1994

Cloning and characterization of a new human collagen locus

Sheryl A. Cammarata
Loyola University Chicago

LOYOLA UNIVERSITY CHICAGO

CLONING AND CHARACTERIZATION OF A NEW HUMAN COLLAGEN LOCUS

A THESIS SUBMITTED TO
THE FACULTY OF THE GRADUATE SCHOOL
IN CANDIDACY FOR THE DEGREE OF
MASTER OF SCIENCE

DEPARTMENT OF BIOLOGY

BY

SHERYL A. CAMMARATA

CHICAGO, ILLINOIS

JANUARY 1994

ACKNOWLEDGEMENTS

I extend a very special thanks to my advisor, Dr. Jeffrey L. Doering, whose encouragement and support has contributed infinitely to this acomplishment. I would also like to thank my committee members, Dr. Howard Laten and Dr. Holden Maecker for their guidance and support. My sincere thanks also to Anne Burket for offering her wisdom and kindness.

Thanks to my families for their unending patience and understanding. Finally, with heartfelt gratitude I thank my husband, John Silverstein, without whom I could never "Imagine".

PUBLICATIONS

Cammarata, S.A., Burket, A.E. and Doering, J.L. (1991) Characterization of a new human collagen locus associated with osteogenesis imperfecta. J. Cell Biol. 115,106a.

Wickens, D.D. and Cammarata S.A. (1986) Response class interference in STM. The Bulletin of the Psychonomic Society. 24(4),266-268

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Osteogenesis imperfecta (OI) is a heterogeneous genetic disease which has been characterized by extremely fragile bones and other connective tissue abnormalities (Sillence et al., 1979a, 1979b; Prockop, 1984). Molecular defects in the Type I collagen genes which affect the synthesis and/or structure of the collagen protein are involved in many cases of OI (Prockop et al., 1989; Byers, 1989; Cole et al., 1989). Most of these mutations alter the repeated Gly-X-Y amino acid sequence in the triple helical domain, leading to the formation of an unstable helix which compromises the tensile strength of mature collagen fibrils (Pope et al., 1989).

The severity of the OI symptoms vary considerably between individuals and no two cases have been found to have exactly the same molecular defect. In cases where individuals have similar mutations their phenotypes can vary greatly (Pope et al., 1985). Furthermore, there are instances where multiple asymptomatic members of a family carry the same defect as the affected individual (Superti-Furga et al., 1989; Sykes et al., 1990). In some cases of OI, particularly the progressively deforming moderately severe form of the disease, or OI Type III, no apparent Type I collagen defects have been found (Byers et al., 1987; Aitchison et al.,

1988). Thus, there are clearly additional genes involved in contributing to the OI phenotype other than the Type I collagen genes.

Previous studies have shown that a pro-α2(I) collagen complementary DNA (cDNA) probe detects two previously-undescribed sites in collagen-like gene(s) other than the already characterized fibrillar collagen genes (Doering et al., 1990). Deletions at the two sites are found to occur much more frequently in OI Type III patients than in normal individuals. At both sites the deletions occur in a spontaneous fashion and are of heterogeneous size (Doering et al., 1987). Site 1 is detected in normal individuals as a 7.4 kb HindIII restriction fragment which completely contains a 3.0 kb PstI fragment. Site 2 is detected as a 4.6 kb HindIII fragment in normal individuals which overlaps with a 5.2 kb PstI fragment. Site 2 has recently been localized to chromosome 17 (Cammarata et al., 1991).

This thesis describes the isolation and characterization of Site 2, specifically the 4.6 kb HindIII fragment. The cloning, mapping and sequencing of its triple helical coding regions are described. Sequence comparisons to known collagen genes, as well as other non-collagenous genes, are made in an effort to further characterize this gene and gain insight to its function. This sequence information will help to identify the structure of the putative collagen-like protein which the sequence encodes and thus suggest how

abnormalities in this protein could contribute to the OI phenotype.

LITERATURE REVIEW

OI and Type I Collagen

Osteogenesis imperfecta is one of the most common connective tissue disorders and occurs in approximately 1 in 5,000 to 10,000 individuals (Byers and Steiner, 1992). In addition to fragile bones which fracture under minimal stress, other symptoms occur to a varying degree, including dentinogenesis imperfecta, thin skin, joint hypermobility, blue sclerae and hearing impairment (Sillence, 1988; Byers, 1989). All of the affected tissues are rich in Type I collagen and many, but not all, OI cases are associated with defects in one of the two Type I collagen genes, COL1A1 or COL1A2 (Byers et al., 1984; Cohn et al., 1988; Bateman et al., 1988). In approximately 85% of those affected with OI, a dominant mutation in one of the two structural genes for Type I collagen is present (Wenstrup et al. 1990).

Collagen is the major structural component of the extracellular matrix or ECM (Kuhn, 1987). Type I collagen accounts for about half the dry weight of bone and over 80% of the dry weight of skin, tendons and ligaments (Prockop and Kivvirko, 1984). It is a heterotrimer composed of two $\alpha 1$ and one $\alpha 2$ polypeptide chains, designated pro-$\alpha 1$(I) and pro-$\alpha 2$(I)

respectively. The two individual chains are coded for by the well-characterized genes COL1A1 and COL1A2, respectively (Chu et al., 1985a; Myers et al., 1983; de Wet et al., 1987). Both chains are synthesized as procollagen precurser molecules, consisting of a central triple helical region made up of a repeated Gly-X-Y amino acid sequence and flanking propeptide regions at the carboxy and amino termini.

Collagen Assembly

Triple-helix formation begins with the aid of sequences at the carboxy end of the chains and proceeds in a zipper-like fashion toward the amino terminus. The small glycine residue in every third position of the central molecule facilitates the folding of the three chains into a characteristic triple helical configuration allowing tight association of the polypeptide chains. Proline and lysine residues are often present in the X and Y positions of the repeat. During helix assembly these residues are hydroxylated and some prolyl residues are further glycosylated. This post-translational modification is important for helix stability and continues until triple helix formation is complete (Prockop, 1984; Byers, 1989; Byers et al., 1987). Therefore, the extent of modification is dependent upon the rate of helix formation (Bateman et al., 1988).

Upon completion of helix formation the procollagen molecule is secreted into the ECM and its carboxy and amino

propeptide regions are cleaved, leaving a mature collagen triple helix with short telopeptides at both ends (Prockop et al., 1989). Within the ECM multiple collagen polypeptides are arranged end on end, stacked in a parallel fashion and crosslinked in a quarter staggered array of collagen microfibrils (Bruns and Gross, 1974; Prockop, 1984). Thus, a compact and rigid triple helix is essential for the tensile strength of mature collagen fibers (Byers et al., 1987; Byers 1989). The propeptide sequences at the carboxy and amino ends are important in chain alignment and secretion of the mature collagen molecules into the extracellular matrix (Sandell and Boyd, 1990; Prockop et al., 1989; Byers et al., 1987). The non-helical telopeptides represent sites where inter- and intra-molecular crosslinking occurs in the mature molecule via disulfide bonding between cysteine residues. Both the appropriate post-translational modification and intermolecular crosslinking are important in stabilizing the molecular arrangement within collagen molecules and between resultant fibrils, affording them thermal stability at body temperatures (Kuivaniemi et al., 1991).

Because the extent of post-translational modification is dependent upon the rate of helix formation, mutations in the Type I collagen genes which result in a decrease in the rate of triple helix formation often result in overmodification (Bateman et al., 1986; Bonadio and Byers, 1985). This can interfere with the folding, processing and formation of

thermally stable collagen molecules. Those unstable molecules are often rapidly degraded resulting in the production of reduced amounts of collagen.

Type I Collagen Defects in OI

Many different kinds of Type I collagen defects associated with OI have been described. Two general classes of mutations have been obeserved: mutations which affect only the amount of Type I collagen produced (Barsh & Byers, 1981) and mutations that alter the structure of the Type I collagen chains. The mildest forms of OI generally result from decreased collagen production due to a non-functional or partially functional allele (Prockop et al., 1989). The more severe phenotypes result from the secretion of structurally defective, overmodified procollagen chains that are incorportated into mature collagen molecules. Subsequently, these interfere with the self-assembly of fibrils (Cohn and Byers, 1991; Kuivaniemi, et al., 1991).

The system of collagen biosynthesis and subsequent fibril formation depends upon the principle of nucleated growth. Such systems begin with a few subunits forming a core nucleus that is then self-propagated to generate a larger structure. This kind of system is highly vulnerable to mutations that cause the synthesis of structurally defective but partially functional molecules. In bone, collagen fibrils are very densely packed. If an abnormal molecule is

incorporated into the fibril the uniform environment which contributes to tissue strength is compromised. Not only is that molecule less stable, but adjacent molecules are also affected. It is also possible that the fibrils which have incorporated abnormal molecules may not be mineralized as efficiently (Byers, 1990). It is not suprising then, that production of structurally defective molecules would cause a more severe phenotype. Thus, in terms of clinical symptoms, it is less detrimental to produce fewer normal collagen chains than to incorporate large numbers of defective ones.

Among the most commonly studied gene defects in OI are point mutations which cause a codon change, most commonly a substitution for a Gly residue (Cohn et al., 1990, Wallis et al., 1990; Cole et al., 1992). These substitutions often result in the disruption of the triplet pattern or cause incorrect exon splicing (Tromp and Prockop, 1988). Deletions can also cause splicing defects (Kuivaniemi et al., 1988; Tromp and Prockop, 1988) or missing coding information. Occasionally insertions have been observed (Byers et al., 1988b; Genovese et al., 1989) and sometimes small insertions or deletions are responsible for frameshift mutations (Willing et al., 1990; Bateman et al., 1989; Pihlajaniemi et al., 1984).

Types of OI

The severity of this disease varies greatly from

individual to individual with regard to frequency of fractures as well as the extent to which the other symptoms are present (Superti-Furga, 1989). Its heterogeneity is reflected at both the clinical and molecular levels but its classification into one of four types is based mainly on clinical symptoms (Sillence, 1979a; Sillence, 1988; Byers et al., 1987; Prockop, 1984). Most research at the molecular level has focused on cases of OI where Type I collagen defects are found.

Severe perinatal lethal OI, or OI Type II, is the best studied form of the disease. It is characterized by neonatal dwarfism, multiple fractures leading to severe bone deformity, and results in death in utero or shortly after birth (Sillence, 1988). It is usually the result of new dominant mutations and involves heterogeneous defects in one of the Type I collagen genes which primarily disrupt helix formation (Byers et al., 1988a; Byers et al., 1988b). For example, point mutations which cause a glycine residue to be replaced by a bulkier residue in the triple helical portions of the collagen molecule not only alter the repeated Gly-X-Y pattern, slowing helix formation, but also result in poor secretion of the overmodified, thermally unstable procollagen molecule into the extracellular matrix (Cohn et al., 1988; Willing et al., 1988). Those structurally defective molecules, with varying degrees of thermal stability, which do get secreted are then incorporated into the microfibril and have consequences for the nucleated growth of mature fibrils. For example, one

lethal variant with a cysteine for glycine substitution in the α1(I) chain did not prevent the folding of the triple helix, but instead it introduced a "flexible kink" in the helix. This delayed fibril formation, reduced the total amount of collagen incorporated into the fibrils, and drastically changed the morphology of the fibrils (Prockop et al., 1989).

The substituted residue in Type II OI cases is often cysteine or arginine (Vogel et al., 1987; Bateman et al., 1987) but substitutions by valine (Lamande et al., 1989; Patterson et al., 1989), serine (Lamande et al., 1989) and aspartate (Baldwin et al., 1989) have also been described. Other mutations such as deletions in the helical coding regions of the Type I collagen genes (Barsh et al., 1985; Chu et al., 1985b; Willing et al, 1988) and point mutations which lead to splicing defects causing missing exons (Tromp and Prockop, 1988) have also been described. Additionally, some recurring cases of Type II OI, originally thought to be due to autosomal recessive inheritance, can be accounted for by germ line mosaicism in one parent (Byers et al., 1988a; Cohn et al., 1990; Constantinou et al., 1990; Wallis et al., 1990).

The mild dominantly inherited form of OI, or Type I, also involves defects in the structure and/or synthesis of Type I collagen. It is characterized by bone fragility which decreases at puberty, short stature, blue sclerae, deafness and sometimes dentinogenesis imperfecta (Sillence, 1988). In a number of studies the phenotype has been shown, by analysis

of restriction fragment length polymorphisms (RFLP), to consistently segregate with one or the other of the Type I collagen genes (Sykes et al., 1986; Tsipouras et al., 1984; Falk et al., 1986). Thus, defects in Type I collagen genes are clearly responsible for the OI Type I phenotype. In these cases however, the triple helix is as thermally stable as normal collagen, or only slightly unstable, but the total amount of collagen secreted is lower than normal (Rowe et al., 1985; Bateman et al., 1989; Byers et al., 1988a). In some cases, it has been suggested that point mutations which cause substitutions in the X or Y positions of the repeat rather than in the glycine position, as is seen in Type II cases, are responsible for greater thermal stability and perhaps the milder phenotype (Bateman et al., 1988). More recently it has been shown that one of the Type I collagen gene alleles is silent, or non-functional, resulting in production of reduced collagen amounts (Willing et. al, 1992).

The more severe dominantly inherited form of OI, or Type IV also involves a variety of Type I collagen defects. It is characterized by moderate clinical severity, variable deformity of long bones and, in contrast to OI Type I, the sclerae of these individuals are pale blue or white (Sillence, 1988). RFLP analysis has shown that in many cases this form of the disease segregates with the COL1A2 gene (Superti-Furga et al., 1989; Sykes et al., 1990). Glycine substitutions in the triple helical domain of the pro-a2(I) chain are, in fact,

associated with many Type IV cases (Wenstrup et al., 1986; Wenstrup et al., 1988). These substitutions are often located in an analogous position in the helix to those α1(I) chain substitutions in Type II OI cases, yet they result in the non-lethal Type IV condition (Cohn and byers, 1991; Superti-Furga et al., 1989).

The progressively deforming, moderately severe form, or OI Type III is one of the most common types of OI and the least well-characterized with respect to molecular defects. Its progressive nature is due to repeated multiple fractures leading to severe deformity of the limbs and spine (Sillence, 1988). Most cases are due to spontaneously-occurring dominant mutations while a few are autosomal recessive, when consanguinous parents are involved. (Pihlajaniemi et al., 1984; Bonadio et al., 1990a). There have been some reports of substitutions and deletions in the Type I collagen genes of OI Type III patients (Bateman et al., 1988; Byers et al., 1988a; Pack et al., 1989; Byers 1983; Pihlajaniemi et al., 1984), but there are many cases where no such defects have been found (Byers et al, 1987; Byers et al., 1988a; Steinman et al., 1988; Beighton et al., 1988; Wenstrup et al., 1990). The variability of the clinical symptoms within this form of OI suggests further genetic heterogeneity within the group. RFLP analysis of several of the recessive pedigrees for Type III OI have shown that this form of OI is often unlinked to <u>either</u> of the Type I collagen genes (Byers et al., 1987;

Aitchison et al., 1988).

It has been suggested that for those cases where mutations in the Type I collagen genes have been excluded, defects in other ECM genes could be responsible for expression of the OI phenotype. Other genes encoding bone-specific proteins which may interact with Type I collagen, are likely candidates (Byers, 1990).

OI Heterogeneity: A Potentially Polygenic Disease

Although large numbers of OI cases have been studied at the molecular level there is still very little understanding of how the specific defects in Type I collagen contribute to the OI phenotype. It has not been possible to correlate specific kinds of molecular defects with particular OI phenotypes. The different forms of the disease cannot always be distinguished by which of the Type I collagen genes contains the defect nor can they always be associated with the location of the defect within certain domains of the collagen propeptide. There are a number of cases in which similar defects, such as glycine substitutions, map very close to each other in the same propeptide chain yet are associated with drastically different phenotypes (Tenni et al., 1988; Constantinou et al., 1990). Glycine substitutions outside the triple-helical domain have been identified in both mild and severe cases (Byers, 1990). There is an example of 3 cases where small insertions or deletions, all located in the C-

propeptide caused frameshift mutations which produce three phenotypes ranging from mild to severe (Willing et al., 1990; Bateman et al., 1989; Pihlajaniemi et al., 1984). There are also many cases where multiple members of a family carry the same Type I collagen defect and show a wide variation of severity, even being asymptomatic (Superti-Furga et al., 1989; Sykes et al., 1990).

It is possible that factors such as local chain structure, domain functions and molecular interactions in the final collagen polypeptide could modify the effect of specific substitutions (Byers, 1990; Cohn et al., 1988). However, as is clearly indicated in the above situations, the variable severity of symptoms cannot be wholly attributed to a lack of understanding of the functional domains of the Type I collagen molecule. Defects in other ECM components, including other collagen types, which interact with the various domains of Type I collagen could clearly affect the OI phenotype.

Other Collagen Types

There are at least 15 other collagen types in addition to Type I collagen. They are composed of multiple distinct polypeptide chains, and thus they are coded for by more than 25 genes (Miller and Gay, 1987). In addition to the structural role the "classical" or "fibrillar" collagens are known to play in providing a scaffolding for underlying tissues, recent discoveries have indicated that other collagen types have

additional properties dictated by the multiple domains in their polypeptides. However, very little is known about how their associations/interactions contribute to the structural integrity of the extracellular matrix. Collagens are defined by a triplet repeat, the presence of the modified amino acids hydroxyproline and hydroxylysine, and by the formation of a triple helical structure (Byers, 1989; Sandell and Boyd, 1990). They can be classified into distinct groups.

The first group consist of the classical or fibrillar collagens (Types I, II, III, V and XI). They all contain three polypeptide chains, with a monomer size greater than 95,000 daltons and long, uninterrupted triple-helical domains, approximately 300 nm in length, containing the characteristic Gly-X-Y amino acid repeat (Kuhn, 1987). They are all expressed in a tissue specific manner and all participate in the formation of the characteristic fibrils known to contribute to the structural framework and strength of connective tissue. Their genes share a highly-conserved exon-intron organization (Chu et al., 1984). The majority of the exons in the triple helical coding regions are 54 bp, or multiples of 54 base pairs and all of them contain integral numbers of the basic Gly-X-Y repeat, beginning with the codon for glycine (Sandell and Boyd, 1990). The highly interrupted coding sequences contain more than 50 exons resulting in genes between 18 and 40 kilobases (de Wet et al., 1987; Chu et al., 1985a; Tsipouras and Ramirez, 1987; Myers et al., 1981). The coding

sequences are divided into a nearly identical number of small exons, and most introns regardless of size, are located in homologous positions in each gene (Ramirez et al., 1985; Bernard et al., 1988). In addition, all of the fibrillar collagens arise from procollagen precursors and undergo proteolytic removal of terminal extensions as was described for Type I collagen above (for a complete review see Kuhn, 1987).

All collagens other than the fibrillar collagen types described above can be referred to as non-fibrillar collagens. These collagens do not form long fibrils and all contain multiple interruptions in the Gly-X-Y repeat. The genes coding for these collagens contain few 54 bp exons (Sandell and Boyd, 1990). They can be divided into subgroups based on their specific structural features and possible functions; however, their placement in one group is not necessarily exclusive.

The large nonfibrillar, minor collagen types IV, VI and VII again have monomer chain sizes that are greater than 95,000 daltons, but they all have at least one interruption in the Gly-X-Y repeat within the triple helical domain (Sandell and Boyd, 1990) and large terminal globular domains. The interruptions allow the molecules flexibility which is not possible for the fibrillar collagens (Hofmann et al., 1984). Although this group in general is not well characterized, analysis of the genes for Type IV, a basement membrane collagen, has shown that the exon sizes and overall intron-

exon patterns are very different from the 54 bp exon pattern of the fibrillar collagens (Soininen et al., 1988; Soininen et al., 1989). In addition, these collagens tend to aggregate through end-to-end association of monomer (Timpl et al., 1981) or dimer units rather than by the lateral, quarter staggered association necessary for fibril formation.

Type VI collagen is a highly disulfide bonded, microfibrillar component present in human blood vessels and in the myocardium (Chung et al., 1976; Bashey et al., 1992) as well as many other "soft" connective tissues (Karkavelas et al., 1988). This molecule appears to be distributed in tissues where continuous contraction and relaxation, such as that seen in the heart, would warrant a very flexible ECM component that would allow the distribution of this sort of stress (Bashey et al., 1992). Type VII collagen, the major protein of anchoring fibrils in epidermal basement membrane, also exhibits properties that could contribute to tissue flexibility (Ryynanen et al., 1991).

The "short-chain" collagen Types VIII and X have monomer chains which are half the length of the fibrillar collagens (Apte et al, 1992) and have a distinct gene organization. These collagens share common structural features in that both contain 8 interruptions of the helix of similar sequence, length and relative locations (Ninomiya et al., 1990; Yamaguchi et al., 1989). The small 5 kb gene for chick α1(X) collagen contains only 3 very large exons and no introns

separating its triple helical coding sequences (Sandell and Boyd, 1990). The 6kb human α1(X) gene contains only 2 exons, one very large and one very small (Thomas et al., 1991). In all of these genes one large exon encodes the entire triple helical coding region (Yamaguchi et al., 1991; Thomas et al., 1991). The genes of the short chain collagens share almost complete sequence identity in some regions, such as the 3′ half of the non-collagenous (NC1) domain (Muragaki et al., 1990). In vitro studies show that Type X collagen forms a hexagonal lattice similar to that formed by Type VIII collagen in Descemet's membrane (Apte et al., 1992). Although the exact function of these collagen types is unknown, Type X collagen, expressed in hypertrophic chondrocytes, is somehow necessary for normal bone development (Thomas et al., 1991).

"FACITs" Collagens

The final subgroup, the fibril-associated collagens with interrupted triple helices or "FACITs" collagens (Olsen et al., 1988), include Types IX,, XII, XIV and XVI (Gordon et al., 1989; Gordon et al., 1991; Pan et al., 1992). Many of these collagens have properties which suggest their ability to interact with other collagens and ECM components. These interactive properties, along with complex regulatory mechanisms, tissue specificity and developmentally regulated gene expression make them particularly interesting candidates for factors which mediate/diversify the function of the major

fibrils in different tissues. It is therefore possible that mutations in genes whose products have properties like these could contribute to the OI phenotype.

All of these collagens, as well as Types VIII and X, have monomer chain sizes less than 95,000 daltons, with the exception of Type XII (Pihlajaniemi and Tamminen, 1987), and have many interruptions in the Gly-X-Y repeat. They are made up of multiple collagenous domains (COL) separated by distinct non-collagenous (NC) regions. Each of these regions is designated by a number indicating its location in the complete polypeptide. The "FACITs" collagens do not undergo proteolytic processing and are thus not secreted as procollagens (Shaw and Olsen, 1991). Although the FACITs all share these common structural features, their sizes and primary structures vary greatly (Pan et al., 1992). This collagen group shows a wide variation in gene structure and overall intron-exon organization. They contain few 54 base pair exons and the exons often do not contain integral numbers of the Gly-X-Y repeat. All of these collagens occur in conjunction with other fibrillar collagen types, and are somehow "associated" with the surface of the fibril.

Type IX collagen is considered a proteoglycan with collagenous properties (Noro et al., 1983). Found in cartilage and corneal stroma, it contains three collagenous and four non-collagenous domains. Three genes code for the heterotrimer (Ninomiya, 1985). The gene for chicken α1(IX) collagen is 100

kb, and contains short exons and multiple large interruptions in its triple helical coding sequences. In contrast, the α2(IX) gene, about 10 kb in length, contains more exons, smaller introns, and therefore a much higher exon density (Ninomiya et al., 1990). The α2 chain also contains the covalently bound glycosaminoglycan (GAG) side chain. Despite these differences, the size and location of the exons coding for the multiple triple-helical regions, as well as most of the non-collagenous regions, are structurally homologous. The recent isolation of cDNA clones of the chick α3(IX) chain indicate an overall structure similar to that of homologous regions of α1(IX) and α2(IX) (Brewton et al., 1992).

