro-Centaur centrifuge for 5 minutes at maximum speed. The supernate was discarded and the pellet rinsed with 70% ethanol, centrifuged for 2 minutes and the 70% ethanol removed. DNA pellets were vacuum-dried and resuspended in an appropriate solution for further manipulation.

iv) Restriction endonuclease digestion of DNA

DNAs at a final concentration of 0.5 mg/ml were incubated in the manufacturers recommended buffer at 37°C for 1 hour, unless otherwise stated. Digests of genomic DNA were incubated for 5 hours at 37°C unless a partial digest was required in which case the necessary period of incubation was determined empirically. Spermidine trichloride was not generally added. If the products of restriction digests were to be subjected to further manipulation the DNA was extracted with phenol/chloroform and recovered by ethanol precipitation. For DNA that was to be electrophoresed 0.1 volume of Orange G Dye (0.1% w/v orange G, 20% (w/v) ficoll 400, 100mM EDTA) was added and the mixture loaded onto an agarose gel of the appropriate size and concentration.
v) Butanol concentration

If necessary DNA solutions were concentrated by extraction of water with butan-2-ol. Solutions were mixed with butan-2-ol and briefly centrifuged. The top phase was discarded and the lower phase re-extracted as required to further reduce the volume of the solution. The final lower phase was extracted 3 times with diethyl ether to remove remaining butan-2-ol. Traces of diethyl ether were removed by gentle passage of air over the solution.

vi) Alkaline phosphatase treatment of DNA

1 µg DNA was digested with appropriate restriction endonucleases. If necessary the required DNA was extracted with phenol/chloroform and recovered by ethanol precipitation; otherwise 15 minutes before the end of the digestion period, 1 unit of calf intestinal phosphatase was added and the incubation continued for 15 minutes at 37°C. The DNA was either electrophoresed on an agarose gel and recovered by one of the methods described in Section 2.6, or recovered by phenol extraction and ethanol precipitation.

vii) Ligation of purified DNA fragments into plasmid

Ligations of linear DNA to plasmid DNA were performed in 1 x ligase buffer (30 mM Tris-HCl, 10 mM MgCl₂, 5 mM dithiothreitol, pH 7.6), 1 mM ATP (100 mM stock stored at -20°C). 0.5 µl of Bio-Labs T4 DNA ligase (equivalent to 0.5 Weiss unit) was added and the mixture incubated at 15°C overnight. Blunt-end ligations were incubated at 4°C. Ligations of cohesive-ends were observed to go to completion within 60 minutes at room temperature, and these conditions were often used.

viii) Ligation of DNA fragments in low gelling temperature agarose

This was modified from the method of Crouse et al. (1983).

DNA to be subcloned was prepared by appropriate restriction endonuclease digestion, followed by electrophoresis on a 0.6% low gelling temperature agarose
gel. Sufficient DNA was digested so that the fragment of interest contained approximately 200 ng. The vector into which this fragment was to be subcloned was prepared in a similar manner. If necessary, vector DNA was dephosphorylated as described before. The fragment and vector DNA were excised on a UV transilluminator, placed into separate pre-weighed microfuge tubes and weighed. The volume of each gel slice was estimated by assuming that the gel has a density of 1 g/ml.

The gels were melted at 70°C for 10 minutes and aliquots were then transferred into fresh tubes which contained the calculated volume of 10 mM Tris-HCl, pH 7.6 in which the ligations were to be performed. Ligations were generally performed at a molar ratio of 1:1 of vector:insert ends. The volume of the ligation reaction was such that the final gel concentration was always less than 0.1%.

The reaction tube was heated at 70°C for 3 minutes and cooled to room temperature. 10 x ligase buffer, ATP and T4 DNA ligase were added as indicated in Section 2.2 (vii). Ligations in low gelling temperature agarose were left overnight at 15-20°C. The ligated DNA was used to transform competent E. coli cells without further treatment.

The unused gel slice was kept at -20°C for future use.

2.3 DNA

DNA was prepared from blood by collecting blood into 10 ml EDTA tubes. The tubes were set in a rack at an angle of 45° and the red blood cells allowed to settle. Plasma, containing white blood cells, was removed from above the red blood cells over a period of 2 hours and transferred to a fresh tube. The cells were pelleted by centrifugation at 400g for 10 minutes at room temperature (1500 rpm in a MSE Centaur). 4 ml of plasma could be collected from 10 ml of whole blood.

The supernate was removed and the pellet was quickly resuspended in 5 ml of distilled water and, exactly 30 seconds later, 5 ml 2 x SSC was added with mixing. This step ensures hypotonic lysis of any residual red blood cells.
White blood cells were pelleted as before and resuspended in 4 ml of lysis buffer (0.3 M sodium acetate, 20 mM Tris-HCl, pH 7.5, 1 mM EDTA) and the addition of 200 μl of 10% SDS with thorough mixing. Lysis was immediate and the solution became viscous. The lysate was vigorously extracted with an equal volume of phenol/chloroform/iso-amyl alcohol/8-hydroxyquinoline for 5 minutes. The aqueous phase was removed with a wide bore pipette after centrifugation at 2000g (3500 rpm in a MSE Centaur), and phenol extracted once more. The aqueous phase from this extraction was extracted with an equal volume of chloroform/iso-amyl alcohol (24:1) to remove residual phenol. The aqueous phase was removed after centrifugation and the DNA precipitated by the addition of two volumes of 100% ethanol. The precipitate was spooled on the sealed end of a glass Pasteur pipette and transferred to a screw capped microfuge tube, washed in 1 ml of 70% ethanol, dried and dissolved in an appropriate volume of TE and stored at 4°C.

DNA recovery was measured by UV light absorbance at 260 and 280 nm or by comparison with known quantities of phage DNA on agarose gels.

DNA from placentae had been prepared previously from tissues obtained from the Royal Infirmary, Leicester.

2.4 PLASMIDS, BACTERIOPHAGE AND BACTERIAL STRAINS

Cosmid vector c2XB was a gift from P. Bates and R. Swift, Michigan State University.

Plasmids pUC13 (Vieira and Messing, 1982), pSP64 (Melton et al., 1984) and Bluescribe (Statagene Cloning Systems, San Diego, U.S.A.) were used as subcloning vectors. The lambda replacement vector λL47.1 (Loenen and Brammar, 1980) was used for the construction of human genomic libraries from size fractionated DNA. The M13 vectors M13mp18 and M13mp19 (Yanish-Perron et al., 1985) were used for the cloning of DNA fragments for sequencing.

Escherichia coli strains used are listed in Table 2.1.
<table>
<thead>
<tr>
<th>Strain</th>
<th>Genotype</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>BHB2688</td>
<td>N205 recA (imm434, cIts, b2, red3, Eam4, Sam7/λ)</td>
<td>1</td>
</tr>
<tr>
<td>BHB2690</td>
<td>N205 recA (imm434, cIts, b2, red3, Dam15, Sam7/λ)</td>
<td>1</td>
</tr>
<tr>
<td>DH1</td>
<td>F-, endA1, hsdR17 (rk-, mk+), supE44, thi-1, λ-, recA1, gyrA96, relA1</td>
<td>2</td>
</tr>
<tr>
<td>DH5</td>
<td>F-, endA1, hsdR17 (rk-, mk+), supE44, thi-1, λ-, recA1, gyrA96, relA1</td>
<td>2</td>
</tr>
<tr>
<td>DH5α</td>
<td>as for DH5 with the additional mutations: (argF-laczya)U169, 80dlaczAM15</td>
<td>3</td>
</tr>
<tr>
<td>ED8910</td>
<td>supE44, supF58, recB21, recC22, hsdS, metB, lacY1, galK2, galT22</td>
<td>4</td>
</tr>
<tr>
<td>JM83</td>
<td>ara, Δ(lac-pro), strA, thi1, 80dlacIg, Z M15</td>
<td>5</td>
</tr>
<tr>
<td>JM83recA</td>
<td>as for JM83 with the additional mutation: rsspSL</td>
<td>6</td>
</tr>
<tr>
<td>JM101</td>
<td>Δ(lac-pro), supE44, thi1, F' traD36, proAB, lacIg, Z M15</td>
<td>7</td>
</tr>
<tr>
<td>WL87</td>
<td>803, supE, hsdRk, hsdMk, tonA, trpR, metB</td>
<td>8</td>
</tr>
<tr>
<td>WL95</td>
<td>P2 lysogen of WL87</td>
<td>8</td>
</tr>
<tr>
<td>1046</td>
<td>recA-, supE, supF, hsd-, met-</td>
<td>9</td>
</tr>
<tr>
<td>C600</td>
<td>F-, thi-1, thr-1, leuB6, lacY1, tonA21, supE44, λ-</td>
<td>10</td>
</tr>
</tbody>
</table>

Reference:

1 Hohn (1979).
3 BRL Focus (1986).
4 Loenen and Brammar (1980).
5 Vieira and Messing (1982).
6 from W.J. Brammar.
7 Messing (1981).
8 from W.J. Brammar.
9 also known as ED8767. Murray et al. (1977).
10 Appleyard (1954).
2.5 AGAROSE GEL ELECTROPHORESIS

i) Test gels

Horizontal agarose gels with 6mm wide loading slots were prepared and run in E buffer (40 mM Tris-acetate, 2 mM EDTA, pH 7.7) containing ethidium bromide at 0.5 µg/ml. The gel size varied with the number of samples to be run, from 5 x 7 cm (minigels) to 20 x 20 cm (mapping gels). The length of the gel was also determined by the extent of the resolution required. The concentration of agarose varied between 0.5-2% (w/v) according to the anticipated molecular weight of the DNA sample(s), which were run alongside a suitable set of known molecular weight marker DNAs, either λ/Hind III or ΦX174 RF/Hae III or both. DNA samples were mixed with 0.1 volume of Orange G dye prior to electrophoresis. Marker DNA samples were normally heated at 65°C for a few minutes before loading onto a gel, this was to dissociate the 23 kb and 4.4 kb fragments of λ/Hind III which associate by virtue of their cohesive ends. Gels were run at room temperature at 60-100 volts for 1-2 hours or overnight at 15-25 volts. Gels composed of low gelling temperature agarose were run at a maximum of 65 volts to prevent melting of the gel during electrophoresis. To visualise the DNA, gels were placed on a Chromato-Vue C-63 UV light transilluminator (Ultra-Violet Products Inc., San Gabriel, California, U.S.A.).

ii) Preparative Gels

Preparation of samples, gel loading and electrophoresis were as described by Jeffreys et al. (1980). The amount of native DNA loaded was adjusted to < 0.5 µg/mm2 of gel slot surface area to avoid overloading. DNA in preparative gels was visualised by the fluorescence of bound ethidium bromide under longwave ultraviolet light.

DNA was recovered from preparative gels by two procedures, as described below.
2.6 RECOVERY OF DNA FROM PREPARATIVE GELS

i) Vertical dialysis membrane

The dialysis membrane was prepared by boiling conveniently sized lengths of tubing for 10 minutes in a large volume of distilled water with 1 mM EDTA. After rinsing in distilled water the tubing was boiled for 10 minutes in water only. Upon cooling the tubing was stored at 4°C totally submerged under water. Before use the tubing was cut along one flattened edge to open it out and rinsed thoroughly in distilled water.

DNA was electrophoresed at high voltage (120V) onto a vertical dialysis membrane inserted into a slot cut into the gel. Another piece of dialysis membrane was used to block the migration of unwanted DNA, of higher molecular weight than the required fraction or fragment, onto the collecting membrane. Once all the DNA to be collected had migrated onto the membrane, the membrane was rapidly transferred, with the current still on, into a 30 ml plastic universal container. One corner of the membrane was trapped by the cap of the bottle which was centrifuged for 1 minute at top speed in a MSE Centaur centrifuge. The DNA-containing buffer was removed carefully from the bottom of the bottle to avoid any pieces of agarose that may have been trapped on the membrane. DNA was recovered by phenol extraction, which efficiently removes ethidium bromide, and ethanol precipitation before further manipulation.

ii) DE81 paper

DNA was electrophoresed onto Whatman DE81 DEAE-cellulose paper, and recovered by a modification of the method of Dretzen et al. (1981). DE81 paper was prepared by soaking in 2.5 M NaCl and rinsing five times in distilled water, and stored in 1 mM EDTA at 4°C. The desired DNA fragments were first separated by agarose gel electrophoresis, then cut out of the gel in small pieces of agarose. The DNA was electrophoresed onto the paper at high voltage (200V). The gel slices were then removed from the paper and unwrapped. The DNA on the DE81 paper was visualised by ethidium bromide.
fluorescence under long wave ultraviolet light and the rest of the paper was cut away and discarded. DNA was recovered by vortexing the paper until it disintegrated in high salt buffer (50 mM Tris-HCl, 1 M NaCl, 1 mM EDTA, pH 7.5), and then by incubating at 37°C for 15 minutes followed by 3 minutes at 65°C. DNA was separated from the fragmented DE81 by centrifugation through a small polyallomer wool column in a MSE Centaur centrifuge at maximum speed for a minute. The DNA was recovered by phenol extraction and ethanol precipitation and redissolved in an appropriate volume of TE buffer.

2.7 TRANSFORMATION OF E. coli

i) Preparation of competent cells

This method was adapted from that of Hutchinson and Halverson (1980), and used primarily to prepare competent JM83 and JM83 recA E. coli. A small volume of overnight culture of the cells was grown at 37°C after inoculating Luria broth with a single colony of freshly streaked cells from a LUA plate. 1 ml of the culture was used to inoculate 50 ml of fresh medium and grown to an OD560 = 0.36-0.44; approximately 90 to 120 minutes. The cells were pelleted by centrifugation at 3 000 rpm for 5 minutes at room temperature in a MSE Centaur bench-top centrifuge, or a SS34 rotor in a Sorvall RC5B centrifuge. The cells were resuspended in 20 ml of ice-cold sterile 0.05 M CaCl₂ and placed on ice for 15 minutes. The cells were pelleted by centrifugation, as above, and resuspended in 5 ml ice-cold sterile 0.05 M CaCl₂, 5% glycerol. The cells were snap frozen in 200 µl aliquots in sterile 1.5 ml microfuge tubes by placing the tubes in a dry ice/ethanol bath. The frozen cells were stored at -80°C. Such cells can remain competent for several months, although the efficiency of transformation drops rapidly, as much as 50% in six days (M. Woodhouse, personal communication).
ii) Transformation

10 ng of plasmid DNA in 100 μl of freshly prepared 10 mM Tris-HCl, 10 mM MgCl₂, 10 mM CaCl₂, pH 7.4, was added to 200 μl of competent cells thawed on ice. The mixture was kept on ice for 25 minutes, the cells were then heat-shocked at 37°C for 90 seconds then held at room temperature for 10 minutes. 1 ml of Luria broth was added and the cells were incubated at 37°C for 60 minutes to allow expression of antibiotic resistance. 20 μl and 200 μl aliquots of the transformed cells were plated onto LUA plates containing the necessary selective antibiotic. If appropriate, BCIG and IPTG were added to a concentration of 50 μg/ml to the plates.

iii) Transformation of commercially prepared competent cells

For critical transformations where high transformation efficiencies were necessary, such as in the generation of mini-libraries of size-selected genomic DNA described in Chapter 3, commercially prepared competent cells E. coli DH5 and DH5α were used.

To 100 μl of freshly thawed competent cells in a chilled polypropylene tube was added 1-10 ng of diluted ligated DNA in ONLY 1 μl. The cells were gently shaken and left on ice for 30 minutes, heat-shocked at 42°C for 45 seconds and placed on ice for 2 minutes. 0.9 ml of S.O.C. at room temperature was added and the tube shaken at 225 rpm at 37°C for 60 minutes. The cells were plated on LUA plates supplemented with the required antibiotic. If DH5α cells were used, BCIG and IPTG were added to the plates as indicated above. The plates were incubated at 37°C overnight. Transformation efficiencies of 1 x 10⁷ per μg DNA were routinely achieved.

2.8 PLASMID DNA

i) Mini-preps

This method is modified from that of Ish-Horowicz and Burke (1981).

A 5 ml culture of the plasmid containing strain of E.
coli was grown, by shaking in a conical flask, at 37°C in Luria broth supplemented with the appropriate antibiotic. 1.5 ml of the culture was harvested by a 20 second spin in a MSE Micro-Centaur microcentrifuge. The supernate was removed and the pelleted cells were resuspended in 100 μl of Solution I (25 mM Tris-HCl, 50 mM glucose, 10 mM EDTA, pH 8.0) and incubated for 5 minutes at room temperature. 200 μl of Solution II (freshly prepared 0.2 N NaOH, 1% SDS) was next added with gentle mixing to disrupt the cells, and left on ice for 5 minutes. To precipitate chromosomal DNA and most of the proteins, 150 μl of Solution III (5 M potassium acetate, pH 4.8) was then added with mixing. After 5 minutes on ice the precipitate was pelleted by a 1 minute spin in a microcentrifuge and 400 μl of supernate was transferred to a fresh tube. Plasmid DNA was precipitated by the addition of 800 μl of 100% ethanol, ensuring there was complete mixing of the solutions, and incubating for 2 minutes at room temperature. The DNA was pelleted by a 1 minute spin in a MSE Micro-Centaur microcentrifuge and rinsed in 500 μl of 70% ethanol. The pellet was spun down, the ethanol carefully removed, and the pellet resuspended in 50 μl TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 7.0).

To remove residual proteins 25 μl of 7.5 M ammonium acetate was added with mixing and left for 5 minutes at room temperature. The contaminating proteins were pelleted by a 2 minute spin in a microcentrifuge. The supernate was transferred to a fresh tube and the plasmid DNA precipitated by the addition of 190 μl of 100% ethanol and incubating at room temperature for 5 minutes. The DNA was rinsed in 70% ethanol as before and redissolved in 50 μl TE buffer at 65°C for 10 minutes.

The addition of ammonium acetate was found to remove efficiently any proteins which can interfere with subsequent manipulations of the plasmid DNA; and is much easier to perform than phenol extraction when processing several samples simultaneously.

Approximately 5-10 μg of plasmid DNA can be prepared by this method. Although containing very little E. coli chromosomal DNA large amounts of RNA are present. This could
largely be removed by the addition of 0.1 volume of 2 μg/ml boiled RNase A during restriction endonuclease digests of the DNA.

ii) Large scale plasmid preps

This method is a scaled up version of the mini-prep method with the addition of a caesium chloride gradient centrifugation purification step.

A 10 ml culture of the plasmid containing E. coli strain was grown with shaking at 37°C in Luria broth overnight, with the appropriate antibiotic to provide selection for the plasmid. 8 ml of the culture was used to inoculate 400 ml of pre-warmed Luria broth plus antibiotics in a 2 l flask. This culture was grown overnight at 37°C with shaking. If, however, the plasmid could be amplified by chloramphenicol, 80 mg chloramphenicol was added when the culture attained an optical density of 1.5 at 600 nm. This culture was then allowed to grow overnight as above.

The bacteria were harvested by centrifugation at 8 200 g for 10 minutes at room temperature in a Sorvall GS3 rotor and Sorvall RC5B High Speed Centrifuge. The supernate was removed and the pellet resuspended in 12 ml Solution I (as above) and left at room temperature for 10 minutes. 24 ml of Solution II was added with gentle mixing and left on ice for 5 minutes, followed by the addition of 12 ml Solution III and a further 15 minute incubation on ice. The mixture was centrifuged at 16 000 g for 10 minutes at 4°C in a Sorvall HB4 rotor in a RC5B centrifuge, to pellet chromosomal DNA and proteins. The supernate was filtered through polyallomer wool into an acid washed measuring cylinder. A 0.6 volume of propan-2-ol was added with mixing to precipitate the nucleic acids. After a 10 minute incubation at room temperature the nucleic acids were pelleted by centrifugation as in the previous step, rinsed in 70% ethanol, spun down, and resuspended in 3 ml TE buffer and pooled. For each ml of DNA solution 1 g CsCl and 0.1 ml ethidium bromide (10 mg/ml) were added. The mixture was thoroughly mixed and centrifuged at 16 000 g for 10 minutes at room temperature in the RC5B centrifuge. Both the resulting pellet and pellicle were discarded.
The density of the solution was checked using a refractometer, and adjusted, if necessary, to a value of 1.3874 with either TE buffer or CsCl solution of density 1.57 g/ml (a refractive index of 1.3874 is equivalent to a density of 1.57 g/ml - which is approximately 1 g CsCl to 1.1 ml TE). The solution was split between two Beckman Quik-Seal polyallomer tubes (16 mm x 76 mm) and topped up with CsCl solution of 1.57 g/ml density. The tubes were sealed and centrifuged in a Beckman 75Ti rotor at 378 000 g overnight at room temperature in a Sorvall OTD 65D Ultracentrifuge.

The plasmid band was removed from the gradient with a syringe and 18 guage needle and transferred to a 15 ml Corex tube. The ethidium bromide was removed by repeated extraction with equal volumes of propan-2-ol saturated with CsCl. The DNA was dialysed extensively against large volumes of TE to remove the CsCl.

The concentration and quality of DNA was determined by measurement of optical density at 260 nm and by agarose gel electrophoresis.

2.9 PREPARATION OF RADIO-LABELLED DNA FRAGMENTS

i) Oligo-labelling of isolated DNA fragments

This method is adapted from that of Feinberg and Vogelstein (1984).

The fragment DNA to be labelled was cleaved from its plasmid by appropriate restriction endonuclease digests and fractionated by electrophoresis on a 0.6% low gelling temperature agarose. The fragment was excised from the gel and the DNA concentration estimated as described in Section 2.2 (viii). The tube was placed in a boiling water bath for 7 minutes to melt the gel and denature the DNA. If the fragment was to be labelled immediately it was placed in a 37°C water bath for 10 minutes. DNA for use as oligo-labelled probe was stored at -20°C, and before each subsequent labelling reboiled for 3 minutes and kept at 37°C for 10 minutes.
Labelling reactions were set up by mixing:

- OLB: 3.0 μl
- BSA (10 mg/ml): 0.6 μl
- DNA (up to 10 ng): x μl
- \( \alpha^\text{32P} \text{dCTP} \) (111 × TBq/mmol, 0.37 MBq/μl): 1.0 μl
- Klenow polymerase (1 unit/μl): 0.6 μl
- H₂O to a final volume of 15 μl.

(Note: 1 Ci is equal to 3.7 × 10^6 MBq)

The reactions were allowed to proceed at room temperature for at least 5 hours, or at 37°C for 1 hour. Fragments of less than 1 kb were allowed to proceed overnight. The reaction was terminated by the addition of 85 μl H₂O before boiling for 3 minutes and adding to the hybridisation mix. OLB is made from the following components and stored at -20°C:

2 volumes Solution A; (625 μl 2 M Tris-HCl, 25 μl 5 M MgCl₂, 18 μl β-mercaptoethanol, 350 μl H₂O, and 5 μl each of dATP, dGTP, dTTP (these were dissolved in 3 mM Tris-HCl, 0.2 mM EDTA, pH 7.0 at a concentration of 0.1 M))

5 volumes Solution B; (2 M HEPES titrated to pH 6.6 with NaOH)

3 volumes Solution C; (hexadeoxyribonucleotides suspended in 3 mM Tris-HCl, 0.2 mM EDTA, pH 7.0 at 90 OD units/ml)

ii) Determination of probe activity

Prior to hybridisation the extent of radioactive nucleotide incorporation into the probe was estimated.

1 μl of the terminated labelling reaction was diluted in 9 μl H₂O. Two 5 μl samples were spotted onto 2 cm x 2 cm
squares of Whatman DE-81 paper, and dried at room temperature for 15 minutes. One sample was washed for 6 x 5 minutes in 0.5 M Na₃HPO₄, rinsed twice for 1 minute in distilled H₂O, and then rinsed twice for 1 minute in 95% ethanol. This process washes off any unincorporated nucleotides.

After drying the radioactivity in each sample was compared by measuring Cerenkov radiation in a Packard Tricarb scintillation counter. If the washed sample contained at least 40% of the radioactivity present in the unwashed sample, the probe was considered to have been labelled to a sufficiently high specific activity for use.

2.10 SOUTHERN BLOTTING

The transfer of DNA from agarose gels to nylon filters (Amersham Hybond-N) is based on the method originally described by Southern (1975). The protocol described is for a 100 ml agarose gel, for gels of other sizes the volumes and other quantities quoted were altered in proportion.

The gel was electrophoresed and photographed, with a ruler placed alongside the gel for subsequent sizing of bands on autoradiographs. The gel was slid onto a glass plate in a tray containing 250 ml of 0.25 M HCl and washed by gentle agitation for 7 minutes. After a brief rinse in distilled water, the gel was transferred into 250 ml denaturing solution (0.5 M NaOH, 1.5 M NaCl) and washed for a further 30 minutes. The gel was again rinsed in distilled water and transferred into 250 ml neutralising solution (0.5 M Tris-HCl, 3 M NaCl, pH 7.4) and washed for 30 minutes.

Gels containing digested genomic DNA were transferred to a blotting apparatus (see Dalgleish, 1987, for a description of this) and the edges of the gel masked with cling film. Gels containing cloned DNA were simply placed on a glass plate. A sheet of nylon filter and a sheet of Whatman No 1 paper were cut to the size of the gel, and wetted in 3 x SSC (0.45 M NaCl, 0.045 M Na citrate, pH 7.0). The nylon was placed onto the gel and overlaid with the Whatman paper. A 5 cm pile of paper towels was placed on top of the Whatman
paper, and on top of this a glass plate with a 0.5 kg weight above it. The blot was allowed to proceed for at least 5 hours for a gel containing genomic DNA, and 2 hours for a gel containing cloned DNA. The paper towels were changed for dry ones as they became wetted, every 5 minutes for the first 30 minutes and then after 15 minute intervals.

After the blot was dismantled, the filter was rinsed in 3 x SSC and blotted dry. The filter was dried thoroughly in a 65°C oven, then wrapped in Saran-Wrap and placed DNA side down on a UV transilluminator. DNA was covalently bound to the filter by exposure to UV light for 5 minutes. Filters were stored at 4°C until required.

2.11 HYBRIDISATIONS

Three methods were generally employed, for most purposes they were interchangeable.

Southern filters were prepared as described in the previous section. Colony filters were prepared by the method of Hanahan and Meselson (1980), described in detail in Section 2.12. Replica filters were normally made of bacteria plated directly onto LUA, if the number of colonies was less than 500 per-82 mm diameter plate; otherwise the bacteria were plated onto filters (Amersham Hybond-N) on plates. Filters for screening M13 and lambda plaques were prepared by the method of Benton and Davies (1977), described in detail in Section 2.13.

i) Denhardt's mix

Filters were washed prior to hybridisation in 3 x SSC, 2 x Denhardt's solution (0.04% Ficoll, 0.04% polyvinyl pyrrolidine, 0.04% BSA, all as w/v in H₂O), 200 μg/ml sonicated and denatured herring sperm DNA (boiled for 10 minutes before adding to the mixture), 6% PEG6000, 0.1% SDS at 65°C for 1 hour.

Hybridisations were carried out in the same conditions as the pre-hybridisation, except that 5 x Denhardt's was used, for 16-18 hours with radio-labelled probe at 0.5 ng/ml
(boiled for 3 minutes before adding to the hybridisation mix).

After hybridisation, the filters were rinsed four times and washed twice, for 10 minutes each, in 3 x SSC, 0.1% SDS at 65°C. This was followed by four 15 minute washes at the required stringency. For probes containing homologous sequence to the DNA on the filter this was 0.1-0.5 x SSC, 0.1% SDS at 65°C. If the sequences being hybridised had only limited similarity, then filters were washed at 1 x SSC, 0.1% SDS at 65°C.

ii) Marvel mix

Filters were prewashed in 1.5 x SSPE (1.5 mM EDTA, 0.27 M NaCl, 15 mM NaH₂PO₄, pH 7.7), 1% SDS, 6% PEG6000, 5 mg/ml dried milk powder (Cadbury's Marvel) at 65°C for 1 hour. It is important that the milk powder is completely dissolved in water before the addition of components of the mix. Hybridisations were carried out in the same conditions for 16-18 hours at a probe concentration of 0.5 ng/ml.

Post-hybridisation washing was as above.

iii) Phosphate mix

This method is adapted from that of Church and Gilbert (1985), and was used predominantly for screening replica filters of colony and plaque lifts. The phosphate mix gave a reduced background, and positives were more intense than with either of the other methods.

Filters were washed in 0.5 M Na₂HPO₄, 7% SDS, 1 mM EDTA, pH 7.2 at 65°C for 30 minutes. Hybridisations were in the same conditions for 16-18 hours at a probe concentration of 0.5 ng/ml.

Post-hybridisation washing of filters was carried out three times in 0.5 M Na₂HPO₄, 1% SDS for 5 minutes at 65°C. Filters were then washed to the stringency required in SSC and SDS as before.

iv) Autoradiography

Filters were blotted to remove excess liquid but were not allowed to dry completely. Filters were wrapped in Saran-Wrap and autoradiographed for an appropriate period, either
at room temperature without intensifying screens or at -80°C with Cawo intensifying screens.

v) Probe stripping
   Filters were kept moist for autoradiography so that the probe could be stripped off and the filters hybridised with another probe. To strip the probe from nylon filters, they were soaked in 0.4 M NaOH at 45°C for 3 minutes, followed by washing in 0.1 x SSC, 0.1% SDS, 0.2 M Tris-HCl, pH 7.5 at 45°C for 30 minutes.

