Volume 15 Number 1 1987 Nucleic Acids Research

The pro α2(V) collagen gene is evolutionarily related to the major fibrillar-forming collagens

Dominique Weil[1,2], Michael Bernard[1], Silvana Gargano[1,3] and Francesco Ramirez[1]

[1]Department of Microbiology and Immunology, Morse Institute of Molecular Genetics, SUNY-Health Science Center at Brooklyn, Box 44, 450 Clarkson Avenue, Brooklyn, NY 11203, USA

Received July 31, 1986; Revised October 2, 1986; Accepted November 17, 1986

ABSTRACT

A number of overlapping cDNA clones, covering 5.2 kb of sequences which code for the human pro α2(V) collagen chain, have been isolated. Analysis of the structural data have indicated a close evolutionary kinship between the pro α2(V) chain and the major fibrillar collagen types. Isolation and analysis of an 8 kb genomic fragment has further supported this notion by revealing a homologous arrangement of nine triple-helical domain exons. These studies have therefore provided conclusive evidence which categorizes the Type V collagen as a member of the Group 1 molecules, or fibrillar-forming collagens.

INTRODUCTION

The vertebrate collagen gene family consists of more than twenty members whose products, the α-chains, are assembled together in a characteristic triple helical conformation. At least eleven such trimeric aggregates (types) have been identified thus far (1). All collagen types are composed of either identical or similar subunits, whose collagenous structures are made up of a series of repeated Gly-X-Y tripeptides (1). The nature of the supramolecular aggregates, the length of the α-chains and the continuity of the Gly-X-Y domains represent three of the criteria used to segregate the different collagen types (1). Accordingly, E.J. Miller (1) has recently suggested that the fibrillar forming collagens, Types I-III, together with the less characterized Types V and K and possibly the still controversial embryonic homotrimer (2), should be classified together as Group 1 collagens, or molecules with long, uninterrupted collagenous domains.

The notion of distinct groups or classes of structurally related collagens has recently been supported by the finding of a

significant heterogeneity at the level of the organization of the genes (for recent reviews see 3-6) (7-11). Parenthetically, these data have provided a novel parameter of categorization of the various collagen types, which is based on the evolutionary features of the genes rather than on the molecular architectures of the proteins.

By far the best characterized are the four genes coding for the subunits of the major fibrillar collagens, Types I-III (3-6). Briefly, the genes exhibit homologous patterns of distribution of their numerous exons, albeit they differ from each other in the coding sequences, the size of the introns, and the chromosomal localization (3-6). The knowledge of a homologous gene organization, besides suggesting a common evolutionary origin of this subfamily of collagens, has provided the basis for defining the prototypical organization of a Group 1 collagen gene or a molecule which satisfies the criteria for forming fibrillar aggregates. To this end, we have extended our previous cloning work to other potential members of the Group 1 collagens, namely Types V and K as well as the embryonic homotrimer. Recently, Myers et al. (12, 13) have reported the first cloning of a pro α 2(V) collagen cDNA. Here we report our findings for the pro α 2(V) collagen gene. The data clearly supports Miller's hypothesis of a close evolutionary kinship between this collagen and the major fibrillar types.

MATERIALS AND METHODS

Cell lines and cDNA libraries

The cells used in these studies were a Type IV collagen producing fibrosarcoma line (HT-1080) and a Type V collagen producing rhabdomyosarcoma line (A204) (14,15). The latter was a kind gift of Dr. G. Todaro. Two cDNAs libraries were used in these studies. The first was previously constructed in pBR322, using as template the RNA extracted from normal human fibroblasts (CRL 1106), whereas the second (a generous gift of Drs. P. Berg and H. Okayama) was generated from the RNA purified from an SV40 transformed human fibroblast line (GM 637) (16,17).

Isolation of the cDNA clones

The cDNA libraries were plated onto nitrocellulose filters at low density. More than 1 x 10^4 recombinants were screened in parallel using labelled single-stranded cDNA synthesized from

either A204 or HT-1080 collagen-enriched poly A^+ RNA (16). Clones believed to carry either A204 or HT-1080 specific sequences were subjected to two additional rounds of positive-negative screening. Approximately one hundred clones were further examined in combined batches by slot and Northern blot hybridizations as previously described (16). Amplification, isolation and characterization of the resulting positive clones were performed following standard procedures (18).

Genomic cloning

The genomic libraries used in these experiments were prepared as follows: Total genomic DNA isolated from normal fibroblasts was digested to completion using the enzyme Hind III and then fractionated by centrifugation on sucrose gradients (18,19). The size of each fraction was estimated by electrophoresis on 0.3% and 0.7% agarose gels. Four DNA pools, averaging 12 to 8 kb, 8 to 6 kb, 6 to 4 kb, and 4 to 2 kb, were ligated to the arms of either Charon 21A or Charon 28 phages and re-infected into E. coli as previously described (19). The resulting genomic libraries were screened using appropriate cDNA sub-fragments as previously described (19).

Nucleic acid hybridizations and DNA sequencing

The conditions used for the transfer of nucleic acids onto nitrocellulose paper, as well as for the hybridization and washing of the filters, have been detailed before (16,19). Both chemical cleavage and dideoxy chain termination techniques were used for the sequencing of the cDNA and genomic clones, respectively (20,21). In the latter case, appropriate genomic fragments were subcloned into pUC18 and sequenced according to a modification of the procedure of Zagursky et al. (22). Sequencing of both strands was performed for the cDNA clones and some of the genomic subclones.

RESULTS AND DISCUSSION

cDNA cloning

The function, composition and structural architecture of the Type V collagen is still controversial due to the fact that the low representation of this protein in many tissues and cell cultures greatly limits biochemical and molecular analyses (23).