Type IX collagen is found to be crosslinked to Type II collagen molecules in cartilage (Vaughan et al., 1988) and is arranged in a periodic manner along the surface of the fibril projecting its third triple-helical, COL3, and NC4 globular domain into the perifibrillar matrix (Olsen et al., 1988; Gordon et al., 1991). The large globular amino-terminal NC4 domain of the cartilage α1(IX) molecule is significantly shorter and very different in structure in embryonic cornea. Tissue-specific promotor usage and alternative splicing is likely responsible for this variation (Gordon et al., 1989). The tissue-specific variation seen in the perifibrillar extension of the Type IX polypeptide affords the major Type II collagen molecule the ability to interact in a tissue specific manner with other collagen types and ECM components.

Type IX collagen also contains an attachment site in the α2 chain for a chondroitin sulfate side chain in its NC3 domain, suggesting its role as a proteoglycan as well as a collagen. This GAG side chain is also in contact with the major Type II fibrils of cartilage, thus perhaps providing an additional way to facilitate interaction between the collagen molecules and other ECM components (Vaughan et al., 1988).

Type XII collagen was originally isolated from embryonic chick tendons (Dublet and van der Rest, 1987; Gordon et al., 1987) a tissue whose major fibril is Type I collagen. It has two short triple-helical domains and three non-collagenous domains (Gordon et al., 1990). Type IX and Type XII are about 50% identical at the protein and gene levels and have multiple conserved cysteine residues in the NC1, COL1 and NC2 domains (Gordon et al., 1987). In contrast to Type IX, Type XII is a homotrimer of three identical α-chains and does not contain a glycosaminoglycan side chain. However, it does contain an amino terminal globular domain, NC3, which is similar to, but much larger than NC4 of α1(IX) (Dublet et al., 1989). The intron/exon structure of portions of α1(XII), specifically COL1, NC2 and COL2, is homologous to that of the α1(IX) and α2(IX) chains (Gordon et al., 1989; Gordon et al., 1990). In addition to these Type IX-like regions (which comprise only 7.7% of the molecule), recent data reveal that the complete Type XII collagen is a chimeric molecule with multiple functional domains including reiterated fibronectin type III

motifs, von Willebrand factor A motifs, and Arg-Gly-Asp cell adhesion sites (Yamagata et al., 1991).

Type XIV collagen fragments were recently isolated from fetal bovine skin and tendon (Dublet & van der Rest, 1991) and embryonic chick skin. This homotrimeric collagen also appears to be another FACITs since the structure of its collagenous domains is similar to that of Type XII, and to a lesser extent Type IX collagen (Gordon et al., 1991; Aubert-Foucher et al., 1992). Electron microscopic studies suggest that Types XII and XIV collagen could be similar in size and shape as well (Dublet and van der Rest, 1990). In Type XIV as in Type XII, the triple helical domains comprise less than 15% of the intact molecule (Aubert-Foucher et al., 1992).

The recently described human Type XVI collagen gene again resembles the structure of the other fibril-associated collagen Types IX, XII and XIV (Pan et al., 1992). Many of its 10 collagenous domains contain interruptions of the Gly-X-Y pattern and these domains are separated by 11 short non-collagenous stretches. Again, the COL1 and the amino terminal NC domain is similar to that of other FACITs collagens. In addition, its cysteine-rich non-collagenous domains resemble those found in the <u>C. elegans</u> cuticle collagen. This similarity, and the expression of Type XVI in epidermal keratinocytes as well as dermal fibroblasts, have implied its possible function in the epidermis. The predicted amino acid sequence also contains three potential N-glycosylation sites

(Asn-X-Thr/Ser) and three Arg-Gly-Asp sites (signal for cell adhesion). It too is thus a multi-domain molecule with both collagenous and non-collagenous properties.

Most mammalian tissues express more than one type of collagen although one is usually predominant. Type V collagen is frequently associated with both Types I and III (Fessler and Fessler, 1987) and immunohistochemical studies have shown close association, in vivo, of Type V fibrils with Type I fibrils and with basal laminae (Martinez-Hernandez et al., 1982; Fitch et al., 1984). It appears that Type V fibrils are somehow involved in regulating the diameter of Type I fibrils (Birk et al., 1988). A similar function has also been suggested for Type XI which is associated with Type II fibrils in cartilagenous tissue (Eyre and Wu, 1987).

The non-fibrillar, FACITs collagen Types IX and XII are are also structurally, genetically and perhaps functionally related (Dublet and van der Rest, 1987; Gordon et al., 1987). It has been suggested, due to their structural similarities, that Type XII may align similarly along the surface of Type I and V fibrils in bone as Type IX does with Type II and XI fibrils in cartilage (Ninomiya et al., 1990). Immunofluorescence studies have indeed found Type XII collagen only in tissues that contain Type I collagen, but not bone (Sugrue et al., 1989). Two Type XII-like collagens, likely to be Type XIV collagen, have been localized to the surface of banded collagen fibrils of human and calf tendon and skin (Keene et

al., 1991). It is possible that Type XIV in bone, and Type XII in non-bone tissues serve similar functions in different Type I collagen-containing tissues.

Once described as a "short-chain" collagen, low molecular weight collagen Type XIII has features of the non-fibrillar Types IV and IX (Pihlajaniemi et al., 1987). Most of the exons of the α1(XIII) gene are not multiples of 54 bp and it encodes three collagenous and four non-collagenous domains (Pihlajaniemi and Tamminen, 1990). The exons all begin with a complete codon for an amino acid, a feature found in the genes for the fibrillar collagens (Tikka et al., 1988). Therefore, this collagen has both fibrillar and non-fibrillar collagen features. This was the first collagen described to undergo complex alternative splicing. As many as 5 alternative splice sites have been identified and at least 2 species of mRNA transcripts are produced by the use of alternative transcription initiation sites as well as alternative splicing (Pihlajaniemi and Tamminen, 1990). This gene also contains one exon which encodes a unique Gly-X-Y repeat extending into a domain that is not triple helical. (Tikka et al., 1988). Therefore, Type XIII collagen has a structure which is different from all other collagen types.

As more unique features of the non-fibrillar collagen genes are discovered, the diverse domains of these collagen-like ECM components imply functions beyond the traditional structural role attributed to fibrillar collagens. There are

likely to be important interactions of these molecules with the fibrillar collagens and with other ECM components in determining the connective tissue type, its basic properties, structural integrity, and capacity to withstand stress.

Additional collagen-like genes continue to be discovered. Recent studies have shown that a pro-α2(I) cDNA probe detects two previously-undescribed loci in a collagen-like gene(s) other than the already-characterized fibrillar collagen genes (Doering et al., 1990). Deletions of heterogenous size have been detected at one or the other of these sites in genomic DNA of Type III OI patients. The deletions are significantly more prevalent in these patients than in the normal population and frequently occur spontaneously in patients who have no clear family history of the disease (Doering et al., 1990). Such correlations suggest that these collagen-like deletion sites could be contributing to the OI phenotype.

One site, detected as a 4.6 kb Hind III restriction fragment in normal individuals, has been localized to chromosome 17 (Cammarata et. al., 1991). The cloning and sequencing of this fragment will allow identification of the gene product of this locus, characterization of the deletion site in OI patients, and thus ultimately an understanding of how mutations at this locus might play a role in the OI phenotype.

MATERIALS AND METHODS

Overall Approach

In order to characterize the new collagen locus, a Charon 21A phage library containing Hind III restriction fragments of human chromosome 17 was screened for a 4.6 kb collagen sequence-containing fragment. Positive phage isolates were plaque purified and the insert-containing phage DNA isolated. The 4.6 kb insert was recloned into a plasmid vector so that a preliminary restriction map could be constructed. This allowed the subcloning of smaller fragments, containing triple helical coding sequences, to be sequenced. Sequence comparisons using Genetics Computer Group software (Devereux et al., 1984) were performed to compare these sequences to other known collagen genes and collagen-like sequences in order to identify the collagen group that the new locus represents.

Library Screening

The previously described pro-α2(I) cDNA clone Hf32 (Myers et al., 1981) was used to screen a human chromosome 17 library (LL17NS02/ ATCC#57759) obtained from American Type Culture Collection, Rockville, MD. This phage library was constructed

in the Charon 21A vector with Hind III digested restriction fragments of human chromosome 17. The E. coli bacterial strain JM109 was infected with the recombinant phage particles (Benton and Davis, 1977) and incubated at 37°C overnight on NZCYM agar plates (Sambrook et al., 1989). Plaque lifts to nylon membranes (NEN, Bethesda, MD) and lysis of the membrane-bound phage were done according to the manufacturer's instructions. A total of 10^6 phage clones were screened. Plaque purification was carried out using previously described methods (Sambrook et al., 1989).

Nick Translation and Hybridizations

The cDNA clone Hf32, digested with EcoRI and PstI, was ^{32}P-labelled by nick translation (Rigby et al., 1977) to an average specific activity of 4.5 x 10^8 cpm/μg. Prehybridization was carried out for a minimum of 5 hours at 37°C (Doering et al., 1982; Rosenthal and Doering 1983) in hybridization solution (50% formamide, 0.9 M NaCl, 50mM Tris pH7.5, 1% SDS, 10ug/ml denatured E. coli DNA). Denatured nick translated probe was added to the prehybridized membranes and incubation continued at 37°C for a minimum of 16 hours. Membranes were washed twice for 10 minutes at room temperature in 2X SSC (1X SSC is 150 mM NaCl, 15mM sodium citrate, 0.1 mM EDTA), twice for 30 minutes at 50°C in 2X SSC, 1% SDS and twice for 30 minutes at room temperature in 0.5X SSC.

Autoradiography was done using Kodak XAR X-ray film and Dupont Lightning-Plus intensifying screens when necessary.

Rehybridization of any additional probes to previously hybridized membranes first required removal of the original probe by very high stringency washes. The membrane was first washed with a 0.4M NaOH solution, then with 0.1XSSC, 0.2M Tris-HCl, 0.1% SDS, in a shaking water bath at 42° C for 30 minutes. If the probe was not sufficiently removed this procedure was repeated and/or a second method was carried out. The additional method required boiling the membrane in a solution of 0.1xSSC, 1% SDS for 30 minutes. This procedure was also repeated as necessary to completely remove the probe.

DNA Isolation

The purified recombinant phage DNA was isolated by the standard plate lysate method (Maniatis et al., 1982) with the following modifications: Harvested phage particles were treated with RNase A and DNase I at a final concentration of 10 μg/ml and 5 μg/ml respectively. The PEG precipitated phage particles were incubated overnight on ice. After addition of SDS/EDTA, lysis was carried out for 1 hour at 68°C. Isopropanol precipitation was carried out overnight at -20°C.

Plasmid purification was carried out using Qiagen Tip-100 anion exchange columns (Chatsworth, CA) according to the manufacturer's instructions.

Restriction Mapping; Alu Repetitive Sequence and Exon Localization

Restriction mapping was carried out by established methods using single, double and triple digests (Doering, 1977; Peterson et al., 1980; Rosenthal et al., 1984). All restriction enzymes were purchased from Bethesda Research Laboratories (Bethesda, MD) or Promega Biotech (Madison, WI) and used in the suppliers' recommended buffers. Restriction digests were stopped by addition of SDS and EDTA to final concentrations of 0.5% and 10mM respectively. Gel electrophoresis was carried out using 1% or 1.2% agarose gels (Rosenthal and Doering, 1983) and the DNA was transferred to NEN (Boston, MA) Gene Screen Plus membranes using the alkaline transfer method (Reed and Mann, 1985). DNA size markers were lambda phage DNA digested with HindIII and phi-X 174 phage DNA digested with HaeIII. Triple helical coding regions were identified using the Hf32 probe and hybridization conditions described above. Localization of Alu repetitive sequences was done using the Alu probe Blur 8 (Deininger et al., 1981). Hybridization conditions were the same as those described for the Hf32 probe above except that the hybridization solution contained 1.0M NaCl instead of 0.9M as for Hf32.

Subcloning and DNA Sequencing

The 4.6 kb phage DNA insert and the 1.6 kb EcoRI/PstI and

1.5 kb EcoRI/SmaI restriction fragments containing triple helical coding sequences were subcloned by standard methods (Sambrook et al., 1989) into the pUC13 (Vieira and Messing, 1982) and pGEM7Zf+ (Promega Biotech, Madison, WI) vectors. These subclones, p4.6EP and p4.6ME respectively, were sequenced by the dideoxy method (Sanger et al., 1977). The double-stranded template and primer were denatured and reannealed as follows prior to extension/termination reactions. Three micrograms of template DNA and 0.5 pmol primer DNA were combined in a total volume of 20.0 μl. NaOH and EDTA were added to a final concentration of 0.2M and 0.2mM respectively and the mixture was incubated for 5-6 minutes at 85°C. The tube was immediately placed on ice, and 6.0 μl of 3M NaOAC, pH 5.2 and 60.0 μl 100% EtOH were added to precipitate the denatured DNA. This solution was placed at -20°C for 20 minutes. The DNA was pelleted by centrifugation at 10,000 rpm in an SHMT rotor (Sorvall/E.I. duPont, Newtown, CT) for 20 minutes and washed carefully with 70% EtOH. The dried DNA pellet was resuspended in distilled H_2O and reannealed (incubated) for 15 minutes at 37°C (p4.6EP) or 5 minutes at 37°C followed by 5 minutes at room temperature (p4.6ME).

The extension and termination reactions were carried out according to the manufacturer's instructions for sequencing double stranded DNA using the "Sequenase" kit (United States Biochemical Corp., Cleveland, OH) and/or the "AmpliTaq" sequencing kit (Perkin-Elmer Cetus, Norwalk, CT). The DNA was

tagged by including ^{35}S-labelled dATP nucleotides during the extension reaction. When obtaining sequence information from inserts in the pUC13 vectors the -40 primer adjacent to the EcoRI site was used (see Figure 7c). Sequence extension began from the EcoRI end of the insert and continued to the left toward the helical coding region. When obtaining information from the pGEM7Zf^{+} clones, the T7 primer was used (Figure 7d). Again, sequence extension was from the Eco RI site, or in the case of the nested deletion clones (Figure 7e), sequences adjacent to this site and proceeding to the left through the helical coding region. The nucleotide analogue deaza-dGTP (Perkin-Elmer Cetus, Norwalk, CT) was also used in the extension reactions, according to the manufacturer's instructions, in conjunction with 40% formamide gels, for confirming sequence in regions with complex secondary structure.

Generation of Nested Deletions

The subclone p4.6ME was used as a template to generate nested deletions by exonuclease III digestion from the Eco RI site. This procedure was carried out by using the "Erase-a-base" system (Promega Corporation, Madison, WI) according to the manufacturer's instructions. Three overlapping clones, p4.6ME-5, p4.6ME-7 and p4.6ME-9 with deletions of appropriate sizes (Figure 7e) were chosen for further sequencing.

Sequencing Gels

Sequencing gels containing either 8% or 6% acrylamide/bis (37.5:1) and 7M urea were used to obtain most of the sequencing data. In order to sequence regions of secondary structure formamide was added to a final concentration of 40% when necessary. After electrophoresis, the sequencing gel was soaked for 15 minutes in 10% acetic acid/10% methanol solution or 10% acetic acid/20% methanol for formamide gels. It was then transferred to Whatman 3MM filter paper by established methods (Kozak et al., 1991), dried in a gel dryer (Bio-Rad Laboratories, Hercules, CA) for a minimum of 2 hours and autoradiographed. The entire sequence was read a minimum of 3 times from independent gel runs in order to confirm its accuracy. Sequence analysis was done with the aid of the VAX network system (Loyola University Chicago) and the Genetics Computer Group sequence analysis program, version 7.1 (Devereux et al., 1984).

RESULTS

Isolation and Confirmation of Cloned Sequences

In order to characterize the new collagen-like Site 2 locus, a Charon 21A lambda phage library containing HindIII restriction fragments from human chromosome 17 was screened for the 4.6 kb collagen sequence-containing fragment. At least 10^6 plaques were screened with the ^{32}P-labelled α2(I) cDNA, Hf32, and 107 positives were identified. The high number of positives was not suprising since a full chromosome's worth of fragments is contained in 1.4 x 10^4 clones. Twelve positives were picked for plaque purification. Purification required plating the phage particles at the appropriate density (approximately 100-200 plaques per plate) so that individual positive plaques were well separated and could be easily identified and isolated. This assures that each positive isolate contains a homogeneous population of phage particles representing a single clone. Five independent isolates were identified and DNA was prepared from them. When digested with HindIII, three of the clones were found to contain a 4.6 kb insert, that hybridized strongly with the α2(I) cDNA probe (Figure 1a). The other two clones contained a 4.4 kb and a 6.0 kb insert respectively, that also

hybridized with the Hf32 probe (Figure 1a).

The blot of the HindIII digested phage isolates was stripped and rehybridized with the α1(I) cDNA probe, Hf677, under higher stringency conditions, (1XSSC/65°C). There was no hybridization of the 4.6 kb cloned inserts but there was hybridization to both the 4.4 and 6.0 kb inserts (Figure 1b). These results confirm that the 4.6 kb fragment is not a member of the already characterized COL1A1 gene which is also located on chromosome 17. Use of high stringency conditions ensures this because a genuine COL1A1 sequence would cross-hybridize even under these high stringency conditions. Positive hybridization of the α1(I) probe, even at high stringency, to the cloned 6.0 kb fragment confirm that this is the same 6.0 kb fragment known to be a member of the COL1A1 gene (Barsh et al., 1984). Although the 4.4 kb insert is also detected by Hf677, there is no fragment of that size in the COL1A1 gene. Therefore, the nature of this clone is undetermined. No further study has been done on this isolate.

The above hybridization results confirm that the 4.6 kb cloned fragment represents a distinct locus and is not part of the COL1A1 gene. The sequences within the 4.6 kb fragment thus represent a new human collagen-like locus on chromosome 17 which has not previously been described.

DNA from an isolate containing a 4.6 kb insert was chosen for further study. It was first used to confirm that this was indeed a clone of the 4.6 kb HindIII fragment characteristic

of Site 2. Hybridization of this phage isolate to HindIII or PstI digested genomic DNA of a normal and an OI III affected individual revealed a smear of hybridization in all lanes under low stringency conditions (Figure 2a). This hybridization smear indicated that the cloned DNA contained repetitive sequences dispersed throughout the genome, as well as the originally identified collagen-like sequences.

When the blot was rewashed at higher stringency (0.1XSSC, 65°C) the smears were greatly diminished and distinct bands could be visualized. In the normal individual the 4.6 kb HindIII and 5.2 kb PstI fragments characteristic of Site 2 are detected by the probe (Figure 2b, lanes 2 & 4). The detection of a doublet in these lanes suggests that this individual is heterozygous for a 300 bp deletion at this locus. These were the only fragments detected in the genomic DNA of the normal individual. In the DNA of the OI affected individual the probe detects only a 4.4 kb fragment in the HindIII digest or a 5.0 kb fragment in the PstI digest. This individual is known to be homozygous for a 200 bp deletion at this locus. The hybridization of the phage isolate probe to only these fragments representing the collagen-like locus confirmed that the 4.6 kb fragment characteristic of this locus had been successfully cloned.

To determine the identity of the repetitive sequences contained within this clone the Southern blot containing HindIII digested phage isolate DNA was hybridized to the Alu

sequence probe Blur 8 (Figure 1c). The positive hybridization indicates that at least some of the repetitive sequences in the 4.6 kb clone are members of the Alu dispersed family.

Restriction Mapping of the HindIII Fragment (p4.6)

In order to facilitate the mapping and characterization of the phage isolate, the 4.6 kb insert was excised from the HindIII site of Charon 21A and recloned at the HindIII site of the pUC13 plasmid vector, forming p4.6. Restriction mapping was carried out by established methods (Doering, 1977; Peterson et al., 1980; Rosenthal et al., 1984). A complete map of the 4.6 kb cloned fragment is shown in Figure 3. Southern blots of restriction digests, hybridized with the Hf32 cDNA probe, allowed localization of the triple helical coding region to an 800 bp region extending from EcoRI to PvuII within a 1.6 kb EcoRI to PstI restriction fragment (Figure 4). Clearly the 1.6 kb EcoRI/PstI fragment hybridized with the probe (Figure 4a, lane 4) indicating that this region contains helical coding sequences. In p4.6 neither the HindIII/PstI fragment to the left of the helical coding region (see Figure 3), nor the HindIII/EcoRI fragment to the right (data not shown) of the 1.6 kb EcoRI/PstI fragment, hybridized with the Hf32 probe. Upon double digestion of p4.6 by the enzymes AvaII and HindIII a doublet is produced at the 1.4-1.5 kb position. The upper fragment of the doublet is a HindIII/AvaII insert fragment shown to the left of the triple helical coding region

in Figure 3. When the Southern blot of this digest was hybridized with Hf32, only the lower, vector fragment of the doublet was detected (Figure 4b, lane 5). There is also a PstI/PvuII subfragment contained within the HindIII/AvaII fragment which does not hybridize (Figure 4b, lane 7). Thus, lack of hybridization to the HindIII/AvaII and PstI/PvuII fragments allowed localization of the helical-coding region solely to the EcoRI/PvuII portion of the EcoRI/PstI fragment. The α1(I) probe Hf677 was also rehybridized to the same Southern blot in order to determine if any other regions contained triple helical coding sequences. No additional regions were identified with this probe (data not shown).

The presence of collagen-like sequences in the 1.6 kb EcoRI/PstI restriction fragment was reconfirmed by its susceptibility to digestion with Sau96I. This approach has previously been used in characterizing other collagenous protein coding regions (Vasios et al., 1987). The recognition sequence of this restriction enzyme is GGNCC. This is the same sequence that codes for the amino acids Glycine (GGN) and Proline (CCA or CCT). Thus, because glycine is present as every third residue and proline is often in the X position of the collagen triplet, triple helical coding regions are extremely suseptible to digestion with this enzyme. The clone, p4.6, was triple digested with Sau96I, EcoRI and PstI so that the effect of Sau 96I on the 1.6 kb EcoRI/PstI fragment could be examined. No sequences of the EcoRI/PstI fragment greater

than 200 bp remained intact (Figure 5, Lane 3), indicating that multiple Sau96I sites were present in this restriction fragment. When p4.6 is digested with EcoRI and PstI only, 4800 bp (see Figure 4a, lane 4) and 750 bp fragments are generated in addition to the 1600 bp fragment. After digestion with Sau96I the 4800 bp fragment remains only as a 2000 bp fragment containing both insert and vector sequences to the right of the EcoRI/PstI fragment (see Figure 3). The pUC13 vector contains multiple Sau96I sites which, upon digestion produce fragments much smaller than 2 kb. Therefore, this fragment must span the vector-insert junction. The distance from EcoRI to HindIII on the right however, is larger than 2000 bp. Thus, there must be additional Sau96I sites close to the EcoRI site in the right-hand EcoRI/HindIII fragment of the clone.

Hybridization of a Southern blot containing various restriction digests of p4.6 to the Alu sequence probe Blur 8 allowed the identification of a minimum of three restriction fragments containing Alu repetitive sequences (Figure 6). These sequences are present in the 1.6 kb EcoRI/PstI helical-coding fragment, and on both sides of this fragment. The flanking fragments were: the 800 bp SstI/HindIII fragment and a more weakly hybridizing 800 bp HindIII/Pst I fragment. The 850 bp PstI/PvuII portion of the 1.6 kb helical-coding region contained all the Alu sequences within this fragment (data not shown). These repetitive sequence regions are indicated by an asterisk in Figure 3.

Subcloning and Sequencing

The helical-coding region of p4.6 remains the focus of the subsequent data in this thesis because its collagen-like nature is of greatest interest. The correlation between deletions in this region and Type III OI (Michaels, unpublished) make it of particular interest. In order to facilitate the fine structure mapping and sequencing of the helical-coding region, the EcoRI/PstI fragment was subcloned into the pUC13 plasmid vector. A schematic diagram illustrating the construction of this subclone (p4.6EP), and subsequent constructs used in sequencing the helical region, is presented in Figure 7. This arrangement simplified restriction digest patterns and placed the EcoRI/PstI fragment directly adjacent to vector sequences for which sequencing primers were available (Figure 7c). Southern blot hybridizations of Hf32 to restriction digests of this subclone confirmed the presence of at least two regions containing one or more AvaII sites (recognition sequence: GGACC). The first site, estimated at a position 150-200 nucleotides to the left of EcoRI, marks the point at which the triple helical coding region begins. At a position 680 bp from EcoRI, the most distal PvuII site defines the approximate endpoint of the triple helical coding region (see Figure 3).