2.12 PACKAGING EXTRACTS FOR COSMID CLONING

Packaging extracts were prepared by the method of Hohn (1979). The E. coli host strains were checked by streaking onto LUA plates, growth should occur at 32°C but not 42°C. This tests for the presence of the mutation of the cl-gene product of the lysogen that renders it temperature-sensitive. The recA function of hosts was checked by testing for UV-sensitivity (see Maniatis et al., (1982) for procedure).

i) Freeze Thaw Lysate (FTL)
   Three 500 ml cultures of E. coli BHB2688 were grown in Luria broth at 30°C. When the cultures had attained an OD600 = 0.3, lysis was induced by placing the flasks in a waterbath at 45°C for 15 minutes. Induction was checked by adding chloroform to a few ml of the culture in a test-tube. Clearing of the culture indicates lysis has occurred. The cultures were shaken hard at 37°C for a further 60 minutes, cooled on ice, and the cells recovered by centrifugation in a Sorvall GS-3 rotor at 9 000 rpm at 4°C for 10 minutes in a Sorvall RC5B centrifuge. The supernate was allowed to drain off on ice and each pellet (250 ml worth of the original culture) was resuspended in 0.5 ml 10% sucrose, 50 mM Tris-HCl, pH 7.5. All the pellets were pooled into two 50 ml polypropylene tubes, to each was added 75 μl fresh lysozyme solution (2 mg/ml in 0.25 M Tris-HCl, pH 7.5). After gentle but thorough mixing, the tubes were quick frozen in liquid
nitrogen, then thawed at room temperature and kept on ice. To each tube 75 μl buffer Ml (10 mM Tris-HCl, 30 mM spermidine, 18 mM MgCl₂, 15 mM ATP, 0.02% (v/v) β-mercaptoethanol, pH 7.5) was added with mixing. The solutions were spun in a SS34 rotor at 17 000 rpm at 4°C for 50 minutes in a Sorvall RC5B. The supernate was distributed into precooled screw-capped microfuge tubes in 25 μl aliquots, and quick frozen in liquid nitrogen and stored at -70°C. During aliquoting the supernate became viscous, it was found that recentrifuging to clear the supernate had no noticeable effect on the efficiency of the packaging reactions.

ii) Sonicated Extract (SE)

A 500 ml culture of BHB2690 in Luria broth was grown at 30°C to an OD600 = 0.3; induced as above, and grown for 60 minutes at 37°C with hard shaking. The cells were cooled on ice and harvested in a GS-3 rotor at 9 000 rpm for 10 minutes at 4°C. The supernate was drained off and each pellet resuspended in 0.5 ml buffer A (20 mM Tris-HCl, 3 mM MgCl₂, 0.05% (v/v) β-mercaptoethanol, pH 8.0). All samples were pooled into one 50 ml polypropylene tube and diluted with 2.6 ml buffer A. This suspension was sonicated, without foaming, until no longer viscous (MSE 150 watt Ultrasonic Disintegrator, Exponential probe end-diameter 1/8 inch, power setting 4). The debris was pelleted by spinning at 6 000 rpm for 6 minutes in SS34 rotor. The supernate was aliquoted into precooled screw-capped microfuge tubes in 25 μl aliquots and quick frozen in liquid nitrogen, then stored at -70°C.

Both FTL and SE can be stored at -70°C for several years without appreciable loss of packaging efficiency. Buffer Ml normally contains 60 mM putrescine. For cloning in cosmids, the absence of putrescine increases the efficiency of packaging of cosmids that contain large inserts (Maniatis et al., 1982).

2.13 COSMID CLONING

The method initially used was that of Bates and Swift.
(1983).

i) Preparation of vector DNA

20 μg vector c2XB DNA, prepared by the large-scale plasmid preparation described in Section 2.7, was digested with Sma I (40 units) and BamH I (20 units) at 37°C for 4 hours to ensure complete digestion. The enzymes were inactivated by the addition of EDTA at final concentration 20 mM, extracted with phenol/chloroform and recovered by ethanol precipitation. The DNA was redissolved in TE buffer at a concentration of 1 μg/μl.

ii) Preparation of insert DNA

Human placental DNA had been prepared previously by Dr. Raymond Dalgleish from fresh placentae by lysis in 0.6% tri-iso-propynaphthalene sulphonic acid, 0.8% butan-2-ol and 0.3% SDS, followed by extraction with phenol/chloroform and precipitation with ethanol. The DNA was further purified by treatment with proteinase K and RNase A before further extraction with phenol/chloroform and a final ethanol precipitation. The quality of the starting material was checked by agarose gel electrophoresis of undigested DNA. The quality of DNA varied considerably between different preparations. Of several preparations checked only those with DNA of an average size of 100 kb or greater were used for cosmid cloning.

iii) Partial digests

Because of its simplicity DNA fragments to be cloned were obtained by partial digestion of high molecular weight DNA with Mbo I, a restriction enzyme that recognises a 4 bp sequence and generates ends that can be ligated to the BamH I cohesive ends of the vector.

Wherever possible, physical manipulation of the insert DNA was minimised to prevent shearing; pipetting was achieved using pipette tips that had had 0.5 cm removed from the ends. The DNA to be cloned was partially digested with Mbo I. The conditions required were determined by serial dilution of Mbo I. A reaction mixture of placental DNA and restriction enzyme
buffer was prepared at a final concentration of 1 µg DNA/15 µl. This was mixed well by inversion of the tube several times. 30 µl was dispensed into one tube and 15 µl aliquots into several other tubes. All tubes were chilled on ice. 2 units of Mbo I were added to the first tube and mixed well. The concentration of the enzyme was, therefore, 1 unit/µg DNA. 15 µl of the mixture was transferred to the second tube. The enzyme concentration was now 0.5 units/µg DNA. The twofold serial dilution was continued through the series of tubes. The tubes were placed in a 37°C water-bath for 60 minutes. The reactions were stopped by heating to 70°C for 15 minutes. To each tube was added 1.5 µl orange dye and the reaction products were analysed by electrophoresis through a 0.3% agarose gel.

The correct partial conditions were selected by masking off the DNA that was greater than 50 kb and less than 35 kb in a photograph of the gel (Figure 4.5). The various gel tracks were compared to estimate the degree of digestion that produced the maximum amount of DNA in the required size range. The correct partial was not the one with the most DNA in this size range, but the next most partial. The rationale for this is discussed by Seed et al. (1982). In subsequent large-scale digestions, half the enzyme concentration that gave maximum fluorescence of 35 to 50 kb was used. The large-scale digests were performed in 15 µl aliquots at a DNA concentration of 1µg/15 µl, exactly as in the test reactions. The reaction was stopped by incubation at 70°C for 15 minutes. The digests were pooled and calf alkaline phosphatase added at a final concentration of 0.012 units per µg DNA. After incubation at 37°C for 30 minutes the phosphatase was inactivated by incubation at 70°C for 90 minutes. The DNA was used in ligations without further treatment.

iv) Recovery of digested DNA

a) Spermine precipitation

Spermine precipitation was performed as described by Hoopes and McClure (1981).
0.1 M spermine-HCl was added to the DNA to a final concentration of 7 mM and incubated for 15 minutes on ice/water. The DNA was pelleted by centrifuging for 10 minutes at 4°C in a Micro-Centaur microfuge. The pellet was washed in 1 ml 75% ethanol, 1 x buffer 2 (0.3 M Na acetate, 0.01 M Mg acetate) for 45 minutes with frequent mixing. This was repeated once before rinsing the pellet in 70% ethanol. The pellet was redissolved in TE buffer at 1 μg/μl.

b) Nensorb 20 purification column

The column was obtained from NEN Research Products, Du Pont Company, Wilmington, Delaware, U.S.A. The column was primed with 2 ml 0.1 M Tris-HCl, 10 mM triethylamine, 1 mM K2EDTA, pH 7.7. DNA was loaded onto the column in the same buffer in a volume of 200-400 μl. The sample was washed with 3 ml of the same buffer to remove salts. DNA was eluted from the column with 1 ml of 20% ethanol, proteins remain bound to the column. The DNA was dried down in a Savant Speed-Vac Concentrator, and redissolved at a concentration of 1 μg/μl.

v) Size selection of DNA on NaCl gradients

As an alternative to dephosphorylating partially digested DNA, and making use of the natural selectivity of in vitro packaging to isolate 35-45 kb fragments, it is possible to physically separate partially digested fragments of the correct size. NaCl gradients were utilised to purify fractions containing partially digested DNA of the required size range, as described by Dillela and Woo (1986).

200 μg of placental DNA was digested as described in Section 2.13 (iii). The reaction was stopped by adding EDTA to 20 mM final concentration. 1 μg of DNA was analysed on a 0.3% agarose gel to ensure the required size distribution had been attained. The DNA was recovered by spermine precipitation and redissolved in 100 μl TE buffer at 4°C for 2 days.

100 μg of DNA was fractionated on a 13 ml NaCl gradient (1.25 - 5 M NaCl in TE buffer). The gradient was centrifuged for 3.5 hours at 39 000 rpm in a Beckman SW40.1 rotor at 18°C. 12 fractions of 1 ml each were collected, each fraction
was diluted in an equal volume of TE buffer and precipitated with ethanol, and redissolved in 400 µl TE. The middle six fractions, as indicated by Dillela and Woo (1986), were centrifuged in the SW40.1 Beckman rotor for 1 hour at 20,000 rpm at 4°C. The pelleted DNA was redissolved in 400 µl of 0.3 M sodium acetate pH 5.5, and ethanol precipitated in a 1.5 ml microcentrifuge tube. The pellet was rinsed with 70% ethanol and redissolved in 20 µl of TE.

vi) Ligations

Ligations of vector to insert DNA were performed in a buffer of 66 mM Tris-HCl, pH 7.5, 10 mM DTT, 5 mM MgCl₂, and 5 mM ATP. This concentration of ATP has been reported to inhibit blunt-end ligation (Ferretti and Sgaramella, 1981), which prevents vector concatamerisation. Ligations were carried out at a mass ratio of vector to insert of 2:1 at a concentration of 300 µg DNA/ml. Ligations were incubated at 15°C overnight before being packaged.

vii) In vitro packaging

Packaging extracts were prepared as described in Section 2.16. Sonicated Extract (SE) and Freeze Thaw Lysate (FTL) were thawed on ice immediately before use, and added to the side of the tube containing the other components of the reaction. The reactions were started by a few seconds of centrifugation in a microfuge. The reactions were as below:

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume (µl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ligated DNA</td>
<td>6.0</td>
</tr>
<tr>
<td>Buffer A</td>
<td>7.0</td>
</tr>
<tr>
<td>Buffer Ml</td>
<td>2.5</td>
</tr>
<tr>
<td>SE</td>
<td>15.0</td>
</tr>
<tr>
<td>FTL</td>
<td>25.0</td>
</tr>
</tbody>
</table>

The reaction mixture was incubated at 25°C for 60 minutes. 100 µl of lambda buffer and a few drops of chloroform were added, and the packaged DNA was stored at 4°C. 6 µl of the packaged cosmids (equivalent to 108 ng insert DNA) was used to transduce 0.2 ml of a saturated culture of E. coli 1046, grown overnight in LB supplemented...
with 0.4% maltose. Phage adsorption was allowed for 20 minutes at 37°C, 1 ml of prewarmed LB was added and incubated for 45 minutes at 37°C to permit expression of antibiotic resistance. Aliquots were plated onto LUB plates supplemented with ampicillin at 50 μg/ml. The plates were incubated for 12 to 15 hours at 37°C.

2.14 SCREENING HUMAN COSMID LIBRARY H2A

Human cosmid library H2A was received from Dr. Alastair Craig, EMBL, Heidelberg. The library was constructed with cosmid vector pcos2EMBL as described by Poustka et al. (1984). After amplification on E. coli DH1 the recombinants had been packaged in vivo by superinfection. The library was provided as a phage lysate, which was titred on E. coli DH5, and estimated to contain 7.5 x 10^5 colony forming units per μl. 1 μl of the phage lysate was used to transduce E. coli DH5, which were plated onto nylon filters (Amersham Hybond-N) on three 20 cm x 20 cm LUB plates with kanamycin at 50 μg/ml. The plates were incubated at 37°C until the colonies were 0.5 mm in diameter. The recombinant colonies were screened by a modification of the method of Hanahan and Meselson (1980), described in detail by Maniatis et al. (1982). Two sets of replica filters were made for each plate using filter-to-filter contact. The replica and master filters were marked around the edge with asymmetrically positioned needle holes to allow future alignment of the autoradiographs to the master filters. After replicas had been made the filters were incubated at 37°C until the colonies were 1-2 mm in diameter. Master filters were stored for a short time at 4°C and for longer at -80°C. The replica filters were transferred to LUB plates supplemented with chloramphenicol at 250 μg/ml, and incubated overnight at 37°C to allow amplification of the cosmids.

The filters were transferred onto Whatman 3MM filters soaked in 1.5 M NaCl, 0.5 M NaOH for 15 minutes to lyse the bacteria and denature the DNA. The filters were then transferred to Whatman 3MM soaked in neutralising solution.
(1.5 M NaCl, 0.5 M Tris-HCl, 1 mM EDTA, pH 7.4) for 3 minutes, this step was repeated once. After rinsing in 2 x SSC (0.3 M NaCl, 30 mM Na citrate) the filters were dried and the DNA covalently fixed to the filters by UV cross-linking. The filters were washed in large volumes of 3 x SSC, 0.1% SDS at 65°C overnight, with frequent changes of the solution, to remove cell debris and so reduce the background caused by non-specific binding of the radio-labelled probe during hybridisation. Pre-hybridisation and hybridisation were carried out using the Denhardt's mix as described in Section 2.10. Autoradiography was for three days at -80°C with intensifying screens.

2.15 PREPARATION OF \( \lambda L47.1 \) ARMS FOR LAMDBA CLONING

i) Blattners

Serial dilutions of a phage stock of \( \lambda L47.1 \) were made in the range \( 10^{-1} \) to \( 10^{-6} \) in a final volume of 100 \( \mu l \) with lambda buffer. Each dilution was mixed with 100 \( \mu l \) of E. coli C600 (grown overnight at 37°C and diluted 1:1 in fresh Luria broth) in a small capped test-tube. The phage and cells were mixed and incubated at room temperature for 15 minutes. 3 ml of BTL, containing 10 mM MgSO\(_4\) and 20 \( \mu g/ml \) thymine, at 50°C was added and the contents plated onto BLA plates; these were incubated at 37°C overnight.

Four lots of 400 ml Luria broth with 10 mM MgSO\(_4\) and 20 \( \mu g/ml \) thymine were seeded with 6 plaques each from the plates. These were grown overnight in 2 l unbaffled flasks. If successful, the culture contains a messy debris of lysed cells. To each flask were added 10 \( \mu l \) 20 mg/ml pancreatic DNase and 10 \( \mu l \) 10 mg/ml pancreatic RNase A, and left for 15 minutes at room temperature. 11.7 g NaCl (to 1M) was added with shaking, to dissolve the salt, and left on ice for 30 minutes. The contents of the flasks were centrifuged at 8000 rpm for 10 minutes at 4°C in a GS3 rotor in a Sorvall RC5B centrifuge. 20 g solid PEG6000 (to 10% v/v) were added slowly, to avoid clumping, to each supernate. The mixture was left on ice for at least 60 minutes, overnight if possible,
to precipitate the phage which were pelleted by centrifugation as above. The phage pellet was drained and resuspended in 3 ml lambda buffer at 4°C, and the phage cleared by spinning at 8000 rpm for 2 minutes at 4°C in a HB4 rotor. The supernate was collected, the pellet was resuspended in 1.5 ml lambda buffer, centrifuged and the supernate pooled with that of the previous step.

To band the phage a CsCl step gradient was used. A stock of 65 g CsCl and 35 ml H2O was made. The gradient comprised CsCl solution of three densities ($\rho$), these were prepared as below:

<table>
<thead>
<tr>
<th>CsCl stock</th>
<th>lambda buffer</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.99 ml</td>
<td>3.00 ml</td>
<td>1.7</td>
</tr>
<tr>
<td>7.77 ml</td>
<td>6.23 ml</td>
<td>1.5</td>
</tr>
<tr>
<td>3.99 ml</td>
<td>6.23 ml</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Step gradients were set up in 13.5 ml SW40 ultracentrifuge tubes with 2 ml of $\rho$ = 1.7 plus 3 ml of $\rho$ = 1.5 and 2 ml of $\rho$ = 1.3. The phage suspension was overlaid, the tube balanced with one containing lambda buffer and spun at 35000 rpm for 60 minutes in a SW40 rotor in a Beckman L5 65 ultracentrifuge at 0°C without brake. The phage band was collected from the interface of $\rho$ = 1.5 and $\rho$ = 1.3 by carefully removing CsCl from above with a syringe. The phage were diluted with lambda buffer to a final volume of 13.5 ml and pelleted by spinning as before, with the brake on. The phage pellet was resuspended in 1 ml lambda buffer.

ii) Preparing lambda DNA

The phage were lysed by the addition of 100 $\mu$l 10% SDS. Protein was extracted with 500 $\mu$l phenol. The aqueous layer was removed and the phenol re-extracted with 300 $\mu$l 10 mM Tris-HCl, pH 7.5. The aqueous phases were pooled and again extracted with phenol. The aqueous phase from this was made 0.2 M with respect to Na acetate, pH 5.6, and 2.5 volumes ethanol were added. The DNA precipitate was removed with a
drawn out Pasteur pipette, rinsed in 70% ethanol and dissolved in 300 µl 10 mM Tris-HCl, pH 7.5. 3 µl 20 mg/ml pancreatic RNase was added and incubated at 37°C for 10 minutes. After phenol extraction the DNA was precipitated four times with ethanol, redissolving each time in 300 µl 10 mM Tris-HCl, pH 7.5. The final pellet was rinsed in 70% ethanol, dried and dissolved in 200 µl 10 mM Tris-HCl, pH 7.5 and stored at 20°C. A yield of 20-100 µg was obtained.

200 µg λL47.1 DNA was digested with 300 units BamH I in 500 µl volume. Samples were removed at 20, 30 and 60 minutes to check, by agarose gel electrophoresis, for complete digestion. BamH I cleaves λL47.1 twice to produce fragments of sizes: 23.6, 10.4 and 6.6 kb. It is the smallest fragment ("stuffer") which is replaced by insert DNA for cloning. To 500 µl of the digested DNA was added 150 µl orange dye mix and loaded into a 1.5 ml slot in a 0.4% agarose gel. The gel was run at 120 volts for 2-3 hours to resolve the three fragments. The gel was viewed under longwave ultraviolet light. The two larger fragments were recovered by electroelution onto dialysis membrane as described in Section 2.6. The DNA was dissolved at a final concentration of 1 µg/µl, about 50 µg DNA could be recovered. To 40 µl λL47.1 arms were added 4 µl Tris-HCl, pH 7.5 and 2 µl 0.2M MgCl₂. The arms were annealed at 42°C for 60 minutes, the efficacy of annealing was checked by gel electrophoresis of two aliquots, one which had been heated at 68°C for 3 minutes.

2.16 ISOLATION OF A MULTI-ALLELIC LOCUS (λKP20.1) FROM A
λL47.1 GENOMIC LIBRARY

i) Construction of a lambda library

Human genomic DNA that had been digested to completion with Sau3A I was generously provided by Dr. Nicola Royle. DNA from several unrelated individuals that had been so treated was pooled and electrophoresed on an agarose gel. The fraction of DNA of size 5 to 20 kb was recovered by electroelution onto dialysis membrane.
100 ng of the fractionated DNA was ligated to 100 ng of λL47.1 arms, prepared as described in section 2.15, in a final volume of 5 µl. After an overnight incubation at room temperature the ligated DNA was precipitated by the addition of 1 µl 2M Na acetate, pH 5.6 and 15 µl 100% ethanol. The precipitate was pelleted by centrifugation for 10 minutes in a Micro-Centaur, washed in 80% ethanol, dried and resuspended on ice in 5 µl packaging buffer A (20 mM Tris-HCl, 3 mM MgCl₂, 0.05% (v/v) β-mercaptoethanol, 1 mM EDTA, pH 8.0).

In vitro packaging of the ligated DNA was accomplished with a commercial packaging kit, Gigapack Plus (Stratagene). 5 µl ligated DNA in packaging buffer was added to 10 µl thawed FTL, to this was added 15 µl of thawed SE. The contents were gently mixed and incubated at room temperature for 2 hours. 500 µl of lambda buffer and a few drops of chloroform were added, the phage were stored at 4°C.

150 µl packaged DNA was used to transduce 150 µl E. coli WL95 cells, the two were mixed in a glass test-tube and incubated at room temperature for 15 minutes. 3 ml of BTL was added and the mixture poured onto BLA plates. Appropriate controls of buffer and cells alone were set up. 5 000 to 10 000 plaques were obtained after overnight incubation of the plates at 37°C. These were screened using the method of Benton and Davies (1977).

ii) Screening of the lambda library

Library plates were chilled at 4°C for 30 minutes. Nitrocellulose filters (Amersham Hybond-C) were placed onto each plate, the filters were marked round the edge with a unique pattern of needle holes. After 5 minutes the filters were removed with a forcep and floated phage side down on denaturing solution (1.5M NaCl, 0.1 M NaOH). After 1 minute the filters were submerged in neutralising solution (2 x SSC, 0.2 M Tris-HCl, pH 7.5) for a further minute. The filters were then blotted dry on Whatman 3MM paper and baked at 80°C for 4 hours. The library plates were stored at 4°C wrapped in aluminium foil.

The nitrocellulose filters were hybridised with radio-labelled probe KP600, isolated from the human α1(I)
procollagen gene; using the Marvel method (see Section 2.10). The radio-active filters were autoradiographed overnight at -80°C with intensifying screens.

A 5 ml overnight culture of E. coli ED8910 was grown with shaking at 37°C in LUB with 10 mM MgSO₄ and 20 µg/ml thymine. Plaques corresponding to positives identified by the first screening were picked from the library plates with a standard plaque picker and put into 500 µl lambda buffer, one drop of chloroform was added. Various dilutions of the phage stock were made with lambda buffer in a final volume of 100 µl and used to inoculate 100 µl of an undiluted overnight culture of E. coli ED8910. The phage and cells were mixed in a glass test-tube and incubated at room temperature for 15 minutes. 3 ml BTL with 10 mM MgSO₄ and 20 µg/ml thymine was added and the contents plated onto BLA plates. These were incubated overnight at 37°C. The plaques were screened as above.

After four rounds of screening, single isolated plaques were obtained.

iii) Mini-preps of phage DNA

100 µl E. coli ED8910 was added to each of three glass test-tubes. 5 µl, 20 µl and 100 µl of phage stock was added to the aliquots of cells. After incubation at room temperature for 15 minutes, 8 ml of LUB with 10 mM MgSO₄ and 20 µg/ml thymine was added to each tube, these were incubated at 37°C overnight on a shaking rack with the tubes at an angle.

The following morning the contents of all three tubes were pooled into one 30 ml Corex tube and centrifuged for 5 minutes at 9 000 rpm in a Sorvall RC5B high speed centrifuge. The supernate was transferred to a fresh 30 ml Corex tube and one-third volume of 10% PEG6000 in 2.5 M NaCl was added with mixing and incubated on ice for 15 minutes. The phage were pelleted by centrifuging at 9 000 rpm for 10 minutes. The supernate was removed and the phage pellet resuspended in 0.5 ml lambda buffer. This was transferred to a microfuge tube and centrifuged in a Micro-Centaur for 2 minutes to pellet any
remaining cells. The supernate was transferred to a fresh microfuge tube and the phage were precipitated by adding 0.3 ml 10% PEG6000 in 2.5 M NaCl and incubating for 15 minutes on ice. The phage were pelleted by centrifuging for 5 minutes and resuspended in 250 μl lambda buffer. 50 μl 10% SDS was added to lyse the phage. 250 μl phenol/chloroform was added and mixed thoroughly and then centrifuged for 3 minutes. The aqueous layer was removed and extracted with phenol/chloroform a further four times. To maximise recovery of the DNA the phenol layer was re-extracted too, the aqueous layer of the re-extraction was added to the phenol layer of the first phenol extraction, and so on. To the final aqueous layer was added 20 μl 5 M Na acetate pH 6 and 500 μl 100% ethanol, and incubated at -80°C for 10 minutes. The DNA was pelleted by centrifuging for 10 minutes. The pellet was rinsed in 70% ethanol, dried and redissolved in 20 μl TE buffer for 10 minutes at 65°C. A yield of 20 ng/μl was obtained.

iv) Subcloning the phage insert into pUC13

10 μl (200 ng DNA) of recombinant phage DNA was digested with Sau3A I for 60 minutes and electrophoresed on a 0.7% agarose gel. The insert, of 7 kb size, was recovered in 60 μl by electroelution onto dialysis membrane; the DNA was precipitated by the addition of one-tenth volume of 5 M Na acetate, pH 6 and two volumes 100% ethanol. The final pellet was resuspended in 5 μl H₂O.

The insert DNA was ligated to 20 ng pUC13 vector DNA that had been digested with BamH I and treated with calf alkaline phosphatase, in a final volume of 10 μl. After an overnight incubation at room temperature the reaction mixture was heated to 65°C for 10 minutes. 5 ng of diluted ligated DNA was used to transform competent E. coli JM83recA cells as described in Section 2.15. The cells were plated on LUA plates supplemented with 50 μg/ml ampicillin, 50 μg/ml BCIG and 50 μg/ml IPTG. White recombinant colonies were picked of the plates and spotted onto nylon filters on LUA plates with ampicillin at 50 μg/ml. The filters were hybridised with probe KP600. Positives identified recombinants that contained
an insert of the repeat sequence. These were grown overnight in liquid culture (Luria broth with 50 µg/ml ampicillin). Plasmid DNA was isolated by the mini-prep method and dissolved in TE buffer at a final concentration of 2 µg/µl.

2.17 M13 SEQUENCING

i) Preparation of insert DNA

Two methods were used routinely for isolating DNA for ligation into M13 vector DNA.

a) Low gelling temperature agarose

Restriction endonuclease fragments, of less than 600 bp size, to be sequenced were isolated by digestion with the appropriate restriction endonuclease(s) and fractionating the products on low gelling temperature agarose. See Section 2.2 (viii) for details. Approximately 100 ng of DNA per fragment was isolated.

b) Sonication of isolated endonuclease fragments

Large restriction endonuclease fragments to be sequenced were isolated by electroelution from agarose gels as described in Section 2.6. The fragments were self-ligated overnight, necessitating the use of restriction enzymes that produce compatible ends to the fragments. Sonication was performed in a sonicating waterbath (Kerry Ultrasonics Ltd) containing 1-2 cm of water. A 1.5 ml micro-centrifuge tube containing 15 µg of isolated restriction endonuclease fragment (previously self-ligated) in a total volume of 30 µl (made up with water) was placed on the bottom of the waterbath for 4 x 30 second bursts of sonication. Between each burst the DNA solution was placed on ice and spun briefly to bring the solution back down to the bottom of the tube. The appearance of a "mist" on the sides of the tube was an indication of successful sonication. A 1 µl aliquot was electrophoresed against φX174 RF/Hae III markers on an agarose gel to check that after sonication the majority of the DNA was a smear between 1.2 and 0.6 kb in size. The
The sonicated DNA was end-repaired incubation overnight at 15°C in the following mixture:

\[
\begin{align*}
\text{DNA (in water)} & \quad 20.0 \ \mu l \\
10 \times \text{ligase buffer} & \quad 3.0 \ \mu l \\
\text{TM buffer} & \quad 3.0 \ \mu l \\
\text{Spermidine 0.1 M} & \quad 1.2 \ \mu l \\
\text{Sequence chase mix} & \\
\text{(0.25 mM solution of each dNTP in TM buffer)} & \quad 2.0 \ \mu l \\
\text{DNA polymerase I} & \quad 2.0 \ \mu l
\end{align*}
\]

10 x ligase buffer is: 500 mM Tris-HCl, 100 mM MgCl₂, 100 mM DTT, pH 7.5. TM buffer is: 100 mM Tris-HCl, 100 mM MgCl₂, pH 7.5.

After end-repair sonicated DNA was electrophoresed in a 1.5% preparative agarose gel against \( \Phi X174 \) RF/Hae III markers. DNA between 200-500 bp and 500-900 bp was collected separately on DE81 paper, as described in section 2.6, and redissolved in 20 \( \mu l \) water.

ii) Preparation of M13 vector DNA

M13mp18 or M13mp19 RF DNA was cleaved at the desired cloning site by the appropriate restriction endonuclease(s). If necessary the DNA was then dephosphorylated by treatment with calf alkaline phosphatase. The DNA was electrophoresed on a 0.6% low gelling temperature agarose gel, and the fragment isolated as described above. Generally 200 ng RF DNA was digested to produce a final concentration of 1.5 ng/\( \mu l \).
iii) Ligation of insert DNA into M13 vector DNA

a) Isolated restriction fragments

Ligation reactions were set up in 50 μl with 10 ng M13 DNA and insert DNA to give an equimolar ratio of ends. Reactions were incubated overnight at 15°C. After heating at 65°C for 10 minutes, 5 μl was used in the transformation reaction.

b) Sonicated DNA

The following ligation reaction mixes were prepared:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>size-selected DNA</td>
<td>1 μl</td>
<td>2 μl</td>
<td>4 μl</td>
</tr>
<tr>
<td>phosphatased M13 vector DNA</td>
<td>2 μl</td>
<td>2 μl</td>
<td>2 μl</td>
</tr>
<tr>
<td>10 mM ATP</td>
<td>1 μl</td>
<td>1 μl</td>
<td>1 μl</td>
</tr>
<tr>
<td>10 x ligase buffer</td>
<td>1 μl</td>
<td>1 μl</td>
<td>1 μl</td>
</tr>
<tr>
<td>H₂O</td>
<td>5 μl</td>
<td>4 μl</td>
<td>2 μl</td>
</tr>
</tbody>
</table>

400 units New England Biolabs T4 DNA ligase (equivalent to 1 Weiss unit) were added to each ligation mixture which were incubated overnight at 15°C.

iv) Transfection of recombinant M13 into E. coli JM101

Competent cells were prepared by a modification of the method of Kushner (1978).