To overcome this problem and to isolate Type V-specific cDNAs, we employed two established human tumor cell lines which have been shown to mainly synthesize one type of collagen. We chose a rhabdomyosarcoma cell line (A204) as the source of Type V collagen RNA and a fibrosarcoma cell line (HT-1080) as the source of Type IV collagen RNA (14,15). To the best of our knowledge, no other minor collagen types have been reported to be present in these cells. The two collagen-enriched poly A^+ RNAs obtained from these cell lines were used as templates to generate labelled single-stranded cDNA probes for the parallel screening of the cDNA libraries. Using this positive-negative screening approach we identified a number of A204 and HT-1080 specific clones whose DNA was in turn used as a probe in filter-bound hybridization analyses of the original RNA populations. This back-hybridization step allowed us to discriminate between those recombinants which hybridized with a pattern consistent with the predicted size, tissue-specificity and quantitative representation of the two collagen types (24). In this initial phase of screening, we isolated two A204 specific clones that satisfied the aforementioned criteria. Each recombinant was derived from a different cDNA library; clone Hf 511 was isolated from the normal fibroblast cDNA library, whereas clone OK 4 was isolated from the SV40-transformed fibroblast cell line (16,17). The latter library has been shown to contain full-size or nearly full-size cDNA transcripts (17).

Although both clones hybridized specifically to A204 RNA, they clearly differed from each other in their restriction maps. Sequencing of Hf 511 and OK 4 confirmed both their differences and their collagenous natures, but failed to determine the identity of OK 4. Notwithstanding this problem, the tissue-specific hybridization of OK 4, as well as the derived amino acid composition of this 4.2 kb cDNA, strongly supported the idea that this clone codes for a Type V-related polypeptide. Currently we are using a specific antibody to further characterize this Type V-related gene product. Hf 511, on the other hand, while shorter than OK 4, provided some structural evidence suggesting that it indeed codes for the human pro α2(V) chain.

At this point, however, and while this work was in progress,

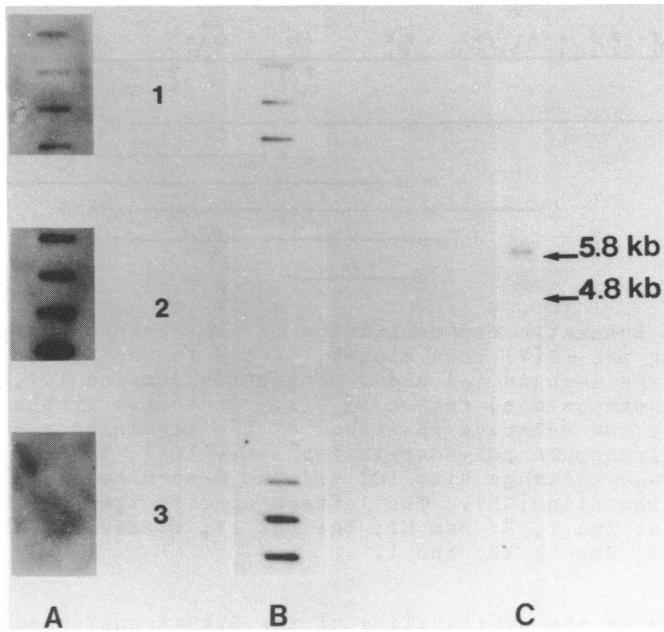


Figure 1: Nucleic acids filter bound hybridization analysis of Type IV and Type V specific collagen clones. Left, slot blot hybridizations of normal (1), HT-1080 (2) and A204 (3) RNAs with α1(IV) collagen cDNA, OK B3 (A) and Hf 511 (B). The four slots in each hybridization contain increasing amounts of RNA, from 0.25 to 1 μg. Right, Northern blotting hybridizations of A204 RNA with Hf 511 (C). The size markers on the right are those derived from the parallel hybridization of a pro α1(I) cDNA probe to normal fibroblast RNA.

Myers et al. (12,13) reached similar conclusions by comparing the amino acid composition of a cyanogen bromide derived peptide, α2(V) CB9, with that deduced from the sequencing of a shorter cDNA clone. In their elegant reports, the authors extensively discussed their structural findings and thus provided the first full account for the determination of the complete primary structure of the C-propeptide and one-fourth of the human α2(V) chain (12,13). In addition, the same group demonstrated the cytological linkage of this gene with the pro α1(III) collagen (25). We and others have later and independently confirmed the same findings (26-28).