The next goal of this project was to sequence this triple helical coding region. Although it was clear that the region did not begin until the first AvaII site, the presence of

multiple AvaII sites very close to each other, and the lack of other convenient restriction sites, made subcloning only the immediate coding region impossible. Therefore dideoxy sequencing using the "Sequenase" kit (U.S. Biochemical Corp.) was carried out beginning at the "-40 universal primer" site adjacent to the EcoRI cloning site (Figure 7c).

At position 230 (230 nucleotides to the left of EcoRI of Figure 3), the first AvaII site was reached. This marked the beginning of the triple helical coding region. Approximately 300 nucleotides of sequence was determined using the Sequenase kit (Figure 8). At that point the ability to resolve sequence using the -40 primer was lost. In addition, the limitations of the Sequenase method prevented the determination of sequence in areas of complex secondary structure near the AvaII sites (see Figure 9a). This sequence not only has a repetitive nature but is also very GC rich. DNA polymerase I, the enzyme employed in the Sequenase kit, is sensitive to the higher temperatures required to sequence a GC rich double stranded region. As a result, reliable sequence extension in these GC rich regions was difficult to obtain in regions greater than 200 nucleotides from the primer. Figure 9a shows a typical sequence "ladder" which results from situations such as this. Use of internal primers based on new sequence information usually allows further sequence determination, and sequence extension closer to these internal primers reduces premature termination due to secondary structure. However, in

this case the identification and use of internal sequencing primers was not possible due to the repetitive nature of the sequence beginning at position 128 (Figure 10).

Due to the difficulties using the "Sequenase" method, all sequence obtained from position 160 to 300 (position numbering starts at EcoRI and moves to the left in Figure 3) was reconfirmed using the "Ampli-Taq" sequencing kit (Perkin-Elmer Cetus). This kit employs Taq polymerase, a very stable enzyme at the higher temperatures (45-72°C) needed to reduce secondary structure in GC rich areas. Although ambiguous sequence was confirmed and some additional sequence was obtained, resolution was lost using this method at position 400.

The majority of the EcoRI/PstI fragment was then isolated and subcloned as a 1.5 kb EcoRI/SmaI fragment (see Figures 3 and 7) into the pGEM7Zf+ vector (p4.6ME), so that nested deletion subclones could be generated for further sequencing.

Creation of Nested Deletion Subclones

The "Erase-A-Base" system (Promega Biotec) employs exonuclease III for the serial removal of nucleotides from the EcoRI digested end of p4.6ME. The extent of digestion is time dependent and allows the generation of a series of clones in which the 3′ sequences are brought closer to the primer. This method allowed the generation of a series of clones containing overlapping sequences covering the remainder of the triple helical coding region (see Figure 7e).

Results Obtained with Sequencing Data Generated

The nucleotide sequence and complete map. The deletion clones p4.6ME-5, p4.6ME-7 and p4.6ME-9 (Figure 7e) were used the generate the remaining 435 nucleotides of sequence which extended 50 nucleotides beyond the most distal PvuII site (see Figure 3). This covered most, if not all, of the helical-coding region homologous to Hf32. The entire 735 nucleotides of sequence generated by the above methods will be subsequently referred to as "NT735". The complete sequence (Figure 8) of NT735 confirms the presence of all restriction sites identified previously (Figure 3). Figure 9b includes a representative example of the acrylamide gels from which the sequence information was obtained. In contrast, Figure 9a shows the sequence "ladder" caused by premature termination of sequence extension due to secondary structure in GC rich regions.

In addition to the site already identified by restriction mapping, two additional PvuII sites are present. That site most distal from EcoRI (farthest left), already determined by restriction digests, is at position 684, approximately 275 nucleotides from the first site (closest to EcoRI). Another site is located equidistant between these two sites at position 408 (see Figure 3 and 8). The presence of this third site at position 408 explains why no 275 bp PvuII restriction fragment could be visualized in restriction digests. Upon

digestion, small fragments (approximately 137 bp each) are generated that are not resolved on a 1 or 1.2% agarose gel and also did not hybridize when restriction mapping and coding region localization were done. Also identified were the exact location of multiple AvaII restriction sites which had been previously difficult to determine due to their proximity to each other (see Figure 3 or 8). Sequencing data also revealed that single Sau96I (GGNCC) sites are always next to each of the PvuII sites (CAGCTG) in addition to the Ava II sites (GGA/TCC). As predicted, there are many (7) Sau96I sites in the helical-coding region of the EcoRI/PstI fragment.

Basic organization. Analysis of the percentage of each of the four nucleotides within this sequence shows that the double stranded sequence contains 60% GC pairs. This GC richness is consistent with other collagen sequences (Bernard et al., 1983). From position 128 to 721 there are at least 31 uninterrupted direct repeats composed of the consensus 18 nucleotide basic repeating unit, AAG ACC AGC AGC CCA GAC, each having a minimum of 83.4% identity to this sequence (Figure 10). There are three brief interruptions. Each interruption is 9 nucleotides long and begins with a PvuII restriction site (indicated by carats and underlined in Figure 10). These interruptions are insertions which disrupt the direct repeats at these locations. However, their presence results in the coding of two perfect triple helical 9-mers with an amino acid sequence of $(Gly\text{-}X\text{-}Y)_2$ in the reverse translation at these

locations. There are also multiple thymidine insertions ("T"-insertions) which interrupt the direct repeat at various locations in the NT735 sequence (see Figure 10).

Although the basic repeating unit itself is not a strict triplet repeat sequence, a single nucleotide substitution, "C" for "A" in the fourth position, makes it so. Interestingly though, very few repeats contain this substitution. When it does occur, a "T"-insertion also occurs in that repeat making it 19 nucleotides long. For those repeats in which the "T" insertions occur subsequent to a PvuII site, two perfect triplet repeats adjacent to each other also occur. Again, these occur in the reverse translation. For example, position 480 to 463 contains two Gly-X-Y 9-mers (refer to Figure 8 or 10). In addition, separated by a single nucleotide at position 462, this occurs again from position 461 to 444. This occurs again at the second PvuII site and in the subsequent repeats with "T"-insertions. All of these perfect triplets occur in the reverse frames. However, because of the single nucleotide separating them, each pair of triplets occurs in a different reading frame.

Higher order repeats. Higher order repeats in NT735 were identified with the aid of the GCG programs "Repeats" and "Find Patterns". In addition to the 18 nucleotide (nt) direct repeats this sequence also contains two 110 nt direct repeats with 93% identity from positions 148-256 and 257-366 with a single nucleotide insertion (Figure 11a). It also contains

three 137 nt direct repeats at positions 291-429, 430-566, and 567-703; each of which contains a PvuII site at homologous positions in the repeat (Figure 11b). These repeats have between 92 and 99.27% similarity. No repeats longer than this were found. Noteworthy however, are four 54 nt repeats which were rejected because their identity (85%) was not significantly different from that of the 18 bp basic unit (83.4%). A repeat organization with higher similarity than the basic unit, such as that seen in the 110 and 137 nucleotide repeats, indicates that these sequences are more recently evolved than those of the early basic unit (Carnahan et al., 1993).

The 18 nucleotide tandem, direct repeat and its higher order repeats are not necessarily a suprising finding considering that most collagen sequences contain a 9 nucleotide triplet-coding basic unit. However, the "T"-insertions and PvuII insertions are present at the same positions in each of the 137 nucleotide repeats, indicating that the insertions occurred prior to the larger repeat (137 nt) duplication event involved in the evolution of these sequences.

Protein translation. Using the GCG program routine "frames" (Figure 12), the analysis of the sequence translation indicates the presence of large open reading frames (greater than 200 nt's long) in the three reverse translation frames (D, E and F). In the forward translation several short open reading frames are confined to the first 300 nucleotides (100

amino acids) of the sequence (Figure 12). Open reading frames stop when an end codon is reached and only begin again if a methionine residue is encoded. Because the reverse frames contain the longer, and more collagen-like reading frames, the remainder of the sequence analysis results will focus on these frames unless otherwise stated.

Amino acid composition. Although the 245 amino acid long protein translation (frame F is longest) does not represent a strict triple-helical coding sequence containing the $(Gly\text{-}X\text{-}Y)_n$ repeat for more than two triplets and does not contain many lysine or proline residues, it does contain glycine at consistent and regular intervals of either every third, sixth or ninth codon in the reverse frames. All of the reverse frames contain patterns similar to this (Figure 13). The GCG program routine "Codon Frequency" was used to examine the protein translation in these frames. In addition to the abundance of glycine residues (16% maximum) all three frames contain nearly 22% of the non-polar residue leucine. The reverse translation frames also include high amounts of alanine (max. 13%), valine (9%) and tryptophan (9%). All are nonpolar residues which contribute to a high degree of hydrophobicity of a molecule. In addition to a 10% serine content, the polar disulfide bonding residue cysteine is present in an abundance equivalent to that of the non-fibrillar collagens (7.7%). Suprisingly, neither proline nor lysine are present in noteably high amounts, since these are

abundant in collagen sequences.

The GCG program routine "Codon preference" is a frame-specific gene finder that recognizes protein coding sequences by examining the specific codon frequency and GC bias or wobble in the translation of a sequence. Codon preference analysis suggests that the NT735 sequence contains expressed protein coding regions (Figure 14). In all frames the average codon preference is much higher (upper third of graph) than that of a random sequence. Although GC bias does not rank as high as codon preference, it is still clearly higher than would be seen for an untranslated or random sequence (bottom third of graph). These data indicate that the coding sequence of the reverse frames is likely to contain preferred codons which occur in genes that are expressed.

Nucleotide sequence comparisons. Nucleotide sequence comparisons to other human collagen sequences in the GCG GenEMBL database indicated that NT735 has the highest similarity to the 3′ end of the cDNA sequence for human pro-α2(I) collagen (Locus=HUMCOLL). The region of optimum identity or "bestfit" is that which includes triple-helical coding sequences of the probe, Hf32 (Figure 15). These results were expected since this was the probe used to isolate the clone. This collagen locus has nearly 52% identity to NT735 in a 690 bp overlap (data not shown). The region of "bestfit" (position 240 to 395 of NT735) within this region has 67.32% identity in a 175 bp overlap with four gaps.

Also having high sequence similarity to NT735 is the 3′ end of pro-α1(I) collagen, represented by the sequences of two cDNA clones, Hf404 & Hf677 (Locus=Humcg1pa1). With 52.5% identity in 621 bp overlap, the "bestfit" region within this was 64.15% in a 122 bp overlap with 3 gaps (Figure 16). This region was contained within sequences of the cDNA clone Hf677. As expected, this probe, previously used as a negative control to confirm that our sequence was not a part of the α1(I) collagen locus, has somewhat less sequence identity to NT735 than the α2(I) locus. Although less similar than sequences of Hf32, it is not unexpected that this locus would also have high sequence similarity to NT735, since the two chains of Type I collagen are very similar to each other.

The above sequence comparisons are consistent with the hybridization results obtained. Under the most stringent conditions used, the minimum similarity necessary for hybridization of Hf32 to occur is in the range of 65-75%. Since there is 67.32% identity between Hf32 and NT735 it is now clear why positive hybridization results were sometimes weak or inconsistent (data not shown). The conditions used approach the minimum similarity necessary for hybridization of the Hf32 probe to NT735. These ranges of similarity were determined by mathematical calculations of T_m values for reassociation in formamide (McConaghy et al., 1969). The upper value for the range is determined by including in the T_m calculation only the GC pairs in the two sequences. The lower

value was calculated by including the total GC content within NT735 in the region of similarity. It is however, difficult to assess the effect of gaps on hybrid stability and equally difficult to quantitate this effect.

The minimum similarity necessary for hybridization of Hf677 to NT735 is in the range of 66.72-79.21%. This value is higher than the actual identity for these sequences (64.15%). According to these figures Hf677 should not hybridize to NT735 under the conditions used, and in fact it did not. This confirms the accuracy of the negative hybridization results with this probe (see Figure 1b).

In the above collagen sequences examined the reverse strand of NT735 has the highest sequence similarities. This is consistent with the analysis of open reading frames (Figure 12) indicating that the reverse frames contained long ORF's and codon preference data (Figure 14) suggesting these frames are likely to contain genuine protein coding sequences.

Other Type I collagen sequences have lower sequence identity to NT735 than Hf32 sequences. For example, the 5' end of both α1 and α2 chains of Type I collagen showed nucleotide sequence identity over similar lengths; however, bestfit results indicated lower optimum identity, and in the case of α2(I), over much shorter lengths (data not shown).

Other collagen types also have high sequence identity with overlaps of up to 550 bp (Table 1). These include the cDNA sequence of the chicken α2(I) collagen gene, and the

alternatively spliced transcripts of the Xenopus laevis α1(II) collagen gene which contains many exons significantly larger than the 54 bp primordial unit. The genomic sequences of the intronless cuticle collagen gene of C. elegans have similarity to the forward strand as does the D. melanogaster collagen gene. The latter is also suggested to belong to the family of basement membrane or cuticle collagens, and contains large exons. Large coding regions such as these are characteristic of the non-fibrillar collagens. Since NT735 also contains large ORF's, this suggests it too has an organization similar to non-fibrillar collagens.

In addition, the human Type XVI FACITs has nearly 75% identity in a 97 bp overlap and chicken Type XIV FACITs has 54% identity over almost a 400 bp overlap (Table 1). Various short stretches of the Type XIV transcript approach 60% similarity when the bestfit is examined. Types IX (also a FACITs) and X collagen of various species also have 50-60% identity over short stretches. Much of the similarity to these non-fibrillar collagen genes is in the forward strand of NT735. This is unexpected given that the Type I and II collagen similarities are in the reverse strands. The reverse frames of NT735 are those with the large, uninterrupted ORF's, a feature characteristic of the FACITs.

Nucleotide sequence comparisons to the entire database showed the highest sequence identity in the forward strand is to the Wheat Glu-1D-1b gene for HMW wheat glutenin subunit 5

(see Table 2). The sequence has as much as 67% identity over 639 bp with 12 gaps. This is the highest identity with the longest region of similarity identified by the GCG program. This identity matches that of human α2(I) collagen but has many more gaps and a longer overlap. If the equivalent number of gaps as α2(I) collagen (4) are allowed, the identity to NT735 is still 50% over 557 bp. Many genes coding the subunits of this wheat seed storage protein have a nucleotide similarity of 55-65%, often over large (400-500) overlaps. Much of this similarity is in the central repetitive coding domain of the glutinen genes. The HMW subunits, such as subunit 5, are of particular importance in the glutinen molecule because the cysteine residues at their termini confer an unusual secondary structure contributing to its elastic properties. (Tatham et al., 1984). The above properties and its highly repetitive nature suggest that the wheat glutinen protein is the equivalent of human elastin, an ECM component which shares some characteristics of the fibrillar and non-fibrillar collagens (Indik et al., 1990).

Because the above information was relevant, an independent sequence comparison of NT735 to elastin coding sequences using the "Fasta" subroutine was done. This comparison indicated a 55% identity over 345 bp. A subsequent "bestfit" analysis indicated 63% similarity over 134 bp (1 gap) to an exon which codes for a glycine rich, hydrophobic domain of the human elastin protein (Table 2).

Fasta comparisons of NT735 to the entire database also indicated high identity in the forward strand for the D. virilis mastermind gene which codes for a large, highly repetitive nuclear protein required for normal CNS development. The D. melanogaster mastermind gene had slightly lower identity. In both species the similarity lies in the region which contains the well characterized motifs called OPA repeats. OPA repeats are defined by a $(CAX)_n$ repeat, where X is an A or G, and $n<31$. These repeats are common to a variety of developmental regulator genes. Other genes which have noteworthy sequence similarity to NT735 are summarized in Table 2. Many of these genes code for very large proteins, contain long ORF's, and/or contain OPA or other homopolymeric repeats. The relative significance of these sequence similarities will become more clear when the coding strand and reading frame of this new collagen-like sequence is determined.

Amino acid sequence comparisons. The first approach to amino acid sequence comparisons employed the GCG routine "Tfasta" to examine the amino acid translations of the genes contained in the GenEMBL database. Consistent with the above results, HUMCOLL, the α2(I) locus of HF32, ranked highest in similarity in the protein translation frame F of NT735. However, the frame in which the similarity is seen is not the triple-helical coding frame of the α2(I) chain. Many amino acids are not identical yet show similarity as "conservative replacements". All of the human collagens with high nucleotide

sequence similarity which appeared in the peptide sequence comparison results showed amino acid similarity in frame D or F, but few genes were represented in the appropriate triple-helical coding frame. The chick collagens showing similarity to frame D and F as well.

The Herpes saimiri collagen-like protein mRNA sequence (Locus:HSVCLS) has 42.5% identity in a 40 amino acid overlap and 88% similarity when amino acids representing conservative replacements are included (Figure 17). This similarity is in the region of NT735, amino acid residues 140-180 of the reverse frame D, in which glycine occurs at every sixth residue. The glycine and tryptophan residues in both sequences are in homologous positions and the cysteine residues in NT735 are always conservative replacements for serine residues. Though this information is interesting its significance is suspect because the protein translation of HSVCLS with this similarity is not the actual reading frame of this gene. However, upon examining the nucleotide sequence of this locus there is another interesting feature in common with NT735. In the region of similarity to NT735, HSVCLS contains an 18 nucleotide repeat which spans its collagen-like open reading frame (Figure 17b). This repeat has 56% similarity to the 18 nucleotide repeat of NT735. This similarity is likely to account for the identity seen between the amino acid translations discussed above. This is interesting because the HSVCLS locus is the only other known example (in addition to

NT735) of a collagen-like gene containing an 18 nucleotide basic repeat in its coding region.

Various loci representing the HMW subunits of Wheat Glutenin also rank high in amino acid similarity in frames D thru F. For example, subunits 5 and 10 have as high as 46-57% identity over lengths of 30-40 amino acids in frames D thru F. Many of the regions of similarity are in the same consensus peptide and central domain sequences indicated in the nucleotide sequence comparisons (Table 2). These regions reside in a highly repetitive region of the molecule where the peptide sequence predicts a secondary structure with an abundance of ß-turns. The cysteine residues located at the termini of these subunits, allow the close association of multiple glutinen monomers by intermolecular disulfide bonding (Forde et al., 1985). The ß-spiral structure predicted by these sequences and the abundance of terminal cysteine residues is suggested to confer elasticity to the wheat glutinen molecule much like that of human elastin (Tatham et al., 1984; Venkatachalam et al., 1981).

The above amino acid sequence comparison results are in agreement with the nucleotide sequence comparison information. However, due to the technical limitations of the GCG "Tfasta" program the peptide translations of the genes identified are often not in the correct coding frame. This leaves their relationship to NT735 and a putitive peptide structure or function ambiguous at best.

In order to overcome the "Tfasta" routine limitations, the peptide database Swissprot was also examined using the GCG routine "Fasta". Each of the reverse frames of the amino acid translations of NT735 was compared to the Swissprot peptide sequence database. The results of these comparisons were somewhat different from those above although some common themes with regard to secondary structure exist. Nearly every peptide with similarity to the NT735 reverse translation frames was described as a membrane-associated molecule with a highly hydrophobic transmembrane region (see Table 3). It is these regions which show the highest similarity to NT735 translations. These peptides included viral envelope glycoproteins, surface antigens and receptors, channel proteins, and structural glycoproteins with carbohydrate attatchment sites. Many of these regions with identity cross the cell membrane multiple times serving as an anchor for large intra- and extra- cellular chains with diverse interactive properties. Table 3 summarizes the key features of some these proteins in an attempt to understand the basic features of the peptide encoded by NT735. Those listed in the table have at least 20% identity over a minimum of 20 amino acids, as has often been reported in other work (Trueb and Trueb, 1992).

Because the Swissprot database search was so informative with regard to the consistent nature of the proteins identified (and of course, in the correct coding frames),

independent protein comparisons were done for both elastin, mentioned earlier in conjunction with wheat glutinen, and the HSVCLS protein. No further informative results were obtained for HSVCLS. However, Swissprot comparisons of NT735 to elastin proved informative. For a number of species, including humans, the range of identity was 39 to 53%, with an overlap of 17 to 30 amino acids. Although this is not a very long overlap, a 30 amino acid length is comparable to some of those proteins listed in Table 3.

Although the specific functional domains of the putative peptide of NT735 are not yet clear, most of the nucleotide and amino acid similarities to other protein coding genes lie in the region of the 18 bp direct repeat. Few matches were seen in amino acids 200 thru 245 (reverse frames) which represents the non-repetitive portion of the nucleotide sequence. Furthermore, much of the similarity identified in the "Swissprot" comparisons is to the left of the first Ava II site (see Figure 3) at nucleotide position 230 (see Figure 8). Consistent with the initial hybridization studies this site was identified as the beginning of the helical coding region for the gene. Although the exact endpoint of the coding region is still not clear, most of the amino acid similarity appears confined to the sequence prior to last Pvu II site (nucleotide position 680). This general pattern is consistent with all my experimental results and the sequence comparison data presented.

Figure 1. Characterization of chromosome 17 phage clones. A pro-α2(I) cDNA clone (Hf32) was used to screen a human chromosome 17 library containing HindIII fragments in lambda phage Charon 21A. Five independent positive phage isolates were digested with HindIII to excise the inserts which were then hybridized with: A- Pro-α2(I) cDNA (Hf32); B- Pro-α1(I) cDNA (Hf677); C- Alu repetitive sequence (Blur 8). Some isolates are present in multiple lanes because several individual plaques from the same original isolate had been chosen for analysis.

Lane 1 & 2, Isolate #1
Lane 3, Isolate #2
Lane 4, Isolate #3
Lane 5 & 6, Isolate #4
Lane 7 & 8, Isolate #5.

Fragment sizes in kb are indicated.