E. coli JM101 was grown overnight with shaking, at 37°C, in Luria broth containing thiamine at 2 g/ml. 0.5 ml of the bacterial culture was diluted 1/100 in the same medium and grown to an OD600 = 0.3. Excess culture was kept at room temperature for later use, while 1.4 ml aliquots of cells (one aliquot per ligation reaction) were pelleted by a 30 second spin in a Micro-Centaur microfuge. The supernate was removed and the cells gently resuspended in 0.5 ml of ice-cold sterile MR (10 mM MOPS, 10 mM RbCl, pH 7.0). The cells were pelleted again and the supernate removed as before. The cells were resuspended in 0.5 ml ice-cold sterile MRC (100 mM MOPS, 10 mM RbCl, 50 mM CaCl₂, pH 6.5) and left on ice for 30 minutes. The cells were then pelleted by another 30 second
spin, resuspended in 0.15 ml MRC and kept on ice. 3 μl of DMSO and 5 μl of the ligation mix were added to each tube of competent cells. The mixture was left on ice for 60 minutes, heat-shocked at 55°C for 30 seconds, cooled on ice for 1 minute, then held at room temperature. Before plating out, 200 μl of log phase JM101 cells (from the original culture kept at room temperature), 25 μl of 25 mg/ml BCIG (in dimethylformamide) and 25 μl of IPTG (in H2O) were added to each tube. The contents of the tube were transferred to glass test-tubes containing 3 ml of LUB soft agar at 50°C, thoroughly mixed and plated onto LUA plates and incubated overnight at 37°C.

White and blue "plaques" develop overnight in the bacterial lawn, corresponding to recombinant and non-recombinant M13 transformants respectively. "Plaque" in this sense refers to an area of reduced bacterial lawn growth due to in vivo M13 replication, rather than the cell death associated with the virulent replication cycle of lambda. White plaques were screened for recombinant sequences by the method of Benton and Davies (1977), see Section 2.10 for experimental details.

v) Preparation of single-stranded DNA

Single white plaques were picked into 1 ml of phage buffer (6 mM Tris-HCl, 10 mM MgSO4, 0.005% gelatin, pH 7.2) and stored at 4°C.

0.1 ml of the phage stock was mixed in a large capped test-tube with 1.5 ml of an overnight culture of E. coli JM101 that had been diluted 1/100 in fresh Luria broth. The mixture was incubated at 37°C with fast shaking for 5.5 hours. The cells were pelleted by two 5 minute spins in a Micro-Centaur microfuge. 1 ml of supernate was carefully removed to a new tube and 300 μl of 10% PEG6000, 2.5 M NaCl was added. The phage were pelleted by centrifugation for 10 minutes. The supernate was removed with care and the phage pellet was resuspended in 100 μl 1.1 M Na acetate, pH 7.

100 μl of phenol/chloroform was added and mixed vigorously before centrifuging for 3 minutes. The aqueous layer was transferred to a fresh tube and 250 μl ethanol was
added and the tube chilled in a dry ice/ethanol bath for 5 minutes. The DNA was pelleted by centrifugation for 10 minutes. The supernate was removed and the pellet (invisible!) was washed with 70% ethanol. The DNA was redissolved in 30 µl H₂O at 60°C for 15 minutes and stored at -20°C.

vi) Sequencing of M13 recombinants

Sequencing of M13 recombinants was based on the method of Biggin et al. (1983) for M13 dideoxyribonucleotide chain-termination using [α³⁵S]-dATP. Quantities quoted are for 15 sequencing templates. Reactions were performed in 1.5 ml polypropylene microfuge tubes and all centrifugations were done in an Eppendorf sequencing centrifuge. This centrifuge holds four racks, each holding two sets of sequencing reactions; the use of this centrifuge permits the simultaneous start of all reactions, all reagents are added to the sides of the tubes and reactions started by centrifugation.

Single-stranded clone DNAs were incubated at 60°C for 10 minutes prior to annealing to ensure they were completely dissolved. Each clone was annealed to the 17-mer universal primer, or a specific synthetic primer, by adding 5 µl of clone DNA to 5 µl of "primer mix" (7.2 µl of 2 µg/ml primer, 8 µl TM buffer, 64 µl H₂O) and incubating at 60°C for 60 minutes, with a quick centrifugation after 30 minutes. The annealed clones were held at room temperature, or kept at -80°C for overnight storage. For each clone, four reaction tubes were prepared, containing 2 µl of annealed clone plus 2 µl of either a "T", "C", "G" or "A" NTP mix. NTP mixes for ³⁵S sequencing were as follows (all volumes in µl):

<table>
<thead>
<tr>
<th></th>
<th>&quot;T&quot;</th>
<th>&quot;C&quot;</th>
<th>&quot;G&quot;</th>
<th>&quot;A&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5 mM TTP</td>
<td>12.5</td>
<td>250</td>
<td>250</td>
<td>250</td>
</tr>
<tr>
<td>0.5 mM dCTP</td>
<td>250</td>
<td>12.5</td>
<td>250</td>
<td>250</td>
</tr>
<tr>
<td>0.5 mM dGTP</td>
<td>250</td>
<td>250</td>
<td>12.5</td>
<td>250</td>
</tr>
<tr>
<td>10 mM dTTP</td>
<td>6.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 mM ddCTP</td>
<td></td>
<td></td>
<td>4.0</td>
<td></td>
</tr>
</tbody>
</table>
10 mM ddGTP 8.0
10 mM ddATP 1.2
TE buffer 500 500 500 250

M13 TE buffer is: 10 mM Tris-HCl, 0.1mM EDTA, pH 8.0.

0.5 mM dNTPs and 10 mM ddNTPs were prepared in M13 TE buffer. 2 μl of freshly prepared "Klenow mix" (114 μl H₂O, 10 units Klenow polymerase, 10 μl [α-³⁵S]dATP (14.8 TBq/mmmole, 0.37 MBq/μl)) was added to the side of each tube. The reactions were initiated by centrifugation, mixed by tapping the tubes and recentrifuging. The reactions were incubated at 37°C for 30 minutes, then 2 μl "sequence chase mix" (0.25 mM each of dATP, dCTP, dGTP and dTTP in M13 TE buffer) was added to each tube, mixed as before and incubated for 40 minutes at 37°C. At this point the reaction tubes were prepared for loading onto a sequencing gel by addition of 4 μl of "dye mix" (stock solution - 10 ml deionised formamide, 10 mg xylene cyanol FF, 10 mg bromophenol blue, 0.2 ml 0.5 M EDTA, pH 8.0). ³⁵S-labelled substrate can be stored for several days at -80°C if the formamide dye is not added.

vii) Buffer gradient sequencing gels

Preparation and running of 40 cm 6% (w/v) polyacrylamide buffer gradient gels were essentially as described by Biggin et al. (1983). A "sharks tooth" comb was used which enabled 15-18 clones to be run on a single gel. The standard mixes were used to prepare two gels are given below.

<table>
<thead>
<tr>
<th></th>
<th>0.5x</th>
<th>2.5x</th>
</tr>
</thead>
<tbody>
<tr>
<td>40% acrylamide stock</td>
<td>27 ml</td>
<td>6 ml</td>
</tr>
<tr>
<td>urea</td>
<td>90 g</td>
<td>20 g</td>
</tr>
<tr>
<td>sucrose</td>
<td>—</td>
<td>2 g</td>
</tr>
<tr>
<td>10x TBE</td>
<td>9 ml</td>
<td>10 ml</td>
</tr>
<tr>
<td>H₂O to a final volume</td>
<td>180 ml</td>
<td>40 ml</td>
</tr>
</tbody>
</table>

40% acrylamide stock solution: 38 g acrylamide and 2 g N,N'-methylenebisacrylamide were dissolved in a final volume.
of 100 ml in \( \text{H}_2\text{O} \). 5 g Amberlite resin was added and the mixture stirred for 30 minutes, the resin was removed by filtration and the 40% acrylamide stored at 4°C.

10 x TBE is: 109 g Tris base, 55 g boric acid, 9.3 g EDTA made up to 1 l with \( \text{H}_2\text{O} \), pH 8.3. Final concentrations: 1 M Tris, 0.83 M boric acid, 30 mM EDTA.

The solutions were filtered through a Whatman Number 1 filter using a Buchner funnel and then the following solutions were added:

<table>
<thead>
<tr>
<th></th>
<th>0.5x</th>
<th>2.5x</th>
</tr>
</thead>
<tbody>
<tr>
<td>10% (w/v) ammonium persulphate (in H(_2)O)</td>
<td>1.4 ml</td>
<td>0.28 ml</td>
</tr>
<tr>
<td>TEMED</td>
<td>96 μl</td>
<td>20 μl</td>
</tr>
</tbody>
</table>

8 ml 0.5x and 12 ml 2.5x solutions were drawn into a syringe and poured into a mould, followed by 60 ml of 0.5x solution. The gels set within 60 minutes and were mounted into the gel apparatus. The top buffer chamber was filled with 0.5x TBE and the bottom chamber with 2.5 x buffer.

Sequencing reactions were boiled for 3 minutes and 2.5 μl samples were loaded onto the gels. Gels were run at 1 400 to 2 000 volts for 3 to 5 hours, the voltage was adjusted to keep the gels hand-hot, approximately 65°C. The high temperature helps to maintain DNA in a denatured form and to prevent formation of secondary structures. Gels were run until the bromophenol blue had run off the end. Under these conditions the bromophenol migrates as a 30 bp fragment, and the xylene cyanol as a 120 bp fragment. After electrophoresis the gels were fixed in 10% (v/v) methanol, 10% (v/v) acetic acid for 15 minutes, transferred to Whatman 3MM paper and dried on a Bio-Rad Gel-Drier. Gels were autoradiographed at room temperature for 16 hours to several days, as necessary.
2.18 MAKING DOUBLE-STRANDED DNA FROM SINGLE-STRANDED M13 RECOMBINANT DNA

This method was used to prepare DNA for oligo-labelling and subsequent use to screen libraries constructed in plasmid or cosmid vectors. Probes are often contaminated with DNA of the plasmid from which they were excised, and so produce a background signal on hybridisation. Subcloning into M13 was used to isolate the DNA for oligo-labelling away from such contamination. Single-stranded M13 DNA had been made as described in Section 2.14. The universal sequencing primer was annealed to the M13 DNA in a reaction set up as follows:

M13 single-stranded DNA  5 μl (1 μg)
H$_2$O  3 μl
M buffer  1 μl
Primer (2 μg/ml)  2 μl

The reaction was incubated at 60°C for 30 minutes then kept at room temperature. For synthesis of the second strand the following reaction was set up:

Annealed reaction mix  11 μl
ACGT mix  10 μl
M13 TE buffer  6 μl
Klenow polymerase  3 units

ACGT mix is 0.125 mM with respect to each of dATP, dCTP, dGTP and dTTP in M13 TE buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0).

This mixture was incubated at 37°C for 30 minutes. The DNA was recovered, after phenol extraction by ethanol precipitation. The DNA was redissolved in 20 μl H$_2$O before digestion with restriction endonuclease(s). The fragment of interest was isolated by excision from a low gelling temperature agarose gel, and radio-labelled as described in Section 2.8.
Culture and harvesting of leukocytes, preparation of slides, and cytogenetic analysis were performed by Dr. Nicola Royle.

0.5 ml whole blood was used to inoculate 7.5 ml of McCoy's 5a Medium (Gibco) supplemented with 20% foetal calf serum and 1.5% phytohaemagglutinin (Wellcome). Cells were grown at 37°C for 72 hours in a CO₂ incubator. BrdU (bromodeoxyuridine, Sigma) was added to 100 μg/ml, and the culture was incubated for a further 17 hours at 37°C. Cells were washed twice in PBS (phosphate buffered saline) and resuspended in 9.5 ml fresh medium. Thymidine was added to a concentration of 10 mM, and the cells incubated for 4 hours 45 minutes at 37°C. The cells were immediately harvested. Slides of the metaphase cells were prepared by standard procedures.

i) Oligo-labelling of pλKP20.1 Dde I insert

The 6.5 kb Dde I insert was isolated from pλKP20.1 by electroelution from an agarose gel following restriction endonuclease digestion. The fragment was extracted with phenol and recovered by ethanol precipitation. 200 ng of DNA was recovered in a final volume of 50 μl.

Oligo-labelling was carried out by adding the DNA, OLB, BSA and Klenow polymerase to freeze-dried tritiated dATP, dCTP and TTP. These were prepared as below:

\[ ^3H\text{-dATP} \quad 0.3 \text{ nmoles (1 μCi/μl)} \quad 19 \text{ μl} \]
\[ ^3H\text{-dCTP} \quad 0.3 \text{ nmoles (1 μCi/μl)} \quad 36 \text{ μl} \]
\[ ^3H\text{-TTP} \quad 0.3 \text{ nmoles (1 μCi/μl)} \quad 18 \text{ μl} \]

The other components were added as follows:

ipλKP20.1 in H₂O \hspace{1cm} 90.0 ng
(boiled for 3 minutes)
5 x OLB \hspace{1cm} 10.0 μl
BSA (10 mg/ml) \hspace{1cm} 2.5 μl
Klenow polymerase (7 units/μl) \hspace{1cm} 1.0 μl
The DNA was in water to give a final volume for the reaction of 50 µl. The reaction was allowed to proceed overnight at room temperature. OLB was exactly as in Section 2.9 with the omission of dATP and TTP.

The reaction was stopped by the addition of 5 µl of 10 x Stop Solution (7.8 mM EDTA, 0.65% SDS, 0.19 mg/ml denatured herring sperm DNA).

The labelled DNA was separated from the unincorporated isotopes on a Sephadex G75 column in TE buffer, pH 8.0. 15 fractions of 50 µl each were collected. 1 µl of each fraction was added to 5 ml of toluene based scintillant (1000 ml toluene, 83.5 ml NCS solubiliser, 4g PPO, 60.5 mg dimethyl POPOP). The samples were counted in a Packard Tri-Carb scintillation counter. The three fractions corresponding to the first peak of radioactivity were pooled, water was extracted with butan-2-ol and the DNA precipitated by the addition of 0.1 volume 2M sodium acetate and 2.5 volumes of ethanol. The DNA was precipitated, after a 20 minute incubation at -80°C, by centrifugation in a Micro-Centaur microfuge for 10 minutes. The pellet was rinsed in 80% ethanol, dried and resuspended in 20 µl H₂O. 0.5 µl was counted in the scintillation counter as before.

Specific activity could not be determined as the quantity of DNA synthesised during the labelling reaction was unknown. The amount of radioactivity was determined as 1.47 x 10⁵ cpm/µl.

ii) Hybridisation

Hybridisations were as described by Harper et al. (1981).

5.3 µl of the labelled fragment was used in a 375 µl hybridisation reaction containing five slides, equivalent to 1.55 x 10⁵ cpm/slide.

Hybridisations were carried out in 50% formamide (twice recrystallised), 2 x SSCP (0.3 M NaCl, 0.03 M Na⁺ citrate, 0.04 M NaH₂PO₄, pH 6.0), 10% dextran sulphate, 0.24 µg/ml denatured herring sperm DNA. The final pH was checked as 7.0 with pH paper. The reactions were incubated overnight at 37°C.
Post-hybridisation washing was at 42°C, three times in 50% formamide, 2 x SSC, pH 7.0; and five times in 2 x SSC, pH 7.0. The slides were dried, dipped in Ilford's Nuclear Research Emulsion K2, and exposed at 4°C for two weeks.

Developing, fixing and G banding were carried out by Dr. Nicola Royle. Banding was as described by Lin et al. (1985).

2.20 COMPUTING

DNA sequence analysis and Chou and Fasman secondary structure predictions were carried out using the DEC VAX computer at Leicester University, using the University of Wisconsin Genetics Computer Group package (Devereux et al., 1984).

Phylogenies based on DNA and amino acid sequences were determined using software written by J. Felsenstein, University of Washington, run on an Olivetti M24 PC.

Statistical tests on DNA sequences were carried out using programs written by Professor Alec Jeffreys, University of Leicester. These were run on a BBC micro-computer.

Word-processing was carried out with Vuwriter Scientific, Version 4.3c, (Vuman Ltd., Manchester) on an ACT Apricot Xi10 micro-computer, and WORD-11 on the DEC VAX.
CHAPTER 3

DNA SEQUENCE ANALYSIS OF cDNA CLONES OF THE HUMAN TYPE III PROCOLLAGEN GENE
3.1 INTRODUCTION

Work by several groups has revealed that the genes encoding the major vertebrate fibrillar procollagens (types I, II and III) exhibit considerable conservation of structural characteristics. In fact, in mammalian and avian genes an almost identical number of exons of similar size and distribution encode the three major domains of fibrillar collagen molecules, namely: N-terminal and C-terminal propeptides and the triple helical region (Tate et al., 1982; Chu et al., 1984).

The C-terminal domain of the procollagen molecule represents an important functional region which is believed to have a prominent role in the specific interactions of the pro-$\alpha$ chains during the formation of the procollagen trimer. Evidence for this comes from several observations: the formation of interchain disulphide bonds within the C-propeptides of types I and II procollagens precedes triple-helix formation (Schofield et al., 1974). Puromycin-induced termination of procollagen chain synthesis results in truncated chains lacking cysteine residues in C-propeptides, and no triple-helix formation (Rosenbloom et al., 1976). Another observation has been that a frameshift mutation in the carboxy terminus of the human $\alpha_2(I)$ C-propeptide prevents heterotrimer formation (Deak et al., 1983; Pihlajaniemi et al., 1984). Additionally, recent work demonstrates that a highly conserved portion of the C-propeptide of fibrillar collagens can repress post-transcriptional biosynthesis of collagen and fibronectin (Aycock et al., 1986).

Although the C-propeptides of the fibrillar collagens have identical function there must exist mechanisms to discriminate between different procollagen chains during assembly; especially where various procollagens are expressed simultaneously necessary. Comparative studies of C-propeptides allow identification of conserved and variable regions of their respective DNA coding sequences. Consequently, it is possible to attempt to relate these findings to the functions of the C-propeptides. These
comparative studies can also provide an insight into the evolution of the procollagen gene family.

Biosynthesis of abnormal type III collagen has been demonstrated in several cases of Ehlers-Danlos syndrome type IV (Pope et al., 1975; Pyeritz et al, 1984; Stolle et al., 1985). An association between this disorder and the type III procollagen gene has also been demonstrated (Tsipouras et al., 1986; Nicholls et al, 1988). A prerequisite, therefore, to an understanding of this disorder at a molecular level is to determine the DNA sequence of the type III procollagen gene.

A study was undertaken to characterise previously isolated cDNA clones of part of the type III procollagen gene. This sequence could then be used in comparative studies and to provide a source of primary information, for use in identifying genetic changes in type III procollagen associated with Ehlers-Danlos syndrome type IV.

3.2 CLONING OF THE HUMAN TYPE III PROCOLLAGEN cDNAs

cDNA clones of human α1(III) procollagen had been isolated previously by Dr. R. Dalgleish (Miskulin et al., 1986). To produce these cDNAs, RNA was extracted from human fibroblast strain HFL-1 (American Type Culture Collection CCL 153) which is known to produce type I and type III collagens, in the ratio of approximately 4:1 (Breul et al., 1980). Poly(A+) RNA was used as a template to prepare double-stranded cDNAs as described by Jaye et al. (1983), and these were cloned into the Pst I site of pBR322 by G-C tailing. The cDNA library was screened with a probe made from a subcloned segment of the chicken α1(III) genomic clone encoding exons 2 and 3 (Yamada et al., 1983a,b). Hybridisations were at 3 x SSC, 2 x Denhardt's solution at 65° C, and post-hybridisation washing was at a final stringency of 2 x SSC, 0.1% SDS at 65° C. Two positive colonies, pIII-21 and pIII-33, each with an insert of 1.9 kb, were selected. Restriction enzyme mapping demonstrated that the two clones are overlapping (Figure 3.1). The clones
Figure 3.1

Sequencing strategy of cDNAs

Restriction endonuclease map of the cDNA clones, pIII-21 and pIII-33, of the 3' end of the human type III procollagen gene. Sites for restriction enzymes EcoR I (E), Hind III (H), Pst I (P), Pvu II (Pv), Sst I (S) and Xho I (X) are indicated. The relationship of the clones to the protein is indicated at the top of the figure.

The sequencing strategy is depicted in the figure. As specific restriction enzyme fragments were cloned into M13 vectors in a predetermined orientation, all sequences obtained in a 5' to 3' direction were of the sense strand and vice versa.
were identified as α1(III) cDNAs by DNA sequencing and comparison to the amino acid sequence of human type III collagen (Seyer and Kang, 1981).

3.3 SEQUENCING

A restriction enzyme map of the two overlapping cDNA clones of the carboxyl end of the pro-α1(III) gene had been determined previously (Miskulin et al., 1986) and using this as a reference restriction fragments were isolated for sequencing (Figure 3.1). The general approach to sequencing the clones was to digest the recombinant plasmid DNA with the appropriate restriction endonuclease(s) and to fractionate the products in low gelling temperature agarose gels. The required fragments were ligated, without further purification, to M13 vector DNA which had been prepared likewise (see Section 2.17 of Chapter 2). An overlapping set of M13 clones was obtained from which the DNA sequence was determined on both strands. Where necessary, synthetic oligonucleotide primers were used to complete the sequence.

3.4 SEQUENCE OF TYPE III PROCOLLAGEN cDNAs

A total of 2561 nucleotides were sequenced, 741 bp of the triple-helical region, 75 bp of carboxyl-telopeptide and 735 bp of carboxyl-propeptide sequence and 1010 bp of 3' untranslated sequence.

The amino acid sequence derived from the nucleotide sequence (Figure 3.2) reveals features that are characteristic of type III collagen. Firstly, two successive cysteine residues are located at the terminus of the α-helical region. These two residues form interchain disulphide bonds in type III collagen (Seyer and Kang, 1981) and are not found in other fibrillar collagen chains (Bernard et al., 1983a and 1983b; Sangiorgi et al., 1985). Secondly, five tandem repeats of the glycine-proline-proline tripeptide are found at the carboxyl end of the triple-
Figure 3.2

Nucleotide and amino acid sequence of COL3A1 cDNAs.

The middle line is the nucleotide sequence of the human type III procollagen cDNA clones pIII-21 and pIII-33, and below it the translated sequence. The sequence of the avian gene (Yamada et al. 1983b) is shown on the top line where this differs from that of the human. The avian sequence extends from nucleotide 690 to 1632. The amino acid sequence is numbered as in Seyer and Kang (1981). The first residue of the telopeptide is numbered 1c, the last residue of the propeptide is 270c. The nucleotide sequence begins at amino acid residue 777 and continues to the polyA tail at the 3' end of the cDNA clone pIII-33.

Deletions in either the human or chicken sequence are denoted by . and the end of the α-chain coding sequence is indicated by and the carboxyl-propeptidase cleavage site by †. The translation stop codon is marked by and cysteine residues by ●. The glycosylation site is shown by a dashed underline. Exon-intron boundaries, as determined by Chu et al. (1985c), are indicated by ▼. The sequences identical to the consensus polyadenylation signal sequence (AATAAA) are underlined twice. Inverted repeat sequences are shown by half-headed arrows and direct repeats by double headed arrows. The nucleotide sequence of the α-chain contains many short direct repeats and imperfect inverted repeats, these are not shown on the figure.
helical domain of both α1(I) and α2(I) chains (Bernard et al., 1983a and 1983b), there are three such tandem tripeptides in the α1(II) chain (Sangiorgi et al., 1985b). There are only two consecutive units lying close to the 3' end of the α-chain of type III collagen.

The translated nucleotide sequence (Figure 3.2) agrees well with the published amino acid sequence of the 3' end of human type III collagen obtained by Edman degradation (Seyer and Kang, 1981), but there are, however, 10 differences within a region of 246 amino acid residues. One of these, histidine in place of serine 983, would result in a charge alteration. Serine has an uncharged R group while histidine has a positively charged R group. The translated nucleotide sequence of a human pro-α1(III) collagen cDNA clone of Loidl et al. (1984) differs from that of Seyer and Kang's in the same way, suggesting a misreporting in the amino acid sequence obtained by the latter group.

At position 1011 Seyer and Kang (1981), Loidl et al. (1984) and Chu et al. (1985b) have reported a proline residue, which we have identified as a serine residue. This difference is due to the presence of a thymine in place of cytosine in the first position of the codon. This region has been sequenced several times using independently isolated M13 recombinant clones, and this difference is certainly genuine (Figure 3.3): although it is possibly an artefact of the cDNA cloning process. Chu et al. (1985b) report a proline residue at position 26c in the propeptide, replacing an aspartate which is present in the sequence of this study, that of Loidl et al. (1984), Seyer and Kang (1981) and also in the chick pro-α1(III) collagen sequence (Yamada et al., 1983a). Residues 25c and 26c comprise the C-propeptidase cleavage site. The presence of alanine at 25c and aspartate at 26c is highly conserved in all four fibrillar procollagens and also the homologous chains in chicken. It is most likely, therefore, that there is an error in the sequence reported by Chu et al. (1985b). Interestingly, 25c is occupied by alanine in all the chains except human and chicken α1(III) procollagen.

The sequence reported in this study differs from that
Figure 3.3

A single amino acid variation in the sequence.

Sequence of the type III procollagen α-chain demonstrating the variation in the sequence of the codon of amino acid residue 1011. At the nucleotide marked by an arrow, the sequences reported by Loidl et al. (1984) and Chu et al. (1985c) have a cytosine, which in the present study was determined to be thymine.
of Loidl et al. (1984) at the third position of two codons, 949 and 22c, and similarly from that of Chu et al. (1985b) at 1015 and 1c. The change at residue 949 results in the shift of a Hae III recognition site by one base; whilst that at residue 1015 would result in the absence of a Hae III site which is present in the sequence determined by Chu et al. (1985b) but is also absent in that of Loidl et al. (1984). Hae III sites, which recognise the palindromic sequence GGCC, are present at a high frequency in the α-helical coding region due to the prevalence of glycine-proline-X tripeptide units encoded by GGNCCNNNN, where the first N is often G or C. Consequently, this polymorphism could not be utilised for restriction fragment length polymorphism analysis by conventional Southern blotting.

3.5 ANALYSIS OF THE α-CHAIN SEQUENCE

Of the 250 amino acid residues of the α-helical region whose sequence has been determined 19 differ from those of the equivalent bovine pro-α1(III) collagen (Allmann et al., 1979). Only two of these changes are not conservative, aspartate in place of tyrosine at 846 and histidine instead of phenylalanine at 953.

The sequence contains the tetrapeptide lysine glycine-histidine-arginine, residues 933 to 937; it also occurs at residues 90 to 94 (Seyer and Kang, 1981). Lysine or hydroxylysine aldehyde derived cross-linkages between the tetrapeptide and the amino or carboxyl terminal non-helical regions would result in the observed quarter stagger array of procollagen molecules in a fibril.

The sequence also contains the tripeptide arginine-glycine-aspartate at position 918 to 920. This is present in all four human procollagens at equivalent positions and similar flanking sequence. This tripeptide has been identified as the cell-attachment site in fibronectin and other extra-cellular matrix proteins (Pierschbacher et al., 1981).

There are many short direct repeats and inverted
repeats present in the nucleotide sequence of the \( \alpha \)-chain (data not shown). These are a function of a sequence encoding the repeating motif of glycine-proline-Y. There are two sets of overlapping inverted repeats which are interesting (see Figure 3.2). One set has a repeat at each end of exon 6 (as defined by Chu et al., 1985b), the other has repeats on the coding sequences flanking the intron between exons 5 and 6. Whether the positioning of these inverted repeats is coincidental or of functional significance cannot be determined. No such repeats are found in the corresponding positions of the other procollagens, and there is no available sequence of the corresponding chicken type III \( \alpha \)-helical region. The inverted repeat which overlaps the junction of the triple-helical region and the telopeptide (Figure 3.2) is not present in the chicken sequence.

### 3.6 ANALYSIS OF TEOLEPTIDE AND CARBOXYL-PROPEPTIDE

The carboxyl-propeptide is believed to represent a very important functional region involved in chain selection and chain association during intracellular assembly of procollagen; and appears to have a role in maintaining the protein in soluble form during transport to the extracellular matrix (Prockop et al., 1979; Bornstein and Sage, 1980). Consequently the sequence of the \( \alpha \)1(III) carboxyl-propeptide was compared to those of the other fibrillar collagen carboxyl-propeptides to identify sequences which might have a common function in all these chains. Conversely, sequences responsible for chain-specific interactions would be expected to vary between the different procollagens but be conserved between species.

#### i) Telopeptide

The carboxyl terminal telopeptide has been suggested to play an important role in the lateral assembly of procollagen molecules (Dion and Myers, 1987).