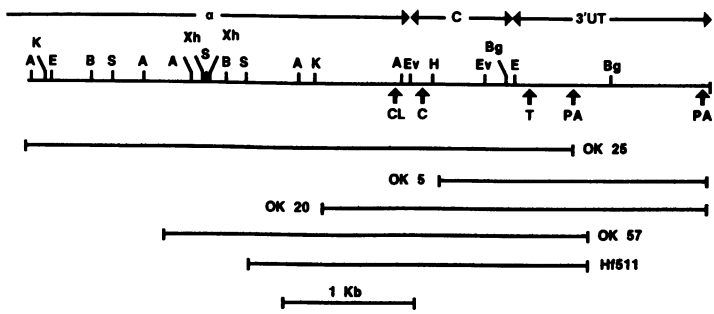
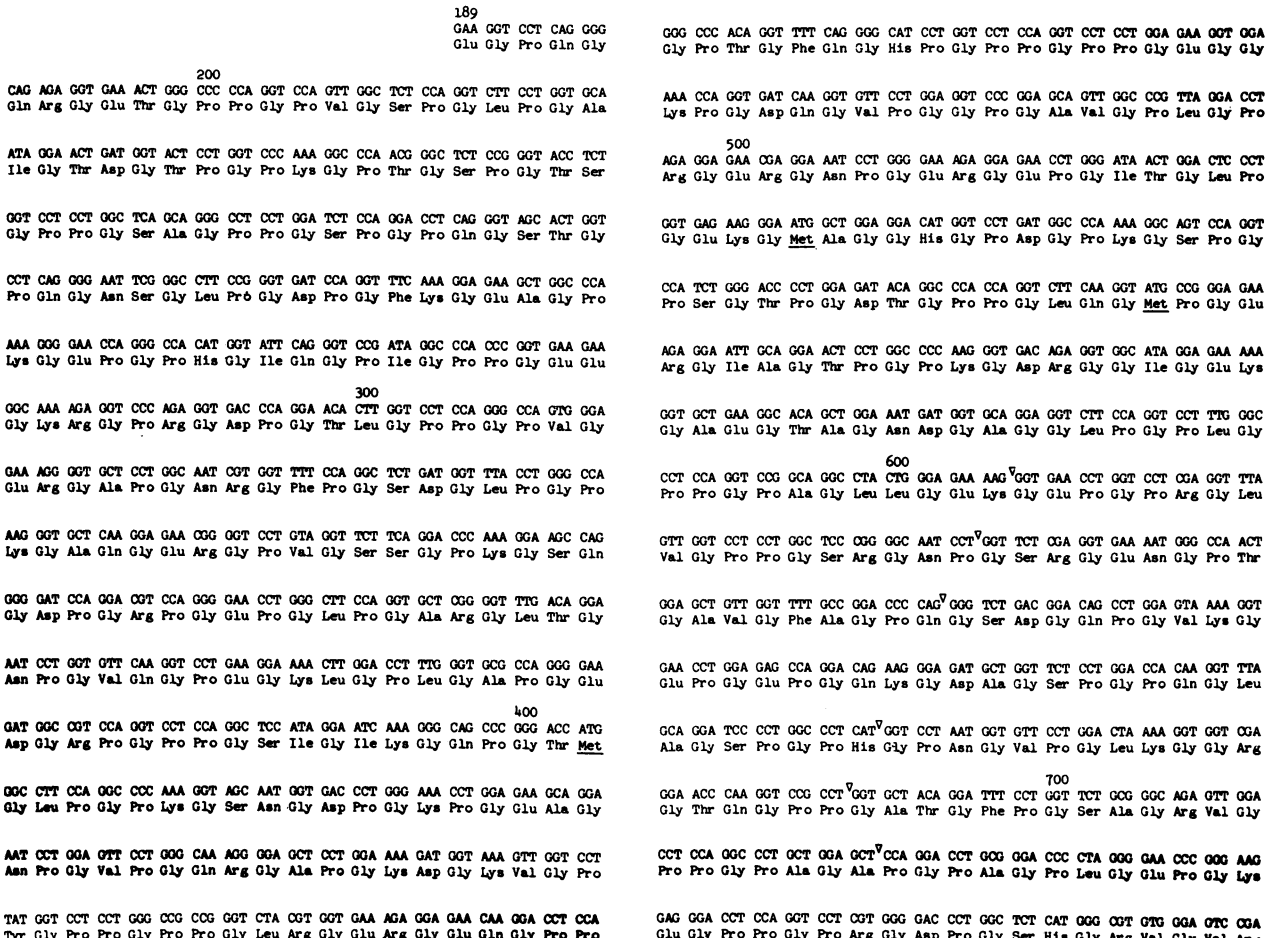
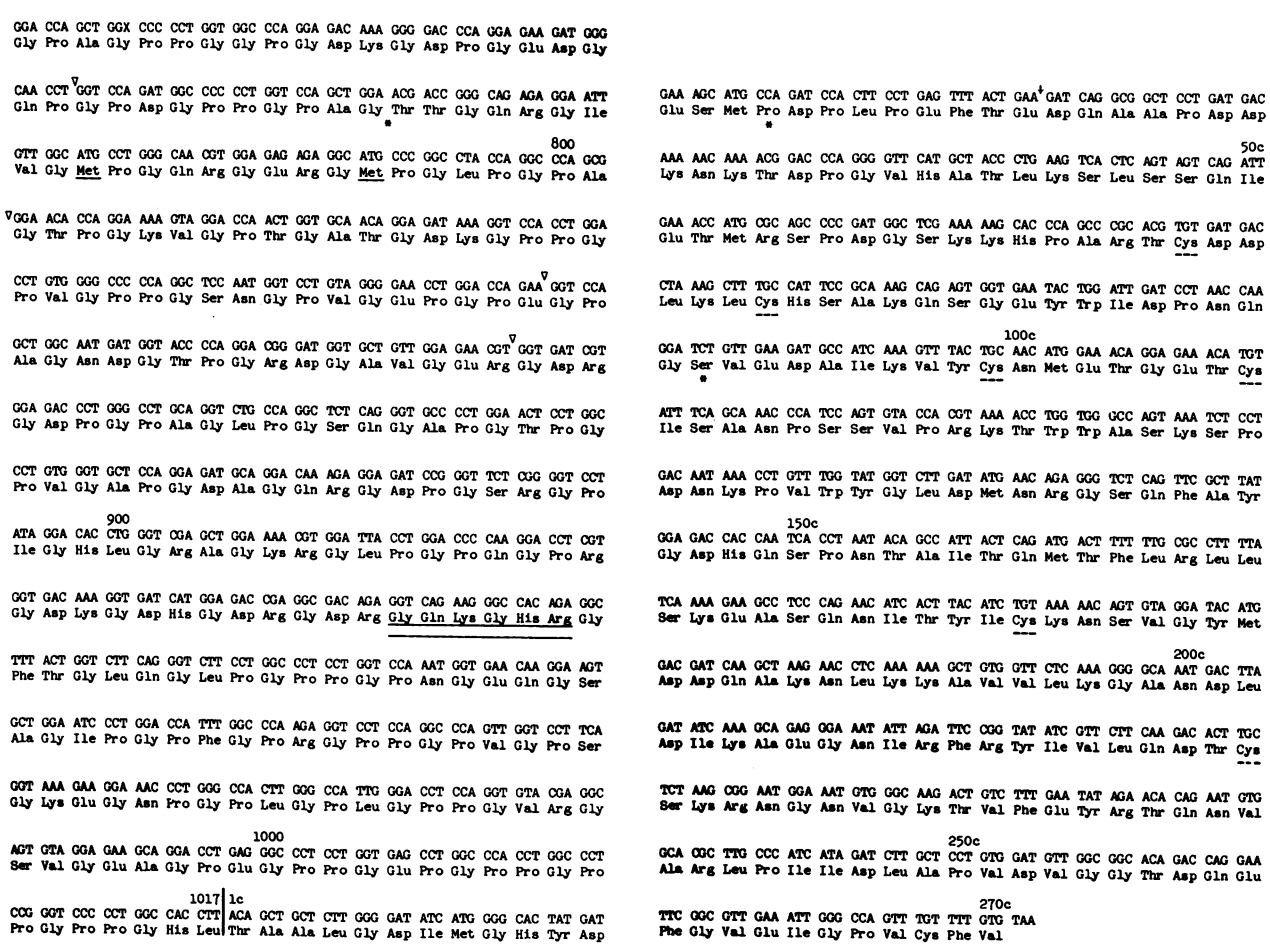