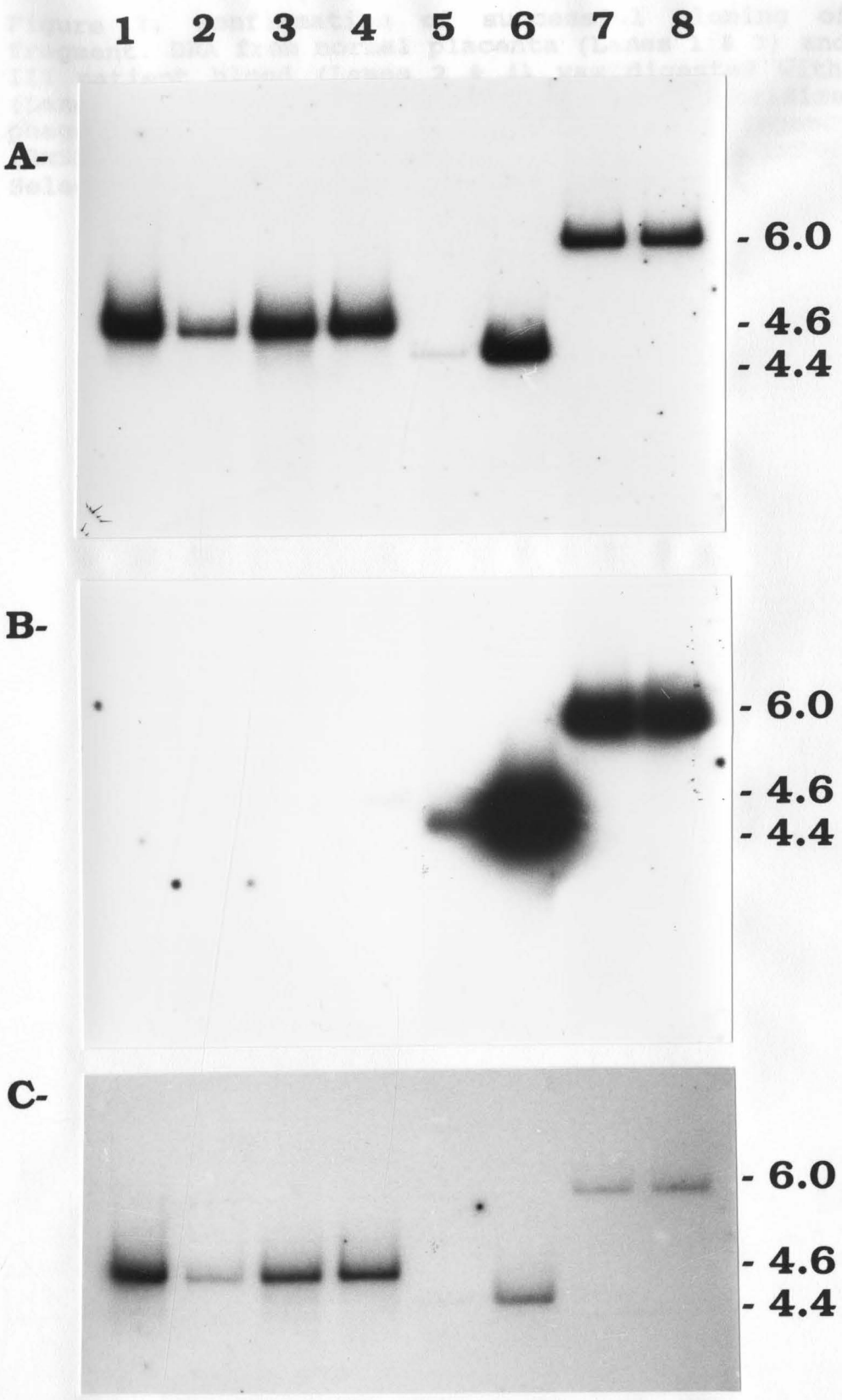
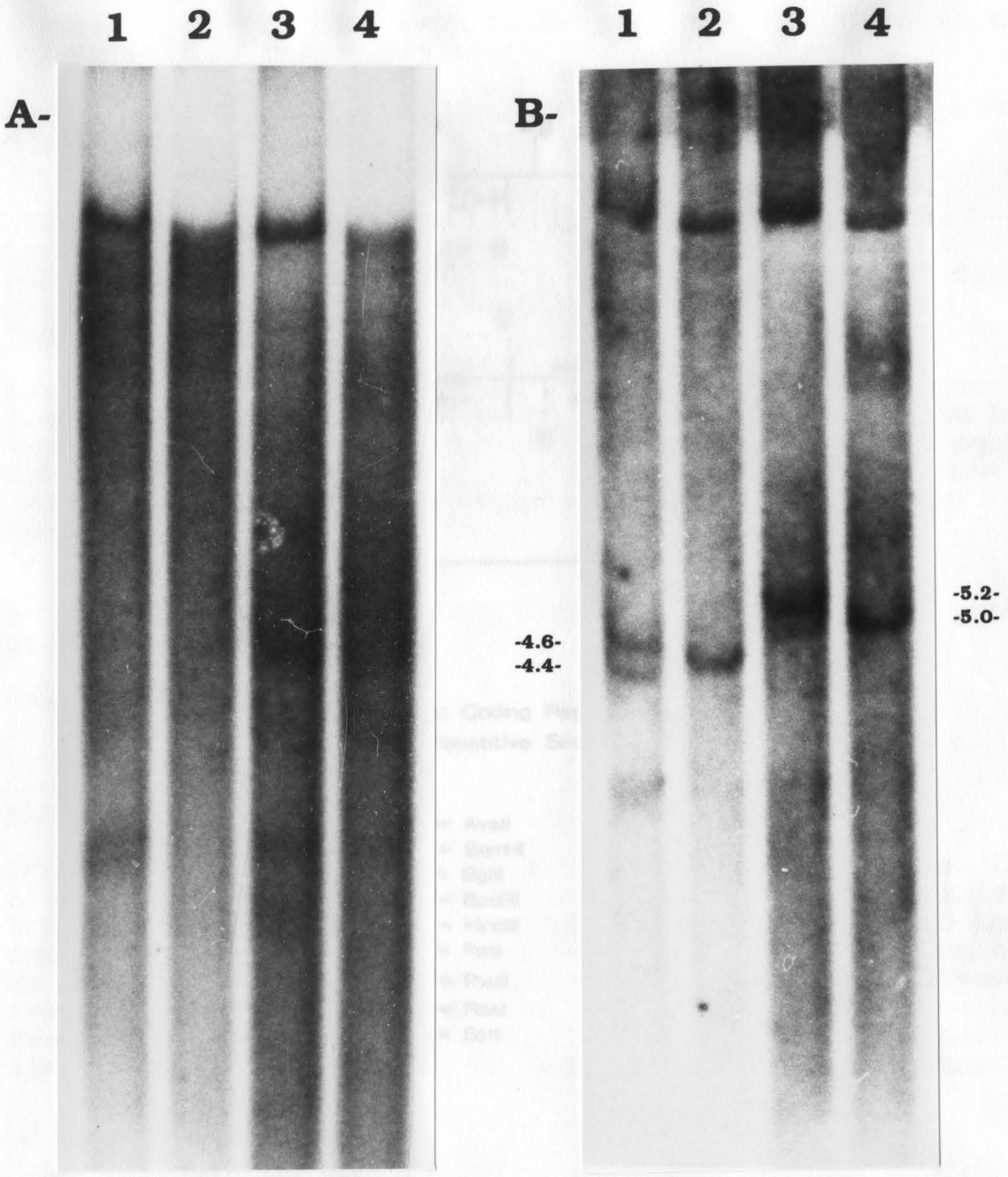
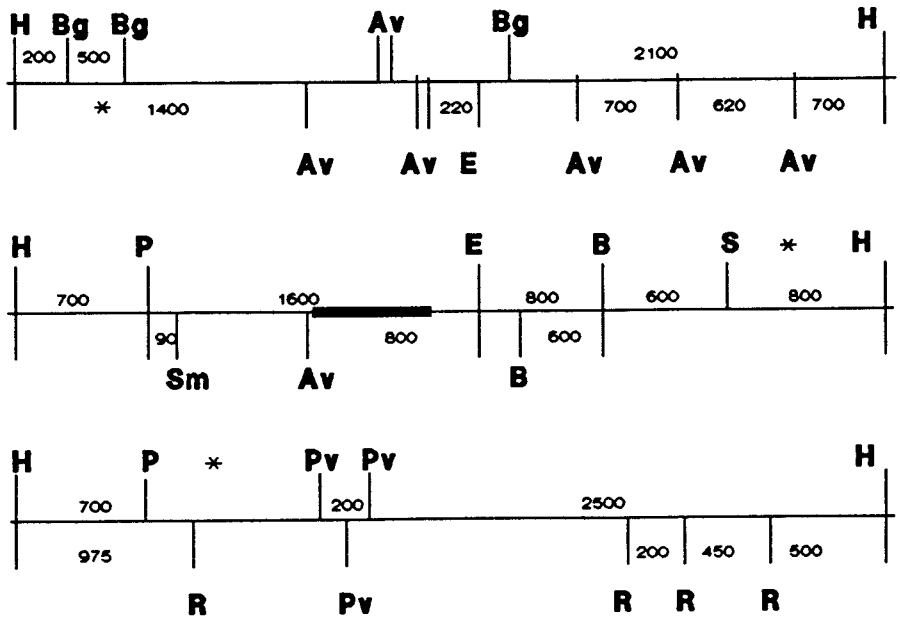

1 2 3 4 5 6 7 8
A-
- 6.0
- 4.6
- 4.4
B-
- 6.0
- 4.6
- 4.4
C-
- 6.0
- 4.6
- 4.4

Figure 2. Confirmation of successful cloning of target fragment. DNA from normal placenta (Lanes 1 & 3) and OI Type III patient blood (Lanes 2 & 4) was digested with HindIII (Lanes 1 and 2) or PstI (Lanes 3 & 4) and hybridized to the phage clone (isolate #1). A- hybridization stringency was low (2xSSC, 50°C). B- stringency was high (0.1xSSC, 65°C). Selected fragment sizes in kb are indicated.


1 2 3 4
A-
1 2 3 4
B-
-4.6-
-4.4-
-5.2-
-5.0-



■ = Helical Coding Region
* = Alu Repetitive Sequence

Av = AvaII
B = BamHI
Bg = BglII
E = EcoRI
H = HindIII
P = PstI
Pv = PvuII
R = RsaI
S = SstI

Figure 3. Restriction map of the 4.6 kb cloned fragment. The new collagen locus on chromosome 17 was cloned from a chromosome specific phage library as a 4.6 kb HindIII restriction fragment. Restriction sites for the indicated enzymes are shown on superimposable lines representing the same HindIII fragment. The locus contains a short triple helical coding region that cross-hybridizes to the pro-α2(I) cDNA, Hf32 (black box). Fragment sizes in base pairs are indicated. The locations of Alu repetitive sequences are also indicated (*). Not all restriction sites are shown.

Figure 4. Restriction mapping of p4.6 and localization of triple helical coding region.

A- Restriction digests of p4.6 were hybridized to the pro-α2(I) cDNA probe Hf32.

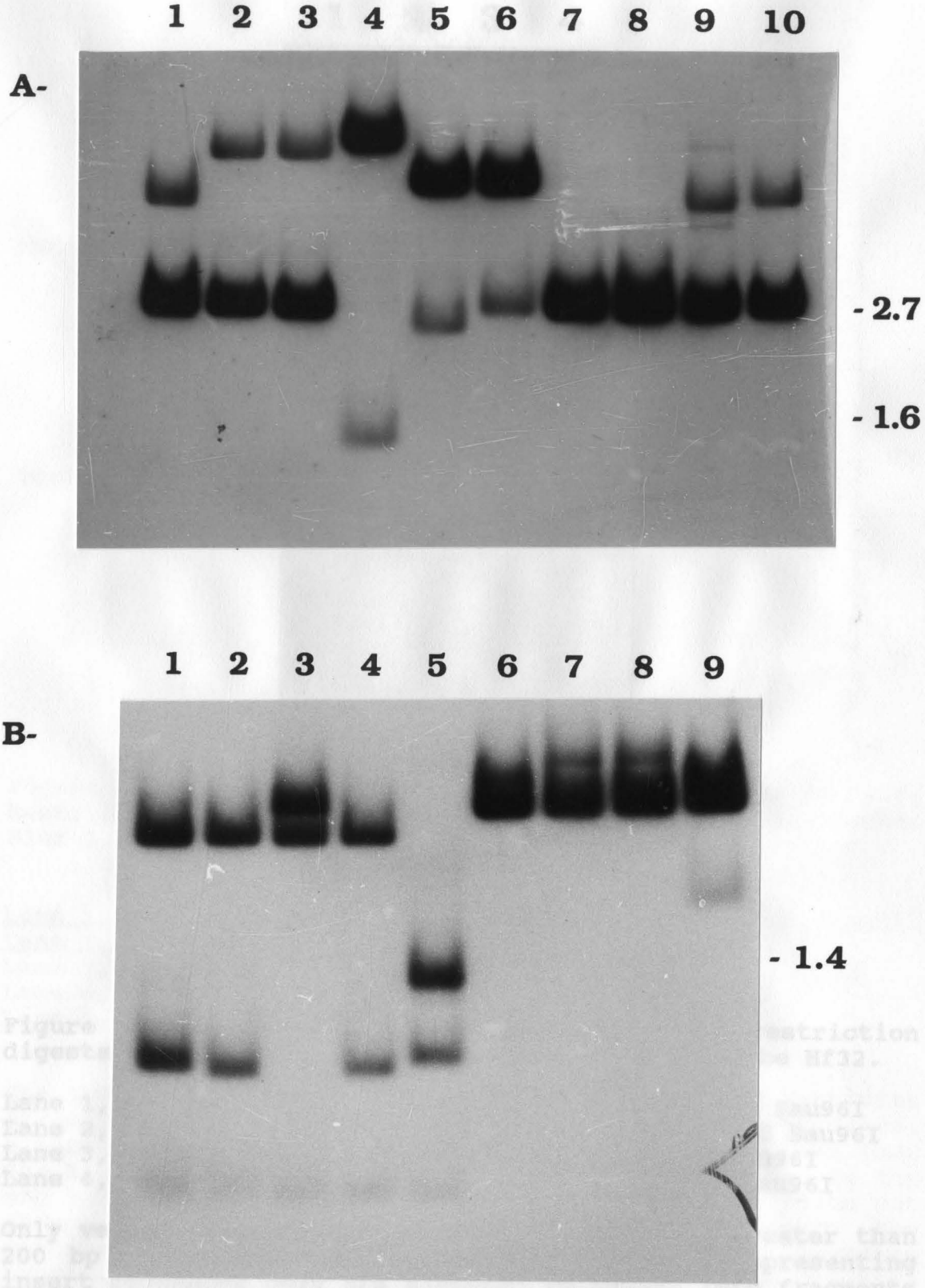
Lane 1, HindIII+PstI
Lane 2, HindIII
Lane 3, HindIII+XbaI
Lane 4, EcoRI+PstI
Lane 5, EcoRI+BamHI
Lane 6, BamHI
Lane 7, HindIII+Bam HI
Lane 8, HindIII+BamHI+SstI
Lane 9, HindIII+SstI
Lane 10, HindIII+SmaI

Selected fragment sizes in kb are indicated. The fragment containing triple helical coding sequences is represented by the 1.6 kb EcoRI/PstI band seen in lane 4. The strongly hybridizing 2.7 kb fragment is the plasmid vector. Lanes indicated by underlining represent the pertinent data as discussed in the text.

B- Restriction digests of p4.6 were hybridized to the pro-α2(I) cDNA probe Hf32.

Lane 1, AvaII+BamHI
Lane 2, AvaII+EcoRI
Lane 3, AvaII
Lane 4, AvaII+SstI
Lane 5, AvaII+HindIII
Lane 6, PvuII+HindIII
Lane 7, PvuII+PstI
Lane 8, PvuII
Lane 9, PvuII+SstI

Selected fragment sizes in kb are indicated. Hybridation to a single (vector) fragment in Lane 6 indicates no hybridization to the upper (1.4 kb) HindIII/AvaII fragment of the doublet at this position. There is also no hybridization to an 800 bp PvuII/PstI subfragment of HindIII/AvaII. These fragments span the PstI end of the triple helical coding fragment (see restriction map). Lanes indicated by underlining represent the pertinent data as discussed in the text.


1 2 3 4 5 6 7 8 9 10
A-
- 2.7
- 1.6
1 2 3 4 5 6 7 8 9
B-
- 1.4

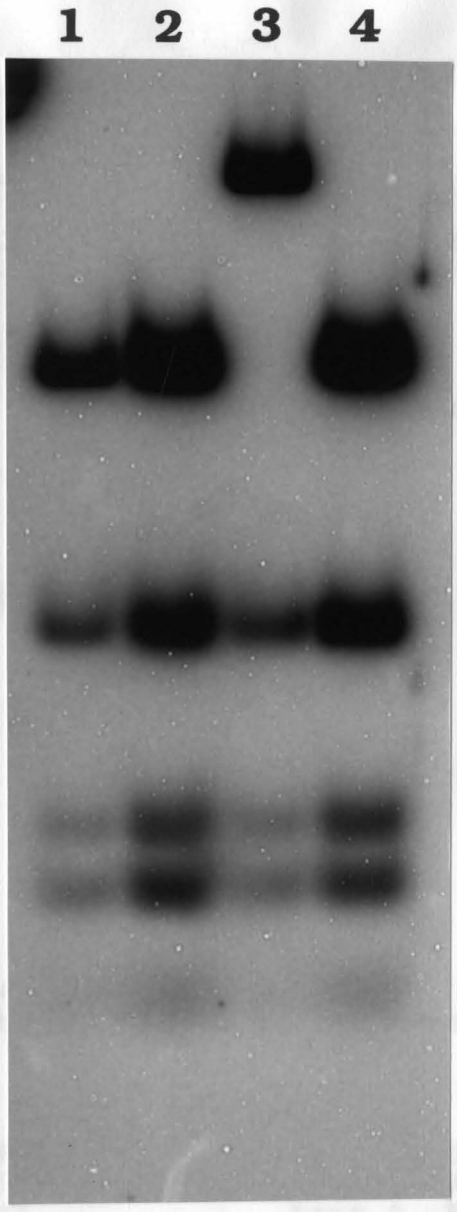


Figure 5. Southern blot representing Sau96I restriction digests of p4.6 hybridized with the pro-α2(I) probe Hf32.

Lane 1, p4.6 DNA digested with HindIII, EcoRI and Sau96I
Lane 2, pUC13 DNA digested with HindIII, EcoRI and Sau96I
Lane 3, p4.6 DNA digested with EcoRI, PstI and Sau96I
Lane 4, pUC13 DNA digested with EcoRI, PstI and Sau96I

Only vector sequences (pUC13 DNA) in fragments greater than 200 bp are detected by the probe. Fragments representing insert sequences only are digested by Sau96I into fragments of less than 200 bp. Sizes in bp are indicated.

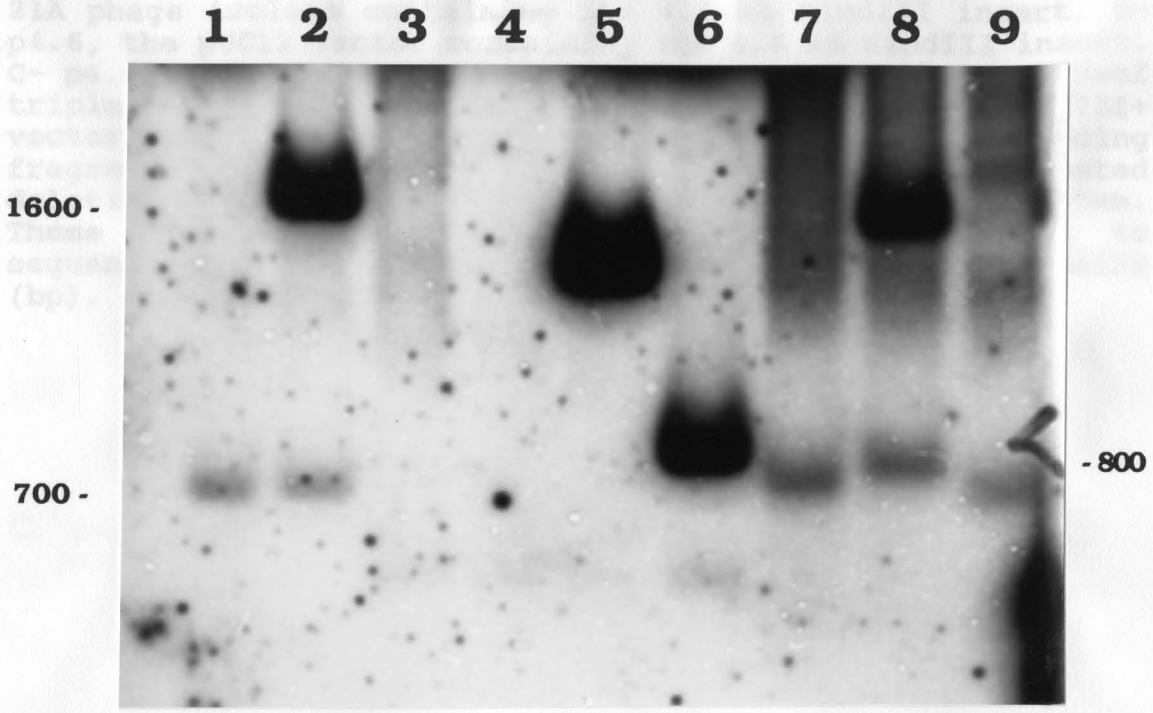
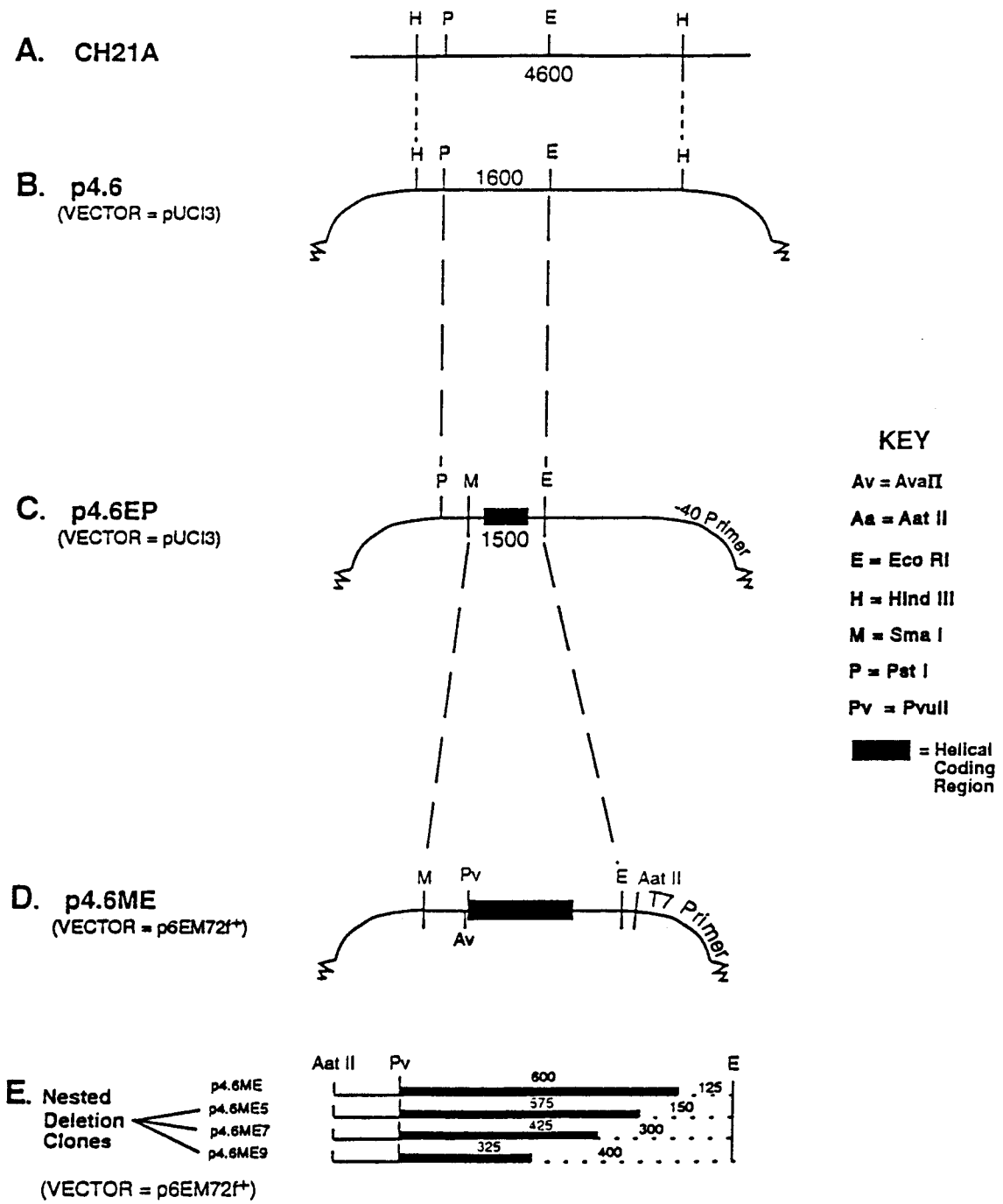


Figure 6. Localization of Alu repetitive sequences in p4.6. Restriction digests of p4.6 were hybridized to the Alu probe, Blur 8.

Lane 1, HindIII+PstI
Lane 2, EcoRI+PstI
Lane 3, EcoRI+BamHI
Lane 4, BamHI
Lane 5, HindIII+BamHI
Lane 6, HindIII+SstI+BamHI
Lane 7, HindIII+SmaI
Lane 8, EcoRI+SmaI
Lane 9, PstI+SmaI

The 1600 bp EcoRI/PstI and the 800 bp HindIII/SstI hybridize strongly to Blur 8. The 700 bp HindIII/PstI band hybridizes, although more weakly. Another Southern blot confirmed that the Alu sequence in the 1600 bp fragment was confined to the 850 bp adjacent to the PstI end of that fragment (data not shown). Lanes indicated by underlining represent the pertinent data as discussed in the text. Vector sequence hybridization and larger fragments which were uninformative are not shown.