A comparison of the telopeptide sequences of the human
fibrillar collagens (Figure 3.4) shows considerable divergence. This could reflect a chain-specific role in the assembly of the procollagens. One prediction of this would be that such specificity should be conserved in the same chain of different species. Pairwise comparisons of the corresponding avian and human sequences (Table 3.1) reveals a high degree of conservation of the \( \alpha_1 \) telopeptide, the one amino acid substitution being conservative. In the \( \alpha_2 \) telopeptide of the six changes two result in alteration of charge, as do two of the five substitutions in the \( \alpha_1 \) chain. While only one of the twelve replacements in the \( \alpha_1 \) telopeptide would cause an alteration of charge, this also affects the recognition site of the carboxyl-propeptidase. Additionally, the chick \( \alpha_1 \) telopeptide has a deletion of three, and insertion of one, amino acid residues. It can be argued that such extensive differences between species favours the notion that the telopeptide represents a non-specific linker segment between the \( \alpha \)-chain and the carboxyl-propeptide. Furthermore, the sequence of the first 30 amino acid residues of the propeptide is highly diverged amongst the fibrillar procollagens. To assess whether the telopeptides could provide a chain specific structural role, the sequence of the telopeptide and the variable amino terminal region of the propeptide of each intersitial procollagen was subjected to an algorithm to predict secondary structure from an amino acid sequence by the method of Chou and Fasman (1978). The results of this analysis are presented in Figure 3.5.

When comparing the human fibrillar procollagens there are obvious similarities, and differences, among the sequences. Each region analysed consists of extensive \( \beta \)-turns and at least two \( \alpha \)-helical segments, except \( \alpha_2 \) which has only one, there are no sequences capable of forming \( \beta \)-sheets. \( \alpha \)-helices are composed of clustered hydrophobic and hydrophilic residues, whereas hydrophobic and hydrophilic amino acids predominate in \( \beta \)-sheets and \( \beta \)-turns (Chou and Fasman, 1978).

For each chain, the amino terminal end of the telopeptide is a continuum of the \( \beta \)-turns of the preceding
Figure 3.4

Comparison of the telopeptide sequences of the human fibrillar procollagens.

The translated nucleotide sequence of each telopeptide and the variable amino terminal sequence of the propeptide are shown. The sequences have been aligned to produce maximum homology, introducing gaps where necessary. The carboxyl propeptidase cleavage site is indicated by i.
Table 3.1

Pairwise comparison between human and chicken telopeptide sequences for each of the fibrillar collagens.

The number of amino acid residues that differ between the human and chicken sequences is given, as well as the proportion of residues that differ. Deletions were counted as replacements. The number of nucleotide positions that differ was also determined, this was divided into those substitutions that produce, or do not produce, an alteration of an amino acid residue.
<table>
<thead>
<tr>
<th></th>
<th>AMINO ACID REPLACEMENTS</th>
<th>NUCLEOTIDES (uncorrected)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Replacement</td>
</tr>
<tr>
<td>COL3A1</td>
<td>12/25 (46%)</td>
<td>29</td>
</tr>
<tr>
<td>COL2A1</td>
<td>5/27 (18%)</td>
<td>5</td>
</tr>
<tr>
<td>COL1A2</td>
<td>6/15 (17%)</td>
<td>7</td>
</tr>
<tr>
<td>COL1A1</td>
<td>1/26 (4%)</td>
<td>2</td>
</tr>
</tbody>
</table>
Figure 3.5

Secondary structure potential of the fibrillar collagen carboxyl propeptides.

The amino acid sequence of each telopeptide was subjected to a computer program designed to predict the secondary structure, according to the principles of Chou and Fasman (1978). The human sequence is prefixed by H and the homologous chicken sequence by C. The straight line with circle depicts the end of the triple helical region; the arrow denotes the carboxyl propeptidase cleavage site. Potential $\alpha$-helical structures are indicated by a wavy line; $\beta$-turns are shown by a zig-zag line.
triple-helical coding region. This is followed by an $\alpha$-helical forming segment, except in the $\alpha_2(I)$ chain where the corresponding sequence has no specific conformation, then several $\beta$-turns and another $\alpha$-helical region. Although caution is required in the interpretation of such analysis, it is possible to suggest that each sequence would adopt a similar conformation despite the differences in primary sequence.

Whether this conformation would be chain specific is less certain. The distribution and number of $\alpha$-helical regions is different for each chain. If this is involved in producing specificity it could be predicted that this arrangement would be conserved across species. Analysis of the corresponding chicken sequences by the Chou and Fasman method (1978) produces results that are equivocal. The $\alpha_1(I)$ and $\alpha_1(II)$ sequences have the potential to produce very similar structures in both species. However, the predicted structure of the $\alpha_2(I)$ sequence is significantly different. The human sequence has a more extensive region of $\beta$-turns at the amino end of the telopeptide and, unlike the chicken telopeptide, this is not followed by an $\alpha$-helical region. The $\beta$-turn region after the propeptidase cleavage site is shorter in the chicken due to the deletion of four amino acid residues. The chicken sequence has a deletion of three amino acid residues in the telopeptide resulting in the formation of only one $\alpha$-helical region, whereas the corresponding human sequence has potential to form two $\alpha$-helical structures separated by $\beta$-turns. Furthermore, except for $\alpha_1(III)$ the propeptidase cleavage site in all the fibrillar procollagens is located within an identical structural motif in humans and chickens.

Earlier studies (Dion and Myers, 1987) which used a similar analysis to conclude that each procollagen telopeptide, and the variable sequence at the amino end of the propeptide, would adopt a chain specific conformation cannot be confirmed with certainty when the analysis is extended to interspecies comparisons. For type III procollagen it can be argued that the cysteine residues at the junction of the triple-helical region and telopeptide
provide this chain specificity, releasing the remainder of the telopeptide and adjoining propeptide from functional constraint. A similar argument cannot be applied to α2(I) procollagen. There must exist, however, mechanisms to discriminate α2(I) chains; not only are these chains incorporated into a type I procollagen heterotrimer, but they must be incorporated in the correct ratio to α1(I) chains.

ii) Carboxyl-propeptide

There is a high degree of sequence conservation between the human and avian α1(III) carboxyl-propeptides. There is 20% amino acid divergence, reflected in conservation of the nucleotide sequences which are 22% diverged. On closer examination this conservation is even more striking as it seems to be localised in three regions within the propeptide. There is very strong homology at residues 54c to 106c, only seven amino acid residues are substituted and none involves a charge alteration. Residues 54c to 106c are highly conserved amongst the four human carboxyl-propeptides. These 52 amino acids encompass the first five (four in α2(I) chain) cysteine residues present in each propeptide. The cysteine residues are involved in inter-chain disulphide bonds which are critical for correct assembly of the molecules.

Previous studies have identified a conserved amino acid sequence around the glycosylation attachment site, asparagine-isoleucine-threonine (171c-173c), in several different procollagens (Yamada et al., 1983a). For positions 168c to 184c there are no amino acid replacements and only seven silent nucleotide substitutions between human and chick type III propeptides. Among the four human procollagens there is only one replacement, valine in place of isoleucine at 180c in the α1(I) propeptide. The conserved sequence is larger than the three residue recognition site required by carbohydrate transferases. Perhaps this region serves additional purposes such as directing the association of propeptides. It has been claimed that the nucleotide sequence for this segment is more highly conserved than in
other segments, even at the third position of codons (Yamada et al., 1983a). The same workers suggested that this conserved sequence could represent a common regulatory element, although there is no obvious symmetrical element in the sequence as is often found in regulatory signals.

To test whether the apparent conservation of nucleotides is significant, the sequence of the carboxyl-propeptide was subjected to a statistical analysis using a computer program designed to give a measure of the significance of conservation or divergence in a confined window space. The details of the program are given in the legend to Figure 3.6a. Briefly, the probability of obtaining the particular sequence divergence in a defined window size is calculated using the Fisher analysis when the expected level of divergence between two sequences is provided. Two composite sequences were constructed, one of the four fibrillar collagen propeptides of the human sequences, and the other of the homologous chicken sequences. As the telopeptides are of variable length, these composite sequences were started at the glutamate residue which occupies position 37c in the human type III sequence; and the corresponding position in each of the other chains.

To create the composite sequences the four sequences for each species were aligned. Nucleotide 1 of the composite was taken from COL1A1, nucleotide 2 from COL1A2, nucleotide 3 from COL2A1, nucleotide 4 from COL3A1, nucleotide 5 from COL1A1 etc. This procedure provides an unbiased statistical test that is not possible if either a chi-squared test or an exact analysis of the number of mutations is performed. The test requires the assumption that the deviations between human and chicken for each of the four genes are independent. To check this the eight sequences of the carboxyl-propeptides were subjected to a matrix method to determine a phylogenetic tree. The results of this analysis are presented in Figure 3.6b. The phylogeny demonstrates that each of the human propeptide sequences is more closely related to its homologue in chicken than other human collagen carboxyl-propeptides; and so fulfils the assumption required for the statistical test.
How significant is the reported sequence conservation around the carbohyderate attachment site of the carboxyl propeptide?

Composite sequences of human and chicken fibrillar collagen propeptides were subjected to a statistical analysis on a BBC microcomputer. In this analysis the window size (sequence search string) was 15 bp. The expected divergence level between the two sequences was given as 30%. The computer totals the number of identities in 15 bp strings and calculates the probability of obtaining this number of identities based on the expected number (70% of 15 BP) using the Fisher analysis:

\[ P_i = (0.30)^{15-N}(0.70)^N \times \text{number of positions the identity can occur}, \]

where \( N \) = number of identities.

If the number of identities is 5, for example, the probabilities of obtaining 5 identities or less are totalled and plotted on a point above the zero significance (i.e. 70% of 15) line. If the number of identities is 13 the probabilities of obtaining 13 or more are totalled and plotted on a point below the zero significance line. Sequences which are more diverged than expected appear as peaks and vice versa.

Plot of the phylogenetic analysis of the DNA sequences used to construct the composite sequences.

The sequences of the four human and four chicken fibrillar collagen carboxyl propeptides were subjected to a matrix method to generate a phylogenetic tree. The method chosen is the average distance of unweighted pair-group method with arithmetic mean (UPGMA) as described by Nei et al. (1983). This is the simplest such method available and is based on the assumption that the expected value of distance is proportional to evolutionary time. The human sequences are prefixed by H and the chicken sequences by C. The positions of cysteine residues are marked by X, and that of the carbohyderate attachment site by ■.

The plot reveals that each human propeptide DNA sequence is more closely related to its chicken homologue than to the sequence of any other fibrillar collagen propeptide.
Figure 3.6a

Figure 3.6b
The results (Figure 3.6a) indicate that the sequence containing the carbohydrate attachment site is not conserved, indeed it lies within a region that shows significant divergence. When the expected level of divergence was set at 0.6, that expected for two unrelated sequences, this region had a mean level of divergence (data not shown). The carboxyl-propeptide seems to consist of two blocks of sequence at either end which are evolving at, or close to, the neutral rate. In the 3' block there is a region which shows considerable conservation (nucleotides 500-520). The statistical level of significance, however, cannot be determined from this analysis. The central block is highly diverged and is evolving much faster than expected for neutral drift. It is possible, therefore, that this region is evolving under positive selection, perhaps to fulfil the chain specific role proposed earlier.

The final third of the carboxyl-propeptide is highly conserved at the amino acid level too. There are 8 substitutions in the region 205c to 270c, all conservative, between the human and avian type III sequences. There is a similar identity of sequence across all four human procollagens. This region contains the last two cysteine residues, involved in intra-chain disulphide bonding. Interestingly, a synthetic peptide of the terminal 22 amino acid residues of the α2(I) propeptide, which are highly conserved in all four fibrillar procollagens, inhibits the post-transcriptional biosynthesis of procollagen and fibronectin in cultured fibroblasts (Aycock et al., 1986).

Comparison of the amino acid sequences of the four fibrillar collagen carboxyl-propeptides reveals that there are two other regions of variability, in addition to the one noted above. These occur at positions corresponding to amino acids 106c-124c (V1) and 143c-156c (V2) in the human type III chain. Pairwise comparisons between the four human human procollagens and between each human and its chicken homologue were made (Table 3.2). Variable region V1 shows substantial divergence between most of the human chains, although α1(III) and α2(I) are only 53% different, and α1(II) and α1(I) are 50% different. Interspecies comparisons
Table 3.2

Variable regions within the carboxyl propeptide.

A comparison of the amino acid sequence of the carboxyl propeptide of the COL3A1 human gene to those of the other fibrillar collagens revealed that there are two short segments that are more variable than others: VI (residues 106c-124c) and V2 (143c-156c). The number of amino acid residues that differ of those compared is given; this is also expressed as a percentage difference to accomodate the different lengths compared.

Similarly the sequences of the human and homologous chicken chains were compared. These are presented in the lower part of the table. The number of amino acid substitutions that produce an alteration in the charge at each position is also presented in parentheses.
<table>
<thead>
<tr>
<th>Comparison</th>
<th>VI Amino acid differences</th>
<th>V2 Amino acid differences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Human</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COL3A1/COL2A1</td>
<td>14/20 70%</td>
<td>9/14 64%</td>
</tr>
<tr>
<td>COL3A1/COL1A2</td>
<td>10/19 53%</td>
<td>13/14 93%</td>
</tr>
<tr>
<td>COL3A1/COL1A1</td>
<td>15/19 79%</td>
<td>12/14 86%</td>
</tr>
<tr>
<td>COL2A1/COL1A2</td>
<td>11/20 55%</td>
<td>13/14 93%</td>
</tr>
<tr>
<td>COL1A2/COL1A1</td>
<td>14/19 74%</td>
<td>10/14 72%</td>
</tr>
<tr>
<td><strong>Human/chicken</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COL3A1</td>
<td>9/19 42% (7)</td>
<td>4/14 29% (1)</td>
</tr>
<tr>
<td>COL2A1</td>
<td>8/19 42% (3)</td>
<td>4/14 29% (2)</td>
</tr>
<tr>
<td>COL1A2</td>
<td>10/19 52% (9)</td>
<td>3/14 21% (1)</td>
</tr>
<tr>
<td>COL1A1</td>
<td>5/19 26% (4)</td>
<td>2/14 14% (0)</td>
</tr>
</tbody>
</table>
revealed some degree of conservation. V2 showed much higher percentage differences between the human procollagens and much greater conservation between species. This observation indicates that these sequences do not represent regions of greater mutability but serve an evolutionary conserved function. When the carboxyl-propeptides were analysed by the method of Chou and Fasman (1978) to predict secondary structures, it was observed that the variable regions, V1 and V2, could adopt chain specific conformations. Furthermore, despite 47% difference in amino acid sequence, the human and chicken α1(III) V1 regions are predicted to have a similar conformation; consisting of hydrophobic and hydrophilic B-sheets and ending in hydrophilic helices. The V2 region consists of hydrophilic B-turns followed by hydrophobic B-sheets. As these regions are adjacent to highly conserved regions containing cysteine residues, it is attractive, therefore, to suggest that these variable regions are involved in providing chain specificity. Hydrophobic and charged amino acid residues present in these regions could modulate interactions between the three carboxyl-propeptides when the procollagen molecule is assembled.

3.7 EVOLUTION OF THE CARBOXYL-PROPEPTIDE

Extensive compilation of protein sequences, particularly those of globins, demonstrate that there is a strong correlation between divergence time and the percentage amino acid difference. The relationship is more evident if, instead of mere percentage differences, the estimated number of amino acid substitutions that have occurred during evolution are used. Assuming that mammals and birds diverged 300 million years ago during the Carboniferous period, the rates of substitution for the the four fibrillar collagen carboxyl-propeptides can be calculated (Table 3.3). The results indicate that the carboxyl-propeptides have evolved at similar rates, approximately $0.2 \times 10^{-9}$ substitutions per site per year, this value is one-fifth that of globins and similar to that
Table 3.3

Estimation of the number of amino acid substitutions that have occurred during the evolution of the carboxyl propeptides.

The number of amino acid substitutions that have occurred since the divergence of mammals and birds was calculated for each of the fibrillar collagen propeptides. Formulae derived by Kimura (1983) were used which assume that amino acid substitutions follow a Poisson distribution. The average number, per site, of amino substitutions between two propeptides is $K$; which is given by $-\ln (1-pD)$. $pD$ is the fraction of amino acid differences. The standard error of $K$ is given by:

$$\sqrt{pD/(1-pD)n};$$

where $n$ is the total number of amino acid residues compared. Deletions were counted as substitutions.

The rate of evolutionary amino acid substitutions per site per year is given by $k = K/2T$. $T$ is the number of years that have elapsed since the evolutionary divergence of the two chains from their common ancestor. The factor 2 in the denominator corresponds to the two branches in a phylogenetic tree. It was assumed that mammals and birds diverged 300 million years ago. The results obtained are for the telopeptide and propeptide combined. The telopeptide is so short that the value of $K$ for it alone would have a significantly large standard error.
<table>
<thead>
<tr>
<th></th>
<th>COL1A1</th>
<th>COL1A2</th>
<th>COL2A1</th>
<th>COL3A1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of amino acids (n)</td>
<td>272</td>
<td>258</td>
<td>273</td>
<td>267</td>
</tr>
<tr>
<td>Number of differences</td>
<td>25</td>
<td>42</td>
<td>34</td>
<td>57</td>
</tr>
<tr>
<td>K</td>
<td>0.096</td>
<td>0.178</td>
<td>0.133</td>
<td>0.240</td>
</tr>
<tr>
<td>$\delta K$</td>
<td>0.019</td>
<td>0.028</td>
<td>0.023</td>
<td>0.032</td>
</tr>
</tbody>
</table>
of cytochrome c.

While changes in replacement sites can easily be acted on by selection, changes in silent sites or non-coding sequences are putatively neutral. On a neutralist basis it would be expected that the rate of appearance of such changes to be very high and the same in all genes, reflecting the mutation rate. To test this the synonymous component of substitution in the data for the carboxyl-propeptides was estimated using a two substitution-type model proposed by Kimura (1983) (Table 3.4). The results suggest that the relationship \( K_2 < K_1 < K_3 \) holds generally. The evolutionary mutant substitutions are most rapid at the third position, followed by the first position, and then the second position.

With respect to amino acid substitutions, \( \alpha_1(I) \), \( \alpha_2(I) \) and \( \alpha_1(II) \) carboxyl-propeptides have all evolved at a similar rate, which is about one-half that of the \( \alpha_1(III) \) propeptide and one-third that of \( \beta \)-globin. Yet with respect to the synonymous component the evolutionary rate is approximately the same for all four propeptides. Furthermore, the \( K'S \) value of the propeptides is roughly equal to that of \( \beta \)-globin and \( \alpha \)-tubulin despite the enormous difference in the rate of amino acid substitutions between these proteins (all regarding mammalian and avian divergence).

3.8 ANALYSIS OF 3' UNTRANSLATED SEQUENCE

The length of the 3' untranslated region (UTR) is 967 nucleotides. The poly(A) tail at the end of the sequence is preceded by a consensus polyadenylation signal sequence (AATAAA) starting at position 2500. Immediately upstream of this signal sequence is an inverted repeat sequence which can be arranged into a stem-loop structure (Figure 3.7). At position 1802 is a sequence that is identical to the consensus poly(A) signal sequence. The separation between this site and that at the end of the clone is 686 nucleotides. This corresponds to the difference in size of
Table 3.4

Evolutionary distance in terms of the number of base substitutions estimated, for comparisons of the human and chicken carboxyl propeptides of the fibrillar collagens.

The number of base substitutions that have occurred during evolution at each of the three positions of codons (K1, K2 and K3) was estimated. The formulae used are those of Kimura (1983) and are presented in detail in the Appendix to Chapter 3. The number of synonymous base substitutions, K's, in other words the number of base substitutions at the third position of codons that have not produced a change in the amino acid coded for, was also calculated. The estimated values of these parameters are presented with their standard errors.
<table>
<thead>
<tr>
<th>Comparison</th>
<th>K1</th>
<th>K2</th>
<th>K3</th>
<th>K'S</th>
</tr>
</thead>
<tbody>
<tr>
<td>COL3A1</td>
<td>0.190 ± 0.036</td>
<td>0.120 ± 0.026</td>
<td>0.788 ± 0.129</td>
<td>0.538 ± 0.129</td>
</tr>
<tr>
<td>COL2A1</td>
<td>0.094 ± 0.021</td>
<td>0.057 ± 0.061</td>
<td>0.439 ± 0.076</td>
<td>0.352 ± 0.048</td>
</tr>
<tr>
<td>COL1A2</td>
<td>0.080 ± 0.020</td>
<td>0.074 ± 0.019</td>
<td>0.386 ± 0.070</td>
<td>0.332 ± 0.048</td>
</tr>
<tr>
<td>COL1A1</td>
<td>0.078 ± 0.019</td>
<td>0.047 ± 0.012</td>
<td>0.426 ± 0.074</td>
<td>0.352 ± 0.048</td>
</tr>
</tbody>
</table>
Figure 3.7

An inverted repeat sequence is present in the 3' untranslated region of the human COL3A1 gene.

A sequence capable of forming a stem-loop structure precedes the most 3' of the polyadenylation signal sequences, which is underlined.
the observed \(\alpha 1(III)\) procollagen transcripts, of 4.8 and 5.4 kb (Miskulin et al., 1986). The signal sequence at 1802 has been demonstrated to be functional by cDNA cloning of the shorter transcript (Chu et al., 1985b).

Several studies have demonstrated that a region 3' of the consensus poly(A) signal sequence is also essential for mRNA 3' end formation. McLauchlen et al. (1985) derived a consensus YGTGTTY from an extensive survey of DNA sequences at the 3' terminii of mammalian genes. Gil and Proudfoot (1987) report that sequence motifs present in the 35 nucleotides immediately 3' of the AATAAA are critical for rabbit \(\alpha\)-globin mRNA 3' end formation. Similar motifs can be identified in the 45 nucleotides immediately 3' to poly(A) signal sequence proximal to the translation termination codon of human \(\alpha 1(III)\) procollagen (Figure 3.8).

3.9 A SEQUENCE ANOMALY

Analysis of the sequence of the 3' UTR demonstrated an anomaly when compared to that reported by Chu et al. (1985b), at the position of the poly(A) signal sequence of the longer transcript. The sequence of the 53 nucleotides reported by those workers to precede the poly(A) site is totally different from that determined in the present study. There are two possible explanations for this observation. First, misreporting of the sequence. The sequence reported in this study was determined by subcloning the Hind III - Pst I fragment at the 3' end of clone pIII-33 and by using oligonucleotide primers to initiate sequencing at the 3' end of the 600 bp terminal Pst I fragment of pIII-33 (Figure 3.1). There is no doubt that this sequence is that at the 3' end of pIII-33. Furthermore, this sequence contains a poly(A) signal sequence that is identical to that of the consensus signal sequence, AATAAA; whereas the sequence reported by Chu et al. (1985b) contains a sequence which they claim to be the polyadenylation signal but which is unlike that of the consensus. To date very few eukaryote poly(A) signal sequences have been identified that are not
Figure 3.8

Consensus sequences near the polyadenylation site.

The sequence downstream of the polyadenylation signal sequence at position 1802 is shown. The site of polyadenylation is indicated. Sequence motifs that have been identified in several studies to be essential for correct 3' end formation in other genes are present in the human COL3A1 gene too. These include CAYTG and GT motifs (Gil and Proundfoot, 1987). The latter has been identified as a YGTGTTYY consensus by McLauchlen et al. (1985).
AATAAAATAACTTTCAACAACACTCTTTATGATAACAACACTGTGTATATTCTTTG

polyA

CAYTG

GT
identical to the consensus (Birnstiel et al., 1985). An alternative explanation is feasible as Chu et al. (1985b) claim to have verified the sequence at the end of each cDNA clone by sequencing of the corresponding genomic DNA. Additionally, they locate the Hind III site, which in the present study is situated 5' of the poly(A) site, to be 3' of the transcription terminus. This raises the possibility of alternative splicing of a transcript. To resolve the issue an attempt was made to clone an appropriate restriction fragment containing the sequence between the Hind III site at the end of the type III cDNA and the proximal EcoR I site, as described in the following section.

3.10 CONSTRUCTION OF A MINI-LIBRARY FOLLOWING FRAGMENT ENRICHMENT

A method to directly clone a DNA fragment for which a flanking restriction enzyme map is known was utilised. This method (Nicholls et al., 1985) requires digestion of genomic DNA with multiple enzymes cutting outside the fragment to be cloned, selection of that fragment by electroelution from an agarose gel, and direct cloning into a plasmid vector. Restriction enzymes that recognise a 6 bp sequence will cut, on average, every 4 kb with a distribution of fragment sizes approximating a normal distribution. Assuming that the distribution of such sites is random, then on digestion with several enzymes the number of fragments, Y, of size, X, in the haploid human genome (3x10^9 bp) will be 

\[ Y = P \times \frac{3 \times 10^9}{X}, \]

where P is the fraction of genome remaining (Table 3.5).

Restriction maps of cDNA clone pIII-33 and Idf17(1.7), a genomic clone of the 1.7 kb EcoR I fragment (Figure 3.9) were used to identify enzymes that could be useful in an attempt to clone the 2.5 kb Pst I fragment; which extends from the most 3' Pst I site in pIII-33 to a site 3' of Idf17(1.7). A shorter fragment of 2.45 kb from the Hind III site to the distal Pst I site was also considered for cloning. Suitability of enzymes was tested by digesting 4 µg of human placental DNA with each. Duplicate blots of
Figure 3.9

Restriction endonuclease map in the 3' region of the human COL3A1 gene.

The restriction endonuclease map was used as the basis in the attempt to clone the region marked by ?, by enrichment of a restriction endonuclease fragment containing this region. The map was obtained by single and double enzyme digests of the cDNA clone pIII-33 (Miskulin et al., 1986) and the genomic clone Idf17(1.7) (Chu et al., 1985b). The symbols for the enzymes depicted are: A (Apa I), B (Bgl II), E (EcoR I), H (Hind III), P (Pst I), X (Xho I) and X2 (Xho II).
agarose gels were made; one filter was hybridised with probe pIII 74-925 and the other with Idf17(1.7). Enzymes that produced a hybridising fragment of identical size, and greater than 2.5 kb, with both probes were deduced not to cut within the fragment required for cloning (Figure 3.10).

Several combinations of multi-enzyme digests were performed and the results analysed by Southern blotting with the two probes mentioned above (Figure 3.11). For preparative digests 350 μg of human placental DNA was digested with five enzymes (BamH I, Bgl I, Kpn I, Pst I, Sst I) for isolation of the 2.5 kb Pst I fragment, or with the same set of five plus Hind III for the 2.45 kb Pst I - Hind III fragment. A third set of digests used the basic set of five enzymes plus Hind III and Pvu II. Digests were allowed to proceed for five hours at 37°C and the products were fractionated on a 0.8% agarose gel for 500 volt.hours. A genomic DNA fraction of 2.0 to 3.0 kb was isolated and used for construction of recombinant libraries in pUC13.

Commercially prepared (BRL) competent cells, E. coli DH5α, were used to guarantee high transformation efficiencies and to allow identification of recombinants on BCIG/IPTG plates.

Several hundred recombinants were screened, using Idf17(1.7) as probe, from the mini-library made using the enriched Pst I - Hind III fraction, however, no positives were detected. The mini-library constructed using the fraction enriched for the 2.5 kb Pst I fragment produced very few recombinants, not surprisingly as Pst I produces fragments with 3' recessed ends that are not amenable to dephosphorylation. At present, therefore, the discrepancy in the observed and published sequences remains to be resolved.

Although this procedure failed to produce a clone of the required fragment, it has been used successfully to isolate a clone of the 1.7 kb Bgl II - Pst I fragment which overlaps Idf17(1.7) (data not shown). It would be the method of choice for the repeated isolation of mutant and polymorphic variants of a single genomic locus for characterisation.
4 µg of placental DNA was digested with each of the restriction endonucleases shown in the figure. Two identical Southern filters were prepared. One was hybridised with radio-labelled pIII-74(925), which is the 925 bp Pst I fragment at the 3' end of the cDNA clone pIII-33 (see Figure 3.1). The other was hybridised with radio-labelled Idf17(1.7). Only those enzymes that produced a hybridising fragment of the same size with both probes were used in the subsequent multi-enzyme digests.

The filters were hybridised and washed in standard conditions, and autoradiographed at -80°C for 4 days. Lambda/Hind III marker DNA was used, the sizes of the marker fragments are indicated in kb.
Figure 3.11

Multi-enzyme digests.

Multi-enzyme digests were performed to enrich for the fractions containing the 2.45 kb Pst I - Hind III or the 2.5 kb Pst I fragments which overlap Idf17(1.7) and the 3' end of pIII-33.

4 µg of placental DNA was digested in 20 µl and the products electrophoresed on agarose gels. As before, two Southern filters were prepared and hybridised with either pIII-74(925) or Idf17(1.7) radio-labelled probes.

The lanes are:

A - BamH I, Bgl I, Kpn I, Sst I, Xho II.
B - BamH I, Bgl I, Kpn I, Sst I, Pst I.
C - BamH I, Kpn I, Sst I, Pst I, Hind III.
D - BamH I, Bgl I, Kpn I, Sst I, Pst I, Hind III, Ava II.
E - BamH I, Kpn I, Sst I, Pst I, Hind III, Ava II, Pvu II.
F - Hind III.
G - Hind III, Pst I.
H - Pst I.
I - Pvu II.