Figure 2: Schematic representation of the restriction map of the overlapping pro α 2(V) cDNA clones. Above is the relative orientation of the α-chain (α) and C-propeptide domains (C), as well as the 3' untranslated region (3'UT). Indicated by the arrows beneath are the relative positions of the termination codon (T), the two polymorphic polyadenylation sites (PA), the potential C-endopeptidase cleavage site (C) and the C-terminal triple helical cross-linking site (CL). The letters specify the following enzymes: A: Apa I; B: Bam HI; Bg: Bgl II; E: Eco RI; Ev: Eco RV; K: Kpn I; S: Sma I; Xh: Xho I.

Because of the availability of the SV40-transformed fibroblast cDNA library, we then attempted to extend the analysis of the α 2(V) chain by isolating longer cDNA clones. This second round of screening was performed using the eucaryotic insert of Hf 511, and it led to the isolation of several positive clones (Fig. 2). Sequencing of the largest recombinant, OK 25, enabled us to nearly complete the determination of the primary structure of the human α 2(V) chain (Fig. 3). Sequencing of OK 5, OK 20 and OK 57 confirmed the polymorphic nature of the two hybridizing mRNA species which, as in other collagen genes, are an indication of multiple polyadenylation sites (Figs. 1 and 2) (4). Several lines of investigation have recently supported the idea that Type V collagen may indeed be a fibrillar-forming molecule also capable of interacting with basement membrane collagen. First, immunohistochemical studies have shown the close association *in vivo* of the Type V fibrils with the interstitial collagen fibrils as well as with the basal laminae (29,30). Secondly, reconstitution experiments have suggested that the ability of Type V collagen to form homo- and heterotypic fibrils *in vitro* may be an indirect indication of the physiological arrangement of this macromolecule in various tissues, alone or in