Figure 7. Schematic diagram of the clones and subclones used to characterize the collagen-like locus. Only a portion of the vectors are shown in the diagrams. A- The original Charon 21A phage isolate containing the 4.6 kb HindIII insert. B- p4.6, the pUC13 vector containing the 4.6 kb HindIII insert. C- p4.6EP, the pUC13 vector containing the 1.6 kb EcoRI/PstI triple helical coding fragment. D- p4.6ME, the pGEM7Zf+ vector containing the 1.5 kb EcoRI/SmaI triple helical coding fragment. E- Linearized p4.6ME and the resulting nested deletion subclones generated by the "Erase-A-Base" system. These deleted subclones were recircularized prior to sequencing. Sizes of fragments are indicated in base pairs (bp).


A. CH21A
H P E H
4600
B. p4.6
(VECTOR = pUC13)
H P E H
1600
C. p4.6EP
(VECTOR = pUC13)
P M E
1500
-40 Primer
D. p4.6ME
(VECTOR = p6EM72f+)
M Pv E Aat II
Av
T7 Primer
E. Nested Deletion Clones
p4.6ME
p4.6ME5
p4.6ME7
p4.6ME9
(VECTOR = p6EM72f+)
Aat II Pv E
600
125
575
150
425
300
325
400
KEY
Av = AvaII
Aa = Aat II
E = Eco RI
H = Hind III
M = Sma I
P = Pst I
Pv = PvuII
= Helical Coding Region

```
      EcoRI
1     GAATTCCAGT CACTCCCTCC AGGAGACACA CCAGCAGTGG TCAGGAAGCA

51    AGTCCACCAC AGCCTTCTGG AAAGGTCTCT AGTCTCAGGA GACACGAGCC
                                                 Sau96I
101   ACTCAAGCCT CCCACATGCC ACCAAAGAAG ATAAGCTGGC CCAGAGAAGA

151   CCAGCAGCCC AGAGAAGACC AGCAGCCTGG ACAAGACCAA CAGCCCAGAC
                            AvaII               AvaII
201   ATAGACCAGC AGCCCAGAAA GGACCAGCAG CCCAGACAGG ACCAGCAGCC

251   CAGGCAAGAC CAGCAGCCTG GACAAGACCA GCAGCCTGGA CAAGACCAAC
                                       AvaII
301   AGCCCAGACA TAGACCAGCA GCCCAGACAG GACCAACAGC CCAGACATAG
                      AvaII
351   ACCAGCAGCC CAGACAGGAC CAGCAGCCCA GACAAGACCA GCAGCCTGGA
          PvuII
401   CAAGAACAGC TGGCCCAGCA GCCCATACAA GACCAGCAGC CCAGATATAG

451   CCCAGCAGCC CAGACATAGC CCAGCAGCCC AGACAAGACC AGCAGCCCAG
                                                    PvuII
501   ACAAGACCAG CAGCCTGGAC AAGACCAGCA GCCAGGATAA GAACAGCTGG

551   CCCAGCAGCC CATACAAGAC CAGCAGCCCA GATATAGCCC AGCAGCCCAG

601   ACATAGCCCA GCAGCCCAGA CAAGACCAGC AGCCCAGACA AGACCAGCAG
                                       PvuII
651   CCCGGACAAG ACCAGCAGCC AGGATAAGAA CAGCTGGCCC AGCAGCCCAT

701   ACAAGACCAG CAGCCCAGAC ATAGCCCAGC AGCCC
```

Figure 8. The complete nucleotide sequence of NT735. Selected restriction sites are indicated above their recognition sequence. Numbers on the left margin indicate the nucleotide position at the start of the line. A space is inserted after every tenth nucleotide for easier reading.

Figure 9. 8% Acrylamide sequencing gels. A- An unreadable sequencing gel generated by the "Sequenase" kit. This illustrates a typical repeating "ladder" pattern resulting from premature termination due to GC richness and 2° structure of the template DNA. This is roughly the same region shown in B. B- A gel from which 108 nucleotides can be read as pictured here (position 324 to 432 of the NT735 sequence). This sequence was generated using the "Ampli-Taq" kit. The AvaII sites at positions 339 and 366 respectively are indicated. The letters at the top indicate the nucleotide represented in the corresponding lane.

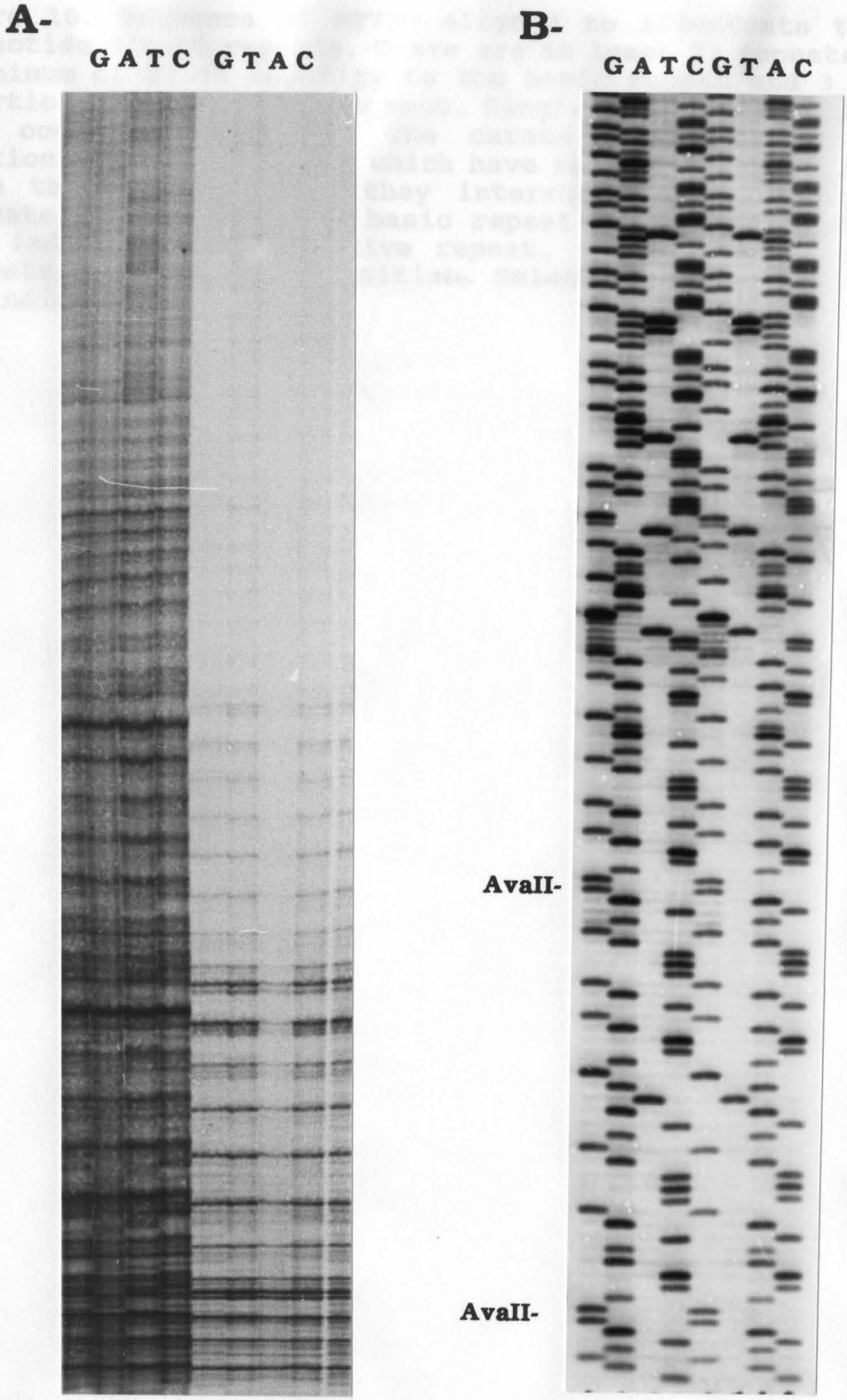
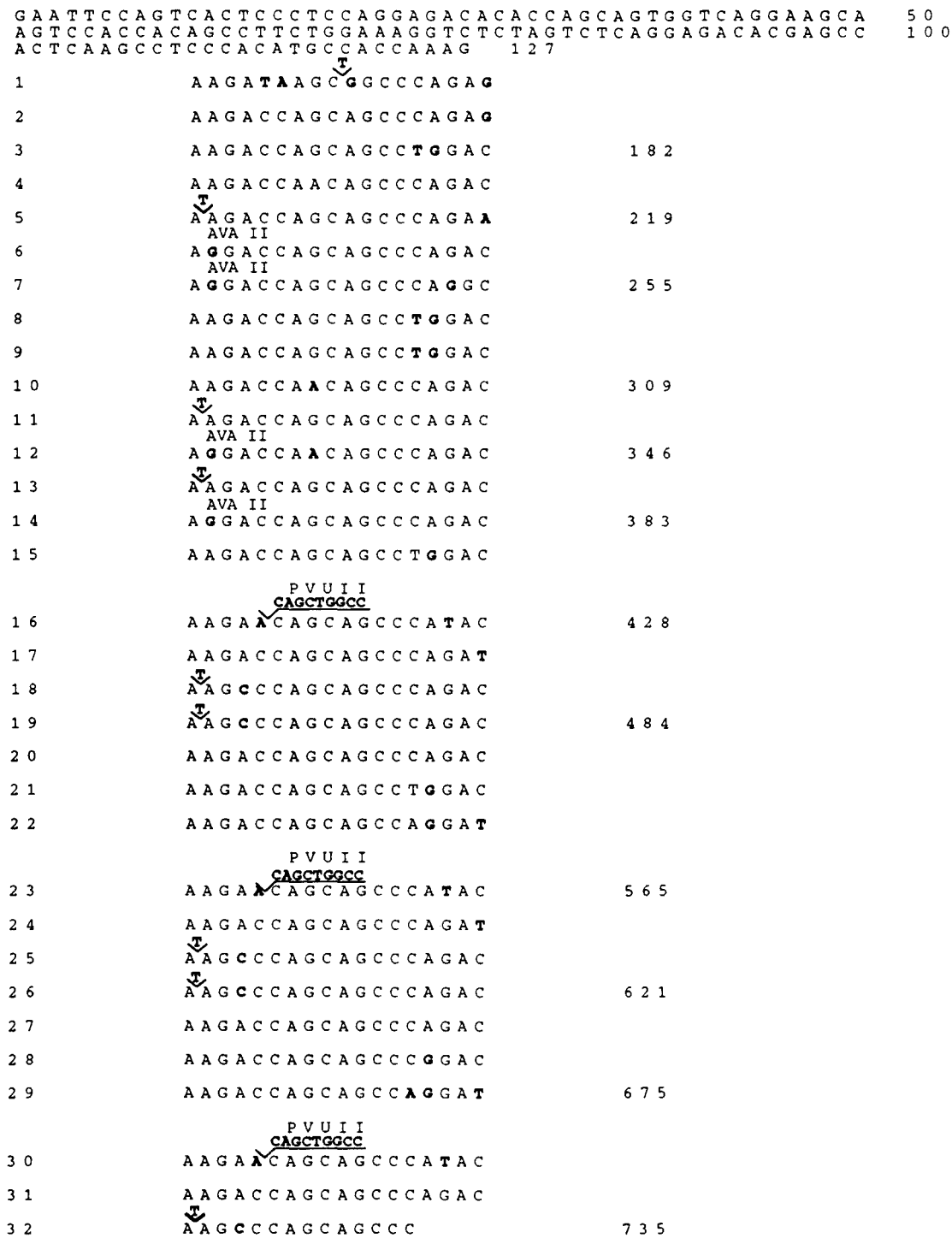

A-
GATC GTAC
B-
GATCGTAC
AvaII-
AvaII-

Figure 10. Sequence of NT735 aligned to illustrate the 18 nucleotide direct repeats. There are at least 31 repeats with a minimum of 83.4% identity to the basic repeat and 3 brief insertions of 9 nucleotides each. Single thymidine insertions also occur periodically. The carats indicate the exact location of the insertions which have been removed and placed above the sequence that they interrupt. Bold nucleotides indicate changes from the basic repeat unit. Numbers to the left indicate the respective repeat. Numbers to the right indicate the nucleotide position. Selected restriction sites are indicated.


GAATTCCAGTCACTCCCTCCAGGAGACACACCAGCAGTGGTCAGGAAGCA 50
AGTCCACCACAGCCTTCTGGAAAGGTCTCTAGTCTCAGGAGACACGAGCC 100
ACTCAAGCCTCCCACATGCCACCAAAG 127
1 AAGATAAGCGGCCCAGAG
2 AAGACCAGCAGCCCAGAG
3 AAGACCAGCAGCCTGGAC 182
4 AAGACCAACAGCCCAGAC
5 AAGACCAGCAGCCCAGAA 219
AVA II
6 AGGACCAGCAGCCCAGAC
AVA II
7 AGGACCAGCAGCCCAGGC 255
8 AAGACCAGCAGCCTGGAC
9 AAGACCAGCAGCCTGGAC
10 AAGACCAACAGCCCAGAC 309
11 AAGACCAGCAGCCCAGAC
AVA II
12 AGGACCAACAGCCCAGAC 346
13 AAGACCAGCAGCCCAGAC
AVA II
14 AGGACCAGCAGCCCAGAC 383
15 AAGACCAGCAGCCTGGAC
PVU I I
CAGCTGGCC
16 AAGAACAGCAGCCCATAC 428
17 AAGACCAGCAGCCCAGAT
18 AAGCCCAGCAGCCCAGAC
19 AAGCCCAGCAGCCCAGAC 484
20 AAGACCAGCAGCCCAGAC
21 AAGACCAGCAGCCTGGAC
22 AAGACCAGCAGCCAGGAT
PVU I I
CAGCTGGCC
23 AAGAACAGCAGCCCATAC 565
24 AAGACCAGCAGCCCAGAT
25 AAGCCCAGCAGCCCAGAC
26 AAGCCCAGCAGCCCAGAC 621
27 AAGACCAGCAGCCCAGAC
28 AAGACCAGCAGCCCGGAC
29 AAGACCAGCAGCCAGGAT 675
PVU I I
CAGCTGGCC
30 AAGAACAGCAGCCCATAC
31 AAGACCAGCAGCCCAGAC
32 AAGCCCAGCAGCCC 735

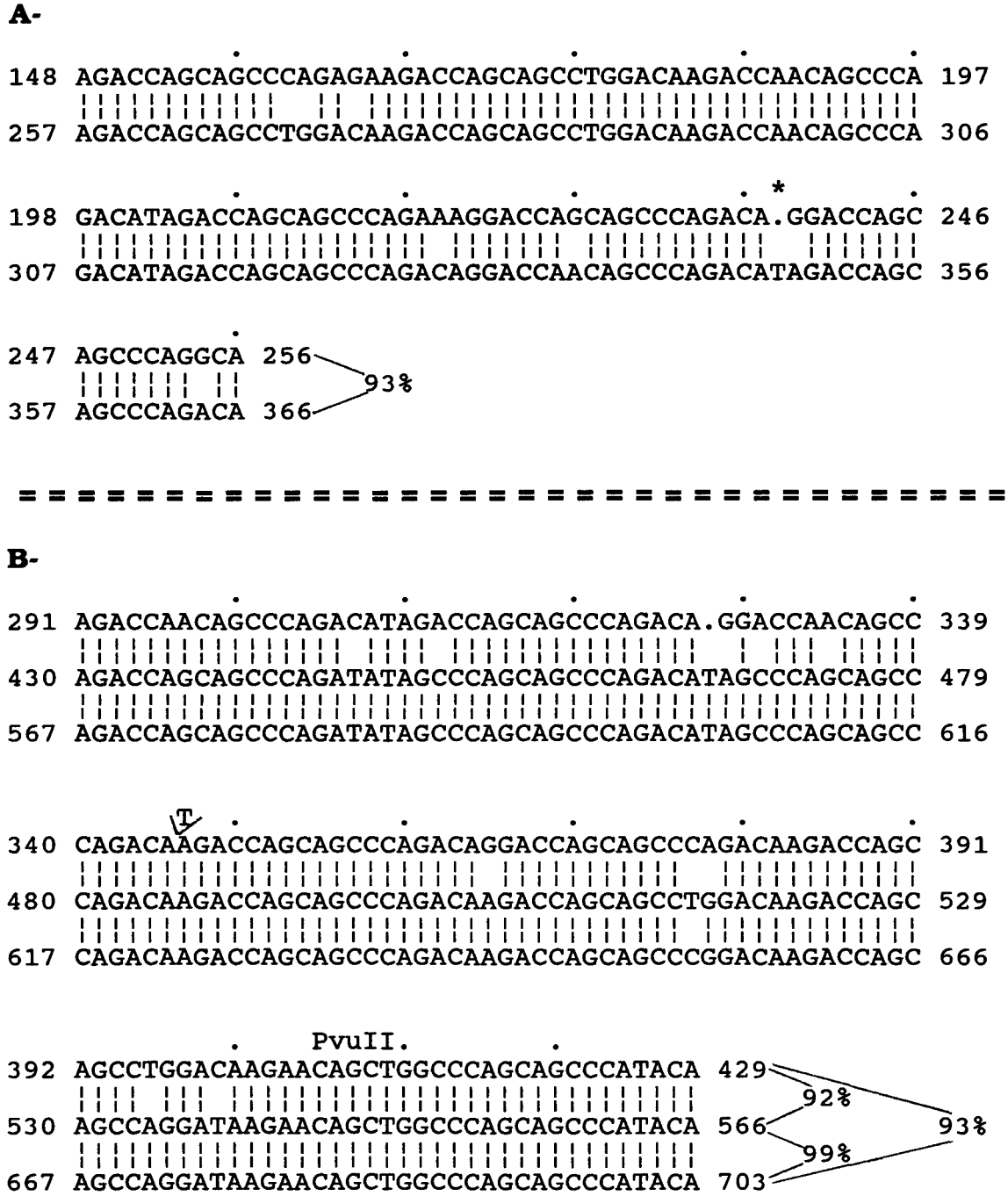


Figure 11. Higher order repeats of NT735. A- Two 110 nucleotide direct repeats. B- Three 137 nucleotide direct repeats. Dashed vertical lines indicate nucleotide identity. Whole numbers indicate nucleotide position numbers. Percentages represent identity between repeats as indicated. Insertions are indicated by an asterisk (*) in A, or carat (^) in B. Dots occur every tenth position for easier reading.

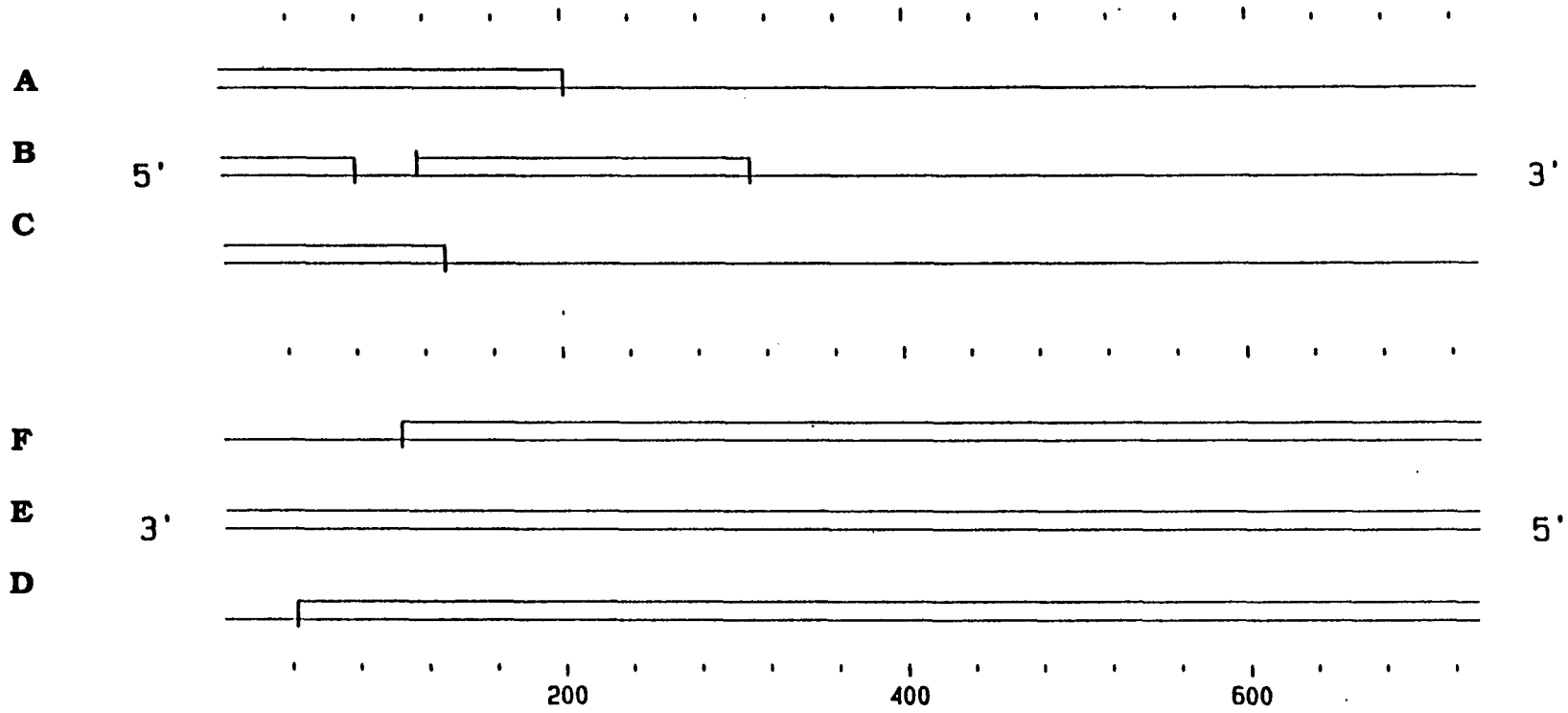


Figure 12. Schematic diagram of the open reading frames of the translated sequence of NT735. Open boxes indicate the open reading frames. Solid lines represent regions which are not open reading frames. Letters to the left indicate the frame, with frame A beginning with nucleotide number 1 and frame F beginning with nucleotide number 735. Numbers on the bottom indicate nucleotide position.