The sizes of the Lambda/Hind III DNA fragments used as marker DNA are indicated in kb.
3.11 DISCUSSION

Sequencing of the type III cDNAs has revealed single base substitutions which may be useful as linkage markers, but cannot be analysed by standard Southern blotting techniques. A total of five mutations were identified when compared to one or more published sequences. One of these variant sites, a cytosine-thymine transition, causes an amino acid polymorphism by replacement of a proline by serine at residue 1011 of the triple-helical region. The other four variant sites occur at the third position of codons and so produce no amino acid replacements. Two of the changes, in codons 949 and 1015 of the helical region, do cause restriction site polymorphisms; the former causes a shift in a Hae III by one nucleotide and the other results in the absence of a Hae III site that is present in another reported sequence (Chu et al., 1985b). These polymorphisms are not amenable to analysis by Southern blotting as Hae III restriction sites occur at high frequency in the helical coding sequence. The remaining three mutations do not affect any known restriction enzyme site. At present two restriction site polymorphisms have been identified in the vicinity of the human type III procollagen gene: an Ava I polymorphism (Dalgleish et al., 1985) and a EcoR I polymorphism (Tsipouras et al., 1986). These two RFLPs are inherited as a pattern (haplotype) of the major allele at each site, only one recombinant was detected in a family of 33 individuals (J.R. Hawkins, personal communication). The two sites are separated by a few kilobases and the expected recombination frequency would be very low. Consequently the polymorphic information available by analysis is lower than might be predicted for two linked markers to the type III gene. Pedigree analysis could be simplified greatly by the availability of detecting multiallelic variation and correspondingly high heterozygosities, as a result critical individuals in a pedigree are heterozygous. The variant sites detected in this study could be utilised to construct haplotypes of polymorphic sites. Studies on linkage of β-thalassaemia mutations and β-globin gene polymorphisms has
revealed that specific mutant genes are strongly linked to certain haplotypes (Orkin et al., 1982). It would be of interest to determine whether this a special case or can be extended to other genes, the detection of variant sites by DNA sequencing makes this possible for the human type III gene.

There are several recently developed techniques by which these polymorphic sites can be detected for pedigree analysis. The RNase A cleavage assay is based on the fact that single base mismatch sites in RNA hybrids with RNA or DNA are cleaved by RNase A. However, there are technical problems associated with this method, only 30-50% of possible base mispairings are cleaved (Myers et al., 1985). Furthermore, partial cleavage at a mismatched site could cause problems in studying heterozygotes; although using a time-course of RNase A treatment could overcome this. The presence of these variant sites in close proximity will make the detection of changes difficult. It may not be possible to discriminate between cleavage at different sites as probes that only overlap a particular site cannot be constructed. An alternative strategy would be to employ a the dot-blot assay for detecting point-mutations in unfractionated genomic DNA (Rabin and Dattagupta, 1987). An oligonucleotide probe complementary to either the mutant or normal DNA is labelled by primer extension and hybridised to denatured genomic DNA. It would be possible to generate a series of probes overlapping each base substitution, and use these to screen a panel of DNAs.

Perhaps the most exact method would be to directly clone amplified segments of genomic DNA into a M13 vector for sequence analysis. The polymerase chain reaction method (Scharf et al., 1986) utilises oligonucleotide primers with restriction endonuclease sites at their 5' ends. The in vitro amplification is based on repeated cycles of primer annealing, and primer extension by DNA polymerase I; resulting in an exponential increase in the copies of the region flanked by primers. The efficiency of this method has been demonstrated by cloning of a 110 bp fragment of the human \( \beta \)-globin gene and a 242 bp fragment of the human DQ \( \alpha \)
locus (Scharf et al., 1986). This technique would be most suitable for studying the clustered variant sites detected in this study.

Analysis of the type III sequence provides information about regions of the protein that have been maintained between different procollagen chains and across species. Comparison of the amino-terminal amino acid sequence of the human type III collagen carboxyl-propeptide with that of the chicken type III gene, and other human procollagen carboxyl-propeptides, showed a striking variability among the different types within this region. What is the significance of this sequence variability? Two extreme possibilities may be considered. One possibility is that the sequence connecting the α-helical domain and the carboxyl-propeptide has little or no structural significance and serves merely as a flexible linker between these two domains. Alternatively, the telopeptide and the first 10 to 12 residues of the propeptide have tissue-specific functions, the sequence variability reflecting chain specificity. Subjecting the amino acid sequence of this variable region to a method of predicting secondary structures (Chou and Fasman, 1978), revealed that, despite the variability in sequence, each chain would adopt a very similar conformation. Comparison of each human procollagen chain with its corresponding chicken sequence suggested that for the α1(I) and α1(II) chains the human and chicken chains would be more similar than different chains in one species, this supports the idea that the amino-terminal variable sequence is chain specific. However, similar comparisons for α2(I) and α1(III) were equivocal. The divergence between the α1(III) chains can be explained by the presence of cysteine residues at the junction of the triple-helical region and the telopeptide. These are involved in interchain disulphide bonds and so release the remainder of the amino-terminal variable region from functional constraint.

Two other regions of sequence variability in the carboxyl-propeptide revealed interspecies conservation, particularly the second of these variable segments. It is attractive to suggest that these serve an evolutionaryl
conserved function. The intraspecies variability could serve as distinguishing features necessary for differential chain association. It must be remembered that the role of such sequences may not be simply structural or in the folding of chains. As mentioned before, the terminal 22 amino acids of the α2(I) carboxyl-propeptide can provide post-transcriptional control of collagen and fibronectin gene expression; these 22 residues are highly conserved in all four fibrillar procollagens.

A statistical test was performed to determine whether the nucleotide sequence around the carbohydrate attachment site is significantly conserved, as was originally claimed by Yamada et al. (1983a). Bernard et al. (1983b) in order to address the same question performed chi-squared and Poisson interval distribution tests on the sequence of the carboxyl-propeptide of the human and chicken COL1A1 and COL1A2 genes. They discovered that the distribution of mutations in the COL1A2 sequence was significantly different from a random distribution, although the same was not found for the COL1A1 sequence. However, both tests cannot identify which region of the sequence is responsible for the nonrandom distribution of mutations, those workers assumed it was the sequence around the carbohydrate attachment site. Furthermore, the Poisson interval distribution test was significantly biased by mutations separated by three nucleotides; in other words, the third base position of adjacent codons. In this study an unbiased test was performed on the sequence of the four fibrillar collagen propeptides of human and chicken. This test revealed that the supposedly highly conserved sequence around the glycosylation site is in fact located within a region which is more diverged than average; although the level of statistical significance cannot be determined by this test. Additionally, another region within the carboxyl-propeptide was shown to be more conserved than expected.

Estimation of the rate of evolution demonstrated that the rate of substitution is highest in the α1(III) sequence, with α1(I) evolving the slowest. The difference in evolutionary rate in different molecules or parts of
molecules is not due to a difference in mutation rate rather a difference in selective constraint. There are two possible interpretations for a molecule or part of a molecule evolving at a faster rate. According to the neutral theory (Kimura, 1983), faster evolving sequences are not functionally important and, therefore, a large fraction of the mutations are neutral and accumulate by random drift. A selectionist interpretation is that a more rapidly evolving part of a molecule has some function and undergoes rapid adaptive improvements by accumulating many slightly advantageous mutations by positive Darwinian selection. Such arguments could be applied to the region of about 200 nucleotides in the central part of the propeptide which, when composite sequences of all four fibrillar collagen propeptides were made, reveals a high degree of divergence.

When the rate of substitution at each position of a codon was estimated, it was found that substitutions are most rapid at the third position, followed by the first position, and then the second position. This can be explained by the neutral theory as follows. Among the three codon positions, base substitutions at the second position tend to produce more drastic changes in the physico-chemical properties of amino acids than those at the first position. For example, the codon for proline is CCN, where N is any of the four bases. If the C at the first position is substituted for U, A or G this results in replacement by serine, threonine and alanine respectively. Similar substitutions at the second position lead to leucine, histidine, glutamine or arginine. Miyata et al. (1979) have proposed a simple distance measure between amino acid pairs based on polarity and volume. In terms of their index, serine, threonine and alanine are, respectively, 0.56, 0.87 and 0.06 units apart from proline. The average distance resulting from substitutions at the second position is 2.5. This means that mutations at the first position are less likely of being harmful, in other words selectively neutral, than changes at the second position. Those at the third position are unlikely to cause amino acid changes. Therefore, at the limit in which all mutations are selectively neutral,
the rate of evolution per site is equal to the mutation rate per site. Kimura (1983) argues that synonymous mutations are not far from this limit and, therefore, the evolutionary rates for synonymous mutations per site should be nearly equal for different molecules.

It must be noted that there might be some selective constraint reflected by 'non-random' synonymous codon usage. Perler (1980) argued that the driving force for fixation is positive selection acting on, or some fraction of, amino acid replacement changes; and that such selective fixation carries along with it neutral alterations—hitch-hiking effect. Evidence against this is provided in the present study. The α1(I), α2(I) and α1(II) carboxyl-propeptides have all evolved at a rate one-half that of the α1(III) propeptide, and one-third that of β-globin. The rate of synonymous base changes occurs at a rate that is equal in all four chains and comparable to replacements in fibrinopeptide, one of the most rapidly evolving molecules. Hitch-hiking cannot explain the observation that evolutionary rates of synonymous substitutions among different proteins are roughly equal, even when the amino acid substitution rates differ a great deal.

Why the α1(III) carboxyl-propeptide should be evolving at a greater rate than the other fibrillar collagens can be explained by either neutralist or selectionist arguments. Sequence comparisons from other species and the determination of the function of variable sequences are required. It has to be remembered too that ideally the carboxyl-propeptide should not be considered in isolation from the rest of the molecule.

The presence of polymorphic mRNAs of the type III gene has been demonstrated previously (Chu et al., 1985b; Loidl et al., 1984). It could be argued that these multiple transcripts represent monogenic products of more than one type III gene. This is an unlikely explanation because genetic and molecular evidence suggests the presence of only one copy of the gene in the human haploid complement (RFLP studies, chromosome studies). The two transcripts are colinear, and the difference in the size of the transcripts
5.4 kb and 4.8 kb - is located within the 3' untranslated region.

Two sequences identical to the consensus hexamer AAUAAA were identified by sequencing, these are the signals for polyadenylation of the transcripts (Proudfoot and Brownlee, 1976), the poly(A) addition sites are 686 nucleotides apart. The necessity of having two transcripts of the type III gene that differ only in the length of the 3' untranslated sequence is unknown. The other fibrillar collagen genes, with the exception of type II, also produce two transcripts with a length polymorphism of the 3' untranslated region (Aho et al., 1983; Chu et al., 1984). A change in the relative ratio of the two mRNA species of the α2(I) collagen gene has been reported to occur during chick embryonic development (Merlino et al., 1983). During the first five days, both RNAs are produced in equal amounts, by day 10 the smaller transcript is more predominant. In cultured bovine aortic smooth muscle cells, the ratio of the two mRNA species transcribed from the α(I) collagen gene showed a change, the larger transcript increasing in proportion up to day 8 (Stepp et al., 1986). No study has yet been made as to whether there is a modulation of such transcripts in tissues differentially expressing any given collagen type.

As many as seven mRNAs that differ in their 3' noncoding region have been observed for dihydrofolate reductase. The fraction of mRNAs that are polyadenylated at different sites shifts between growing and stationary cells (Kaufman and Sharp, 1983). This observation suggests that the metabolic state of the cell is important in determining either the efficiency of polyadenylation at various sites, or the stability of mRNA polyadenylated at various sites.

The two human α-globin genes code for identical proteins. The entire nucleotide sequences of the two mRNAs are identical except for the 3' untranslated region, in which the divergence is 17.2%. There is, however, a three-fold difference in the translation efficiency of the two mRNAs (Liebhaber and Kan, 1982). These results suggest that the difference in translation efficiency is determined by the sequences of the 3' noncoding region.
Analysis of the RNA and protein produced in cells transfected with various chimaeras of v-fos and c-fos genes, brought to conclusion the idea that the 3' untranslated region interacts with other parts of the c-fos mRNA to inhibit the synthesis of the fos protein in transfected cells (Miller et al., 1984). It is possible that 3' noncoding sequences interact with specific factors that affect RNA transport or stability; and such sequences may be differentially recognised during development or in different tissues.

Interestingly, the signal sequence for the shorter type III collagen transcript was found to be preceded by an inverted repeat that can form a hairpin loop. The corresponding hexamer for the larger transcript is immediately preceded by an inverted repeat that can be arranged into a stem-loop structure. It is possible that these sequences may participate in the formation of a secondary structure in the primary transcript, which may control the pattern of splicing etc. Stem-loop formation of terminal sequences was seen as a possible regulatory mechanism in the expression of the bovine growth hormone gene, a sequence like the AAUAAA hexamer occupies a single-stranded hairpin loop that is followed by a single-stranded region containing the polyA site (Woychik et al., 1984). The non-polyadenylated mRNAs of histones require the formation of a RNA hairpin structure involving sequences at the 3' terminus of the RNA, (Birchmeier et al., 1983). Association of this region with U7 RNA allows formation of a stem-loop structure downstream of the palindrome; the loop sequence being highly conserved from fruit fly to man. The association with U7 RNA suggests a role for these terminal sequences in RNA processing.

It is interesting to note a previously reported sequence for the end of the large transcript of type III collagen (Chu et al., 1985b) is different to that determined in the present study. Unless there has been a gross error in the reporting of that sequence, the possibility of differential splicing of alternative polyadenylation sites is feasible. The sequence has reportedly been checked by
sequencing of the corresponding genomic clone. An attempt to verify this by cloning a restriction endonuclease fragment containing this region failed, and so the question remains to be resolved. Chu et al. (1985b) identify the sequence UAUAAU as the polyadenylation signal sequence for the end of the large transcript. Sequencing of many genes has revealed that AAUAAA is by far the most predominant sequence, although minor specific nucleotide substitutions are tolerated. A survey of the literature reveals that the sequences AAUGAA, AAUAGA, AAUAUA, AAUAAC, CAUAAA, AAUUAU, UAUAAG, AAUAUA are found at an appropriate site from the polyA addition site of many genes. These may also be functional hexamers, though not necessarily with the same efficiency. The study of the late SV40 mutants injected into frog oocytes showed that the AAUAUA mutant can support cleavage/polyadenylation at 10%, and AACAAA or AAUGAA mutants at 2%, of wild-type level (Wickens and Stefenson, 1984).

In this context, it is of interest to note that the polyadenylation signal sequence of the RNA transcript of human α1(II) collagen gene is AUUAAA; and not either of the AAUAAA hexamers present in the 3' untranslated region (Elima et al., 1987).

Resolution of these anomalous results requires cloning and sequencing of the relevant DNA segment from genomic DNA.
To investigate how homologous sites differentiate from each other starting from a common ancestor $T$ years ago, the model proposed by Kimura (1983) was used. This model utilises the rate of evolutionary base substitutions. As it takes into account the different rates of transitions and transversions it is a more sensitive model than that of Perler's (1980). Fitch (1986) presented a method that estimates the number of nucleotide substitutions since the common ancestor of two nucleotide sequences, with no assumption as to the proportion of transition and transversion substitutions, except that it is constant over time. The method gave the same result as Kimura's, thus verifying the mathematical correctness of each.

Kimura's model is as follows:

\[
\begin{align*}
\text{Pyrimidines} & \quad \text{Purines} \\
\alpha & \quad \beta \\
\beta & \quad \beta \\
\alpha & \quad \beta \\
\end{align*}
\]

Where $\alpha$ is the rate of transitions and need not be the same as $\beta$, the rate of transversions.

The total rate of substitutions per site per year is $K = \alpha + 2\beta$. The total number of substitutions which separate two sequences, and therefore involves two branches each of length $T$, is given by $2TK$

\[
K = 2TK = 2\alpha T + 4\beta T
\]

There are 12 combinations of base changes. $P$ is the probability of homologous sites showing transition type substitutions, and $Q$ is the probability of homologous sites showing transversion differences.

\[
\frac{dP}{dT} = 2\alpha - 4(\alpha + \beta)P - 2(\alpha - \beta)Q
\]
\[ \frac{dQ}{dT} = 4\beta - 8\beta Q \]

\( P = Q = 0 \) at \( T = 0 \), the sequences are identical at the start.

\[
P = \frac{1 - \frac{1}{4} e^{-4(\alpha + \beta)T} + \frac{1}{4} e^{-8\beta T}}{2} + \frac{1}{2}
\]

\[
Q = \frac{1 - e^{-8\beta T}}{2}
\]

From which

\[
4(\alpha + \beta)T = -\log_e(1 - 2P - Q)
\]

and

\[
8\beta T = -\log_e(1 - 2Q)
\]

As \( K = 2Tk \),

then \( K = -\frac{1}{2} \log_e \left[ (1 - 2P - Q) \frac{1}{\sqrt{1 - 2Q}} \right] \)

\[
P = \frac{n_1}{n} \quad \text{and} \quad Q = \frac{n_2}{n}
\]

\( n_1 \) and \( n_2 \) are number of sites at which the two sequences differ with respect to transitions and transversions respectively. \( n \) is the total number of sites compared.

If \( \alpha = \beta \), then,

\[
P = Q = \left[ 1 - \frac{\frac{1}{2} e^{-8\beta T}}{4} \right]
\]

Therefore, \( K = -\frac{3}{4} \log_e (1 - 4\lambda) \)

Where \( \lambda = P + Q = 3Q \), the fraction of sites at which the two sequences differ from each other.
To estimate the synonymous component:

\[ K' = -\frac{1}{2} \log_e (1-2P-Q) \]

The rate of synonymous substitutions was calculated as below, the following model being used.

\[ K = 2Tk = 2(\alpha + \beta + \gamma)T \]

\( P \) is the probability of transition substitutions: UC, CU, AG, GA. \( Q \) is the probability of homologous sites showing UA, AU, CG or GC substitutions. \( R \) is the probability of homologous sites being occupied by UG, GU, CA or AC.

The evolutionary distance, in terms of base substitutions is given by:

\[ K = -\frac{1}{2} \log_e \left( \frac{(1-2P-2Q)(1-2P-2R)(1-2Q-2R)}{(1-2P-2Q)(1-2P-2R)} \right) \]

\[ 2\alpha T = -\frac{1}{4} \log_e \left( \frac{(1-2P-2Q)(1-2P-2R)}{(1-2P-2Q)} \right) \]

\[ 2\beta T = -\frac{1}{4} \log_e \left( \frac{(1-2P-2Q)(1-2Q-2R)}{(1-2P-2R)} \right) \]

\[ 2\gamma T = -\frac{1}{4} \log_e \left( \frac{(1-2P-2R)(1-2Q-2R)}{(1-2P-2Q)} \right) \]

If \( \gamma = \beta \) then

\[ K = -\frac{1}{2} \log_e \left( \frac{(1-2P-Q') \sqrt{1 - 2Q'}}{2} \right) \]

Where \( Q' = Q + R \), the total fraction of transversions.

If \( T \) is known, the rate of base substitutions per site per year is \( k = K \)

\[ 2T \]
The synonymous component is given by,

\[
K_s' = \frac{-1}{4} \log_e [(1-2P-2Q)(1-2P-2R)]
\]

Formulae for determining error variance due to sampling were employed also

\[
K = \frac{1}{\sqrt{n}} \left[ a^2 P + b^2 Q - (aP + bQ)^2 \right]
\]

\[
K'S = \frac{[4P + Q - (2P + Q)^2]}{2(1-2P-Q) \sqrt{n}}
\]

where \( n \) is the total number of nucleotides compared.

\[
a = \frac{1}{1-2P-Q} \quad \text{and} \quad b = \frac{1}{2} \left[ \frac{1}{1-2P-Q} + \frac{1}{(1-2Q)} \right]
\]
CHAPTER 4

ATTEMPTED COSMID CLONING OF THE HUMAN TYPE III PROCOLLAGEN GENE
4.1 INTRODUCTION

Type III collagen accounts for 10-50% of the total collagen in adult tissues such as large arteries, muscle, lung, liver, and skin. Its tissue distribution is similar to that of type I collagen but it is absent from bone and tendon. The function of type III collagen appears to be related to maintaining the integrity of distensible tissues, and in its absence, large vessels and viscera lack the strength to withstand normal stresses. It is this feature that characterises the phenotypes of the connective tissue disorder Ehlers-Danlos syndrome IV (EDS IV).

Biochemical studies of tissues or cultured cells from several patients have demonstrated a deficiency of type III collagen. There is, however, biochemical heterogeneity in the EDS IV phenotype. The deficiency of type III collagen is due to defects in the rate of synthesis, in the secretion, or the stability of this protein (Pope et al., 1975, 1980; Byers et al., 1979; Pyeritz et al., 1984; Stolle et al., 1985; Tsipouras et al., 1986). Recently, the cosegregation of EDS IV with the COL3A1 gene has been demonstrated by use of RFLP linkage analysis (De Paepe et al., 1986; Tsipouras et al., 1986). The ultimate answer as to the nature of the mutations which cause EDS IV will be obtained with the cloning of mutant alleles and the characterisation of the mutations at the molecular level.

Collagen has also been implicated in acquired disorders such as scleroderma and a variety of fibrotic conditions (Prockop et al., 1979). Type III collagen is especially important as it occurs in rapidly proliferating tissues, not only embryonic but also cancerous ones.

Several studies have demonstrated that types I and III collagen genes are expressed in a coordinated fashion; in systems including cultured human fibroblasts (Miskulin et al., 1986), chick embryos (Merlino et al., 1983), and bovine aortic smooth muscle cells (Stepp et al., 1986).

Interestingly, the 5' ends of all three mRNAs for types I and III chicken collagens share several structural features not commonly seen in other eukaryotic mRNAs. Around the
translation start site of these genes a conserved sequence of almost 50 bp is found. This sequence contains an inverted repeat of 8-9 bases capable of forming a stem and loop structure. In each gene two AUG codons are found in the inverted repeat (Yamada et al., 1983); the AUG distal to the promoter being the one used for translation initiation. The remarkable conservation of this inverted sequence suggests that it plays an important role in the control of expression of these genes. More recently, it has been demonstrated that this conserved sequence can form stable intermolecular self antisense dimers in the presence of ribosomal eluate (Rossi and de Crombrugghe, 1987a). It is possible, therefore, that modulation of the equilibrium between monomers and dimers influences the rate of translation of these mRNAs.

An enhancer element in the first intron of the mouse α1(I) procollagen gene has been described (Rossi and de Crombrugghe, 1987b). This enhancer displays cell-specificity since it is functional in fibroblasts but is completely inactive in lymphoid cells. An enhancer that exhibits chondrocyte specific transcriptional control was reported to exist in the first intron of the rat α1(II) procollagen gene (Horton et al., 1987). If such elements are important for directing cell-specific expression of collagen genes, it is necessary to determine whether such elements are also present in the type III gene.

A project to isolate the intact human type III procollagen gene was initiated. To date it has been observed that the size of a particular fibrillar collagen gene is conserved across species. The size of the human type III gene has been estimated as 38 kb on the assumption that it is the same size as the cloned chicken type III procollagen gene (Yamada et al., 1983). This requires a large insert of DNA in the vector if the gene is to be isolated intact. Studies of gene expression and linkage may require cloning of extensive regions of flanking sequences.

The lambda bacteriophage cloning systems have been used most frequently in the past to generate genomic libraries. In this system, the upper size limit of inserted DNA is approximately 20 kb (Maniatis et al., 1978). The primary
system used to clone large DNA fragments efficiently is that of a cosmid vector (Collins and Hohn, 1978). Cosmids are modified plasmids which contain a plasmid replicon, a selectable drug resistance marker, and the lambda cos site which enables in vitro packaging of DNA into phage heads (Hohn and Murray, 1977). Recombinant cosmids can, therefore, be transduced into *E. coli* with high efficiency. The bacteriophage lambda packaging system selects DNA molecules about the size of the lambda genome (37-52 kb). Since the cosmid vector can be small (5-10 kb), nearly all of the packaging capacity of the lambda phage head can be used to package genomic DNA.

There are, however, three major problems in the use of cosmids: vector concatamerisation resulting in packaging of several tandem ligated vector molecules lacking inserts; recombinational rearrangements caused by multiple non-contiguous insert fragments ligated into a single cosmid; and differential growth of cosmid clones causing misrepresentation of sequences in amplified cosmid libraries.

Bates and Swift (1983) described a cosmid vector, c2XB, which avoids the problems of insert rearrangement and cosmid concatamers. This was accomplished by constructing a vector containing two cos elements of lambda separated by a blunt-end restriction enzyme site (Sma I). The use of two cos sites on a single plasmid eliminates the need to prepare separately two cosmid arms, each of which contains only a single cohesive end (Ish-Horowicz and Burke, 1981). After cleavage at the blunt-end restriction enzyme site, cosmid concatamerisation does not occur when a high ATP concentration is employed in the subsequent ligation step. This method is claimed (Bates and Swift, 1983) to allow the rapid construction of a new cosmid library for each individual screening, thereby eliminating any amplification steps which can result in the loss of clones.

As the human type III procollagen gene is estimated to be 38 kb in size, cosmid cloning offered an attractive method to isolate the gene intact, and possibly with flanking sequences too.
The method used to construct a cosmid library has been described in Section 2.13 of Chapter 2, and is presented in Figure 4.1. The procedure employed by Bates and Swift (1983) is claimed to yield $1 \times 10^5$ colonies per $\mu g$ insert DNA. About 350,000 transformants must be generated and maintained to achieve a 99% probability that a particular, single-copy sequence of human DNA will be represented in a library of approximately 40 kb size inserts. This is derived from the equation of Clarke and Carbon (1976):

$$N = \frac{\ln (1-P)}{\ln (1-f)}$$

where $N$ is the number of recombinants in the library, $P$ is the probability of having a given single-copy sequence, and $f$ is the fraction of the total genome represented by each recombinant. $f$ is calculated by dividing the average fragment size by the haploid genome size.

A more detailed account of the attempts to construct a sufficiently large cosmid library is presented below; together with experiments which attempted to determine where problems were occurring.

### 4.2 PREPARATION OF COSMID DNA

i) Vector DNA was prepared as described in Section 2.13 of Chapter 2. After simultaneous digestion with Sma I and BamHI, the DNA was extracted with phenol/chloroform and recovered by precipitation with ethanol. The DNA was redissolved in TE buffer at final concentration of 1 $\mu g/\mu l$.

ii) Ferretti and Sgaramella (1981) demonstrated that 2.5 mM ATP effectively inhibits blunt-end ligations but does not affect cohesive-end ligation. It was discovered empirically that a nominal ATP concentration of 10 mM was required to
Cosmid cloning strategy.

This procedure is that described by Bates and Swift (1983). Following digestion of c2XB DNA with BamHI and SmaI, the vector fragments are ligated to 35-45 kb fragments of human DNA generated by partial digestion with MboI. High ATP concentrations are used to prevent blunt-end ligation of vector molecules. The ligated DNA is used as a substrate for in vitro packaging of lambda bacteriophage particles. Following introduction into E. coli the cosmid DNA recircularises and replicates in the form of a large plasmid. The cosmid contains the β-lactamase gene that confers resistance to ampicillin, allowing for selection of transformants.
target genomic DNA

(1) Partial Mbo I digest
(2) Phosphatase

mix, ligate

30-45 kb

Sma (Blunt) cos Ap\textsuperscript{R} rep cos

37-52 kb

ONLY PACKAGEABLE MOLECULES

Package \textit{in vitro}
inhibit blunt-end ligation. Under these conditions only BamHI cohesive ends can ligate to produce fragments of 11.4, 6.8 and 2.2 kb and residual unligated fragments of 5.7 and 1.1 kb. These same size fragments would be expected if blunt-end ligation occurs; this would result in ligated fragments with cohesive ends at their free ends. These could ligate too and so a ladder of varying sized molecules will be produced. This is what was observed at ATP concentrations of less than 10 mM. Furthermore, as would be predicted, at 10 mM ATP vector fragments ligated to produce fragments of 11.4, 6.8 and 2.2 kb, which were cut by BamHI to restore the 5.7 and 1.1 kb fragments but were refractive to digestion by SmaI (Figure 4.2).

Ligations were performed in the same conditions described by Ferretti and Sgaramella (1981). As 10 mM ATP was required to suppress the ligation of blunt ends, it is possible that the concentration of the stock ATP solution was inaccurate, or that the ATP had hydrolysed on storage, so reducing the effective concentration. To ensure that SmaI ends did not ligate, and so allow the formation of vector concatamers, an alternative method of preparing c2XB DNA was used. The vector DNA was digested by SmaI alone and then subjected to dephosphorylation before digesting with BamHI. The efficacy of dephosphorylation was assessed by subjecting SmaI cut c2XB DNA to ligation conditions (Figure 4.3). Vector DNA cut with SmaI but not treated with phosphatase was observed to self-ligate. DNA that had been dephosphorylated did not self-ligate.

In all subsequent ligations of vector and insert DNA, c2XB DNA that had been prepared by dephosphorylation of the blunt ends was used. The use of a high ATP concentration to inhibit the ligation of blunt ends was not deemed to be effectively reproducible.

4.3 PREPARATION OF INSERT DNA

1) Quality of the starting DNA

The human DNA used for cloning was prepared by partial
10 mM ATP inhibits blunt-end ligation.

c2XB DNA was cleaved with Sma I and then BamH I. 600 ng of DNA was subjected to ligation conditions for 2 hours at room temperature and then 15 hours at 4°C; 10 mM ATP was used in this reaction. The DNA was recovered by ethanol precipitation following phenol/chloroform extraction. 200 ng of the ligated DNA was treated with Sma I (lane: ligated + Sma I) and 200 ng was treated with BamH I (lane: ligated + BamH I). These were run alongside the input and ligated c2XB DNAs.
Figure 4.3

Testing the efficiency of dephosphorylation of the vector DNA.

20 µg of c2XB DNA was cut with Sma I for 4 hours in a volume of 100 µl. 2.5 units of calf intestinal alkaline phosphatase were added and the reaction incubated for 30 minutes at 37°C. DNA was recovered by phenol/chloroform extraction and ethanol precipitation.