189
GAA GGT CCT CAG GGG
Glu Gly Pro Gln Gly
200
CAG AGA GGT GAA ACT GGG CCC CCA GGT CCA GTT GGC TCT CCA GGT CTT CCT GGT GCA
Gln Arg Gly Glu Thr Gly Pro Pro Gly Pro Val Gly Ser Pro Gly Leu Pro Gly Ala
ATA GGA ACT GAT GGT ACT CCT GGT CCC AAA GGC CCA ACG GGC TCT CCG GGT ACC TCT
Ile Gly Thr Asp Gly Thr Pro Gly Pro Lys Gly Pro Thr Gly Ser Pro Gly Thr Ser
GGT CCT CCT GGC TCA GCA GGG CCT CCT GGA TCT CCA GGA CCT CAG GGT AGC ACT GGT
Gly Pro Pro Gly Ser Ala Gly Pro Pro Gly Ser Pro Gly Pro Gln Gly Ser Thr Gly
CCT CAG GGG AAT TCG GGC CTT CCG GGT GAT CCA GGT TTC AAA GGA GAA GCT GGC CCA
Pro Gln Gly Asn Ser Gly Leu Pro Gly Asp Pro Gly Phe Lys Gly Glu Ala Gly Pro
AAA GGG GAA CCA GGG CCA CAT GGT ATT CAG GGT CCG ATA GGC CCA CCC GGT GAA GAA
Lys Gly Glu Pro Gly Pro His Gly Ile Gln Gly Pro Ile Gly Pro Pro Gly Glu Glu
300
GGC AAA AGA GGT CCC AGA GGT GAC CCA GGA ACA CTT GGT CCT CCA GGG CCA GTG GGA
Gly Lys Arg Gly Pro Arg Gly Asp Pro Gly Thr Leu Gly Pro Pro Gly Pro Val Gly
GAA AGG GGT GCT CCT GGC AAT CGT GGT TTT CCA GGC TCT GAT GGT TTA CCT GGG CCA
Glu Arg Gly Ala Pro Gly Asn Arg Gly Phe Pro Gly Ser Asp Gly Leu Pro Gly Pro
AAG GGT GCT CAA GGA GAA CGG GGT CCT GTA GGT TCT TCA GGA CCC AAA GGA AGC CAG
Lys Gly Ala Gln Gly Glu Arg Gly Pro Val Gly Ser Ser Gly Pro Lys Gly Ser Gln
GGG GAT CCA GGA CGT CCA GGG GAA CCT GGG CTT CCA GGT GCT CGG GGT TTG ACA GGA
Gly Asp Pro Gly Arg Pro Gly Glu Pro Gly Leu Pro Gly Ala Arg Gly Leu Thr Gly
AAT CCT GGT GTT CAA GGT CCT GAA GGA AAA CTT GGA CCT TTG GGT GCG CCA GGG GAA
Asn Pro Gly Val Gln Gly Pro Glu Gly Lys Leu Gly Pro Leu Gly Ala Pro Gly Glu
400
GAT GGC CGT CCA GGT CCT CCA GGC TCC ATA GGA ATC AAA GGG CAG CCC GGG ACC ATG
Asp Gly Arg Pro Gly Pro Pro Gly Ser Ile Gly Ile Lys Gly Gln Pro Gly Thr Met
GGC CTT CCA GGC CCC AAA GGT AGC AAT GGT GAC CCT GGG AAA CCT GGA GAA GCA GGA
Gly Leu Pro Gly Pro Lys Gly Ser Asn Gly Asp Pro Gly Lys Pro Gly Glu Ala Gly
AAT CCT GGA GTT CCT GGG CAA AGG GGA GCT CCT GGA AAA GAT GGT AAA GTT GGT CCT
Asn Pro Gly Val Pro Gly Gln Arg Gly Ala Pro Gly Lys Asp Gly Lys Val Gly Pro
TAT GGT CCT CCT GGG CCG CCG GGT CTA CGT GGT GAA AGA GGA GAA CAA GGA CCT CCA
Tyr Gly Pro Pro Gly Pro Pro Gly Leu Arg Gly Glu Arg Gly Glu Gln Gly Pro Pro
GGG CCC ACA GGT TTT CAG GGG CAT CCT GGT CCT CCA GGT CCT CCT GGA GAA GGT GGA
Gly Pro Thr Gly Phe Gln Gly His Pro Gly Pro Pro Gly Pro Pro Gly Glu Gly Gly
AAA CCA GGT GAT CAA GGT GTT CCT GGA GGT CCC GGA GCA GTT GGC CCG TTA GGA CCT
Lys Pro Gly Asp Gln Gly Val Pro Gly Gly Pro Gly Ala Val Gly Pro Leu Gly Pro
500
AGA GGA GAA CGA GGA AAT CCT GGG GAA AGA GGA GAA CCT GGG ATA ACT GGA CTC CCT
Arg Gly Glu Arg Gly Asn Pro Gly Glu Arg Gly Glu Pro Gly Ile Thr Gly Leu Pro
GGT GAG AAG GGA ATG GCT GGA GGA CAT GGT CCT GAT GGC CCA AAA GGC AGT CCA GGT
Gly Glu Lys Gly Met Ala Gly Gly His Gly Pro Asp Gly Pro Lys Gly Ser Pro Gly
CCA TCT GGG ACC CCT GGA GAT ACA GGC CCA CCA GGT CTT CAA GGT ATG CCG GGA GAA
Pro Ser Gly Thr Pro Gly Asp Thr Gly Pro Pro Gly Leu Gln Gly Met Pro Gly Glu
AGA GGA ATT GCA GGA ACT CCT GGC CCC AAG GGT GAC AGA GGT GGC ATA GGA GAA AAA
Arg Gly Ile Ala Gly Thr Pro Gly Pro Lys Gly Asp Arg Gly Gly Ile Gly Glu Lys
GGT GCT GAA GGC ACA GCT GGA AAT GAT GGT GCA GGA GGT CTT CCA GGT CCT TTG GGC
Gly Ala Glu Gly Thr Ala Gly Asn Asp Gly Ala Gly Gly Leu Pro Gly Pro Leu Gly
600
CCT CCA GGT CCG GCA GGC CTA CTG GGA GAA AAG▽GGT GAA CCT GGT CCT CGA GGT TTA
Pro Pro Gly Pro Ala Gly Leu Leu Gly Glu Lys Gly Glu Pro Gly Pro Arg Gly Leu
GTT GGT CCT CCT GGC TCC CGG GGC AAT CCT▽GGT TCT CGA GGT GAA AAT GGG CCA ACT
Val Gly Pro Pro Gly Ser Arg Gly Asn Pro Gly Ser Arg Gly Glu Asn Gly Pro Thr
GGA GCT GTT GGT TTT GCC GGA CCC CAG▽GGG TCT GAC GGA CAG CCT GGA GTA AAA GGT
Gly Ala Val Gly Phe Ala Gly Pro Gln Gly Ser Asp Gly Gln Pro Gly Val Lys Gly
GAA CCT GGA GAG CCA GGA CAG AAG GGA GAT GCT GGT TCT CCT GGA CCA CAA GGT TTA
Glu Pro Gly Glu Pro Gly Gln Lys Gly Asp Ala Gly Ser Pro Gly Pro Gln Gly Leu
GCA GGA TCC CCT GGC CCT CAT▽GGT CCT AAT GGT GTT CCT GGA CTA AAA GGT GGT CGA
Ala Gly Ser Pro Gly Pro His Gly Pro Asn Gly Val Pro Gly Leu Lys Gly Gly Arg
700
GGA ACC CAA GGT CCG CCT▽GGT GCT ACA GGA TTT CCT GGT TCT GCG GGC AGA GTT GGA
Gly Thr Gln Gly Pro Pro Gly Ala Thr Gly Phe Pro Gly Ser Ala Gly Arg Val Gly
CCT CCA GGC CCT GCT GGA GCT▽CCA GGA CCT GCG GGA CCC CTA GGG GAA CCC GGG AAG
Pro Pro Gly Pro Ala Gly Ala Pro Gly Pro Ala Gly Pro Leu Gly Glu Pro Gly Lys
GAG GGA CCT CCA GGT CCT CGT GGG GAC CCT GGC TCT CAT GGG CGT GTG GGA GTC CGA
Glu Gly Pro Pro Gly Pro Arg Gly Asp Pro Gly Ser His Gly Arg Val Gly Val Arg