Translation frame "F"

1 GLLGYVWAAG LVWAAGPAVL ILAAGLVRAA GLVWAAGLVW AAGLCLGCWA

51 ISGLLVLYGL LGQLFLSWLL VLSRLLVLSG LLVLSGLLGY VWAAGLYLGC

101 WSCMGCWASC SCPGCWSCLG CWSCLGCWSM SGLLVLSGLL VYVWAVGLVQ

151 AAGLVQAAGL AWAAGPVWAA GPFWAAGLCL GCWSCPGCWS SLGCWSSLGQ

201 LIFFGGMWEA *VARVS*D*R PFQKAVVDLL PDHCWCVSWR E*LEF

Translation frame "E"

1 GCWAMSGLLV LYGLLGQLFL SWLLVLSGLL VLSGLLVLSG LLGYVWAAGL

51 YLGCWSCMGC WASCSYPGCW SCPGCWSCLG CWSCLGCWAM SGLLGYIWAA

101 GLVWAAGPAV LVQAAGLVWA AGPVWAAGLC LGCWSCLGCW SMSGLLVLSR

151 LLVLSRLLVL PGLLVLSGLL VLSGLLVYVW AVGLVQAAGL LWAAGLLWAS

201 LSSLVACGRL EWLVSPETRD LSRRLWWTCF LTTAGVSPGG SDWN

Translation frame "D"

1 AAGLCLGCWS CMGCWASCSY PGCWSCPGCW SCLGCWSCLG CWAMSGLLGY

51 IWAAGLVWAA GPAVLILAAG LVQAAGLVWA AGLVWAAGLC LGCWAISGLL

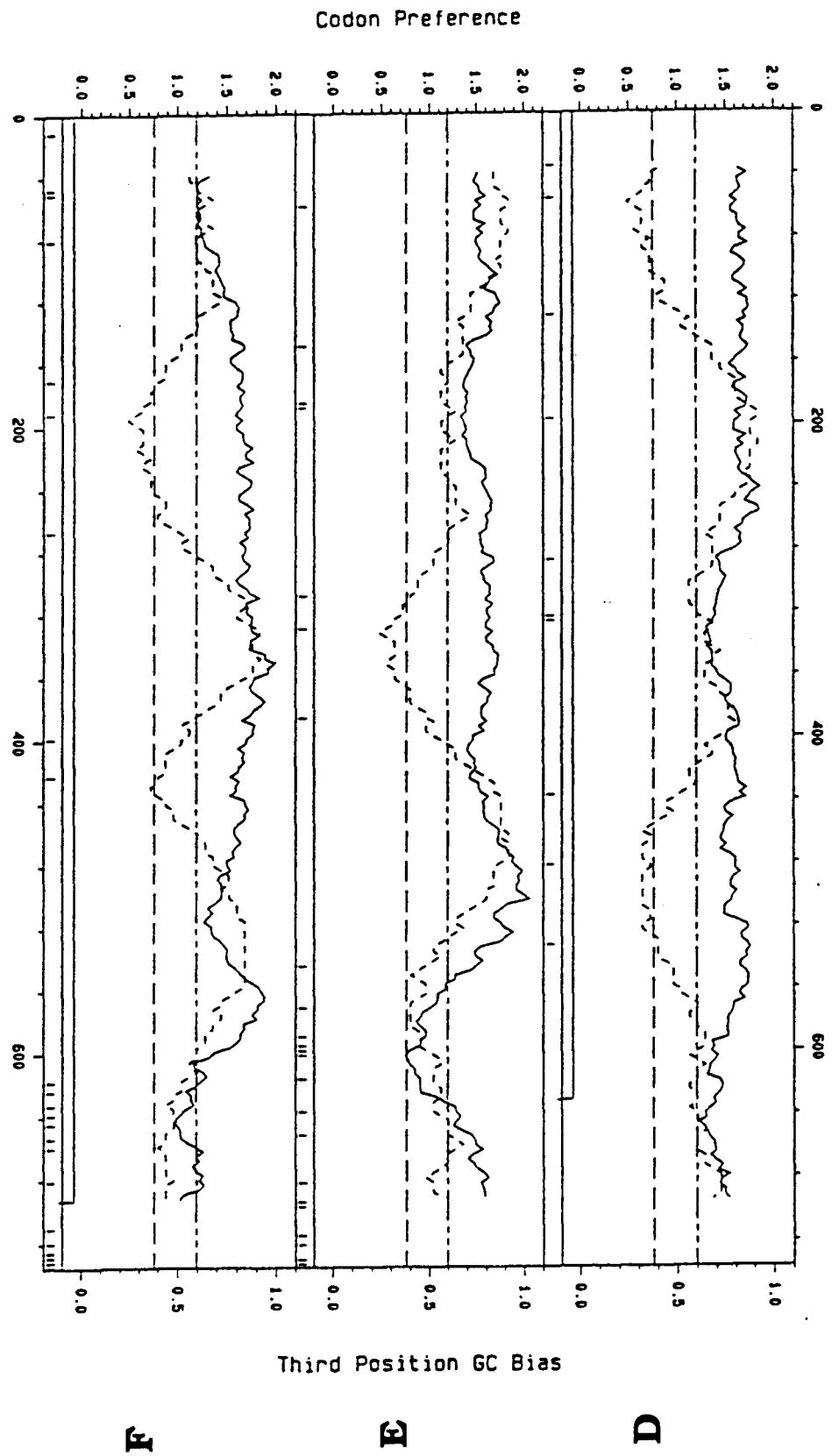
101 VLYGLLGQLF LSRLLVLSGL LVLSGLLVYV WAVGPVWAAG LCLGCWSCPG

151 CWSCPGCWSC LGCWSCLGCW SFLGCWSMSG LLVLSRLLVF SGLLVFSGPA

201 YLLWWHVGGL SGSCLLRLET FPEGCGGLAS *PLLVCLLEG VTGI

Figure 13. Amino acid sequence translations of the reverse frames of NT735. A space is inserted after every tenth residue to facilitate readability. Some glycine residues are highlighted in bold in order to demonstrate the regular interval pattern in which they occur. Asterisks (*) indicate the presence of a "stop" codon.

Figure 14. Codon preference data for reverse translation frames of NT735 (D, E and F). The letters on the left indicate the reading frame represented by the corresponding graph. The numbers on the top and bottom of the figure indicate approximate nucleotide positions. The solid line represents codon preference. The dashed line represents GC bias. The Y-axis is split in three sections. A curve lying in the upper region (1.2-2.0) indicates a strong pattern of codon choices similar to that of expressed genes. One lying in the lower region indicates no pattern of similarity to expressed genes (0-0.75). The middle region represents a pattern which cannot be accurately assessed based on the information available. Nearly all codon preference curve lies in the upper third of the graph. Much of the GC bias curve lies in the upper two-thirds of the graph; sometimes in the upper third coinciding with codon preference in select regions of the sequence. This data indicates that the coding sequence in the reverse frames is likely to contain preferred codons which occur in genes which are expressed. The boxes beneath each graph correspond to the open reading frame and the small vertical ticks below them represent the occurance of a "rare" codon relative to the frequency of the codons for that frame.


Codon Preference
Third Position GC Bias
F
E
D

NT735 x Humcoll

DEFINITION: Human collagen pro-α2(I) chain mRNA.

Percent Identity: 67.320 Length: 175 Gaps: 4

```
                .         .         .         .         .
395 GCTGCTGGTCTTGTCTGGGCTGCTGGTCCTGTCTGGGCTGCTGGT..... 351
    |||  ||||  |||  | ||  |||||||||    | ||  | |||
164 GCTCATGGTGCTGTAGGTGCCCCTGGTCCTGCTGGAGCCACAGGTGACCG 213

                .         .         .         .         .
350 ...CTATGTCTGGGCTGTTGGTCCTGTCTGGGCTGCTGGTCTATGTCTGG 304
       | | |   ||||||| ||||||||   |  |||||||||    || ||
214 GGGCGAAGCTGGGGCTGCTGGTCCTGCTGGTCCTGCTGGTC...CTCGGG 260

                .         .         .         .         .
303 G.....CTGTTGGTCTTGTCCAGG......CTGCTGGTCTTGTCCAGGCT 265
    |     ||| ||  | || | |||      |||||||  |    |     |
261 GAAGCCCTGGTGAACGTGGCGAGGTCGGTCCTGCTGGCCCCAACGGATTT 310

                .         .
264 GCTGGTCTTGCCTGGGCTGCTGGTC 240
    |||||||  ||  | ||||||||||
311 GCTGGTCCGGCTGGTGCTGCTGGTC 335
```

Figure 15. GCG "bestfit" comparisons of NT735 (top) to the Hf32 (bottom) region of HUMCOLL locus. Dashed vertical lines indicate nucleotide identity. Whole numbers indicate nucleotide position. Dots are placed at every tenth position above the sequence.

NT735 x Humcg1pa1

DEFINITION: Human pro-α1(I) chain of type I collagen mRNA.

Percent Identity: 64.151 Length: 122 Gaps: 3

```
                .         .         .         .         .
342  GGGCTGTTGGTCCTGTCTGGGCTGCTGGTCTATGTCTGGGCTGTTGGTCT 293
     | ||| :|||| || :  | ||  |||| |          | |||||  |
1865 GTGCTCNTGGTGCTCNTGGTGCCCCTGGCC..........CCGTTGGCCC 1904

                .         .         .         .         .
292  TGT...CCAGGCTGCTGGTCTTGTCCAGGCTGCTGGTCTTGCCTGGGCTG 246
     ||    | ||  || || || ||   ||    ||||| ||| |  | | |
1905 TGCTGGCAAGAGTGGTGATCGTGGTGAG...ACTGGTCCTGCTGGTCCCG 1951

                .         .
 245 CTGGTCCTGTCTGGGCTGCTGG 224
     | ||||| ||| |  | |||||
1952 CCGGTCCCGTCGGCCCCGCTGG 1973
```

Figure 16. GCG "bestfit" comparison between NT735 (top) and Hf677 region of Humcg1pa1 locus. Dashed vertical lines indicate nucleotide identity. Whole numbers indicate nucleotide position. Dots are placed at every tenth position above the sequence.

A-

DEFINITION Herpesvirus saimiri collagen-like protein
ACCESSION M31964
KEYWORDS collagen-like protein

42.5% identity in 40 aa overlap
88% similarity over 40 aa overlap

```
            120        130        140        150        160
Nt735dLLVLSGLLVLSGLLVYVWAVGPVWAAGLCLGCWSCPGCWSCPGCWSCLGCWSC
                                    |::  :|||:|:: |:|:: |:|:::|||:|::
HsvclsXNYDSKQXVCYKQSWQSXRSWRSWGSWRSLGSWRSLGSWRSLGSWRSLGSWRS
                   440        450        460        470        480

         170        180        190        200
Nt735dLGCWSFLGCWSMSGLLVLSRLLVFSGLLVFSGPAYLLWWHVGGLSGSCLLRLE
        ||:|: ||:|
HsvclsLGSWRSLGSWRSWRSLRSSISCFFSWISXVWLACHFLSSVFVLLLLAVSCVTL
               490        500        510        520        530
```

B-

```
NT735    ACAAGACCAGCAGCCCAG
           |   |||| | ||||
HSVCLS   GGACCTCCAGGACCCCAA
```

Figure 17. Sequence comparisons between NT735 and the collagen-like sequences of H. saimiri virus. A- Amino acid sequence comparison results from the GCG "Tfasta" subroutine. Dashed vertical lines (¦) indicate identity. Double dot vertical lines (:) indicate that the residues have similarity as conservative replacements. Whole numbers indicate amino acid position. B- Nucleotide sequence comparison between the 18 nucleotide repeats of NT735 and H. saimiri. Dashed vertical lines (¦) indicate identity. The two repeats have 56% identity.

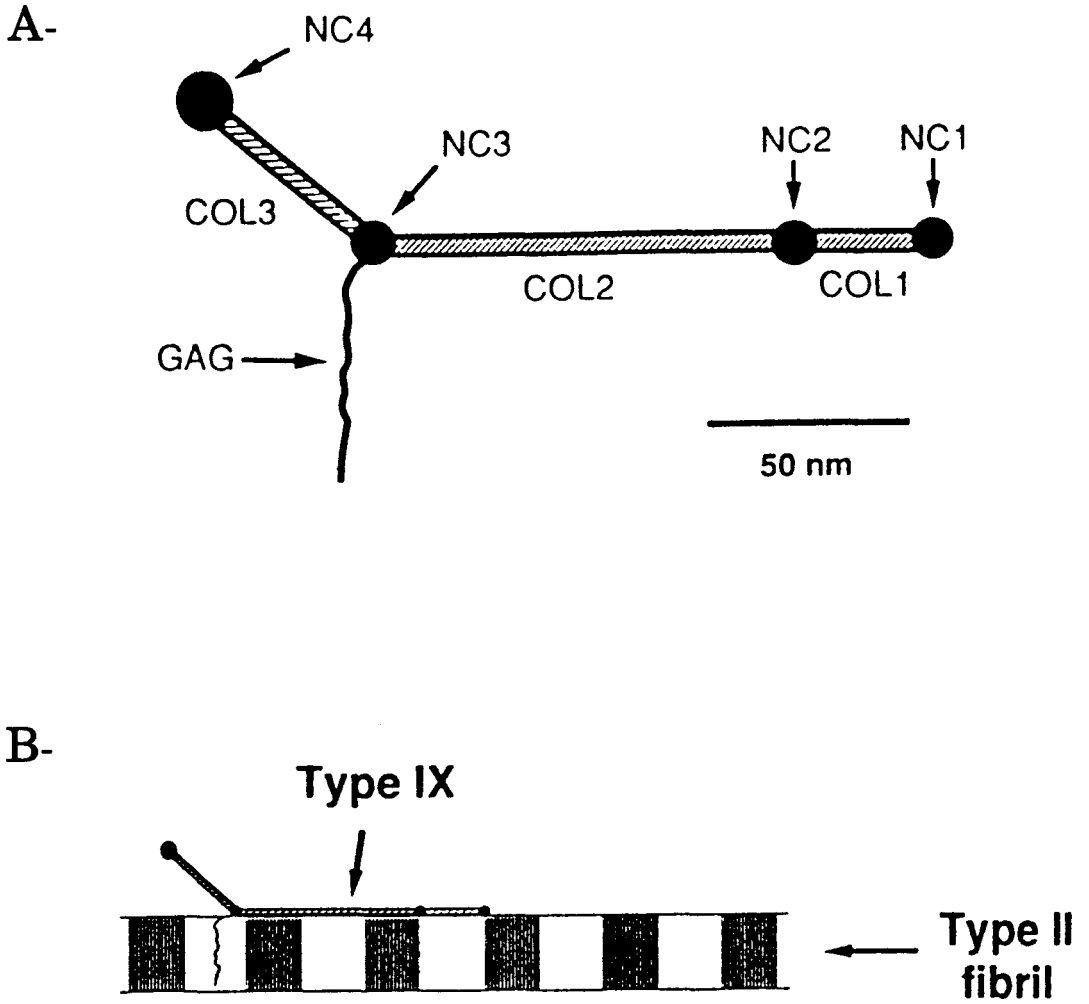


Figure 18. Schematic diagrams of the domains of the Type IX collagen molecule (A) and its alignment on the surface of the Type II fibril in cartilage (B). At the hinge region the COL3 and NC4 domains project into the matrix, while the GAG side chain associates in the gap region of the fibril. Reproduced from Biomaterials Vol. 11, p. 30 (van der Rest et al., 1990).

Table 1. Collagen genes with nucleotide similarity to NT735. This table summarizes those collagen genes identified with the GCG fasta program having 50% or greater similarity to NT735. Genes with less than a 100 nucleotide overlap are not included. The column labelled "Strand" indicates whether the similarity is in the forward (/for) or reverse (/rev) strand of NT735. In the "GAPS" column an asterisk (*) indicates a "fasta" comparison result, therefore no significant gaps are present. Fasta comparisons optimize the length of overlap and minimize gaps to only a few nucleotides, whereas "bestfit" comparisons optimize identity by finding that portion of the sequence with highest identity (forfeiting overlap length and inserting gaps). When two bestfit identities are reported for the same gene, they represent the same region of comparison with a different number of gaps allowed.

GENE TYPE	LOCUS	STRAND	%IDENTITY	OVERLAP (bp)	GAPS	DESCRIPTION
Human Collagens	HUMCOLL	/rev	67.32 52.0	175 690	4 *	α2(I) Collagen Locus of Hf32
	HUMCG1PA1	/rev	64.15 52.5	122 621	3 *	α1(I) Collagen Locus of Hf677
	HUMCOL16A	/for	67.6 74.6	179 97	6 3	α1(XVI) collagen Bone FACITs
Other Collagens	CELCOL	/for	54 63	550 164	* 4	C. Elegans cuticle collagen
	CHKCOL	/rev	54	402	*	Chicken α2(I) collagen
	XELCOL2A1	/rev	53	549	*	X. Laevis α1(II) collagen
	DROCOL	/for	59 56	131 308	1 *	Drosophila cuticle or B.M. collagen gene
	CHKCOL14A1	/for	54	387	*	Chicken α1(XIV) FACITs collagen
Type X collagen		/for	50-60	100-200	n/a	Multiple Species
Type IX collagen		/for	50-60	100-300	n/a	Multiple Species

Table 2. Non-collagen genes with nucleotide similarity to NT735. This table summarizes the nucleotide sequence comparison results for those non-collagen loci with high sequence similarity over long (>200 nt) stretches. The column labelled "STRAND" indicates whether the similarity is in the forward (/for) or reverse (/rev) strand of NT735. In the "gap" column an asterick (*) indicates a "fasta" comparison result, therefore no significant gaps are present. Fasta comparisons optimize the length of overlap and minimize gaps to only a few nucleotides, whereas "bestfit" comparisons optimize identity by finding that portion of the sequence with highest identity (forfeiting overlap length and inserting gaps). The column labelled "OPA" indicates if the gene does (Y=Yes) or does not (N=No) contain OPA repeats.

GENE	LOCUS	STRAND	%ID	OVERLAP	GAPS	DESCRIPTION	OPA
Wheat Glutinen	WHTGL1D1B	/for	66.6 60	639 557	12 4	Wheat equivalent of Elastin	N
Human Elastin	HSELASRNA	/rev	54.8 63.2	345 134	* 1	Human elastin gene -Hydrophobic exon	N
Drosophila Mastermind	DROMAS	/for	67.7	196	4	OPA repeat of D.virilis Mastermind gene	Y
Mouse OPA	MUSOPA	/for	60.9	417	*	OPA repeat of M.musculis	Y
Drosophila Prospero	DROPROS	/for	60	225	*	Large, nuclear, DNA binding protein of D.melanogaster	Y
Drosophila FSH	DROFSH	/for	53.7	566	*	D.melanogaster homeotic gene -Membrane prot.	Y
Drosophila Elav	DROELAVG	/for	59.2	255	*	D.virilis possible RNA binding protein	N

Table 3. Summary of proteins with similarity to the NT735 reverse translations. Comparisons were done using the "Swissprot" database. The region of similarity in the peptide translation of NT735 is indicated by frame and amino acid position. Comments regarding structural features of the identified proteins are limited to the regions with identity.

PROTEIN	Region of similarity within NT735 translation (amino acid position)	FRAME	PROTEIN FEATURES
Latent membrane protein of Epstein-Barr virus	10-40	E	Transmembrane region -Hydrophobic Viral membrane protein -Leucine rich
Kell G-protein	25-45	E	Corresponds to exact transmembrane region (30-45) of Kell -Same as in VGLM
Uukuniemi virus membrane glycoproteins G1 and G2	30-95	D	A.K.A. VGLM Viral lipo-membrane region adjacent to G2 -Hydrophobic
Rat heart Na+ channel isoform	40-80	D	Transmembrane region of Na+ channel
Chicken Elastin	150-180	F	Collagen-like nature -Hydrophobic region
Hepatitis C virus (HCV) envelope protein	135-185	E	Clustered glycosylation sites -Hydrophobic
X.laevis cystic fibrosis tranductance regulator	175-215	E	Overlaps two transmembrane regions and spans junction
Large regions of overlap:			
Latent membrane protein of Epstein Barr virus	66-208	D	Spans membrane five times! (25% ID, 52% Similarity)
Woodchuck Hepatitis virus Surface Antigen	100-210	D	Nearly entire Surface Ag chain! (23% ID, 57% Similarity)

DISCUSSION

This work involves the isolation and characterization of a new human collagen-like locus associated with osteogenesis imperfecta. The locus, represented by a 4.6 kb HindIII fragment, is located on chromosome 17 (Cammarata et al., 1991). The goal in this study was to isolate the 4.6 kb HindIII fragment and characterize the triple helical coding region within it. Restriction mapping and sequencing was the approach chosen as a first step to understanding what molecular organization might contribute to the deletions seen in OI patients at this site. Determining the detailed molecular organization of this locus is also the necessary initial step in characterizing the structure, and ultimately the function, of the putative protein for which the new locus codes.

Identification of Cloned Sequences

The initial hybridization studies (Figure 1) confirmed that sequences within this locus have similarity to triple helical coding regions of the COL1A2 gene. These sequences are distinct from those of the COL1A1 gene, which also resides on chromosome 17. Since there are no other reported

collagen genes on chromosome 17 (Byers, 1989), this locus clearly contains sequences of a new human collagen-like locus. These studies also demonstrate that the clone, p4.6, is more similar to helical coding sequences of COL1A2, than COL1A1. This information was later confirmed through direct sequence analysis.

Hybridization of the clone back to human genomic DNA confirmed that this locus does, in fact, represent the deletion-prone fragment previously shown to be associated with the OI phenotype (Figure 2).

Repeat Organization and Deletion Mechanisms

Although the presence of Alu direct repeat family sequences is not unusual in collagen genes (Barsh et al., 1981; Chu et al., 1984; Sandell and Boyd, 1990), their exceptionally high density in regions adjacent to the helical coding sequences of this new locus is noteworthy. Many instances have been described where Alu-Alu recombination has caused clinical pathologies due to the deletion of coding information (Lehrman et al., 1987, Markert et al., 1988). The evolutionary divergence of two related genes due to homologous recombination at Alu repeats has also been suggested (Kudo et al., 1989). Both intra- and inter-chromosomal unequal recombination and interchromatid unequal recombination have been suggested to contribute to polymorphisms seen at Alu sites (Vnencak-Jones et al., 1988).

Repetitive sequences with relatively short basic repeat units (2-25 nucleotides) can also be responsible for deletions which lead to length polymorphisms. There are reports of a region adjacent to the COL2A1 gene which exhibits a high degree of length polymorphism due to the variation in the number of copies of a simple tandemly repeating dinucleotide sequence. The tandem repeats which are adjacent to COL2A1 are suggested to cause the polymorphisms via a mechanism known as "slipped-strand mispairing" (Tiller et al., 1990; Sykes et al., 1985; Stoker et al., 1985). Such a mechanism can cause the looping out and excision of intervening DNA when one direct repeat misaligns with another during DNA replication. (Richards et al., 1992; Efstratiadis et al., 1980). This same event also occurs in some cases of Alu-Alu recombination except that the Alu repeats reside in the introns of a gene and the intervening or excised coding DNA does not necessarily contain a directly repeating motif (Lehrman et al., 1987).

The organization of the 18 nucleotide tandem repeat of NT735 is much like that of the direct repeats which cause polymorphisms seen in other collagen and non-collagen genes (Stoker et al., 1985; Richards et al., 1992; Zuliani et al., 1990). Since there is high identity between the repeats of NT735, and they extend for at least 600 base pairs, misalignment during DNA replication not only seems possible, but highly probable. It is therefore likely that this direct repeat organization is the cause of the length variations

seen at the Site 2 locus.

Recent studies in our lab have, in fact, mapped the deletion site in an OI Type III patient to the EcoRI/PvuII restriction fragment in the sequenced region containing the direct repeats (Michaels, unpublished). Future studies analyzing the molecular structure of this region in affected individuals will clarify the endpoints of those deletions which are, in fact, associated with the Type III OI phenotype.

Although the possibility may exist, the information obtained to this point is not consistent with the involvement of the Alu repeats as a cause of the deletions at this site.

The detection of 4.6 kb and 4.3 kb alleles of a normal individual, in addition to the single 4.4 kb deletion fragment of a homozygous, OI-affected individual (Figure 2) indicates that not all deletions at Site 2 result in an obvious OI phenotype. It is possible that individuals heterozygous for small deletions may be less affected, if at all, than individuals who are either heterozygous for larger deletions or homozygous for deletions at this locus. The specific location and/or extent of the deletion within Site 2 could also play a role in the expression of the OI phenotype.

Slipped-strand mispairing can also result in the duplication or amplification of repeated sequences. The amplification of direct repeats has been associated with other heritable disorders such as fragile X syndrome and

myotonic dystrophy (Caskey et al., 1992). In these disorders premutation alleles contain triplet-repeat sequences which, when amplified beyond a crucial threshold number, result in events that cause the expression of the disease (Caskey et al., 1992). Due to the same molecular mechanism (slipped-strand mispairing), the sequence amplification that results in these pathologies is the direct complement to the deletion-caused condition in p4.6 associated with Type III OI.

Not only can the slipped-strand mispairing mechanism contribute to detrimental events resulting from duplication or deletion but it can also serve a function in the evolution of repetitive sequence families and gene families which contain repetitive motifs (Levinson and Gutman, 1987; Kudo et al., 1989; Saitta et al., 1991). The structure of the higher order repeats in NT735 suggests that tandem duplication events are likely the mode of evolution in this gene. The nature of those duplications can be assessed by examining the relationship between similar repeats. Generally, those repeats with higher identity to each other are more closely related and more recently amplified than their counterparts of lesser identity (Carnahan et al., 1993). For example, in Figure 11b, the 137 nucleotide higher order repeats which occur at a location represented by nucleotide positions 291-429 and 430-566 have 92% identity, whereas the later repeats, positions 430-566 and 567-703 have 99% identity. This repeat organization indicates that the last two repeats, with 99%

identity, are more recently amplified than repeats earlier in the sequence. Thus, based on the extent of the identity between repeats, the likely evolutionary order of these sequences is from the 18 nt repeats (approximately 84% identity), to the more recent 110 nt (93% identity) intermediate. Finally, the 137 nt repeats with the highest identity (99%) must be most recently evolved.