0.2 µg of phosphorylated DNA was electrophoresed before and after ligation (lanes: P and P ligated, respectively). Similarly 0.5 µg of DNA subjected to dephosphorylation was also electrophoresed (lanes: deP and deP ligated). Only the phosphorylated DNA is observed to self-ligate.
digestion with Mbo I using the procedures detailed below. Because large fragments of DNA are required for cloning, the DNA must be very large (> 100 kb) before digestion. The placental DNA had been prepared from fresh placentae by lysis in 0.6% tri-iso-propynaphthalene sulphonic acid, 0.8% butan-2-ol and 0.3% SDS, followed by extraction with phenol/chloroform and precipitation with ethanol. The DNA was further purified by treatment with proteinase K and RNase A before further extraction with phenol/chloroform and a final ethanol precipitation. The quality of the starting material was checked by agarose gel electrophoresis of undigested DNA (Figure 4.4). The quality of DNA varied considerably between different preparations. Of several preparations checked only those DNAs that had an average size estimated as 100 kb or greater were considered suitable for cosmid cloning.

ii) Partial digests

Because of its simplicity, DNA fragments to be cloned were obtained by partial digestion of high molecular weight DNA with a restriction enzyme that recognises a 4 bp sequence and generates ends that can be ligated to the BamH I cohesive ends of the vector.

Wherever possible, physical manipulation of the insert DNA was minimised to prevent shearing; pipetting was achieved using pipette tips that had had 0.5 cm removed from the ends. The DNA to be cloned was partially digested with Mbo I. The conditions required were determined by serial dilution of Mbo I, see Section 2.13 Chapter 2. A reaction mixture of placental DNA and restriction enzyme buffer was prepared at a final concentration of 1 µg/15 µl. The reaction products were analysed by electrophoresis through a 0.3% agarose gel. The correct partial conditions were selected by masking off the DNA in a photograph of the gel that was greater than 50 kb and less than 40 kb (Figure 4.5). The various gel tracks were then compared to estimate the degree of digestion that produces the maximum amount of DNA of the desired size. The intensity of the fluorescence is related to the mass distribution of the DNA. To obtain the maximum number of molecules in this size range, the correct partial was not the
Figure 4.4

Testing the quality of DNA to be used for cloning.

Three samples of placental DNA (1, 2 and 3) are shown after electrophoresis on a 0.3% agarose gel for 24 hours at 10 volts. Molecular weight markers are of intact lambda DNA (\(\lambda\)), lambda DNA cut with either Apa I (\(\lambda/{\text{Apa I}}\)) or Sal I (\(\lambda/{\text{Sal I}}\)). The latter two have intact lambda DNA present too. The sizes of the markers are given in kilobases alongside the gel.
Figure 4.5

Establishing conditions for partial digestion of high molecular weight DNA.

1 μg of placental DNA was digested with varying amounts of Mbo I for 1 hour at 37°C in a reaction volume of 15 μl. The samples were analysed by electrophoresis using a series of lambda DNA fragments as molecular weight markers. The sizes of these are given in kilobases. The digestion conditions that gave the greatest mass of DNA fragments in the 35-45 kb range were determined from the intensity of fluorescence.
one with the most DNA in this size range, but the next most partial. The rationale for this is discussed by Seed et al. (1982). The DNA resulting from the appropriately determined partial digestions is shown in Figure 4.6:

iii) Recovery of DNA after dephosphorylation

The insert DNA was not at first recovered after dephosphorylation, the intention being to minimise the physical manipulation of the partially digested DNA which could result in shearing. However, initial attempts to generate a cosmid library produced very poor results and it was possible that the presence of salt or unsuitable pH of the phosphatase buffer was inhibiting effective ligation of the insert DNA to vector. That the phosphatase buffer does inhibit ligations was demonstrated by subjecting lambda DNA, digested with BamH I, to ligation conditions in the absence and presence of phosphatase buffer (data not shown).

Subsequently, the partially digested DNA was recovered after dephosphorylation. Different methods of doing this were assessed using DNA that had been partially digested to give average sized fragments of approximately 20 kb. Any changes that occur to the quality of the DNA, particularly degradation, are easier to assess on material of this size than on DNA of 45 kb size. The DNA was either extracted with phenol/chloroform, with gentle or vigorous mixing, before precipitation with ethanol; or precipitated with 7 mM spermine and rinsed in ethanol as described by Hoopes and McClure (1981); or eluted from a Nensorb 20 purification column (see Chapter 2 Section 2.13). Extraction with phenol/chloroform followed by ethanol precipitation produced a very poor yield of material; also the region of maximum fluorescence was of lower molecular weight than the starting material indicating that it had sheared making it unsuitable for cloning (Figure 4.7). Recovery of DNA from the Nensorb column was very high; however, degradation was again observed. Spermine precipitation also produced a high recovery of DNA and without any observable deterioration in the size quality of the DNA. Spermine precipitation was, therefore, employed to recover the partially digested DNA.
Figure 4.6

Partially digested DNA for cloning.

Half the number of units of Mbo I per μg of DNA that gave the greatest fluorescence in the required size range were used for the large-scale digestions. Lanes A and B represent digests performed with 1/64 and 1/128 Mbo I units per μg DNA respectively. For each digest two separate DNAs are shown, to demonstrate that the conditions were consistently reproducible.

The products of the digest were electrophoresed on a 0.3% agarose gel for 24 hours at 15 volts.
Figure 4.7

Recovery of DNA after dephosphorylation.

1 μg of placental DNA was digested with Mbo I to give an average fragment size of 20 kb. The volume was made up to 100 μl by the addition of 85 μl TE buffer and kept on ice before performing the extraction.

Lanes 1 to 3 represent DNA recovered by phenol/chloroform extraction and ethanol precipitation. The extractions were performed with gentle mixing, lane 1, to more vigorous shaking, lanes 2 and 3. Lane 4 represents DNA recovered by precipitation with 7 mM spermine, as described by the protocol of Hoopes and McClure (1981). Lane 5 contains DNA recovered from a Nensorb 20 purification column. DNA that was not treated after the digest but frozen is in lane 6. Lambda DNA cut with Hind III was used as molecular weight marker, the sizes of the fragments are given in kb.
used for cloning.

iv) Efficiency of dephosphorylation

Partially digested insert DNA that had been dephosphorylated was recovered by precipitation with spermine and subjected to ligation. Whilst DNA that was phosphorylated did self-ligate, producing an upward shift in the mobility of the DNA; DNA that had been dephosphorylated did not self-ligate. The results are shown in Figure 4.8, three DNA preparations of differing mean molecular weight were used for each treatment. It is known that calf intestinal phosphatase not only removes terminal phosphates but can also produce damaged sticky ends. It was necessary to check that the ends of the insert DNA were not damaged by this procedure. The ability of partially digested DNA, before and after phosphatase treatment, to ligate to vector DNA was tested. In both cases the insert DNA was observed to ligate (Figure 4.9); though to a lesser extent for the dephosphorylated material indicating that a proportion of the ends were damaged, this obviously would reduce the efficiency of cloning. The insert DNA used for this experiment had an average molecular weight of 15 kb. When partially digested DNA with an average size of 20-30 kb was treated in a similar manner the phosphorylated DNA alone seems to show an upward shift in the distribution of fluorescence; also the 'tail' has reduced in intensity (Figure 4.10). The phosphatased DNA does not self-ligate. When ligated to vector DNA the treated DNA does seem to show a reduction in the intensity of the 'tail', but it is still not clear if ligations was been successful. The high molecular weight of the input material would mask such a change, or, alternatively, ligation had not occurred. From the results obtained from the smaller size DNA, it had to be assumed that dephosphorylation occurs effectively and without damage to the ends of the DNA to be used for cloning.

Partially digested DNA of the required size range for cloning was treated with calf intestinal phosphatase and ligated to c2XB DNA; and packaged into lambda heads as described in Section 2.13 of Chapter 2. In ligations, several
1 µg of placental DNA was digested for 1 hour at 37°C. Three sets of digests were prepared, and for each set three different concentrations of Mbo I were used. These were 1, 1/2 and 1/4 units, shown as lanes 1, 2 and 3 respectively. All three digests were recovered by precipitation with 7 mM spermine. One set was then frozen (INPUT), and another prepared for ligation (LIGATED). The third set (deP + LIGATED) was treated with 2.5 units of calf intestinal phosphatase for 30 minutes at 37°C, recovered as before and set up for ligation.

Ligations were in 20 µl volume at 16°C overnight, in the presence of 0.25 Weiss units of T4 ligase.
Figure 4.9

Does phosphatase damage the ends of DNA fragments?

1 µg placental DNA that had been digested with 1/4 unit of Mbo I was subjected to ligation before (lane: B) and after (lane: C) treatment with calf intestinal phosphatase. To test whether the ends of DNA had been damaged by phosphatase, they were ligated in the presence of c2XB DNA (lane: E). A control lane of DNA that had not been treated with phosphatase was also included (lane: D).
Dephosphorylation and ligation of 20-30 kb size DNA fragments.

1 µg of placental DNA was digested with 1/16 unit of Mbo I, to produce an average sized fragment of 20-30 kb. The DNA was electrophoresed after recovery with 7 mM spermine and after (lane: 1) or before (lane: 2) ligation. DNA that had been treated with calf intestinal phosphatase was also electrophoresed after (lane: 3) or before (lane: 4) ligation. Placental DNA that had been dephosphorylated was ligated in the presence of vector DNA (lane: 5). A control ligation of c2XB DNA alone was included (lane: 6). The input c2XB vector is present in lane 7.

Molecular weight markers were M1 (λ/Sal I), M2 (λ/Apa I) and M3 (intact λ). The sizes of the fragments are indicated in kilobases.
different mass ratios of vector:insert were used, from 1:1 to 10:1. Different dilutions of the packaged reactions were used to transduce _E. coli_ 1046.

The results obtained were highly variable and not consistently reproducible. The best results were obtained using vector:insert ratios of 1:1 and 2:1 in ligations; and using 6 µl (108 ng DNA) of packaged DNA to transduce _E. coli_ 1046. The efficiency of transduction obtained with these conditions was 0.3 - 0.9 x 10^3 colonies per µg of chromosomal DNA. Controls using DNA of a cosmid clone of the human type II procollagen gene, cosHcol.1 (Weiss et al., 1982), produced transduction efficiencies of 1 x 10^4 per µg DNA. Wild type lambda was packaged and transduced at efficiencies of 2 x 10^7 plaques per µg DNA.

v) Size selected DNA

As an alternative to dephosphorylating partially digested DNA, and making use of the natural selectivity of in vitro packaging to isolate 35-45 kb fragments, it is possible to physically separate partially digested fragments of the correct size. NaCl gradients were utilised to purify fractions containing partially digested DNA of the required size range, as described by Dillela and Woo (1986), see Chapter 2, Section 2.13.

100 µg of DNA was fractionated on a 13 ml NaCl gradient (1.25 - 5 M NaCl in TE buffer). 1 µl, nominally 2.5 µg DNA, was electrophoresed on an agarose gel. The recovery of the DNA was very poor for the required size range, estimated at approximately 10% of input material. There was also a 'tail' of low molecular weight material present (Fig. 4.11). Upon ligating to c2XB DNA and transducing appropriately prepared _E. coli_ 1046 no recombinant colonies were obtained, despite several attempts.

No further attempts to construct a cosmid library were made. All effort was directed to screening a cosmid library that had been donated by Dr. Alastair Craig, EMBL, Heidelberg. The results of that screening are presented below.
200 µg of placental DNA was digested with a concentration of Mbo I units previously determined to give the required fragment size distribution required for cloning. 100 µg of DNA was fractionated on each 13 ml NaCl gradient (1.25-5 M). The middle 6 fractions of 12 were electrophoresed on a 0.3% agarose gel at 15 volts for 15 hours, lanes 1 to 6. Each lane nominally contains 2.5 µg DNA.

Molecular weight markers are: M1 (λ + λ/Hind III), M2 (λ/Apa I) and M3 (λ/Sal I). The sizes of the fragments are in kilobases.
4.4 SCREENING OF COSMID LIBRARY

Human cosmid library H2A was constructed in cosmid vector pcos2EMBL. The insert DNA was prepared from human white blood cells. After amplification on *E. coli* DH1 the recombinants had been packaged in vivo by superinfection as described by Poustka et al. (1984). The library was provided as a phage lysate, this was titred on *E. coli* DH5, and estimated at 7.5 x 10⁵ colony forming units per µl. 1 µl of the phage lysate was used to transduce *E. coli* DH5, which were plated onto nylon filters (Amersham Hybond-N) on three 20 cm x 20 cm LUA plates with kanamycin at 50 µg/ml. The recombinant colonies were screened as described in section 2.14 of Chapter 2, with an oligo-labelled probe of pIII-55, a cDNA clone of the type III gene (see Chapter 3). pIII-55 was used as it overlaps the highly conserved carboxyl-propeptide region. It was hoped, therefore, that the probe might detect other collagen genes in addition to the type III gene.

After three rounds of screening, four clones were finally isolated from the library. DNA was prepared by the plasmid mini-prep method (Chapter 2 section 2.8). Preliminary digests of these DNAs with restriction endonucleases (BamH I, EcoR I and Hind III) identified three of the clones as being identical and the fourth having an extensive overlap with these.

Genomic clones of the 1(I), 2(I) and 1(II) procollagen genes have been isolated and characterised previously (Barsh et al., 1984; Myers et al., 1983; Weiss et al., 1982). During the period of this work the isolation of a partial lambda genomic clone overlapping the 3' half of the human α1(III) procollagen gene was reported (Chu et al., 1985). It was possible, therefore, to predict the sizes of restriction endonuclease fragments that would be produced by digestion of cosmid clones that overlapped the C-propeptide region of any of these genes. The maps are presented in Figure 4.12. The preliminary digests indicated that the cosmid clones isolated from the H2A library had restriction fragments, when cleaved with BamH I, EcoR I and Hind III, common to those of the previously isolated cosmid clone.
(CG103) of the \( \alpha_1(1) \) procollagen gene (Barsh et al., 1984).

i) Hybridisation of cosmid clone DNAs to gene specific probes

Restriction endonuclease digests of the cosmid clone DNAs were blotted and probed with gene-specific probes isolated from each of the four fibrillar collagen genes. The physical relationship of these probes to their respective genes is indicated in Figure 4.12. The figure shows that probe Hf677, a cDNA clone of the \( \alpha_1(1) \) gene that overlaps one-third of the \( \alpha \)-helical region, the carboxyl-propeptide and 3' untranslated sequence, hybridised most strongly to the cosmid DNAs. Probes for the \( \alpha_2(1) \) and \( \alpha_1(11) \) genes hybridised weakly to cosmid DNAs. Cross-hybridisation is expected as these probes contain collagen sequences. The gene-specific probe of the \( \alpha_1(111) \) gene, Idf17(1.7), produced a very weak signal. This probe consists of 3' flanking sequences and would not cross-hybridise to collagen containing sequences. This weak signal was demonstrated to be the result of contamination of probe Idf17(1.7) DNA with pBR322 vector DNA. A probe consisting solely of plasmid pBR322 hybridised to the same fragments. Furthermore, when subcloned into M13 vector (Chapter 2, Section 2.18 for method) before radio-labelling, to isolate it from contaminating plasmid DNA, Idf17(1.7) produced no signal (data not shown). The possibility that the cosmid clones were isolates of the \( \alpha_1(111) \) gene could not be discounted, the clones might terminate upstream of the 3' untranslated region contained in Idf17(1.7). However, although Hf677 also overlaps the highly conserved carboxyl-propeptide, the signal produced was more intense than that obtained with pIII-55, suggesting that the cosmid clones are isolates of the \( \alpha_1(1) \) gene.

ii) Identity of cosmid clones

Restriction digests of cosH2A-5, one of the isolates from the H2A library, and of CG103, a previously isolated cosmid clone of the \( \alpha_1(1) \) gene (Barsh et al, 1984), were blotted and hybridised with probe Hf677. From the published data it was possible to predict which restriction fragments
Figure 4.12

Relationship of the restriction endonuclease maps of the human fibrillar procollagen genes and the carboxyl propeptide of each gene.

For each of the four human fibrillar procollagen genes the restriction endonuclease map in the 3' region of each gene is shown. The maps are based on published data, they are as follows: COL1A1 (Barsh et al., 1984; Chu et al., 1984), COL1A2 (Myers et al., 1983), COL2A1 (Sangiorgi et al., 1985a; Weiss et al., 1982) and COL3A1 (Chu et al., 1985b). The restriction endonuclease fragments used as radio-labelled probes, to screen the cosmid clones isolated from the H2A library, are shown below each map. Hf677, unlike the others, is a cDNA clone; and the coding sequences contained within this clone are indicated below the map of the COL1A1 gene. The DNA sequences which contain the exons coding for the carboxyl propeptide of each gene are indicated by a shaded box.
of digested CG103 would hybridise to Hf677, these were compared to restriction fragments of cosH2A-5 that hybridised to Hf677 (Figure 4.13 and Table 4.1).

In both single and double enzyme digestes, it was observed that several restriction fragments were common to CG103 and cosH2A-5. For instance, the 20 kb and 5.7 kb EcoR I fragments, the 5.7 kb Hind III fragment and the 3.7 kb Xho I fragment. No fragments of these sizes, for each respective restriction endonuclease digest, are predicted for the other three fibrillar collagen genes. Furthermore, double digests produce fragments that are consistent with the restriction enzyme map of the α1(I) gene, in the region overlapping the carboxyl-propeptide. For example, the large 20 kb EcoR I fragment is cut by Xho I to produce 3.7 kb and 1.4 kb fragments that hybridise to Hf677. BamH I cleaves this EcoR I fragment to produce 5.7 and 4.2 kb hybridising fragments. The 5.7 kb Hind III fragment is cut by Xho I to produce 3.7 and 1.7 kb hybridising fragments. This was considered conclusive evidence that the cosmid clones isolated from the H2A library were isolates of the α1(I) gene.

The end-points of the cosH2A-5 clone were determined, by use of radio-labelled probes corresponding to the vector pcos2EMBL. As indicated in Figure 4.14, pcos2EMBL comprises: pBR322 sequence 0-1408 (with the BamH I cloning site at 375), 863 bp EcoR I - BstE II fragment containing the cos sequences, and 3858 bp of R6K sequence containing the origin of replication (ori) and the kanamycin resistance gene. Radio-labelled λ/Hind III, pBR322, and the 3.8 kb EcoR I fragment of pcos2EMBL containing the R6K ori and kanamycin resistance sequences, were used to identify restriction endonuclease fragments that contained specific vector sequences (Figure 4.14).

All three vector probes hybridised to a 6 kb BamH I fragment. This is the intact vector molecule, indicating that the BamH I cloning site has been restored at either end of the vector. Hf677 hybridised to a 12 kb BamH I fragment as predicted from the known data. Probe Hf677 also hybridised to 23 kb and 5.7 kb EcoR I fragments. The 23 kb fragment was cleaved by Kpn I to produce a 15 kb hybridising fragment.
A restriction map was compiled from the data of Barsh et al. (1984) and Chu et al. (1984). The DNA sequence containing the gene is shown in the upper box, the shaded area represents the carboxyl propeptide coding region. Underneath is shown the region of the COL1A1 gene covered by the cDNA clone Hf677. At the bottom are shown the regions contained within the cosmid clones CG102 and CG103 (Barsh et al., 1984) and the region that was determined to contain the clone cosH2A-5.
COL1A1

C-propeptide

Hf677

BamHI

HindIII

EcoRI

XhoI

KpnI

CG102

CG103

COSH2A-5
Table 4.1

Comparison of restriction endonuclease map of COSH2A-5 and CG103, a cosmid clone of the COL1A1 gene.

The table summarizes the data obtained when restriction endonuclease digest of DNA of cosH2A-5 and CG103 were performed. After electrophoresis the products of the digests were blotted by Southern transfer. These were hybridised with radio-labelled Hf677, a cDNA clone of the COL1A1 gene. Post-hybridisation washing was at 0.1 x SSC, 0.1% SDS at 65°C.

All the restriction fragments that were observed are listed. Those that hybridised with Hf677 are underlined. The enzymes were: B(BamH I), E (EcoR I), H (Hind III), K (Kpn I), X (Xho I).
### COSH2A-5

<table>
<thead>
<tr>
<th>K</th>
<th>E</th>
<th>H</th>
<th>X</th>
<th>K+E</th>
<th>K+X</th>
<th>E+X</th>
<th>H+X</th>
<th>H+E</th>
<th>B+E</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;20</td>
<td>20</td>
<td>9.6</td>
<td>18.0</td>
<td>&gt;20</td>
<td>20</td>
<td>9.6</td>
<td>5.5</td>
<td>9.5</td>
<td>9.5</td>
</tr>
<tr>
<td>6.6</td>
<td>6.0</td>
<td>5.0</td>
<td>5.7</td>
<td>6.6</td>
<td>9.6</td>
<td>6.6</td>
<td>4.0</td>
<td>6.2</td>
<td>6.1</td>
</tr>
<tr>
<td>5.7</td>
<td>5.0</td>
<td>3.8</td>
<td>3.8</td>
<td>5.0</td>
<td>5.0</td>
<td>5.7</td>
<td>5.7</td>
<td>5.0</td>
<td>5.6</td>
</tr>
<tr>
<td>4.4</td>
<td>4.0</td>
<td>2.9</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>3.7</td>
<td>4.5</td>
<td>4.2</td>
<td>4.2</td>
</tr>
<tr>
<td>4.0</td>
<td>3.5</td>
<td>2.5</td>
<td>2.3</td>
<td>4.0</td>
<td>3.0</td>
<td>4.0</td>
<td>3.2</td>
<td>2.5</td>
<td>2.7</td>
</tr>
<tr>
<td>1.5</td>
<td>2.0</td>
<td>1.7</td>
<td>1.5</td>
<td>3.0</td>
<td>2.5</td>
<td>3.0</td>
<td>2.5</td>
<td>2.0</td>
<td>2.5</td>
</tr>
<tr>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>0.9</td>
<td>1.3</td>
<td>1.2</td>
<td>0.9</td>
<td>0.3</td>
<td>0.3</td>
</tr>
</tbody>
</table>

### CG103

<table>
<thead>
<tr>
<th>K</th>
<th>E</th>
<th>H</th>
<th>X</th>
<th>K+E</th>
<th>K+X</th>
<th>E+X</th>
<th>H+X</th>
<th>H+E</th>
<th>B+E</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>14.0</td>
<td>10.0</td>
<td>36</td>
<td>15.0</td>
<td>15.0</td>
<td>20</td>
<td>14.5</td>
<td>10.0</td>
<td>6.0</td>
</tr>
<tr>
<td>9.0</td>
<td>9.5</td>
<td>9.5</td>
<td>3.7</td>
<td>6.3</td>
<td>3.7</td>
<td>6.3</td>
<td>4.0</td>
<td>8.4</td>
<td>5.4</td>
</tr>
<tr>
<td>6.3</td>
<td>5.7</td>
<td>5.7</td>
<td>2.3</td>
<td>5.0</td>
<td>3.0</td>
<td>4.3</td>
<td>3.7</td>
<td>3.0</td>
<td>4.0</td>
</tr>
<tr>
<td>1.4</td>
<td>1.4</td>
<td>1.4</td>
<td>1.4</td>
<td>1.4</td>
<td>1.4</td>
<td>1.4</td>
<td>1.4</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>3.8</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
</tr>
<tr>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>1.6</td>
<td>1.6</td>
<td>1.6</td>
<td>1.6</td>
<td>1.6</td>
<td>1.6</td>
<td>1.6</td>
<td>1.6</td>
<td>1.6</td>
<td>1.6</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
</tbody>
</table>
This too could be predicted from the known map and indicates that cosH2A-5 overlaps the Kpn I site present in the $\alpha_1(I)$ gene, as previously identified by Barsh et al. (1984).

Xho I produced two fragments, of 20 kb and 3.7 kb, that hybridised to Hf677. The 20 kb fragment is a junction fragment, containing both insert and vector sequences. The Xho I digest also permitted the determination of the orientation of the vector, which in turn allows more accurate determination of the end-points of the insert. Both $\lambda$ and pBR322 probes hybridised strongly to the 20 kb Xho I fragment, while probe R6K hybridised only weakly. pcos2EMBL has a single Xho I site; one vector Xho I - BamH I fragment contains 813 bp of R6K sequence, 863 bp of cos sequence, and 1033 bp of pBR322 sequence. The other Xho I - BamH I fragment comprises 2941 bp of R6K sequence and 375 bp of pBR322 sequence. The strong hybridisation signal observed with probe $\lambda$, and the weak signal obtained with probe R6K, suggest that at the right end of the clone the vector is oriented as shown in Figure 4.14. Probes $\lambda$ and pBR322 hybridised to 4.5 kb EcoR I and 5.5 kb Hind III fragments. These are junction fragments too, and so the terminus of the clone with respect to the positions of these restriction sites was estimated.

In the EcoR I single digest radio-labelled pBR322 hybridised with a 20 kb fragment, a Xho I fragment of 12 kb, and a fragment of 9 kb produced by the double digest with EcoR I and Xho I. R6K also hybridised with the 12 kb Xho I fragment; but not the same EcoR I or EcoR I - Xho I fragments as there is an EcoR I site at the junction of R6K and pBR322 sequences in the vector. None of these fragments hybridised with radio-labelled $\lambda$, confirming that these fragments are at the "left" end of the clone. In the Hind III and Kpn I - EcoR I digests pBR322 hybridised with 4 kb and 5.6 kb fragments respectively. These allow siting of the end of the clone with respect to EcoR I, Hind III, Kpn I and Xho I restriction sites, as shown in Figure 4.14.
Figure 4.14

Determining the end-points of the cosmid clone cosH2A-5.

The restriction map of the 3' terminus of cosH2A-5 (RIGHT END) is shown in the upper part of the diagram; and that of the 5' terminus (LEFT END) in the lower part. Restriction endonuclease sites are: B (BamH I), R (EcoR I), H (Hind III), K (Kpn I), X (Xho I). The boxed areas represent the vector pcos2EMBL sequence. The regions of the vector comprising pBR322, lambda cos, or R6K DNA sequences are labelled. Below each restriction map are depicted the restriction endonuclease fragments that hybridised to the different components of the vector. At the side of each such fragment are shown those probes which hybridised with that fragment.

A map of pcos2EMBL is also shown. R6K DNA sequences comprise the R6K ori and kanamycin resistance (kan) regions. The pBR322 tetracycline (tet) sequence is contained within an EcoR I - BstE II fragment. The lambda cos sequences are within another EcoR I - BstE II fragment.
4.5 DOES THE COSMID LIBRARY CONTAIN TYPE III COLLAGEN SEQUENCES?

To check whether there were any clones of the type III procollagen gene present in the H2A library, but which may have been missed during the screening, DNA was isolated from the library. 1 µl of the phage lysate was used to transduce E. coli DH5 cells, as for the original screening. The recombinant bacteria were allowed to grow overnight at 37°C on LUA plates with kanamycin. The following day the bacteria were harvested by scraping off the plates into LUB. A scaled-up version of the plasmid mini-prep isolation procedure was used to prepare DNA. A human cosmid library constructed in vector pAVCV7007 (Choo et al., 1986), that had been transduced and amplified in E. coli DH5, was received from Professor Alec Jeffreys. DNA was prepared from this library in a similar manner.

Restriction endonucleases that would produce fragments of known size, and overlap with Idf17(1.7), were identified from the known map of the 3' end of the human 1(III) collagen gene and previous genomic blots hybridised to this probe. DNAs prepared from the two cosmid libraries and placentae were digested with the designated restriction fragments, fractionated by agarose gel electrophoresis and Southern blotted following standard procedures. The filters were hybridised with radio-labelled Idf17(1.7). This had been isolated from contaminating pBR322 DNA by cloning into M13 and preparing double-stranded DNA from a single-stranded template as described in Chapter 2 Section 2.18.

The pAVCV007 library DNA produced no positively hybridising restriction endonuclease fragments. It was concluded that there were no clones that overlapped the 3' end of the type III procollagen gene present in the library.

A positively hybridising EcoR I restriction endonuclease fragment of 1.7 kb was produced with DNA prepared from the H2A library. This corresponds to the size of the Idf17(1.7) probe, which is an EcoR I fragment. However, no other hybridising restriction endonuclease fragments were obtained that corresponded to those predicted, from the published data.
(Chu et al., 1985) and blots of digested placental DNA. More DNA was prepared from the H2A library as described above. This failed to produce a positively hybridising EcoRI fragment. It was concluded that the initial result was spurious; and that there were no clones of the 3' end of the type III procollagen gene in the H2A library.

4.6 DISCUSSION

The ideal vector for cloning DNA from higher eukaryotes would have a cloning capacity that was as large as possible, so that the number of overlapping clones required to isolate a large gene is kept to a minimum. Cosmid vectors are potentially the ideal vectors for cloning collagen genes which are of 30-38 kb size. However, the attempts to construct a cosmid library of a sufficiently large size failed; and due to the difficulty of carrying out control experiments, it was not easy to determine the causes of this. The original method of Bates and Swift (1983) utilised high ATP concentrations in ligations to prevent the generation of vector concatamers. However, this approach was not found to be reproducible at the quoted ATP concentrations; 10 mM ATP was required to prevent blunt-end ligation compared to the reported concentration of 2.5 mM (Ferretti and Sgaramella, 1981). This difference could have been due to errors in preparation of the ATP stock, or to hydrolysis of the ATP effectively reducing the concentration. Solid nucleoside triphosphates are subject to disproportionation reactions (1-2% decomposition per day at room temperature). This gives rise to increasing amounts of di- and tetraphosphates. Disproportionation is due to the lability of the glycosidic bond between carbon atom 1 of the pentose and nitrogen atom 9 of the purine base (adenine). A more effective approach to prevent vector concatamerisation was to prepare vector DNA by dephosphorylation of the blunt ends before cleaving with BamHI.