GGA CCA GCT GGX CCC CCT GGT GGC CCA GGA GAC AAA GGG GAC CCA GGA GAA GAT GGG
Gly Pro Ala Gly Pro Pro Gly Gly Pro Gly Asp Lys Gly Asp Pro Gly Glu Asp Gly
CAA CCT GGT CCA GAT GGC CCC CCT GGT CCA GCT GGA ACG ACC GGG CAG AGA GGA ATT
Gln Pro Gly Pro Asp Gly Pro Pro Gly Pro Ala Gly Thr Thr Gly Gln Arg Gly Ile
800
GTT GGC ATG CCT GGG CAA CGT GGA GAG AGA GGC ATG CCC GGC CTA CCA GGC CCA GCG
Val Gly Met Pro Gly Gln Arg Gly Glu Arg Gly Met Pro Gly Leu Pro Gly Pro Ala
GGA ACA CCA GGA AAA GTA GGA CCA ACT GGT GCA ACA GGA GAT AAA GGT CCA CCT GGA
Gly Thr Pro Gly Lys Val Gly Pro Thr Gly Ala Thr Gly Asp Lys Gly Pro Pro Gly
CCT GTG GGG CCC CCA GGC TCC AAT GGT CCT GTA GGG GAA CCT GGA CCA GAA GGT CCA
Pro Val Gly Pro Pro Gly Ser Asn Gly Pro Val Gly Glu Pro Gly Pro Glu Gly Pro
GCT GGC AAT GAT GGT ACC CCA GGA CGG GAT GGT GCT GTT GGA GAA CGT GGT GAT CGT
Ala Gly Asn Asp Gly Thr Pro Gly Arg Asp Gly Ala Val Gly Glu Arg Gly Asp Arg
GGA GAC CCT GGG CCT GCA GGT CTG CCA GGC TCT CAG GGT GCC CCT GGA ACT CCT GGC
Gly Asp Pro Gly Pro Ala Gly Leu Pro Gly Ser Gln Gly Ala Pro Gly Thr Pro Gly
CCT GTG GGT GCT CCA GGA GAT GCA GGA CAA AGA GGA GAT CCG GGT TCT CGG GGT CCT
Pro Val Gly Ala Pro Gly Asp Ala Gly Gln Arg Gly Asp Pro Gly Ser Arg Gly Pro
900
ATA GGA CAC CTG GGT CGA GCT GGA AAA CGT GGA TTA CCT GGA CCC CAA GGA CCT CGT
Ile Gly His Leu Gly Arg Ala Gly Lys Arg Gly Leu Pro Gly Pro Gln Gly Pro Arg
GGT GAC AAA GGT GAT CAT GGA GAC CGA GGC GAC AGA GGT CAG AAG GGC CAC AGA GGC
Gly Asp Lys Gly Asp His Gly Asp Arg Gly Asp Arg Gly Gln Lys Gly His Arg Gly
TTT ACT GGT CTT CAG GGT CTT CCT GGC CCT CCT GGT CCA AAT GGT GAA CAA GGA AGT
Phe Thr Gly Leu Gln Gly Leu Pro Gly Pro Pro Gly Pro Asn Gly Glu Gln Gly Ser
GCT GGA ATC CCT GGA CCA TTT GGC CCA AGA GGT CCT CCA GGC CCA GTT GGT CCT TCA
Ala Gly Ile Pro Gly Pro Phe Gly Pro Arg Gly Pro Pro Gly Pro Val Gly Pro Ser
GGT AAA GAA GGA AAC CCT GGG CCA CTT GGG CCA TTG GGA CCT CCA GGT GTA CGA GGC
Gly Lys Glu Gly Asn Pro Gly Pro Leu Gly Pro Leu Gly Pro Pro Gly Val Arg Gly
1000
AGT GTA GGA GAA GCA GGA CCT GAG GGC CCT CCT GGT GAG CCT GGC CCA CCT GGC CCT
Ser Val Gly Glu Ala Gly Pro Glu Gly Pro Pro Gly Glu Pro Gly Pro Pro Gly Pro
1017 1c
CCG GGT CCC CCT GGC CAC CTT ACA GCT GCT CTT GGG GAT ATC ATG GGG CAC TAT GAT
Pro Gly Pro Pro Gly His Leu Thr Ala Ala Leu Gly Asp Ile Met Gly His Tyr Asp
GAA AGC ATG CCA GAT CCA CTT CCT GAG TTT ACT GAA GAT CAG GCG GCT CCT GAT GAC
Glu Ser Met Pro Asp Pro Leu Pro Glu Phe Thr Glu Asp Gln Ala Ala Pro Asp Asp
50c
AAA AAC AAA ACG GAC CCA GGG GTT CAT GCT ACC CTG AAG TCA CTC AGT AGT CAG ATT
Lys Asn Lys Thr Asp Pro Gly Val His Ala Thr Leu Lys Ser Leu Ser Ser Gln Ile
GAA ACC ATG CGC AGC CCC GAT GGC TCG AAA AAG CAC CCA GCC CGC ACG TGT GAT GAC
Glu Thr Met Arg Ser Pro Asp Gly Ser Lys Lys His Pro Ala Arg Thr Cys Asp Asp
CTA AAG CTT TGC CAT TCC GCA AAG CAG AGT GGT GAA TAC TGG ATT GAT CCT AAC CAA
Leu Lys Leu Cys His Ser Ala Lys Gln Ser Gly Glu Tyr Trp Ile Asp Pro Asn Gln
100c
GGA TCT GTT GAA GAT GCC ATC AAA GTT TAC TGC AAC ATG GAA ACA GGA GAA ACA TGT
Gly Ser Val Glu Asp Ala Ile Lys Val Tyr Cys Asn Met Glu Thr Gly Glu Thr Cys
ATT TCA GCA AAC CCA TCC AGT GTA CCA CGT AAA ACC TGG TGG GCC AGT AAA TCT CCT
Ile Ser Ala Asn Pro Ser Ser Val Pro Arg Lys Thr Trp Trp Ala Ser Lys Ser Pro
GAC AAT AAA CCT GTT TGG TAT GGT CTT GAT ATG AAC AGA GGG TCT CAG TTC GCT TAT
Asp Asn Lys Pro Val Trp Tyr Gly Leu Asp Met Asn Arg Gly Ser Gln Phe Ala Tyr
150c
GGA GAC CAC CAA TCA CCT AAT ACA GCC ATT ACT CAG ATG ACT TTT TTG CGC CTT TTA
Gly Asp His Gln Ser Pro Asn Thr Ala Ile Thr Gln Met Thr Phe Leu Arg Leu Leu
TCA AAA GAA GCC TCC CAG AAC ATC ACT TAC ATC TGT AAA AAC AGT GTA GGA TAC ATG
Ser Lys Glu Ala Ser Gln Asn Ile Thr Tyr Ile Cys Lys Asn Ser Val Gly Tyr Met
200c
GAC GAT CAA GCT AAG AAC CTC AAA AAA GCT GTG GTT CTC AAA GGG GCA AAT GAC TTA
Asp Asp Gln Ala Lys Asn Leu Lys Lys Ala Val Val Leu Lys Gly Ala Asn Asp Leu
GAT ATC AAA GCA GAG GGA AAT ATT AGA TTC CGG TAT ATC GTT CTT CAA GAC ACT TGC
Asp Ile Lys Ala Glu Gly Asn Ile Arg Phe Arg Tyr Ile Val Leu Gln Asp Thr Cys
TCT AAG CGG AAT GGA AAT GTG GGC AAG ACT GTC TTT GAA TAT AGA ACA CAG AAT GTG
Ser Lys Arg Asn Gly Asn Val Gly Lys Thr Val Phe Glu Tyr Arg Thr Gln Asn Val
250c
GCA CGC TTG CCC ATC ATA GAT CTT GCT CCT GTG GAT GTT GGC GGC ACA GAC CAG GAA
Ala Arg Leu Pro Ile Ile Asp Leu Ala Pro Val Asp Val Gly Gly Thr Asp Gln Glu
270c
TTC GGC GTT GAA ATT GGG CCA GTT TGT TTT GTG TAA
Phe Gly Val Glu Ile Gly Pro Val Cys Phe Val