An Expressed Collagen-Like Gene

From the hybridizations (Figures 1 and 4) and nucleotide sequence translations (Figure 13) it is clear that the Site 2 locus contains a collagen-like sequence with a repetitive motif. The level of GC richness (60%) seen in the NT735 sequence is similar to that of other collagen and connective tissue (ECM) genes. Analysis of potential open reading frames, codon preference, and GC bias (Figures 12 and 14) suggests that the reverse translation frames contain coding information likely to occur in a gene that is expressed. Recently, Northern blot and dot blot hybridizations of total RNA from normal human fibroblasts to p4.6 sequences confirmed that this locus is, in fact, transcribed (Breslin, unpublished). Future studies will determine the correct reading frame and the exact regions that are transcribed.

Abnormal Type I collagen incorporated into fibrils has a drastic effect on tissue integrity, resulting in the OI phenotype (Prockop et al., 1989). Since the Site 2 collagen-like gene is associated with the OI phenotype (Doering et

al., 1987), it is likely that defects in its protein product, which result from misalignment and deletion, will also contribute to the OI phenotype. Because the exact structure of the protein product has not yet been identified, it is difficult to say specifically how these deletions affect its structure.

Possible Gene Families and Protein Product

The lack of "pure" triple helical coding sequence (Figure 13), along with similarity to collagen genes with exons larger than the 54 bp primordial unit, and a high similarity to sequences of the FACITs Types IX, XIV and XVI (Table 1), all indicate that this gene may encode a non-fibrillar or fibril-associated (FACITs) molecule. Imperfections in the triplet pattern of non-fibrillar collagens are often the result of a missing glycine so that the amino acid sequence is GLY-X-Y-X-Y. These sorts of imperfections are seen in many of the FACITs collagens (Gordon et al., 1991). Such imperfections afford the otherwise rigid protein more flexibility at that site. Also characterized by triplet sequence interruptions, the non-collagenous (NC) domains of the FACITs often contain few, if any, glycine residues. The pattern in the Site 2 collagen gene is also an interrupted triplet sequence which contains a substitution for a glycine rather than a missing glycine so that the amino acid sequence (in some regions) is Gly-X-Y-Z-X-Y (see Figure 13). This pattern also suggests that this

molecule has the potential for flexibility while still maintaining a collagen-like α-helical structure.

The nucleotide similarity between NT735 and all three of the Type IX genes is in the region coding for the COL3 domain of the Type IX molecule (Table 1). This region is adjacent to the "flexible hinge" which allows this domain and the globular NC4 domain to protrude into the perifibrillar matrix (Figure 18). This flexibility imparts a unique, interactive capacity to the Type IX collagen molecule. The arrangement is ideal if the NC4 and GAG domains (see Figure 18) are to facilitate an interactive function with other cellular or extracellular matrix components. Although COL3 is collagenous in nature, it does not align in parallel with the Type I fibril. Thus, though a somewhat rigid extension is needed in order to project the NC4 domain into the matrix, a strict triple helix is less crucial, and perhaps in some cases impracticable, in this kind of region. The structure of the putative protein coded by p4.6 may serve a similar function, since it too has the potential flexibility of Type IX collagen, while maintaining a somewhat rigid form.

Also adjacent to the COL3 domain of Type IX collagen, at the site of the hinge, is the GAG side chain attachment site (see Figure 18). This is the domain that imparts the proteoglycan characteristic and function to the otherwise collagenous nature of the Type IX molecule. It is this molecular architecture that makes the COL3 domain a bridge between the interactive components of the Type IX molecule.

It is possible that the same kind of region in p4.6 may serve a similar function.

The NT735 translation (reverse frames) indicate other features in common with Type IX collagen and other FACITs, such as a high serine and leucine content and the abundance of cysteine residues (Ninomiya et al., 1990). Cysteine residues allow inter- and intra-molecular crosslinking of the collagen polypeptide and are important in determining the structure of the terminal globular domains of the Type IX molecules (Ninomiya et al., 1985). The abundance of these residues suggests that the p4.6 gene product has the capacity for forming other higher order structures in addition to a typical α-helix.

In addition to its common features with the collagens and their genes, the high sequence identity of NT735 to the elastin-like wheat glutinen genes (Table 2), suggests the possibility that NT735 codes for an ECM gene which contributes tissue strength through an elastic type of flexibility. The sequence identity to human elastin (Table 2) supports this conclusion. The human elastin gene, like p4.6, contains high numbers of glycine and other hydrophobic residues as well as an abundance of Alu repetitive sequences (Indik et al, 1990). Two elastin exons with similarity to NT735 code for very hydrophobic regions of the molecule (Tatham et al., 1984). This agrees with codon frequency data indicating that NT735 codes a highly hydrophobic peptide.

Amino acid sequence comparisons using the "Swissprot"

database indicate a nearly exclusive similarity of the NT735 translation (reverse frames) with proteins containing hydrophobic transmembrane domains (Table 3). Much of this identity is with viral envelope glycoproteins and other transmembrane or cell/viral surface proteins. For example, there are long stretches of similarity to Hepatitis C virus envelope glycoprotein, Woodchuck hepatitis virus surface antigen, latent membrane protein of Epstein-Barr virus, and a shorter stretch of the Uukuniemi virus transmembrane portion of the G_1 and G_2 membrane glycoproteins. There are also multiple stretches of similarity to other transmembrane or membrane associated proteins such as the sodium channel of the rat heart (Cinc), the cystic fibrosis transmembrane conductance regulator of X. laevis, which has 77% identity to the human CF gene, and the surface exposed human Kell blood group glycoprotein.

Most of these proteins are membrane-associated glycoproteins containing hydrophobic, α-helical transmembrane regions rich in the nonpolar amino acids Gly, Ala, Leu, Ilu, Val and Arg; the same features identified in the the putative peptide of NT735. Some are involved in membrane transport, for example, Na^+ transport system and the CFTR channel activity (Rogart et al., 1989; Tucker et al., 1992). Others are viral envelope glycoproteins with effector functions (Moorthy et al., 1990, Kato et al., 1990; Kato et al., 1991), and yet others are transmembrane glycoproteins with a

structural and/or interactive role in matrix association (Ronnholm and Petterson, 1987).

Many transmembrane glycoproteins are multidomain molecules also containing cytoplasmic and extracellular extensions with various functions (Nicholls, 1978). Other ECM associated receptor molecules have been described which contain such an architecture. For example, the ECM-associated membrane glycoproteins Syndecan and CD44 are ligand binding molecules which contain large interactive extracellular extensions anchored by an α-helical transmembrane domain which crosses the cell membrane multiple times (Sandell and Boyd, 1990). A short cytoplasmic domain may additionally associate intracellularly with the cytoskeleton. Both of these molecules are able to bind the intersitial collagens as well as other smaller ECM molecules (Jalkanen et al., 1991).

The above sequence comparisons have identified high similarity of the NT735 sequence and its structural characteristics to a number of ECM components and transmembrane glycoproteins, which are important in directing the composition and structure of the extracellular matrix and effecting cellular interaction with the matrix (Alberts et al., 1989). This supports a function in which the putative protein encoded by the p4.6 locus could modulate the interaction of Type I collagen fibrils with the cell by acting as a communication bridge between them (Shaw and Olsen, 1991; Gordon et al., 1989). This is possible via a signaling function which directs the secretion of ECM

components or in a structural capacity, somehow connecting or associating the major fibril with other cell-secreted ECM components (Vaughan et al, 1988; Muller-Glauser et al., 1986).

If the p4.6 gene product does modulate Type I fibril interactions, then it has the potential to affect the strength and structural integrity of Type I collagen-containing tissue. Thus, defects in p4.6 which disrupt this interactive capacity, could in turn effect the expression of the OI phenotype where either a Type I collagen defect is or is not present. For example, there are many cases described where substitutions for glycine in similar positions of the same polypeptide chain can produce drastic differences in severity between individuals. It is possible that a second mutation, which modulates the effect of a Type I collagen mutation, could cause a more severe condition, such as that seen in Type III OI. Alternatively, a single defect at this new collagen-like locus, could itself result in the Type III OI phenotype.

RFLPs that are linked to the Type I procollagen genes have been shown to segregate with many of the autosomal dominant forms of OI (Sykes et al., 1990; Superti-Furga et al., 1989). However, the fact that multiple individuals carrying the same Type I collagen defect exhibit substantial variability in the severity of their OI clinical symptoms, even being asymptomatic, complicates diagnosis and genetic counseling. This situation can be greatly improved by the

detailed knowledge of other connective tissue genes, such as the p4.6 locus, that might modulate the OI phenotype.

Sequencing of this "normal" p4.6 allele provides the basis for examining the structure of deleted alleles in the normal population in comparison to those of asymptomatic parents and their affected offspring. This sequence information will permit pinpointing the specific deletion boundaries in OI patients which may suggest the molecular basis for the deletions. Ultimately, further characterization of this locus might permit its use as a marker for individuals predisposed to having a Type III OI child.

Determining the detailed molecular organization of this locus has also been the necessary initial step in characterizing the structure, and ultimately the function, of the putative protein(s) for which the new locus codes. Whatever its product, whether it be a true structural collagen, a related ECM protein, or a transmembrane glycoprotein with extracellular extensions that influences matrix assembly, this new locus is of importance since previous studies of OI patients and their families demonstrate that deletions at this locus are clearly associated with the disease.

REFERENCES

Aitchison, K., Ogilvie, D., Honeyman, M., Thompson, E. and Sykes, B. (1988). Homozygous osteogenesis imperfecta unlinked to collagen I genes. Hum. Genet. 78,233-236.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D. (1989). Molecular Biology of the Cell. (Garland Publishing Inc., New York). pp. 577-584.

Apte, S., Seldin, M., Hayashi, M. and Olsen, B. (1992). Cloning of the human and mouse Type X collagen genes and mapping of the mouse Type X collagen gene to chromosome 10. Eur. J. Biochem. 206,217-224.

Aubert-Foucher, E., Font, B., Eichenberger, D., Goldschmidt, D., Lethias, C. and van der Rest, M. (1992). Purification and characterization of native Type XIV collagen. J. Biol. Chem. 267,15759-15764.

Baldwin, C., Constantinou, C., Dumars, K. and Prockop, D. (1989). A single base mutation that converts glycine 907 of the α2(I) chain of Type I procollagen to aspartate in a lethal variant of osteogenesis imperfecta. The single amino acid substitution near the carboxy terminus J. Biol. Chem. 264,3002-3006.

Barsh, G. and Byers, P. (1981). Reduced secretion of structurally abnormal Type I procollagen in a form of osteogenesis imperfecta. Proc. Natl. Acad. Sci. USA 78,5142-5146.

Barsh, G., Roush, C., Bonadio, J., Byers, P. and Gelinas, R. (1985). Intron-mediated recombination may cause a deletion in an α1 Type I collagen chain in a lethal form of osteogenesis imperfecta. Proc. Natl. Acad. Sci. USA 82,2870-2874.

Bashey, R.I., Martinez-Hernandez, A., Jimenez, S.A. (1992). Isolation, Characterization, and localization of cardiac carilage collagen Type VI: associations with other extracellular matrix components. Circulation Research. 70,1006-1017.

Bateman, J., Chan, D., Lamande, S., Mascara, T. and Cole, W. (1986). Collagen defects in lethal perinatal

osteogenesis imperfecta. Biochem. J. 240,699-708.

Bateman, J., Chan, D., Lamande, S., Mascara, T. and Cole, W. (1988). Biochemical heterogeneity of Type I collagen mutations in osteogenesis imperfecta. Ann. N.Y. Acad. Sci. 543,95-105.

Bateman, J., Chan, D., Walker, I., Rogers, J. and Cole, W. (1987). Lethal perinatal osteogenesis imperfecta due to the substitution of arginine for glycine at residue 391 of the α1(I) chain of Type I collagen. J. Biol. Chem. 262,7021-7027.

Bateman, J., Lamande, S., Dahl, H.-H., Chain, D., Mascara, T. and Cole, W. (1989). A frameshift mutation results in a truncated nonfunctional carboxylterminal pro-α1(I) propeptide of Type I collagen in osteogenesis imperfecta. J. Biol. Chem. 264,10960-10964.

Beighton, P., Wallis, G., Viljoen, D. and Versfeld, G. (1988). Osteogenesis imperfecta in Southern Africa. Diagnostic categorization and biomolecular findings. Ann. N.Y. Acad. Sci. 543,40-46.

Benton, W. and Davis, R. (1977). Screening lambda gt recombinant clones by hybridization to single plaques in situ. Science 196,180-182.

Bernard, M., Myers, J., Chu, M.-L., Ramirez, F., Eikenberry, E. and Prockop, D. (1983). Structure of a cDNA for the pro-α2 chain of human Type I procollagen. Comparison with chick cDNA for pro-αI(I) identifies structurally conserved features of the protein and the gene. Biochemistry 22,1139-1145.

Bernard, M., Yoshioka, M., Rodriguez, E., van der Rest, M., Kimura, T., Ninomiya, Y., Olsen, B.R., and Ramirez, F. (1988). Cloning and sequencing of pro-α1(XI) reveals that the expression of the gene is not restricted to cartilagenous tissue. J. Biol. Chem. 263,17159-17166.

Birk, D., Fitch, J., Babiarz, J. and Linsenmayer, T. (1988). Collagen Type I and Type V are present in the same fibril in the avian corneal stroma. J. Cell. Biol. 106,999-1008.

Bonadio, J. and Byers, P. (1985). Subtle structural alterations in the chains of Type I procollagen produce osteogenesis imperfecta type II. Nature 316,363-366.

Bonadio, J., Ramirez, F. and Barr, M. (1990a). An intron mutation in the human α1(I) collagen gene alters the

efficiency of pre-mRNA splicing and is associated with osteogenesis imperfecta type II. J. Biol. Chem. 265,2262-2268.

Bonadio, J., Saunders, T., Tsai, E., Goldstein, S., Morris-Wiman, J., Brinkley, L., Dolan, D., Alschuler, R., Hawkins, J. and Bateman, J., et al. (1990b). Transgenic mouse model of the mild dominant form of osteogenesis imperfecta. Proc. Natl. Acad. Sci. USA 87,7145-7149.

Brewton, R., Ouspenskaia, M., van der Rest, M. and Mayne, R. (1992). Cloning of the chicken α3(IX) collagen chain completes the primary structure of Type IX collagen. Eur. J. Biochem. 205,443-449.

Bruns, R. and Gross, J. (1974). High resolution analysis of the modified quarter-staggered model of the collagen fibril. Biopolymers 13,931-941.

Byers, P. (1989). Inherited disorders of collagen gene structure. Am. J. Med. Genet. 34,72-80.

Byers, P. (1990). Brittle bones- fragile molecules: disorders of collagen gene structure and expression. TIG 6,293-300.

Byers, P., Bonadio, J., Cohn, D., Starman, B., Wenstrup, R.J. and Willing, M.C. (1988a). Osteogenesis imperfecta: The molecular basis of clinical heterogeneity. Ann. N.Y. Acad. Sci. 543,117-128.

Byers, P., Bonadio, J. and Steinmann, B. (1984). Osteogenesis Imperfecta: Update and Perspective. Am. J. Med. Genet. 179,429-435.

Byers, P., Shapiro, J., Rowe, D., David, K. and Holbrook, K. (1983). Abnormal α2-Chain in Type I collagen form a patient with a form of osteogenesis imperfecta. J. Clin. Invest. 71,689-697.

Byers, P., Starman, B., Cohn, D. and Horwitz, A. (1988b). A novel mutation causes a perinatal lethal form of osteogenesis imperfecta. An isertion in one α1(I) collagen allele of COL1A1. J. Biol. Chem. 263,7855-7861.

Byers, P. and Steiner, R. (1992). Osteogenesis Imperfecta. Annu. Rev. Med. 43,269-82.

Byers, P., Wenstrup, R., Bonadio, J., Starman, B. and Cohn, D. (1987). Molecular basis of inherited disorders of collagen biosynthesis: Implications for prenatal

diagnosis. Curr. Probl. Derm. 16,158-174.

Cammarata, S.A., Burket, A.E. and Doering, J.L. (1991). Characterization of a new human collagen locus associated with osteogenesis imperfecta. J. Cell. Biol. 115,106a

Carnahan, S., Palamidis-Bourtsos, E., Musich, P. and Doering, J. (1993). Characterization of an evolutionarily old human alphoid DNA. Gene 123,219-225.

Caskey, C.T., Pizzuti, A., Fu, Y-H., Fenwick, R.G., Nelson, D.L. (1992). Triplet repeat mutations in human disease. Science 256,784-789.

Chu, M.-L., de Wet, W., Bernard, M., Ding, J.-F., Morabito, M., Myers, J., Williams, C. and Ramirez, F. (1984). Human pro-α1(I) collagen gene structure reveals evolutionary conservation of a pattern of introns and exons. Nature 310,337-340.

Chu, M.-L., de Wet, W., Bernard, M. and Ramirez, F. (1985a). Fine structural analysis of the human pro-α1(I) collagen gene. J. Biol. Chem. 260,2315-2320.

Chu, M.-L., Gargiulo, V., Williams, C.J. and Ramirez, F. (1985b). Multiexon deletion in an osteogenesis imperfecta variant with increased Type III collagen mRNA. J. Biol. Chem. 260,691-694.

Chung, E., Rhodes, K. and Miller, E. (1976). Isolation of three collagenous components of probable basement membrane origin from several tissues. Biochem. Biophys. Res. Commun. 71,1167-1174.

Cohn, D. and Byers, P. (1991). Cysteine in the triple helical domain of the pro-α2(I) chain of Type-I collagen in nonlethal forms of osteogenesis imperfecta. Human Genetics 87,167-172.

Cohn, D., Starman, B., Blumberg, B. and Byers, P. (1990). Recurrence of lethal osteogenesis imperfecta due to parental mosaicism for a dominant mutation in a human Type I collagen gene (COL1A1). Am. J. Hum. Genet. 46,591-601.

Cohn, D., Wenstrup, R. and Willing, M. (1988). General strategies for isolating the genes encoding Type I collagen and for characterizing mutations which produce osteogenesis imperfecta. Ann. N.Y. Acad. Sci. 543, 129-135.

Cole, W., Jaenisch, R. and Bateman, J. (1989). New insights into the molecular pathology of osteogenesis imperfecta. O. J. Med. 261,1-4.

Cole, W., Patterson, E., Bonadio, J., Campbell, P. and Fortune, D. (1992). The clinicopathological features of three babies with osteogenesis imperfecta resulting from the substitution of glycine by valine in the pro-α1 chain of T.ype I procollagen. J. Med. Genet. 29,112-118.

Constantinou, C., Pack, M., Young, S. and Prockop, D. (1990). Phenotypic heterogeneity in osteogenesis imperfecta: The mildly affected mother of a proband with a lethal variant has the same mutation substituting cysteine for a1-glycine 904 in a Type I collagen gene. Am. J. Hum. Genet. 47,670-679.

Deininger, P., Jolly, D., Rubin, C., Friedmann, T. and Schmid, C. (1981). Base sequence of 300 nucleotide renatured repeated human DNA clones. J. Mol. Biol. 151,17-33.

Devereux, J., Haeberli, P. and Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12,387-395.

de Wet, W., Bernard, M., Benson-Chanda, V., Chu, M.-L., Dickson, L., Weil, D. and Ramirez, F. (1987). Organization of the human pro-α2(I) collagen gene. J. Biol. Chem. 26,16032-16036.

Doering, J. (1977). The structure of *X. borealis* oocyte and somatic 5s DNAs. Carnegie Inst. Wash. Yrbk. 76,102-105.

Doering, J.L., Burket, A.E. and Vogel, L.C. (1987). Collagen gene deletions in osteogenesis imperfecta patients. J. Cell Biol. 105,213a.

Doering, J.L., Burket, A.E. and Vogel, L.C. (1990). Collagen gene deletions associated with osteogenesis imperfecta. First regional meeting, The American Society for Cell Biology. 96a.

Doering, J.L., Jelachich, M.L. and Hanlon, K.M. (1982). Identification and genomic organization of human tRNAlys genes. FEBS Lett. 146,47-51.

Dublet, B., Sugrue, S.B., Gordon, M.K., Gerecke, D.R., Olsen, B.R. and van der Rest, M. (1989). The structure of avian Type XII collagen. α1(XII) chains contain 190 kDa non-triple-helical amino-terminal domains and form

homotrimeric molecules. J. Biol. Chem. 22,13150-13156.

Dublet, B. and van der Rest, M. (1987). Type XII collagen is expressed in embryonic chick tendons. J. Biol. Chem. 262,17724-17727.

Dublet, B. and van der Rest, M. (1990). Comparison between chicken type XII collagen and bovine homologues. Ann. N.Y. Acad. Sci. 580,436-439.

Dublet, B. and van der Rest, M. (1991). Type XIV Collagen, a New Homotrimeric Molecule Extracted from Fetal Bovine Skin and Tendon, with a Triple Helical Disulfide-bonded Domain Homologous to Type IX and Type XII Collagens. J. Biol. Chem. 266,6853-6858.

Efstratiadis, A., Posakony, J.W., Maniatis, T. (1980). The structure and evolution of the human ß-Globin gene family. Cell. 21,653-668.

Eyre, D. and Wu, J.-J. (1987) Type XI or 1α2α3α collagen. In Structure and Function of Collagen Types. Mayne, R. and Burgeson, R. eds. (Academic Press, New York). pp. 261-281.

Falk, C.T., Schwartz, R.C., Ramirez, F. and Tsipouras, P. (1986). Use of molecular haplotypes specific for the human pro-α2(I) collagen gene in linkage analysis of the mild autosomal dominant forms of osteogenesis imperfecta. Am. J. Hum. Genet. 38,269-279.

Fazio, M., Olsen, D., Kauh, E., Baldwin, C., Indik, Z., Ornstein-Goldstein, N., Yeh, H., Rosenbloom, J. and Uitto, J. (1988). cloning of full-length elastin cDNAs from a human skin fibroblast recombinant cDNA library: Further elucidation of alternative splicing utilizing exon-specific oligonucleotides. Soc. Invest. Derm. 91,458-464.

Fessler, J. and Fessler, L. (1987) Type V collagen. In Structure and Function of Collagen Types. Mayne, R. and Burgeson, R., eds. (Academic Press, New York). pp. 81-97.

Fitch, J., Gross, J., Mayne, R., Johnson-Wint, B. and Linsenmayer, T. (1984). Organization of collagen Type I and Type II in the embryonic chicken cornea: Monoclonal antibody studies. Proc. Natl. Acad. Sci. USA 81,2791-2795.

Forde, J., Malpica, J-M., Halford, N.G., Shewry, P.R., Anderson, O.D., Greene, F.C., Miflin, B.J. (1985). The

nucleotide sequence of a HMW glutenin subunit gene located on chromosome 1A of wheat (*Triticum aestivum L.*). Nucleic Acids Res. 13,6817-6833.

Genovese, C., Brufsky, A., Shapiro, J. and Rowe, D. (1989). Detection of mutations in human Type I collagen mRNA in osteogenesis imperfecta by indirect RNase protection. J. Biol. Chem. 264,9632-9637.

Gordon, M., Castagnola, P., Dublet, B., Linsenmayer, T., van der Rest, M., Mayne, R. and Olsen, B. (1991). Cloning of a cDNA for a new member of the class of fibril-associated collagens with interrupted triple helices. Eur. J. Biochem. 201,333-338.

Gordon, M., Gerecke, D., Dublet, B., van der Rest, M., Sugrue, S.P. and Olsen B.R. (1990). The Structure of Type XII Collagen. Ann. N.Y. Acad. Sci. 580,8-16.

Gordon, M., Gerecke, D., Nishimura, I., Ninomiya, Y. and Olsen, B. (1989). A new dimension in the extracellular matrix. Conn. Tissue Res. 20,179-186.

Gordon, M., Gerecke, D. and Olsen, F. (1987). Type XII collagen: Distinct extracellular matrix component discovered by cDNA cloning. Proc. Natl. Acad. Sci. USA 84,6040-6044.

Hoffman, H., Voss, T. and Kuhn, K. (1984). Localization of flexible sites in thread-like molecules from electron micrographs: Comparison of interstitial, basement membrane and intima collagens. J. Mol. Biol. 172,325-343.