On ligation to insert DNA only simple concatamers can be formed with vector DNA prepared in such a way, the insert
being flanked by vector arms. Only half of the molecules produced by ligating c2XB and insert DNA have the two cos sites in the same orientation; namely, if there are 5.7 kb and 1.1 kb vector fragments flanking the insert. If the insert is flanked by either two 5.7 kb or two 1.1 kb vector arms the cos sites are in opposite orientation and cannot serve as substrates for the packaging system. This further reduces the efficiency of cloning, but applies equally to both cosmid and lambda cloning. Furthermore, only concatamers with the cos sites separated by about 50 kb can be packaged into lambda heads.

The efficiency of cosmid packaging and transduction was determined using a previously isolated cosmid clone of the human \( \alpha_1(II) \) procollagen gene (Weiss et al., 1982). An efficiency of \( 1 \times 10^4 \) recombinants/\( \mu \)g DNA was attained, about 2000-fold lower than the value for lambda. cosHcol.I DNA was present in circular and supercoiled form. It is known that monomeric circles are packaged poorly. Hohn (1975) found that monomeric circles are packaged with a 13-fold lesser efficiency than linear monomers and concatamers. A more appropriate control for the packaging of cosmids would have been to use linearised cosHcol.I. Despite that, it is clear that packaging of cosmids is significantly less efficient than that of lambda with the in vitro packaging system used. This indicates that a factor in the failure to construct a sufficiently large cosmid library was the poor packaging and transduction efficiency of cosmids. Bates and Swift (1983), quoted a value of \( 1 \times 10^5 \) colonies per \( \mu \)g of insert DNA, Dillela and Woo (1985) cited a value of \( 1-8 \times 10^5 \) transformants/\( \mu \)g chromosomal DNA. Little (1987) reports that using phosphatased insert DNA gives a maximum efficiency of \( 5 \times 10^4 \) transformants/\( \mu \)g DNA, and often much less. The maximal value obtained for the DNA ligated to c2XB was one-tenth that for the cosHcol.I, the cosmid clone of the type II gene; suggesting that other factors were also responsible for the poor efficiency.

Perhaps the most critical component in the success or failure of library construction is the quality of the DNA used to make the cosmid library. Although the size quality of
the starting material was checked and estimated as greater than 100 kb, it has been reported that 35-45 kb partially digested DNA isolated from molecules of starting size 100 kb has only one-third of its molecules with restriction endonuclease sites at both ends (Little, 1987). These broken molecules compete in the ligation reaction and make a successful library difficult to construct. The method of isolating the starting material from placentae is similar to that described by Dillela and Woo (1985) to prepare DNA from white blood cells. This involves several phenol extractions followed by ethanol precipitations. Repeated physical manipulation of high molecular weight DNA results inevitably in shearing of large molecules. The drying of DNA subsequent to ethanol precipitation has been reported to cause DNA denaturation (Svaren et al., 1987). Little (1987) recommends a method which avoids ethanol precipitation and RNase treatment of the DNA; the DNA is extensively dialysed after extraction with phenol.

Isolation of DNA that had been size selected on a salt gradient produced even worse results. It was noted that there was an extensive 'tail' of low molecular weight DNA in the fraction containing 35-45 kb size DNA. This was either due to poor fractionation of DNA on the gradient or degradation of such DNA subsequent to that step. In either case this would counteract the objective of preparing the DNA in this way. This phenomenon has also been reported by Little (1987).

To perform control experiments DNA of an average molecular weight of 15-20 kb was used. This was to allow detection of relatively small changes in the size distribution of the DNA following various treatments. With DNA of a molecular weight of approximately 45 kb it was not possible to easily observe changes of 5 kb or less. By this method it was demonstrated that the insert DNA could be effectively dephosphorylated, albeit with some degradation, and could ligate to the vector DNA. It had to be assumed that similar treatment of partially digested DNA to be cloned would also produce the same result. This inability to perform control experiments directly on the DNA to be cloned made it difficult to identify the roots of the problems encountered.
in generating a cosmid library of the required size.

The H2A library was screened with an oligo-labelled probe of a cDNA clone of the 3' end of the human \( \alpha_1(III) \) procollagen gene. For a cosmid library with an average insert size of 40 kb, 75 000 clones are required to cover the human haploid genome. 750 000 recombinants, 10 genome equivalents, were screened. Two different, overlapping, clones of the \( \alpha_1(I) \) procollagen gene were isolated. However, the library had been amplified previously, a consequence of the \textit{in vivo} packaging step which permits rapid and efficient transfer between hosts of different genetic backgrounds. No clones of the other fibrillar collagen genes were isolated, despite the strong homology in the carboxy-propeptide, the region which is overlapped by the probe used. It had to be concluded that amplification had resulted in the loss of a substantial proportion of recombinants producing an unrepresentative library.

The cosmid library constructed in the vector pAVCV007 was plated onto LUA and the colonies harvested to make cosmid DNA. This was hybridised, after digestion with suitable restriction endonucleases, with the probe Idf17(1.7). This contains the 3' flanking region immediately adjacent to the \( \alpha_1(III) \) procollagen gene. Such intergenic sequences are not conserved, and so the probe should hybridise specifically to its complementary sequence. No signal was detected however. This library had been amplified previously by growth on LUA plates. Little and Cross (1985) demonstrated that for coIE1-derived cosmids (such as pAVCV007), small molecules, probably generated by \texttt{recA} independent illegitimate recombination, will inevitably dominate a population by virtue of their increased copy number.

It is not known whether the R6K origin of replication, used in the construction of pcos2EMBL, also leads to overgrowth of smaller molecules.

Coulson et al. (1986) discovered that 5% of clones in a primary cosmid library were of ribosomal sequences, compared to the 0.05% expected. One possible reason for this may be the absence of EcoK sites from the ribosomal sequences, since EcoK activity has been demonstrated in standard packaging
extracts. Rosenberg (1985) found that packaging systems using *E. coli* K-12 cells produce a 2 to 7-fold loss in the titre of DNA unmodified by Eco K restriction when compared to modified DNA. The Eco K restriction may depress recovery of hybrid molecules and specifically bias libraries against large inserts containing many Eco K sites. If the type III procollagen gene contains several Eco K sites this may explain the absence of clones of this gene in either of the two libraries screened. A better strategy would have been to use an in vitro packaging system such as SMR10 which is derived from *E. coli* C cells. These cells lack the Eco K restriction system. It has been reported that the SMR10 system produces a ratio of 1.3 when comparing the titre of recombinants obtained with unmodified DNA to that of modified DNA (Rosenberg, 1985).

Although the efficiency of in vitro packaging is independent of the concentration of the exogenous DNA; i.e., the number of colonies is linearly dependent on the amount of DNA added (Hohn, 1975); packaging extracts have been reported to inhibit the growth of, and even kill, bacteria (Little, 1987). It is, therefore, necessary to keep the ratio of packaging extract to bacteria low. Partly for this reason only 6 μl of packaged DNA was used to transduce 200 μl of bacteria. It was not possible, therefore, to use a large inoculum of packaged material to infect the cells and so overcome the low transduction efficiency.

Screening of two cosmid libraries failed to produce positive clones of the type III procollagen gene. Both libraries had been amplified prior to receipt. It has been demonstrated that 8.9% of the recombinant phages in lambda Charon 30 grown in recB, recC and sbcB hosts failed to grow in *E. coli* hosts commonly used (Wyman et al., 1985); indicating that most of the currently used cosmid and lambda libraries are not representative of the whole genome, especially after amplification. The failure to clone part of the human thyroglobulin gene was attributed to the presence of sequence which preclude the growth of recombinants in *E. coli* (Baas et al., 1986).

In retrospect, it has to be concluded that the attempt
to clone the type III procollagen gene should not have been attempted in a cosmid vector, because of the technical difficulties of working with high molecular weight DNA. Lambda cloning would be recommended as the method of first choice.
CHAPTER 5

A VARIABLE LENGTH REPEAT SEQUENCE IN THE COL1A1 GENE?
5.1 INTRODUCTION

The segregation analysis of the \( \alpha_1(I) \) procollagen gene in human connective tissue disorders, by means of restriction endonuclease site dimorphisms, has been hampered by the lack of detectable variation within, or linked to, the gene. The first report of RFLPs linked to the COL1A1 locus was published recently (Sykes et al., 1986). A variable Msp I site was mapped 26 kb upstream of the 5' end of the gene, and a Rsa I variant site was found towards the 5' end of the gene. In both cases, the frequency of the minor allele is low, 23% for Msp I and 14% for Rsa I. Consequently, 69% of the four possible haplotypes are of the major allele at each side. This gives a polymorphic information content of 0.43 at the locus.

An investigation of the reported Rsa I polymorphism produced an interesting observation. Rsa I digested human genomic DNA was probed with a 4.8 kb Hind III fragment of cosmid clone CG102, which contains part of the human \( \alpha_1(I) \) collagen gene (Barsh et al., 1984). This Hind III fragment contains a 4.5 kb Hind III - Mbo I subfragment of the 9.6 kb Hind III fragment that overlaps the \( \alpha_1(I) \) gene. It was observed that not only were the two alleles of the Rsa I variant site detected, but also a background of variable sized bands was seen (R. Dalgleish, personal communication). These bands varied from one individual to another (Figure 5.1). It was also noted that the background was only produced if hybridisations were carried out in Marvel or phosphate hybridisation solutions, but not in hybridisation solutions containing herring sperm competitor DNA (data not shown). This suggested that there was a repeat sequence, possibly polymorphic, within the probe. Similar repeat sequences in the herring sperm DNA would compete out hybridisations of this repetitive element.

A polymorphic marker for the \( \alpha_1(I) \) gene would serve as a useful marker in linkage analysis. Particularly as there are technical difficulties in detecting the Rsa I RFLFP. The variant site is often refractive to digestion, resulting in partial digests. A study was initiated to identify and
A repeat sequence is present in the human COL1A1 gene.

Ten samples, each of 4 µg, of placental DNA were digested with Rsa I and electrophoresed on a 0.8 % agarose gel for 15 hours at 15 volts. A Southern blot of the gel was probed with the radio-labelled 4.8 kb Hind III fragment isolated from the cosmid clone CG102 (Barsh et al., 1984). The filters were washed in 0.1 x SSC, 0.1% SDS at 65°C, and autoradiographed at -80°C for 24 hours. The variant Rsa I fragment of 3.6 kb and the common one of 2.6 kb are indicated. Other bands are also seen and they vary in size between different individuals.
characterise this repeat sequence in the \( \alpha_1(\text{I}) \) procollagen gene, as a potential polymorphic marker for linkage analysis.

5.2 ISOLATION OF THE REPEAT SEQUENCE

The approach taken to isolate and characterise the repeat sequence was to identify the smallest suitable restriction endonuclease fragment which could be sub-cloned into M13 for sequencing.

The 4.8 kb Hind III fragment used in preliminary studies is a junction fragment and contains the terminal 4.5 kb Hind III - Mbo I fragment of the cosmid clone CG102, and 0.3 kb of the vector (pHC79). The 0.3 kb vector sequence consists of pBR322 sequence 31-374. The 4.5 kb insert fragment is itself a subfragment of the 9.5 kb Hind III fragment that overlaps the \( \alpha_1(\text{I}) \) gene (Figure 4.13).

From the published data it was possible to predict that there are two BamH I restriction sites within the 4.5 kb fragment. These sites were used as a basis for identifying suitable restriction endonuclease fragments for isolation of the repeat sequence. As predicted, BamH I cleaved the 4.8 kb fragment twice, the BamH I cloning site at the junction of the insert and the cosmid vector having being destroyed upon cloning. Three fragments of 2.6 kb, 1.4 kb and 0.8 kb size were produced. The order of these fragments could be predicted from the published restriction enzyme maps. According to these the 1.4 kb fragment should consist of 1.1 kb insert DNA and 0.3 kb of cosmid vector DNA, and so provide the 3' flank to the central 2.6 kb fragment. The 0.8 kb fragment, therefore, is predicted to be the 5' flank of that 2.6 kb fragment.

However, when these fragments were used as radiolabelled probes and hybridised with DNA of cosH2A-5, a cosmid clone containing the human \( \alpha_1(\text{I}) \) procollagen gene (see Chapter 4), the results were not in agreement with the predictions. The results are summarised in Table 5.1. The 2.6 kb fragment hybridised to all the expected fragments. The 1.4 kb fragment hybridised to 9.6 kb EcoR I, 12 kb Xho I and 5 kb BamH I.
Determining the correct orientation map of the 4.8 kb Hind III fragment isolated from CG102.

The 4.8 kb Hind III fragment isolated from CG102 was digested with BamH I. Three fragments (0.8, 1.4 and 2.6 kb) were isolated and used as radio-labelled probes to hybridise with digested DNA of the COL1A1 cosmid clone cosH2A-5. According to the published map (Barsh et al., 1984) the 1.4 kb fragment should be the terminal junction fragment of the clone, and hence contain vector DNA. The fragment extends to the Hind III site in the pHC79 cosmid vector of CG102, this contains pBR322 DNA of nucleotides 31 to 374. This is supported by the map published by Chu et al. (1984).

A table is shown summarising those restriction enzyme fragments of cosH2A-5 that hybridised with each of the radio-labelled probes. Also shown are the fragments of cosH2A-5 that had been determined previously (Chapter 4) to contain pBR322 sequence 31 to 374. It was observed that the 0.8 probe hybridised with fragments expected to do so with probe 1.4, and vice versa. Furthermore, the 0.8 probe also contains the pHC79 DNA, as it hybridises with those fragments that contain the vector sequence.
<table>
<thead>
<tr>
<th>Enzyme</th>
<th>0.8</th>
<th>2.6</th>
<th>1.4</th>
<th>pBR322 DNA 31-374</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Expected</td>
<td>Observed</td>
<td>Expected</td>
<td>Observed</td>
</tr>
<tr>
<td>KpnI</td>
<td>42</td>
<td>&gt;20</td>
<td>42</td>
<td>&gt;20</td>
</tr>
<tr>
<td>KpnI+EcoRI</td>
<td>15.0</td>
<td>15.0+6.0</td>
<td>15.0</td>
<td>15.0</td>
</tr>
<tr>
<td>EcoRI</td>
<td>20.0</td>
<td>20+20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>EcoRI+XhoI</td>
<td>9.0</td>
<td>9.0+2.3</td>
<td>5.5+2.5</td>
<td>9.6+2.3</td>
</tr>
<tr>
<td>XhoI</td>
<td>12.0</td>
<td>12.0+2.3</td>
<td>12.0+2.5</td>
<td>12.0+2.3</td>
</tr>
<tr>
<td>KpnI+XhoI</td>
<td>3.0</td>
<td>9.0+2.3</td>
<td>3.0+2.5</td>
<td>3.0+2.3</td>
</tr>
<tr>
<td>EcoRI+HindIII</td>
<td>9.5</td>
<td>9.5+4.0</td>
<td>9.5</td>
<td>9.5</td>
</tr>
<tr>
<td>HindIII</td>
<td>9.5</td>
<td>9.5+4.0</td>
<td>9.5</td>
<td>9.5</td>
</tr>
<tr>
<td>HindIII+BamHI</td>
<td>0.8</td>
<td>2.4+0.3</td>
<td>2.6</td>
<td>2.6</td>
</tr>
<tr>
<td>BamHI</td>
<td>4.8</td>
<td>6.0+2.4</td>
<td>2.6</td>
<td>2.6</td>
</tr>
</tbody>
</table>
fragments, which were not as expected.

Furthermore, the 1.4 kb fragment should also hybridise with fragments that contain the pBR322 sequence 31-374. Such fragments had been identified in a previous study (see Chapter 4, Section 4.4). The 0.8 kb BamH I-Hind III fragment was identified as containing pBR322 sequence. Additionally, the second hybridising fragment in each track was not of the size predicted. In fact, the fragments predicted to hybridise with the 0.8 kb and 1.4 kb fragments could be interchanged. It would seem, therefore, that there are inaccuracies in the relative siting of restriction endonuclease sites in the published maps. The correct arrangement of the 4.8 kb Hind III fragment consists of a central BamH I fragment of 2.6 kb, with a 5' flanking fragment of 1.4 kb and a 3' flanking fragment of 0.8 kb. The latter consisting of 0.5 kb insert DNA and 0.3 kb vector DNA.

The source of the variable pattern was localised to a 600 bp Kpn I - Pst I fragment (KP600), subcloned from the 1.4 kb Hind III - BamH I fragment. The physical relationship of these fragments is shown in Fig. 5.2. KP600 was isolated from each of three separate clones of the \( \alpha_1(I) \) procollagen gene; viz: CG102, CG103 (Barsh et al., 1985) and cosH2A-5 (Chapter 4, Section 4.3). The rationale being that if repeat sequence exhibits length polymorphism, this would be observed when this region is sequenced from clones representing different alleles. All three clones contained a 2.6 kb Rsa I fragment (data not shown); in other words, this represented the same allele with respect to the Rsa I restriction fragment length polymorphism. It could be argued that alleles of the polymorphic repeat, if it is polymorphic, would be in strong linkage disequilibrium with a given Rsa I allele due to the very close proximity of the variant Rsa I site to the repeat. However, length polymorphisms are probably generated by gene conversion or replication slippage events. These would occur independently of inheritance of the Rsa I alleles.

KP600 was subcloned into M13mp 18 and M13mp 19 vectors and sequenced using the chain-termination dideoxy method (Chapter 2, Section 2.17). Fragments produced by partial
A map of the 4.8 kb Hind III fragment isolated from CG102 was determined by restriction endonuclease digests and by using isolated radio-labelled probes to Southern filters of digested cosH2A-5 DNA. The shaded box within the 0.8 kb BamH I - Hind III fragment is pBR322 DNA (31 to 374) from the pH79 cosmid vector. The relationship of the 600 bp Kpn I - Pst I fragment (KP600) to the 4.8 kb Hind III fragment is indicated. Restriction endonuclease sites are as follows: B (BamH I), H (Hind III), K (Kpn I), P (Pst I). The sizes of the restriction fragments are indicated above the line for each endonuclease.
digestion of KP600 with Hind II were also subcloned into a M13 vector and sequenced.

A repeat sequence element was observed in the sequence of KP600. The element consists of four sets of GGGCCT, followed by two sets of GGGCT and one of GGT, see Figure 5.3. The sequence of the repeat in all three separate isolates of KP600 was verified by using custom-synthesised oligonucleotide primers. The repeat motif is exactly the same in all three clones. It was concluded that the repeat sequence is monomorphic and not of use in studying variation in the human α1(I) procollagen gene.

The sequence of KP600 was used to search for similar sequences in the EMBL and GENBANK databases. Many sequences containing two to three repeats similar to GGGCCT (principally GGGCT) were identified. One sequence showed striking similarity. This was the sequence of exons 48-51 of the human α1(I) procollagen gene.

Alignment of the KP600 sequence to the sequence of cDNA clones of the COLIA1 gene encoding exons 46-51 revealed that KP600 overlapped exons 48, 49 and 50; and the repeat motif is contained in the intron between exons 49 and 50; namely the second intron from the 5' end of the gene.

5.3 KP600 DETECTS A POLYMORPHIC LOCUS

As KP600 was identified as a monomorphic repeat sequence, the variable band pattern seen in the original Rsal digest must be due to cross-hybridisation of the KP600 repeat with sequences elsewhere in the genome. A panel of fourteen restriction enzymes, including those have recognising four, five and six base palindromes, was tested using placental DNA of three unrelated individuals. All enzymes, except three, produced several hybridising fragments when KP600 was used as a radio-labelled probe (see Figure 5.4). The four cutters Alu I, Rsa I and Sau3A I, either singly or in combination, and the five cutters Ava II, Dde I and Hinf I, produced the most clear results. It was obvious that there was variation in the size of at least one
Figure 5.3

The DNA sequence of KP600.

The nucleotide sequence of KP600 is presented. The GGGCCT repeat motif is underlined with solid arrows, and the GGGCT repeat by a dashed arrow. A computer search of the EMBL sequence database identified KP600 as having an overlap with exons 48-50 of the human COL1A1 gene, accession number X00820 (Chu et al., 1984).
GGTACCATGACCGGACGTGTGGAACCGAGCCCTGCAAGATCTCGTCGACGCAAA

EXON 50
CGGCAAGGTGTGCAGATCGAGTGTCGACGAGACCAAGAATGCCCAGCAGC
CGAAGTCCCCGAGGGCGAGTGTCTGGCTCGTCTCGCCTGACGGCTCAGtgcgtctgc

tcctggagcttcgctctgcacgagasgagcatctgcagctgctctctac

EXON 49
gcgcgtctctctctaggAGTCACCCACCGACCAAGAAAAACCCGCGGCTCGAGct
atctcctgcccctgacatggccactgcgccccgtgaactctcagtcctccctt
ctctaacctggccttttttctttcctccaaatccacagGGACCCAAGGGAGACAC

EXON 48
TGGCCCGAGGCCGAAGGgcctaacc
Variation in the human genome is detected by KP600.

4 µg each of three different placental DNAs (46, 49 and 56) was digested with various combinations of restriction endonucleases for 3 hours at 37°C. Southern blots were prepared after electrophoresis on 0.8% agarose gels, 15 volts for 24 hours. These were hybridised with radio-labelled KP 600. Post-hybridisation washing was in 3 x SSC, 0.1% SDS at 65°C, and autoradiography was at -80°C for 3 days. λ/Hind III DNA was used as a molecular weight marker, the sizes of the fragments are shown in kilobases.

The tracks are as follows: A (Alu I), B (Hae III), C (Rsa I), D (Sau3A I), E (Alu I + Rsa I), F (Alu I + Sau3A I), G (Sau3A I + Rsa I), H (Alu I + Rsa I + Sau3A I), I (Ava II), J (BstN I), K (D de I), L (Hinf I), M (Nci I), N (Sau96 I), O (Ava I), P (Hinc II), Q (Hind III), R (BamH I) and S (BamH I + Hind III).
DNAs 49 and 56 had a 9 kb Hinf I fragment, while DNA 46 had two bands, of lower intensity, of 20 kb and 9 kb. Such an observation can be produced by partial digestion. However, a corresponding band pattern is seen with other enzymes too. This is strong evidence that KP600 is detecting a polymorphic locus due to length variation, and not due to a variant restriction enzyme site.

As mentioned above, three enzymes did not produce hybridising fragments; namely, Hae III, BstNI and Sau96I. All three recognise and cleave palindromic sequences rich in GC, and would cut many times sequences similar to the KP600 repeat.

To characterise the polymorphic locus, a panel of 18 placental DNAs of unrelated individuals was prepared by digestion with MboI and DdeI (MboI is an isochizomer of Sau3AI, the two can be interchanged, except for cloned DNA where MboI sites can be subjected to E.Coli dam methylation). Several alleles, possibly 7 or 8, were seen, varying in size between 4.6 and 15 kb (Fig. 5.5). By determining the proportion of heterozygotes present in the test panel, a rough approximation of heterozygosity was estimated at 78%.

A variable pattern was also observed for bands of about 2 kb size. This variation, more evident with DdeI than MboI, could represent a second locus; or restriction fragment length polymorphisms due to variant enzyme sites within the larger locus. The variation of the lower set of bands seems to be independent of the high molecular set. For example, individuals 46, 50 and 53 have the same sized upper set of bands (9.6 and 5.0 kb in DdeI digested DNA), however they have different sized fragments of less than 2 kb.

This suggests that the lower set of bands represents a separate locus. A formal test of this was to follow segregation of these bands within a family. This would also provide a test of Mendelian inheritance. A family group of eleven individuals was tested. The family was tested for correct paternity/maternity by use of a cloned "minisatellite", λMS8, provided by Professor A. Jeffreys, Leicester (Wong et al, 1987). The post-hybridisation washings
KP600 detects a multi-allelic locus.

4 μg of placental or blood DNA from each of 19 unrelated individuals was digested with either Mbo I or Dde I. The products were electrophoresed on 0.8% agarose gels at 20 volts for 20 hours. Southern filters prepared and hybridized with radio-labelled KP600. The filters were washed at a stringency of 3 x SSC, 0.1% SDS at 65°C, and autoradiographed for 3 days at -80°C. Heterozygosity was determined by counting the number of heterozygotes present as a proportion of the total. The results of the second track were ignored, this individual is probably a heterozygote but exhibits four bands, a consequence of partial digestion with both enzymes.
were performed at a low stringency of 3XSSC, 0.1% SDS to produce a "fingerprint", as seen in Figure 5.6. For each offspring, 50% of the bands were assigned to each of the nominal parents, thus verifying correct maternity and paternity.

When tested with the same hybridising probe, all individuals from the paternal parent failed to react for all bands in that strain, indicating a real difference in the four offspring in the subtelomeric alleles from their father B.

A second pedigrees was constructed, and this pattern of inheritance was more consistent for genuine. A three generation family was examined, and individuals were obtained from CEPH (courtesy of Dr. A. Jeffreys), and whose paternity and maternity had been verified, was analysed. The result is presented in Fig. 5.7. Here the two bands are seen to segregate in a typical autosomal Mendelian pattern.

It was not possible to determine the inheritance of the lower intensity and obscure bands.

3.4 CLONING OF THE HUMAN alpha 1-antitrypsin gene

The polymorphic long terminal repeat (LTR) was used as a detailed procedure is given in reference 20. Size-fractionated human placental DNA from individuals, that had been determined to be negative with the 10000 plaques were obtained. These were hybridised with radio-labelled probe XP5600. After four rounds of screening
were performed at a low stringency of 3XSSC, 0.1% SDS to produce a "fingerprint", as seen in Figure 5.6. For each offspring, 50% of the bands were assigned to each of the nominal parents, thus verifying correct maternity and paternity.

When tested with KP600, the same family pedigree demonstrated that all offspring inherited the same allele from the paternal parent. This is shown in Figure 5.7a. The hybridising fragment of individual A6 is slightly higher than for other individuals. As this anomalous migration is seen for all bands in that track, it reflects abnormal mobility of DNA, possibly due to contamination with protein, rather than a real difference in fragment size. The probability that all four offspring in the subgroup A should inherit the same allele from their father in \((1/2)^4\), and \((1/2)^3\) for subgroup B.

A second pedigree was also tested to determine whether this pattern of inheritance was pure coincidence or genuine. A three generation family of seventeen individuals obtained from CEPH (courtesy of Professor A. Jeffreys), and whose paternity and maternity had been verified, was analysed. The result is presented in Fig. 5.7b. Here the two bands are seen to segregate independently, characteristic of an autosomal Mendelian pattern of inheritance.

It was not possible to determine the pattern of inheritance of the lower set of bands; they were of low intensity and obscured by a background of other, non variable, bands.

5.4 CLONING OF THE HIGHLY VARIABLE LOCUS

The polymorphic locus detected by KP600 was cloned. The detailed procedure is given in Section 2.16 of Chapter 2. Size-fractionated human placental DNA, from several unrelated individuals, that had been digested to completion with Sau3A I was used to construct a library in the \(\lambda\)47.1 vector; 5 - 10 000 plaques were obtained. These were screened with radio-labelled probe KP600. After four rounds of screening
2 μg of blood DNA of each individual from a pedigree of eleven was digested with Hinf I. The DNA was electrophoresed on an agarose gel until the 0.5 kb marker fragment had been run off the gel. A previously isolated locus-specific hypervariable clone, λMS8 (Wong et al., 1987), was used as a radio-labelled probe to a Southern filter of the gel; post-hybridisation washing was in 3 x SS C, 0.1% SDS at 65°C. Autoradiography was at room temperature for 5 hours.

For each offspring, half the bands present can be assigned to one parent and half to the other. This verifies that these are the true biological parents. As several loci are scored by this "fingerprint" there is only a remote chance that they are not the true parents but close relatives.
Figure 5.7

Mendelian inheritance of the multi-allelic locus

In figure 5.7a the Southern filter of the family whose paternity/maternity has been previously checked was tested with the isolated radio-labeled Sau3AI fragment of pXKP20.1, in order to check the inheritance of the multi-allelic locus. The filter was washed in 0.01 x SSC, 0.05% SDS at 65°C and autoradiographed for 3 days at -80°C. All the offspring were observed to inherit the same allele from the paternal parent. This is not typical of autosomal inheritance. To check this, a three generation family pedigree of 17 individuals was also tested for inheritance of this locus. The results in the lower panel show that the fragments are inherited independently of gender.
several isolates were obtained. These all contained a 7 kb Sau3A I insert fragment. One of these clones (λKP20.1) was used for further characterisation.

5.5 CHARACTERISATION OF THE CLONE λKP20.1

The 7 kb Sau3A I fragment was isolated from the phage vector and subcloned into pUC13, this clone was known as φλKP20.1.

Restriction enzyme mapping by single and double digests was not simple. The fragment was refractive to digestion by most commonly used enzymes. As expected, Hae III and other enzymes that recognise and cleave a palindromic sequence rich in GC cut the fragment into minute unresolvable fragments. Surprisingly, Alu I did the same, this contrasts with the result obtained when Alu I was used to digest genomic human DNA (see Figure 5.4). This raises the possibility that the uncloned DNA is modified to make Alu I sites at this locus refractive to digestion.

The enzyme cleavage sites were clustered in a 1 kb region at either end of the 7 kb Sau3A I fragment. This must represent the presence of non-repeated sequences at either end of the fragment.