combination with Type I and III collagens (31). In line with this, it is believed that one of the most important requirements for the stabilization of collagen fibrils is the presence of four homologous intermolecular cross-links between lysyl derived aldehyde and hydroxylysine residues (32). In the major fibrillar collagen types, two of these sites are located at the N- and C-terminal ends of the α-chains and two at their N- and C- terminal telopeptides. The α2(I) chain shows some differences in that it lacks the C-terminal aldehyde site and it slightly differs for the location and composition of the canonical sequence Gly-X-Lys-Gly-His-Arg at the C-terminal triple helical cross-linking site. A search for similar sequences in OK 25 revealed that this gene product similarly lacks the latter site but has maintained the location and the high degree of sequence homology of the former (Fig. 3). Because our cDNA clones did not extend further 5', we were unable to ascertain the presence of the two N-terminal cross-linking loci. Notwithstanding this problem, the structural data are in agreement with the aforementioned experimental evidence that suggests a fibrillar role for Type V collagen.

The deduced amino acid sequence of more than 80% of the α2(V) chain enabled us to establish the order of most of the cyanogen bromide derived peptide fragments (Fig. 4). Interestingly, the DNA data revealed the existence of an additional small peptide, CB0, previously undetected by protein analysis (33).

Figure 3: DNA and deduced amino acid sequences of the coding regions of the pro α2(V) cDNAs. Underlined are the cysteinyl (dotted line) and the α-chain methionyl residues (continuous line), as well as the C-terminal triple helical cross-linking sites (double line). The vertical bar separates the triple helical domain from the C-terminal telopeptide, whose first residue is numbered as 1c. Asterisks beneath the sequence indicate the regions that in the fibrillar collagen chains contain the following structural elements discussed in the text: the vertebrate collagenase cleavage site (position 775) the eighth cysteine (position 90c), and the C-terminal telopeptide cross-linking site (position 16c). The arrow signifies the position of the potential C endopeptidase cleavage site, whereas the triangles demarcate the exons sequenced in DMC 28. Numbering of the α-chain residues is based on the homology with the major fibrillar chains (3-6).

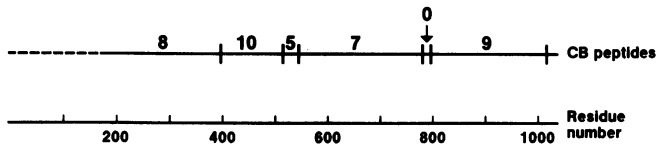


Figure 4: Partial distribution of the methionyl residues of the human α 2(V) chain. The CB peptides are numbered according to Rhodes et al., (33). The small peptide previously undetected by protein analysis is numbered with a 0.

Genomic cloning

In their reports, Myers et al. (12,13) have argued that Type V collagen may belong to a distinct subfamily of genes which arose before those encoding the fibrillar collagens. In order to test this hypothesis and to ascertain the evolutionary relationship between the Type V and Types I-III collagen genes, we used our clones to screen several genomic libraries constructed by partial digestion of human DNA. At first, and despite our numerous attempts, we failed to isolate any positive clones from these "total-representation" libraries. There is now in the literature experimental evidence in support of the anecdotal reports of "unclonable" sequences under particular choices of vector and host (34). To circumvent this problem we constructed a number of "size-selected" Hind III genomic libraries. The resulting recombinants were screened with different cDNA segments, aided by our prior knowledge of the expected length of the corresponding genomic fragments (Fig. 5). Three clones were isolated: DMC 2, DMC 21 and DMC 28. Of the three, only DMC 28 contained exclusively α chain sequences, and it was therefore subjected to further and more extensive analyses (Fig. 5). The rationale of this choice stemmed from the consideration that the distribution of the exons coding for the triple helical domain is the most distinctive feature of the collagen genes (3-11). Appropriate genomic fragments, mostly spanning from restriction sites common to the cDNA, were subcloned into pUC18 and sequenced by the Sanger method (21,22). The location and size of six exons was directly determined by DNA sequencing, while the size of three additional exons was inferred by comparison with the cDNA sequence (Fig. 5). When the pattern of distribution of these