Indik, Z., Yeh, H., Ornstein-Goldstein, N. and Rosenbloom, J. (1990) Structure of the elastin gene and alternative splicing of elastin mRNA. In Extracellular Matrix Genes. Sandell, L.J. and Boyd, C.D. eds. (Academic Press, New York, NY). pp. 221-250.

Jalkanen, M., Jalddanen, S. and Bernfield, M. (1991). Binding of Extracelllular Effector molecules by cell surface proteoglycans. In: Receptors for Extracellular Matrix. McDonald, JA and Mecham, RP, eds. (Academic Press, New York). pp. 1-38.

Karkavelas, G., Kefalides, N., Amenta, P. and Martinez-Hernandez, A. (1988). Comparative ultrastructural localization of collagen Types III, IV, VI and laminin in rat uterus and kidney. J. Ultrastruct. Mol. Struct. Res. 100,137-155.

Kato, N., Hijikata, M., Nakagawa, M., Ootsuyama, Y., Muraiso, K., Ohkoshi, S. and Shimotohno, K. (1991). Molecular structure of the Japanese hepatitis C viral genome FEBS Lett. 280,325-328.

Kato, N., Hijikata, M., Ootsuyama, Y., Nakagawa, M., Ohkoshi, S., Sugimura, T. and Shimotohno, K. (1990). Molecular cloning of the human hepatitis C virus genome from Japanese patients with non-A, non-B hepatitis. Proc. Natl. Acad. Sci. USA 87,9524-95528.

Keene, D., Lunstrum, G., Morris, N., Stoddard, D. and Burgeson, R. (1991). Two Type XII-like collagens to the surface of banded collagen fibrils. J. Cell. Biol. 113,971-978.

Kozak, K., Foster, L. and Ross, I. (1991). A method for transferring sequencing gels from glass to absorbent filter paper. BioTechniques 11,54

Kudo, S. and Fukuda, M. (1989). Structural organization of glycophorin A and B genes: Glycophorin B evolved by homologous recombination at Alu repeat sequences. Proc. Natl. Acad. Sci. USA 86,4619-4623.

Kuhn, K. (1987). The classical collagens: Types I, II, and III. In Structure and Function of Collagen Types. Mayne, R. and Burgeson, R.E., eds. (Academic Press, New York). pp. 1-42.

Kuivaniemi, H., Sabol, C., Tromp, G., Sippola-Thiele, M. and Prockop, D. (1988). A 19-base pair deletion in the pro-α2(I) gene of Type I procollagen that causes in-frame RNA splicing from exon 10 to exon 12 in a proband with atypical osteogenesis imperfecta and in his asymptomatomatic mother. J. Biol. Chem. 263,11407-11413.

Kuivaniemi, H., Tromp, G. and Prockop, D. (1991). Mutations in collagen genes: causes of rare and some common diseases in humans. FASEB 5,2052-2060.

Lamande, S., Dahl, H., Cole, W. and Bateman, J. (1989). Characterization of point mutations in the collagen COL1A1 and COL1A2 genes causing lethal perinatal osteogenesis imperfecta. J. Biol. Chem. 264,15809-15812.

Lehrman, M., Russell, D.G., JL, B. and MS, (1987). Alu-Alu recombination delete splice acceptor sites and produces secreted low density lipoprotein receptor in a subject with familial hypercholesterolemia. J. Biol. Chem. 262,3354-3361.

Levinson, G. and Gutman, G.A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. 4,203-221

Maniatis, T., Fritsch, E. and Sambrook, J. (1982) Molecular Cloning. A Laboratory Manual. (Cold Spring Harbor Press, Cold Spring Harbor, NY). pp. 371-372.

Markert, M., Hutton, J., Winginton, D., States, J. and Kaufman, R. (1988). Adenosine Deaminase (ADA) deficience due to deletion of the ADA gene between two Alu elements. J. Clin. Invest. 81,1323-1327.

Martinez-Hernandez, A., Gay, S. and Miller, E. (1982). Ultrastructural localization of Type V collagen in rat kidney. J. Cell Biol. 92,343-349.

Mayne, R. and Burgeson, R. (1987). Structure and Function of Collagen Types. Mayne, R. and Burgeson, R.E., eds. (Academic Press, New York). 58-100.

Miller, E. and Gay, S. (1987) Collagen: An overview. In Methods in Enzymology. L.W. Cunningham, ed. (Academic Press, New York). pp. 3-41.

M^{c}Conaghy, B.L., Laird, C.D. and M^{c}Carthy, B.J. (1969). Nucleic acid reassociation in formamide. J. Biochem. 8,3289-3295.

Moorthy, R. and Thorley-Lawson, D.A. (1990). Processing of the epstein-barr virus-encoded latent membrane protein p63/LMP. J. Virol. 64,829-837.

Muller-Glauser, W., Humbel, B., Glatt, M., Strauli, P. and Winterhalter, K. (1986). On the role of Type IX collagen in the extracellular matrix of cartilage: Type IX collagen is localized to intersections of collagen fibrils. J. Cell Biol. 102,1931-1939.

Muragaki, Y., Kimura, T., Ninomiya, Y. and Olsen, B. (1990). The complete primary structure of two distinct forms of human α1(IX) collagen chains. Eur. J. Biochem. 192,703-708.

Myers, J.C., Chu, M.-L., Faro, S.H., Clark, W.J., Prockop, D.J. and Ramirez, F. (1981). Cloning a cDNA for the pro-α2 chain of human Type I collagen. Proc. Natl. Acad. Sci. USA 78,3516-3520.

Nicholls, P. (1978). Membrane Proteins. (Pergamon Press, New York) 45, pp. 25-30.

Ninomiya, Y., Castagnola, P., Gerecke, D., Gordon, M., Jacenko, O., LuValle, P., McCarthy, M., Muragaki, Y., Nishimura, I., Oh, S., Rosenblum, N., Sato, N., et al. (1990). The molecular biology of collagens with short triple-helical domains. In Extracellular Matrix Genes. Sandell, L.J. and Boyd, C.D. eds. (Academic Press, New York). pp. 80-114.

Ninomiya, Y., van der Rest, M., Mayne, R., Lozano, G. and Olsen, B. (1985). Construction and characterization of cDNA encoding the α2 chain of chick Type IX collagen. Biochemistry 24,4223-4229.

Noro, A., Kimata, K., Oike, Y., Shinomura, T., Maeda, N., Yano, S., Takahashi, N. and Suzuki, S. (1983). Isolation and characterization of a third proteoglycan (PG-Lt) from chick embryo cartilage which contains disulfide bonded collagenous polypeptide. J. Biol. Chem. 258,9323-9331.

Olsen, B., Ninomiya, Y., Lozano, G., Konomi, H., Gordon, M., Green, G., Parsons, J., Seyer, J., Thomson, H. and Vasios, G. (1985). Short-chain collagen genes and their expression in cartilage. Ann. N.Y. Acad. Sci. 460,141-153.

Olsen, B., Nishimura, I., Gordon, M., Gerecke, D., Muragaki, Y. and Ninomiya, Y. (1988). The fibril associated collagens with interrupted triple-helices (FACITs). J. Cell Biol. 107,232a.

Pack, M., Constantinou, D., Kalia, K., Nielsen, K. and Prockop, D. (1989). Substitution of serine for α1(I)-glycine 844 in a severe variant of osteogenesis imperfecta minimally destabilizes the triple helix of Type I procollagen. J. Biol. Chem. 264,19694-19699.

Pan, T., Zhang, R., Mattei, M., Timpl, R. and Chu, M. (1992). Cloning and chromosomal location of human α1(XVI) collagen. Proc. Natl. Acad. Sci. USA 89,6565-6569.

Patterson, E., Smiley, E. and Bonadio, J. (1989). RNA Sequence Analysis of a Perinatal Lethal Osteogenesis Imperfecta Mutation. J. Biol. Chem. 264,10083-10087.

Peterson, R.C., Doering, J.L. and Brown, D.D. (1980). The characterization of two Xenopus somatic 5s DNAs and one minor oocyte-specific 5s DNA. Cell 20,131-141.

Pihlajaniemi, T., Dickson, L.A., Pope, F.M., Korhonen, V.R., Nicholls, A., J., P.D. and Myers, J.C. (1984).

Osteogenesis imperfecta: Cloning of a pro-α2 I collagen gene with a frameshift mutation. J. Biol. Chem. 259,12941-12944.

Pihlajaniemi, T., Myllyla, R., Seyer, J., Kurkinen, M. and Prockop, D. (1987). Partial characterization of a low molecular weight human collagen that undergoes alternative splicing. Proc. Natl. Acad. Sci. USA 84,940-944.

Pihlajaniemi, T. and Tamminen, M. (1990). The α1 chain of Type XIII collagen consists of three collagenous and four noncollagenous domains, and its primary transcript undergoes complex alternative splicing. J. Biol. Chem. 265,16922-16928.

Pope, F., Daw, S., Narcisi, P., Richards, A. and Nicholls, A. (1989). Prenatal diagnosis and prevention of inherited abnormalities of collagen. Inher. Metab. Dis. 121,135-173.

Pope, F., Nicholls, A., McPheat, J., Talmud, P. and Owen, R. (1985). Collagen genes and proteins in osteogenesis imperfecta. J. Med. Genet. 22,466-478.

Prockop, D. (1984). Osteogenesis Imperfecta: Phenotypic heterogeneity, protein suicide, short and long collagen. Am. J. Hum. Genet. 34,60-67.

Prockop, D., Constantinou, D., Dombrowski, K., Hojuma, Y., Kadler, K., Kuivaniemi, H., Tromp, G. and Vogel, B. (1989). Type I procollagen: The gene-protein system that harbors most of the mutations causing osteogenesis imperfecta and probably more common heritable disorders of connective tissue. Am. J. Med. Genet. 34,60-67.

Prockop, D.J. and Kivirikko, K.I. (1984). Heritable diseases of collagen. N. Engl. J. Med. 311,376-386.

Ramirez, F., Bernard, M., Chu, M., Dickson, L., Sangiorgi, F., Weil, D., de Wet, W., Junien, C. and Sobel, M. (1985). Isolation and characterization of human fibrillar collagen genes. Ann. N.Y. Acad. Sci. 4460,117-129.

Reed, K.C. and Mann, D.A. (1985). Rapid transfer of DNA from agarose gels to nylon membranes. Nucleic Acids. Res. 13,7207.

Richards, A.J., Lloyd, J.C., Narcisi, P., Nicholls, S.C., DePaepe, A. and Pope, F.M. (1992). A 27-bp deletion from one allele of the Type III collagen gene (COL3A1) in a

large family with Ehlers-Danlos syndrome Type IV. Hum. Genet. 88,325-330.

Rigby, P.W.J., Diekmann, W., Rhodes, C. and Berg, P. (1977). Labelling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. J. Mol. Biol. 113,237-251.

Rogart, R.B., Cribbs, L.L., Muglia, L.K., Kephart, D.D. and Kaiser, M.W. (1989). Molecular cloning of a putative tetrodotoxin-resistant rat heart Na+ channel isoform. Proc. Natl. Acad. Sci. USA 86, 8170-8174.

Ronnholm, R. and Petterson, R.F. (1987). Complete Nucleotide Sequence of the mRNA segment of Uukuniemi Virus encoding the membrane glycoproteins G1 and G2. Virology. 160,191-202.

Rosenthal, D. and Doering, J. (1983). The genomic organization of dispersed tRNA and 5S RNA genes in *Xenopus laevis.* J. Biol. Chem. 258,7402-7410.

Rosenthal, D.S., Doering, J.L., Fokta, F.J. and Jeske, J.B. (1984). Two new tDNA families in *Xenopus laevis.* J. Cell Biol. 99,253a

Rowe, D.W., Shapiro, J.R., Poirier, M. and Schlesinger, S. (1985). Diminished Type I collagen synthesis and reduced α1(I) collagen messenger RNA in cultured fibroblasts form patients with dominantly inherited Type I osteogenesis imperfecta. J. Clin. Invest. 76,604-611.

Ryynanen, J., Sollberg, S., Olsen, D. and Uitto, J. (1991). Transforming growth factor-ß up-regulates Type VII collagen gene expression in normal and transformed epidermal keratinocytes in culture. Biochem. Biophys. Res. Comm. 180,2,673-680.

Saitta, B., Wang, Y.-M., Renkart, L., Zhang, R.-Z., Pan, T.-C., Timpl, R. and Chu, M.-L. (1991). The exon organization of the triple-helical coding regions of the human α1(VI) and α2(VI) collagen genes is highly similar. Genomics 11,145-153.

Sambrook, J., Fritsch, E. and Maniatis, T. (1989). Molecular Cloning, A Laboratory Manual. Second edition. (Cold Spring Harbor Press, Cold Spring Harbor, NY). pp. 2.60-2.111.

Sandell, L. and Boyd, C. (1990). Conserved and divergent sequence and functional elements within the collagen

genes. In Extracellular Matrix Genes. Sandell, L.J. and Boyd, C.D. eds. (Academic Press, New York). pp. 1-56

Sanger, F., Nicklen, S. and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74,5463-5467.

Shaw, L.M. and Olsen, B.R. (1991). FACIT Collagens: diverse molecular bridges in extracellular matrices. TIBS. 16, 191-194.

Sillence, D.O. (1988). Osteogenesis imperfecta nosology and Genetics. Ann. N.Y. Acad. Sci. USA 543,1-15.

Sillence, D., Rimoin, D. and Danks, D. (1979a). Clinical variability in osteogenesis imperfecta - Variable expressivity or genetic heterogeneity. Birth Defects: Original Article Series XV,113-129.

Sillence, D.O., Senn, A.S. and Danks, D.M. (1979b). Genetic heterogeneity in osteogenesis imperfecta. J. Med. Genet. 1,101-116.

Soininen, R., Huotari, M., Ganguly, A., Prockop, D. and Tryggvason, K. (1989). Structural organization of the gene for the a1 chain of human Type IV collagen. J. Biol. Chem. 264,13565-13571.

Soininen, R., Huotari, M., Hostikka, S., Prockop, D. and Tryggvason, K. (1988). The structural genes for a1 and a2 chains of human Type IV collagen are divergently encoded on opposite DNA strands and have an overlapping promoter region. J. Biol. Chem. 263,17217-17220.

Steinmann, B., Superti-Furga, A. and Royce, P.M. (1988). Imperfect collagenesis in osteogenesis imperfecta. The consequences of cysteine-glycine substitutions upon collagen structure and metabolism. Ann. N.Y. Acad. Sci. 543,47-61.

Stoker, N.G., Cheah, K.S.E., Griffin, J.R., Pope, F.M. and Solomon, E. (1985). A highly polymorphic region 3′ to the human Type II collagen gene. Nucl. Acids Res. 13,4613-4622.

Sugrue, S., Gordon, M., Seyer, J., Dublet, B., van der Rest, M. and Olsen, B. (1989). Immunoidentification of Type XII collagen in embryonic tissues. J. Cell Biol. 109,939-945.

Superti-Furga, A., Pistone, F., Romano, C. and Steinmann, B. (1989). Clinical Variability of Osteogenesis Imperfecta

Linked to Col1A2 and Associated with a Structural Defect in the Type I Collagen Molecule. J. Med. Genet. 26,358-362.

Sykes, B., Ogilvie, D. and Wordsworth, P. (1990). Consistent linkage of dominantly inherited osteogenesis imperfecta to the Type I collagen loci: COL1A1 and COL1A2. Am. J. Hum. Genet. 46,293-307.

Sykes, B., Ogilvie, D., Wordsworth, P., Anderson, J. and Jones, N. (1986). Osteogenesis imperfecta is linked to both Type I collagen structural genes. Lancet ii,69-72.

Sykes, B.C., Ogilvie, D.J. and Wordsworth, B.P. (1985). Lethal osteogenesis imperfecta and a collagen gene deletion. Length polymorphism provides an alternative explanation. Hum. Genet. 70,35-37.

Tatham, A.S., Shewry, P.R., Miflin, B.J. (1984). Wheat gluten elasticity: a similar molecular basis to elastin? FEBS 177,205-208.

Tenni, R., Cetta, G., Dyne, K., Valli, M., Zanaboni, G. and Castellani, A. (1988). Severe nonlethal osteogenesis imperfecta: Biochemical heterogeneity. Ann. N.Y. Acad. Sci. USA 460,73-82.

Thomas, J., Cresswell, C., Rash, B., Nicolai, H., Jones, T., Solomon, E., Grant, M. and Boot-Handford, R. (1991). The human collagen X gene: Complete primary traslated sequence and chromosomal location. Biochem. J. 280,617-623.

Tikka, L., Pihlajaniemi, T., Henttu, P., Prockop, D. and Tryggvason, K. (1988). Gene structure for the α1 chain of a human short-chain collagen (Type XIII) with alternatively spliced transcripts and translation termination codon at the 5′ end of the last exon. Proc. Natl. Acad. Sci. USA 85,7491-7495.

Tiller, G., Rimoin, D., Murray, L. and Cohn, D. (1990). Tandem duplication within a Type II collagen gene (COL2A1) exon in an individual with spondyloepiphyseal dysplasia. Proc. Natl. Acad. Sci. USA 87,3889-3893.

Timpl, R., Wiedemann, H., van Delden, V., Furthmayer, H. and Kuhn, K. (1981). A network model for the organization of Type IV collagen molecules in basement membranes. Eur. J. Biochem. 120,203-211.

Tromp., G. and Prockop, D.J. (1988). Single base mutation in the pro-α2(I) collagen gene that causes efficient

splicing of RNA from exon 27 to exon 29 and synthesis of a shortened but in-frame pro-α2(I) chain. Proc. Natl. Acad. Sci. USA 85,5254-5258.

Trueb, J. and Trueb, B. (1992). Type XIV collagen is a variant of undulin. Eur. J. Biochem. 207, 549-557.

Tsipouras, P., Borresen, A.-L., Dickson, L.A., Berg, K., Prockop, D.J. and Ramirez, F. (1984). Molecular heterogeneity in the mild autosomal dominant forms of osteogenesis imperfecta. Am. J. Hum. Genet. 36,1172-1179.

Tsipouras, P. and Ramirez, F. (1987). Genetic disorders of collagen. J. Med. Genet. 24,2-8.

Tucker, S.J., Tannahill, D. and Higgins, C.F. (1992). Identification and developmental expression of the *Xenopus laevis* cystic fibrosis transmembrane conductance regulator gene. Hum. Mol. Genet. 1,77-82.

van der Rest, M. and Garrone, R. (1990). Collagens as multidomain proteins. Biochimie 72,473-484.

Vasios, G., Ninomiya, Y. and Olsen, B. (1987). Analysis of collagen structure by molecular biology techniques. In Structure and Function of Collagen Types. Mayne, R. and Burgeson, R.E., eds. (Academic Press, New York). pp. 283-309.

Vaughan, L., Mendler, M., Huber, S., Bruckner, P., Winterhalter, K., Irwin, M. and Mayne, R. (1988). D-periodic distribution of collagen Type IX along cartilage fibrils. J. Cell Biol. 106,991-997.

Venkatachalam, A. and Urry, D.W. (1981). Development of a linear helical conformation from its cyclic correlate-ß-spiral model of elastin poly (penta peptide). Macromolecules 14,1225-1229.

Vieira, J. and Messing, J. (1982). The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene 19,259-268.

Vnencak-Jonhes, C.L., Phillips III, J.A., Chen, E.Y. and Seeburg, P.H. (1988). Molecular basis of human growth hormone gene deletions. Proc. Natl. Acad. Sci. USA 85,5615-5618.

Vogel, B.E., Minor, R.R., Freund, M. and Prockop, D.J. (1987). A point mutation in a Type I procollagen gene converts glycine 748 of the α1 chain to cysteine and destabilizes the triple helix in a lethal variant of osteogenesis imperfecta. J. Biol. Chem. 262,14737-14744.

Wallis, G., Starman, B., Zinn, A. and Byers, P. (1990). Variable expression of osteogenesis imperfecta in a nuclear family is explained by somatic mosaicism for a lethal point mutation in the α1(I) gene (COL1A1) of type I collagen in a parent. Am. J. Hum. Genet. 46,1034-1040.

Wenstrup, F., Willing, M., Starman, B. and Byers, P. (1990). Distinct biochemical phenotypes predict clinical severity in nonlethal varients of osteogenesis imperfecta. Am. J. Hum. Genet. 46,975-982.

Wenstrup, R.J., Cohn, D.H., Cohen, T. and Byers, P.H. (1988). Arginine for glycine substitution in the triple-helical domain of the products of one α2(I) collagen allele COL1A2 produces the osteogenesis imperfecta Type IV phenotype. J. Biol. Chem. 263,7734-7740.

Wenstrup, R.J., Tsipouras, P. and Byers, P.H. (1986). Osteogenesis imperfecta Type IV. Biochemical confirmation of genetic linkage to the pro-α2(I) gene of Type I collagen. J. Clin. Invest. 78,1449-1455.

Willing, M., Cohn, D. and Byers, P. (1990). Frameshift mutation near the 3′ end of the COL1A1 gene of Type I collagen predicts an elongated pro-α1(I) chain and results in osteogenesis imperfecta Type I. J. Clin. Invest. 85,282-290.

Willing, M.C., Cohn, D.H., Starman, B., Holbrook, K.A., Greenberg, C.R. and Byers, P.H. (1988). Heterozygosity for a large deletion in the α2(I) collagen gene has a dramatic effect on Type I collagen secretion and produces perinatal lethal osteogenesis imperfecta. J. Biol. Chem. 263,8398-8404.

Willing, M., Pruchno, C., Atkinson, M. and Byers, P. (1992). Osteogenesis imperfecta type I is commonly due to a COL1A1 null allele of Type I collagen. Am. J. Hum. Genet. 51,508-515.

Yamagata, M., Yamada, K., Yamada, S., Shinomura, T., Tanaka, H., Nishida, Y., Obara, M. and Kimata, K. (1991). The Complete Primary Structure of Type XII Collagen Shows a Chimeric Molecule with Reiterated Fibronectin Type III Motifs, VWD Motifs, a Domain Homologous to a Noncollagenous Region of Type X Collagen. J. Cell Biol.

115,209-221.

Yamaguchi, N., Benya, P., van der Rest, M, and Ninomiya, Y. (1989). The cloning and sequencing of α1(VIII) collagen cDNAs demonstrate that Type VIII collagen is a short chain collagen and contains triple-helical and carboxy terminal domains similar to Type X collagen. J. Biol. Chem. 264,16022-16029.

Yamaguchi, N., Mayne, R. and Ninomiya, Y. (1991). The α1(VIII) collagen gene is homologous to the α1(X) collagen gene and contains a large exon encoding the entire triple helical and carboxyl-terminal non-triple helical domains of α1(VIII) polypeptide. J. Biol. Chem. 266,4508-4513.

Zuliani, G. and Hobbs, H.H. (1990). A high frequency of length polymorphisms in repeated sequences adjacent to alu sequences. Am. J. Hum. Genet. 46,963-969.

VITA

The author, Sheryl Anne Cammarata, is the daughter of Richard and Joan Cammarata. She was born January 4, 1964 in Framingham, Massachusetts.

In September, 1981 she entered Colorado State University and received a Bachelor of Science with a major in Biology and a minor in Psychology in December, 1985.

In September, 1990 she entered the Department of Biology of the Graduate School at Loyola University Chicago and was granted a fellowship in biology enabling her to complete the Master of Science degree in June 1993.

In 1991, Sheryl Cammarata presented a poster at the American Society for Cell Biology meeting in Boston. She also presented at the Graduate Research Forum sponsored by the Loyola University Chapter of Sigma Xi, where she was awarded 3rd place for oral presentations.

THESIS APPROVAL SHEET

The thesis submitted by Sheryl A. Cammarata has been read and approved by the following committee:

Dr. Jeffrey L. Doering, Director
Associate Professor, Biology
Loyola University Chicago

Dr. Howard Laten
Associate Professor, Biology
Loyola University Chicago

Dr. Holden Maecker
Assistant Professor, Biology
Loyola University Chicago

The final copies have been examined by the director of the thesis and the signature which appears below verifies the fact that any necessary changes have been incorporated and that the thesis is now given the final approval by the committee with reference to content and form.

The thesis is therefore accepted in partial fulfillment of the requirements for the degree of Master of Science.

11-29-93
Date

Jeffrey L. Doering
Director's Signature