Sequencing of the Sau3A I insert of φλKP20.1 was achieved by subcloning fragments, obtained either by sonication or by partial digestion, into M13mp18. The majority of clones obtained from sonicated DNA were the same. This clone characteristically contains an AT rich region greater than 100 bp (Figure 5.8 a), this was not seen with any other clone. Further clones were prepared following partial digestion of the Sau3A I fragment with Hae III or Alu I. A representative set of sequences obtained from different clones is presented in Figure 5.8b. Gaps have been introduced to produce maximal sequence identity. A GC rich consensus is present, with minor variations, in most clones. The tandemly repeated motifs based on GGGGCT are embedded within a larger repeat of approximately 50-60 bp. These blocks are separated by apparently unique sequences of undetermined length.
DNA sequencing was performed by sub-cloning restriction endonuclease fragments of the 7 kb Sau3A I insert of pλKP20.1. These fragments were generated by either sonication or partial digestion with Hae III or Alu I. In Figure 5.8a is shown the nucleotide sequence of the 'AT-rich' clone which was isolated repeatedly. There are 34 consecutive nucleotides consisting of alternating A and T. There are no motifs similar to the GGGCCT repeat of KP600, which was used to clone λKP20.1.

In Figure 5.8b are presented the sequences of five M13 sub-clones of the Sau3A I insert of pλKP20.1. The GGGGCT and GGGCT motifs are underlined with arrows. These motifs presumably provide the homology by which λKP20.1 hybridises with KP600. The sequence GCTGCTGG, which is similar to the chi recombination sequence, is indicated by double underlining. This sequence occurs approximately every 60 nucleotides. A repeating block of 50 - 60 nucleotides is identified with a vertical bar. Although the five sequences are highly similar there are differences, not only in certain positions but also in the number of repeats present, indicating that they are independently derived clones.
5.6 MENDELIAN INHERITANCE

Mendelian inheritance of the locus was confirmed using the cloned Sau3A I fragment as radio-labelled probe. It was also noticed that the lower set of bands, previously seen with probe KP600, was not produced. This indicates that the probe is hybridising specifically to one locus, and the lower set of bands must correspond to a separate locus. The presence of unique sequences in the probe, and presumably this other locus, would destabilise any heterologous hybrids at the high stringencies used, 0.001 x SSC, 0.05 % SDS at 65°C.

5.7 CHROMOSOMAL ASSIGNMENT AND LOCALISATION

Southern blot filters of a panel of human/hamster somatic cell hybrid DNAs, digested with Hinf I, were generously provided by Dr. N. Royle (Leicester). This panel had been donated by Dr. S. Povey (University College, London) and is fully described by Wong et al. (1987). The filters were hybridised with the radio-labelled probe of the 7 kb Sau3A I fragment of KP20.1, using the phosphate hybridisation conditions (see Chapter 2, Section 8). Only one DNA of the panel produced a positive signal. This DNA, MOG2C2, is the only one to contain human chromosome 19 and gave 100 % concordance for that chromosome.

This was verified by analysing a further somatic cell hybrid DNA provided by Dr. S. Povey. POT B2/B2 contains only human chromosomes 12, 17 and 19. The results are shown in Figure 5.9.

Chromosomal localisation by in situ hybridisation to metaphase chromosomes was essentially as described by Harper and Saunders (1981), the detailed procedure is given in Section 2.19 of Chapter 2.

The number of silver grains over spreads of high quality, with normal chromosome number and satisfactory
Segregation of the multi-allelic locus in relation to human chromosomal content in 25 human-hamster somatic cell hybrids.

The following somatic cell hybrids were typed: 1, DUR4.3; 2, FIR5; 3, C4A; 4, DUR4R3; 5, 3W4CL5; 6, clone 21; 7, WILF.1; 8, F4SC13CL12P; 9, SIF4A31; 10, FST9/10; 11, HORL 411B6; 12, HORP 9.5; 13, FQ10; 14, PCTBA18; 15, 1a9498; 16, SIF4A2E1; 17, SIF15P5; 18, CTP34B4; 19, TWIN19 D12; 20, TWIN19 F9; 21, TWIN19 F6; 22, TWIN19 C5; 23, MOG2C2; 24, MOG2E5, 25, POTB2/B2. All except POTB2/B2 are described fully by Wong et al. (1987). POT B2/B2 was provided by Dr. Sue Povey (University College, London) and contains human chromosomes 12, 17 and 19.

The radio-labelled Sau3A I insert of pλKP20.1 hybridised to MOG2C2 and POTB2/B2, these are the only hybrids to contain human chromosome 19. Post-hybridisation washing was at high stringency, 0.001 x SSC, 0.05% SDS at 65 C. The results were subjected to an algorithm written by Professor Alec Jeffreys which estimates the percentage concordance for the presence or absence of the intact chromosome and the multi-allelic locus. This analysis produced 100% concordance for the presence of the locus on chromosome 19. Mouse (m), rat (r) and two human (51 and 52) genomic DNAs were used as controls to estimate the degree of cross-hybridisation, if any, of the probe to rodent DNA. A weak cross reaction was seen with rat DNA. Hamster DNA was not available unfortunately.
distribution was recorded. The deletion was scored in 28 cells, 30 were normal, and 19p13.3. The pattern was as shown in Figure 5.10.


5.6 DISCUSSION

The study of human chromosome abnormalities is exemplified by the detection of chromosome abnormalities in conceptions used in antenatal diagnosis, the detection of abnormal nucleotide sequences, and the analysis of recurrent mutations.

Single-copy probes are used to detect restriction fragments of DNA, most of which result from single base substitutions, which restrict endonuclease cleavage.

Many examples of disease have been studied using probes or random cloned DNA sequences. Reviewed by Cooper and Schaidt, 1986, have shown that are not uniformly dispersed. The HLA gene cluster are rich in DNA sequence homology. Other regions, for example, the X and Y (1984), are markedly deficient in homology.

The essential role of chromosome abnormalities in the pathogenesis of inherited connective tissue disorders was discussed in Chapter 1. The detection of and abnormalities is closely linked to, a type I collagen gene. The prenatal diagnosis of such abnormalities by analysis of DNA, using the COLIA1 gene (Williams et al., 1986), and Sykes et al. (1986) reported the use of a 192bp region to detect abnormalities. However, for technical reasons, it has proved difficult to use in segregation analyses. For reasons that remain unclear, the
distribution was recorded. A total of 172 grains were scored in 28 cells. 30% were on chromosome 19. 21% on 19pter and 19p13.3. The pattern of silver grain distribution is shown in Figure 5.10.

The locus was assigned to chromosome 19p13.3-pter. A human gene mapping number, D19S25, has been assigned by Human Gene Mapping Library, Howard Hughes Medical Institute, New Haven, Connecticut, USA (Mankoo and Dalgleish, 1987).

5.8 DISCUSSION

The study of human genetics has been revolutionised by the detection of DNA polymorphisms. They have been used in antenatal diagnosis, mapping of human linkage groups, and the analysis of recombination in inherited cancer.

Single-copy human DNA probes are used normally to detect restriction fragment length polymorphisms (RFLPS) most of which result from small-scale changes in DNA, usually base substitutions, which create or destroy specific restriction endonuclease cleavage sites.

Many examples of RFLPS detected by human gene probes or random cloned DNA segments have been reported (reviewed by Cooper and Schmidtle, 1984). The variable sites are not uniformly dispersed; some regions such as the HLA gene cluster are rich in RFLPS (Wake et al., 1982), whereas other regions, for example thyroglobulin (Baas et al., 1984), are markedly deficient in variation.

The essential role of type I collagen in the genesis of inherited connective tissue disorders has been discussed in Chapter 1. The detection of an RFLP within, or tightly linked to, a type I collagen gene is essential for prenatal diagnosis of such disorders. Despite extensive analysis, it has proved difficult to detect RFLPS associated with the COL1A1 gene (Williams, G., personal communication). Sykes et al. (1986) reported the existence of Rsa I and Msp I dimorphisms in close linkage to this gene. However, for technical reasons the Rsa I RFLP has been difficult to use in segregation analyses. For reasons that remain unclear, the
Figure 5.10

Regional mapping of D19S25 on chromosome 19 by in situ hybridisation.

A diagrammatic representation of the distribution of 172 labelled sites in 28 metaphase cells. The number of grains per chromosomal unit is indicated on the ordinate, while each chromosome is shown on the abscissa. A significant clustering of grains on 19p13.3 – 19pter was observed.
Number of grains observed.
variant Rsa I site is difficult to cleave in a reproducible manner (Dalgleish, R., personal communication). During this analysis it became evident that the radio-labelled probe being used to detect this RFLP was also hybridising other DNA fragments of varying lengths. It was possible that these sequences represented variable length repeat sequences in the COL1A1 gene.

Most RFLPs are dimorphic, with a heterozygosity determined by allele frequencies that cannot exceed 50%. As a result, all such RFLPs are uninformative in pedigree analysis where critical individuals are homozygotes. The more alleles a locus exhibits the more likely it is that an individual will be heterozygous. Therefore, the informativeness of a marker system for a disease locus is much greater for regions of DNA showing multiallelic variation and correspondingly high heterozygosities.

Many such highly variable regions (HVRs) have been isolated in the past few years. For example, near the human insulin gene (Bell et al., 1982), the C-Ha-ras-I oncogene (Capon et al., 1983), the α-globin gene (Jarman et al., 1986), and many such loci which are defined by anonymous DNA clones (Wong et al., 1987; Nakamura et al. 1987). In each case, the variable region consists of tandem repeats of a short sequence; and polymorphism results from allelic differences in the number of repeats. These probably arise by mitotic or meiotic unequal crossovers or by slippage of DNA during replication (Jeffreys et al., 1985). The resulting length variation can be high, with many fold differences in the lengths of a large number of alleles, and heterozygosity frequencies can approach 100%. The detection of HVRs is not dependent on the restriction endonuclease used, as long as it does not cleave the minisatellite repeat unit.

Consequently, the possibility that there is a HVR within the human COL1A1 gene offered great potential for use in linkage analysis. Analysis of the repeat sequence that was identified, however, indicated that the repeat is monomorphic. It is possible that the three clones utilised represented the same allele. Additionally if there are several alleles in the population they may differ by single
or a few repeat units (GGGCCT) only. These variation are too small to be detected by conventional Southern blotting. It could be possible to amplify this region of DNA from total genomic DNA by use of polymerase chain reaction (Scharf et al., 1986). This technique would make it feasible to analyse a large group of individuals, sequencing of the amplified DNA would permit identification of any polymorphisms.

A minisatellite of four tandem repeats of a 33 base pair sequence was observed in an intron of the human myoglobin gene (Weller et al., 1984). Although this minisatellite is itself monomorphic, it detected HVRs elsewhere in the human genome (Jeffreys et al., 1985). Similarly, the repeat sequence in the COL1A1 gene hybridised to other polymorphic loci in the human genome. At high stringencies, two such loci hybridised most strongly. These loci were not detected if the human DNA was digested with Hae III or BstN I, both of which cleave the repeat sequence of the minisatellite in the COL1A1 gene.

KP600 was used to isolate clones of the HVR with alleles of higher molecular weight. This was achieved by using a size-selected fraction of DNA digested to completion with Sau3A I. This HVR was demonstrated to show Mendelian inheritance, and heterozyosity was approximately determined at 78 %. A more accurate estimation of heterozyosity would have been to perform a nearest neighbour analysis to estimate the level of allele sharing between individuals (Wong et al., 1987).

Sequencing studies proved to be difficult. Subcloning into M13 vectors fragments obtained by sonication or by partial digestion repeatedly produced a singular clone. This could be due to the instability of the repeated sequences when propagated in E. Coli JM101 which is not recombination deficient. A more suitable strategy would be to determine the sequence directly from the plasmid clones. Those clones that were sequenced, and they may represent a biased subpopulation from the locus, showed a high level of homogeneity in the repeated GGGGCT motif. There was variation in the sequence of these clones, indicating they are different clones. Jeffreys et al. (1985) reported that highly polymorphic HVRs all show
high repeat copy number with substantial sequence homogeneity of repeats. This, they suggested, is an indication that HVRs are actively and repeatedly engaging in unequal exchange. The rate of unequal exchange in HVRs was determined by the same authors as $10^{-4}$ per kb of sequence. This contrasts to the rate of recombination at meiosis in human DNA at $10^{-5}$ per kb. Thus, it is possible that HVRs are hotspots of recombination. HVRs isolated using the myoglobin minisatellite show sequence similarity to the chi recombination signal of E. Coli (Jeffreys et al., 1985). HVRs isolated by others (Jarman et al., 1986; Nakamura et al., 1987) have core sequences which bear some, but not striking, similarity to the chi sequence. Nevertheless, it is possible that several different sequences promoting recombination could exist in eukaryotes. The repeat motif in the XKP20.1 clone does contain the sequence GCTGCTGG which is very similar to that of chi, GCTGGTGG. Whether such sequences do mediate recombination will require in vitro recombination assays.

The use of somatic cell hybrid DNAs and in situ hybridisation to metaphase chromosome permitted the localisation of a HVR locus hybridising to \textit{XKP20.1}. Interestingly, the most informative of all RFLPs tend to cluster near the ends of chromosome linkage maps (Donis-Keller et al., 1987). These authors have identified that 39% of all RFLPs showing a heterozygosity greater than 0.70 lie within the terminal 5% at either end of the genetic linkage map. It has been speculated (N. Royle, personal communication) that the increased polymorphism in the most telomeric regions of chromosomes may be related to the increased recombination in these regions. Support for this comes from the fact that the pseudoautosomal regions of the sex chromosomes have highly increased recombination and abundant polymorphism (Cooke et al., 1985; Goodfellow et al., 1986).

The main interest in the HVR D19S25 will be its use in linkage studies in genetic disorders. There are three interesting loci which map within the 19p13.2 - 19p13.3 region. Component 3 of complement is probably within 19p13.2 - p13.11, as determined by Lusis et al. (1986) using
translocations between chromosomes 19 and 1. C3 has important functions in the immune system and deficiencies of C3 lead to increased susceptibility to infection (Alper et al., 1987). Yang-Feng et al. (1985) assigned the insulin receptor gene to 19p13.2 - 19p13.3. This site is involved in a nonrandom translocation in pre-B cell leukaemia (Williams et al., 1984).

Familial hypercholesterolaemia is due to a basic defect in the cell membrane receptor for LDL. In most populations the frequency of heterozygotes is not less than 1 in 500. Thus the disease is the most frequent Mendelian disorder, more so than cystic fibrosis. The LDLR gene has been localised to 19p13.1 - 19p13.3 by in situ hybridisation (Lindgren et al., 1985). Obviously a highly polymorphic marker for this gene would be most useful for both pre-natal and post-natal diagnosis, permitting the use of corrective therapy at an early age. The potential use of the HVR identified by p\( KP20.1\) will require further investigation.
REFERENCES
Nucleic Acids Res. 11: 5443-5450.

Nucleic Acids Res. 12: 6117-6125.


J. Biol. Chem. 258: 2653-2661.


Nature. 298: 240-244.
   Genetics. 39: 440-452.


   J. Biol. Chem. 262: 17144-17148.


BARSH, G.S. and BYERS, P.H. (1981) Reduced secretion of structurally abnormal Type I procollagen in a form of osteogenesis imperfecta.


J. Biol. Chem. 259: 14906-14913.


J. Biol. Chem. 262: 7021-7027.

Gene. 26: 137-146.

Science 196: 180-182.


BERNARD, M.P., CHU, M-L., MYERS, J.C., RAMIREZ, F., EIKENBERRY, E.F. and PROCKOP, D.J. (1983b) Nucleotide sequence of the complementary deoxyribonucleic acids for the proα1 chain of human type I procollagen. Statistical evaluation of structures that are conserved during evolution.

BIRCHMEIER, C.N FOLK, W. and BIRNSTIEL, M.L. (1983) The terminal RNA stem-loop structure and 80 bp of spacer DNA are required for the formation of 3' termini of sea urchin H2A mRNA.
Cell. 35: 433-440.


Trends Genet. 3: 342-347.

BIRNSTIEL, M.L., BUSSLINGER, M. and STRUB, K. (1985) transcription termination and 3' processing: the end is in site!
Cell. 41: 349-359.


Cell. 38: 9-16.


Gene. 32: 277-286.


CHU, M-L., GARGIULO, V., WILLIAMS, C.J. and RAMIREZ, F. (1985a) Multiexon deletion in an osteogenesis imperfecta variant with increased Type III collagen mRNA.


COOPER, D.N. and SCHMIDTKE, J. (1986) Diagnosis of genetic disease using recombinant DNA.

CORNISH-BOWDEN, A. (1985) Are introns structural elements or evolutionary debris?


Methods Enzymol. 101: 78-89:

DALGLEISH, R., HAWKINS, J.R. and KESTON, M. (1987) Exclusion of the α2(I) and α1(III) collagen genes as the mutant loci in a Marfan syndrome family.


  J. Biol. Chem. 258: 15192-15197.

to autosomal dominant Ehlers-Danlos syndrome type IV in a large Belgian pedigree.

DEVEREUX, J., HAEBERLI, P. and SMITHIES, O. (1984) A comprehensive set of sequence analysis programs for the VAX.


BRL Focus. 7.2: 1-5.


Biochemistry. 18: 5686-5692.

Cell. 51: 319-337.

DRETZEN, G., BELLARD, M., SASSONE-CORSI, P. and CHAMBON, P.

EMBO J. 4: 1025-1032.


J. Biol. Chem. 262: 17724-17727.

DUNCAN, K.G., FESSLER, L.I., BACHINGER, H.P. and FESSLER, J.H.
J. Biol. Chem. 258: 5869-5877.

Cell. 27: 413-415.

(1985) Human α1(III) and α2(V) procollagen genes are located on the long arm of chromosome 2.

EMANUEL, B.S., SELLINGER, B.T., GUDAS, L.J. and MYERS, J.C.
(1986) Localisation of the human procollagen α1(IV) gene to
chromosome 13q34 by in situ hybridisation.


   Nucleic Acids Res. 9: 3695-3705.


FESSLER, J.H., SHIGAKI, N. and FESSLER, L. (1985) Biosynthesis and properties of procollagens V.


J. Biol. Chem. 262: 13323-13332.


FOCHT, R.J. and ADAMS, S.L. (1984) Tissue specificity of Type
I collagen gene expression is determined at both transcriptional and post-transcriptional levels.


   Cell. 49: 399-406.


   Science. 228: 823-824.

   J. Biol. Chem. 262: 16724-16729.


   J. Biol. Chem. 262: 9844-9849.


GROBLER-RABIE, A. F., BREBNER, D. K., VANDENPLAS, S., WALLIS, G.,

EMBO J. 4: 1745-1748.


HOFMANN, H., FIETZEK, P.P. and KUHN, K.J. (1980) Comparative analysis of the sequences of the three collagen chains \( \alpha 1(1) \), \( \alpha 2 \) and \( \alpha 1(111) \).

HOHN, B. (1983) In vitro packaging of \( \lambda \) and cosmid DNA.


Nucleic Acids Res. 9: 5493-5504.


HULMES, D.J.S. and MILLER, A. (1981) Molecular packing in
collagen quasi-hexagonal model.

Gene. 8: 267-278.

IGNOTZ, R.A., ENDO, T. and MASSAGUE, J. (1987) Regulation of fibronectin and type I collagen mRNA levels by transforming growth factor-B.
J. Biol. Chem. 262: 6443-6446.

Nucleic Acids Res. 9: 2989-2998.

Cell. 26: 141-142.

Biochemistry. 23: 3675-3681.

JARMAN, A., NICHOLLS, R.D., WEATHERALL, D.J., CLEGG, J.B. and HIGGS, D.J. (1986) Molecular characterisation of a hypervariable region downstream of the human α-globin gene
EMBO J. 5: 1857-1863.

JAYE, M., de la SALLE, H., SCHAMBER, F., BALLARD, A., KOHLI, V., FINDELI, A., TOLSTOSHEV, P. and LECOCQ, J-P; (1983) Isolation of a human anti-haemophilic factor IX cDNA clone using a unique 52-base synthetic oligonucleotide probe deduced from the amino acid sequence of bovine factor IX.
Nucleic Acids Res. 11: 2325-2335.


Cell. 21: 555-564.


KADLER, K.E., HOJIMA, Y. and PROCKOP, D.J. (1987) Assembly of collagen fibrils de novo by cleavage of the type I pC-collagen with procollagen C-proteinase.

Biochemistry. 25: 3930-3937.


Cambridge University Press. Cambridge.


J. Biol. Chem. 259: 13668-13673.


LIAU, G., YAMADA, Y. and de CROMBRUGGHE, B. (1985) Coordinate regulation of the levels of Type III and Type I collagen mRNA in most but not all mouse fibroblasts.


Gene. 10: 249-259.
Cell. 38: 597-607.

Nucleic Acids Res. 12: 9383-9394.


LYONS, B.L. and SCHWARZ, R.I. (1984) Ascorbate stimulation of PAT cells causes an increase in transcription rates and a decrease in degradation rates of procollagen mRNA.
Nucleic Acids Res. 12:2569-2579.

Nucleic Acids Res. 15:9205.

Dev. Biol. 59: 75-85.


Cell. 27: 45-55.


McKEON, C., SCHMIDT, A. and de CROMBRUGGHE, B. (1984a) A sequence conserved in both the chicken and mouse α2(I) collagen promoter contains sites sensitive to S1 nuclease.
J. Biol. Chem. 259: 6636-6640.

Nucleic Acids Res. 12: 3491-3502.

McLAUCHLAN, S., GAFFNEY, D., WHITTON, J.L. and CLEMENTS, J.B. (1985) The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3' termini.


Nucleic Acids Res. 12: 7035-7056.

J. Biol. Chem. 258: 10041-10048.


Cell. 36: 51-60.

Methods in Enzymol. 82: 3-32.

with the type I collagen genes during fibroblast growth.

MIYATA,T., MIYAZAWA,J. and YASUNAGA,T. (1979) Two types of
amino acid substitution in protein evolution.

MODESTI,A., KALEBIC,T., Sgarpa,S., TOGO,S., GROTENDORST,G.,

MONSON, J.M., FRIEDMAN, J. and MCCARTHY, B.J. (1982a) DNA
sequence analysis of a mouse pro-α 1(I) procollagen gene:
evidence for a mouse B1 element within the gene.

MONSON, J.M., NATZLE, J., FRIEDMAN, J. and MCCARTHY, B.J.
(1982b) Expression and novel structure of a collagen gene in
Drosophila.

MORRIS, N.P. and BÄCHINGER, H.P. (1987) Type XI collagen is a
heterotrimer with the composition (1α, 2α, 3α) retaining non
triple-helical domains.
  J. Biol. Chem. 262: 11345-11350.

MORRIS, N.P., KEENE, D.R., GLANVILLE, R.W., BENTZ, H. and
BURGESON, R. (1986) The tissue form of type VII collagen is
an antiparallel dimer.

MULLER-GLAUSER, W., HUMBEL, B., GLIATT, M., STAULI, P.,
IX collagen in the extracellular matrix of cartilage: type
IX collagen is localised to intersections of collagen
fibrils.

J. Biol. Chem. 258: 10128-10135.


J. Biol. Chem. 258: 11696-11699.

Cytogenetic location and expression of collagen-like genes in Drosophila.

Biochemistry. 25: 3184-3188.


Lancet. 1: 1193.


Biochemistry. 24: 4223-4229.

J. Biol. Chem. 259: 14170-14174.
OBERBÄUMER, I., LAURENT, M., SCHWARZ, V., SAKURAI, Y.,
YAMADA, Y., VOGELI, G., VOSS, T., SIEBOLD, B., GLANVILLE, R.W.
and KUHN, K. (1985) Amino-acid sequence of the non-
collagenous globular domain (NCI) of the \( \alpha_1(IV) \) chain of
basement membrane collagen as derived from cDNA.

Structural diversity and domain composition of a unique
collagenous fragment (intima collagen) obtained from human
placenta.

OGILVIE, D.J., WORDSWORTH, B.P., PRIESTLEY, L.M., DALGLEISH, R.,
SCHMIDTKE, J., ZOLL, B. and SYKES, B.C. (1987) Segregation of
all four major fibrillar collagen genes in the Marfan
syndrome.

OHKUBO, H., VOGELI, G., MUDRYJ, M., AVVEDIMENTO, E.V.,
SULLIVAN, M., PASTAN, I. and de CROMBRUGGHE, B. (1980)
Isolation and characterisation of overlapping genomic clones
covering the chicken 2 (Type I) collagen gene.

PAGLIA, L.M., WIESTNER, M., DUCHENE, M., OVELETTE, L.A.,
procollagen peptides on the translation of Type II messenger
ribonucleic acid and on collagen biosynthesis in
chondrocytes.
Biochemistry. 20: 3523-3527.

Cell. 41: 343-345.

PELTONNEN, L., PALOTIE, A. and PROCKOP, D.J. (1980) A defect in
the structure of type I procollagen in a patient who had
osteogenesis imperfecta: excess mannose in the COOH terminal
propeptide.


Hum. Genet. 75: 120-122.


ROSENBERG, S.M. (1985) Eco K restriction during in vitro packaging of c oJtphage lambda DNA.

Nucleic Acids Res. 15: 8935-8956.


SANGIORGI, F.O., BENSON-CHANDA, V., de WET, W.J., SOBEL, M.E. and RAMIREZ, F. (1985b) Analysis of cDNA and genomic clones coding for the pro-α1 chain of Type II collagen.


J. Biol. Chem. 258: 9504-9509.

SCHMIDT, A., YAMADA, Y. and de CROMBRUGGHE, B. (1984) DNA sequence comparison of the regulatory signals at the 5' end of the mouse and chick α2 Type I collagen genes.
J. Biol. Chem. 259: 7411-74165.

Cell. 41: 339-340.


imperfecta.

TSIPOURAS, P., BYERS, P.H., SCHWARTZ, R.C., CHU, M-L., WEIL, D.,
syndrome IV: cosegregation of the phenotypes to a COL3A1
allele of type III procollagen.

TSIPOURAS, P. and RAMIREZ, F. (1987) Genetic disorders of
collagen.

VOGELI, G., OHKUBO, H., AVVEDIMENTO, V.E., SULLIVAN, M.,
YAMADA, Y., MUDRYJ, M., PASTAN, I. and de CROMBRUGGHE, B. (1980)
A repetitive structure in the chick α2 collagen gene.

mp7-derived system for insertion mutagenesis and sequencing
with universal primers.

Regulation of types I, III, and IV procollagen mRNA
synthesis in glucocorticoid-mediated intestinal development.
J. Biol. Chem. 262: 10814-10818.

and complexity of the genes for HLA-DR B-chains - direct
analysis by DNA-DNA hybridisation.

proα2(V) collagen gene is evolutionarily related to the
major fibrillar-forming collagens.

WEIL, D., MATTEI, M-G., PASSAGE, E., VAN CONG, N., PRIBULA-


analysis by DNA-DNA hybridisation.

proα2(V) collagen gene is evolutionarily related to the
major fibrillar-forming collagens.

WEIL, D., MATTEI, M-G., PASSAGE, E., VAN CONG, N., PRIBULA-
CONWAY, D., MANN, K., DEUTZMANN, R., TIMPL, R. and CHU, M-L.
encoding the three chains of type VI collagen.

WEISS, E.H., CHEAH, K.S.E., GROSVELD, F.G., DAHL, H.H.D.,
SOLOMON, E. and FLAVELL, R.A. (1982) Isolation and
characterisation of a human collagen α1(I)-like gene from a
cosmid library.

Organization of the human myoglobin gene.
EMBO J. 3: 439-446.

de WET, W., BERNARD, M., BENSON-CHANDA, V., CHU, M-L.,
the human pro-α2(I) collagen gene.
J. Biol. Chem. 262: 16032-16036.

AAUAAA sequence: four AAUAAA point mutations prevent mRNA 3'
end formation.

WIESLANDER, J., BARR, J.F., BUTKOWSKI, R.J., EDWARDS, S.J.,
antigen of the glomerular basement membrane. Localisation to
noncollagenous regions of type IV collagen.
WILLIAMS,C.J. and PROCKOP,D.J. (1983) Synthesis and processing of a Type I procollagen containing shortened pro- $\alpha_1(I)$ chains by fibroblasts from a patient with osteogenesis imperfecta.
J. Biol. Chem. 258: 5915-5921.


Nature. 300: 401-402.


J. Biol. Chem. 262: 16376-16385.


requires mutant *Escherichia coli* hosts.


YAMADA, Y., KUHN, K. and de CROMBRUGGHE, B. (1983) A conserved nucleotide sequence coding for a segment of the C-propeptide is found at the same location in different collagen genes. Nucleic Acids Res. 11: 2733-2744.


The collagens are a family of structural proteins which function as an extracellular framework in eukaryotic organisms. They are characterised by a unique protein conformation which consists of three polypeptide chains in a triple helix. At least twelve collagen types encoded by at least twenty non-allelic genes have been identified in vertebrates.

There is considerable evidence that each member of the family is represented by a single copy gene. These genes constitute a multi-gene family with a common evolutionary origin. Some of the genes are known to be clustered on certain human chromosomes.

The collagens have an extremely important role in development. Alterations in collagen genes can result in a heterogenous group of heritable diseases of connective tissue. There is accumulating evidence that similar phenotypes are due to similar mutations or location of mutations.

DNA sequencing studies of cDNA clones of the human type III procollagen gene revealed single base polymorphisms, and one amino acid polymorphism, by comparison with published data. These may be used to generate haplotypes at this locus and increase the polymorphic content for genetic analysis. An attempt to isolate the human type III procollagen gene by cosmid cloning failed. Due to the technical difficulties of performing control experiments, the reasons for this failure could not be identified with certainty. It is possible that this gene is not amenable to cloning using the commonly used cloning reagents.

A study of the human type I procollagen gene, COL1A1, revealed the presence of a monomophic repeat sequence in an intron. This repeat sequence was used to identify a multi-allelic locus in the terminal region of the short arm of chromosome 19. The usefulness of this locus in linkage studies to disorders that map close to this region remains to be analysed.