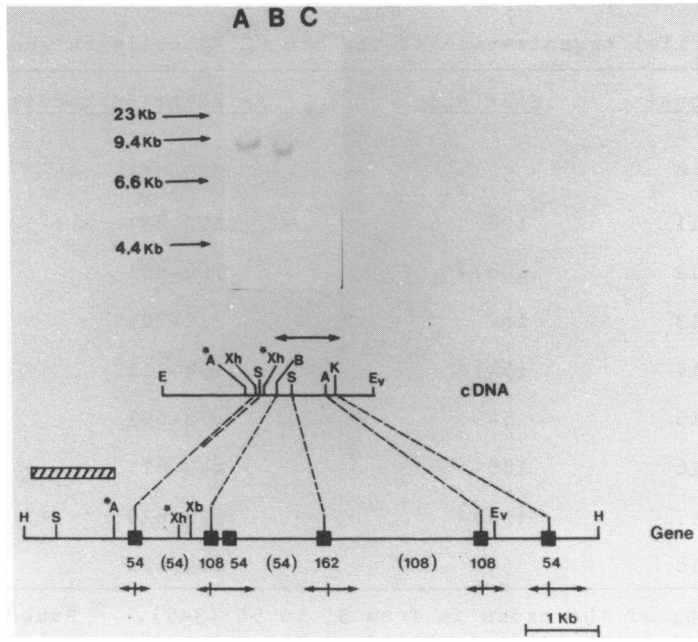


Figure 5: Restriction map of the genomic clone DMC 28. The boxes indicate the position of the exons sequenced with the size in bp indicated below. In parentheses are the exons whose sizes were inferred by comparison with the cDNA sequence. Above is the partial restriction map of the cDNA with the common restriction sites connected to the genomic clone by the dotted lines. The asterisks seen above two of the common restriction sites identify the relative position of two exons, number 17 and 19, which were not sequenced in these studies. The double pointed arrow above the cDNA indicates the fragment used for the isolation of DMC 28, as well as for the Southern blot hybridizations shown on top, Hind III (A), Bam HI (B) and Eco RI (C). The sizes of the λ DNA Hind III markers are shown on the left. Letters identify the following restriction enzymes: A: Apa I, B: Bam HI, E: Eco RI, Ev: Eco RV, K: Kpn I, S: Sma I, Xb: Xba I; Xh: Xho I. The cross hatched box indicates the position of repetitive sequences. The arrows underneath DMC 28 signify the extent and the direction of the sequencing procedures.

nine exons of the pro α2(V) gene was compared to that derived for the homologous region of a prototypical Group 1 collagen gene (3-6), complete identity was observed (Table I). Hence, the analysis of this large genomic fragment led to the conclusion that the pro α2(V) gene is evolutionarily related to the major fibrillar collagens.

TABLE I
Partial organization of the pro α 2(V) collagen gene

Exons[1]	Size(bp)	Amino Acids Encoded[2]
10	54	838-855
11	108	802-837
12	(108)[3]	766-801
13	162	712-765
14	(54)[3]	694-711
15	54	676-693
16	108	640-675
17	(54)[3]	622-639
18	54	604-621

[1] Numbering of the exons is from 3' to 5' (3-5). [2] Numbering of the encoded amino acids is based on the numeration derived from the homologous region of the other fibrillar collagen genes. [3] The parentheses signify exons not sequenced.

In addition, this observation enabled us to make the following two predictions, which are based upon the structural features of the Types I-III genes. First, assuming that the coding sequences of the remaining C-terminal portion of the α 2(V) chain are distributed among five exons (numbers 5 to 9) in a manner identical to that of the fibrillar collagen genes, we suggest that the triple helical domain of the pro α 2(V) chain is 1017 amino acids long (3-6). Secondly, because the coding capacity of exons 5 to 9 can only account for 144 amino acids, we extrapolate that the last 18 residues (54 bp) of the α 2(V) chain should make up the collagenous segment of exon 4. Such an arrangement of sequences within exon 4, the C-propeptide junction exon, is consistent with the organization of the fibrillar collagen genes where there are either 45 or 54 bp coding for the triple helical segment of exon 4 (3-6).

Finally, it is also of interest to note that the loss of the vertebrate collagenase cleavage site, which distinguishes the Type V collagen from the Types I-III (35,36), appears to be the

TABLE II

Sequence divergencies of the C-terminal propeptides

	Amino acids	Nucleotides
pro α1(II)/pro α1(I)	33%	28%
pro α2(I)/pro α1(I)	39%	35%
pro α1(III)/pro α1(I)	40%	40%
pro α2(V)/pro α1(I)	48%	42%

result of a number of nucleotide changes rather than a gene rearrangement (Fig. 3).

Further analysis of the structural data

Once the cloning experiments conclusively demonstrated the close relationship between the pro α2(V) chain and the fibrillar-forming collagens, we attempted to place this collagen within the already established evolutionary tree of the fibrillar chains. Accordingly, the four genes evolved from an ancestral, multiexon gene by processes of duplication and transposition on different chromosomes, where they began to diverge, albeit maintaining identical organizations (4). Based on the pairwise comparisons of the protein and DNA data, it is believed that the duplications of the pro α1(III) gene preceded that of the pro α2(I), pro α1(II) and pro α1(I) genes, in the order indicated (1,4). When such an analysis was extended to the pro α2(V) collagen gene, its relative level of divergence was found to be greater than that of the pro α1(III) (Table II). Similar conclusions have been previously reached by Myers et al. (12,13).

In the past, several investigators have shown that the fibrillar collagen genes exhibit a distinct pattern of codon usage for the more abundant amino acids of the α-chains, namely glycine, proline and alanine (3,5,16). In this respect, we have noted that the Type III differs from the other fibrillar collagen genes in the second nucleotide choice in the wobble position of all three codons (16). The same distinct pattern of preference was seen in the pro α2(V) gene (Table III), as well as in the

TABLE III
Codon Usage of the Group I collagen chains

Amino Acid	Third Base	α1(I)	α1(II)	α2(I)	α1(III)	α2(V)
Glycine	U	50%	37%	53%	47%	34%
	C	30%	37%	21%	17%	24%
	A	18%	24%	21%	31%	34%
	G	2%	2%	5%	5%	8%
Proline	U	57%	53%	62%	65%	50%
	C	40%	38%	22%	14%	8%
	A	3%	9%	14%	21%	38%
	G	0%	0%	2%	0%	4%
Alanine	U	74%	58%	83%	59%	46%
	C	21%	33%	10%	16%	3%
	A	5%	9%	7%	25%	36%
	G	0%	0%	0%	0%	9%

Type V-related clone, OK 4 (data not shown). Based on these observations, we therefore concluded that the pro α2(V) gene evolved from the same branch of the pro α1(III) and not prior to it.

A final point we wish to discuss in support of this idea is the pattern of selective loss of the cysteinyl residues of some of the C-propeptide domains. One paramount feature of the fibrillar collagen genes is the phylogenetic retention of the location of eight cysteines and their surrounding nucleotide sequences in the four exons coding for the C-propeptide domains (16,37-42). The first four of these cysteines are believed to be involved in interchain disulfide bonding, whereas the last four are required for intrachain linkages. The elimination of one of these functional residues in an osteogenesis imperfecta variant has been shown to alter the normal assembly of the Type I collagen

heterotrimer and to lead to the pathological phenotype (43). On the other hand, in some cases chain-specific changes have been observed, namely the loss of the third or second cysteine in the pro α2(I) and pro α2(V) chains, respectively (12,37,38). In our studies, we have confirmed the latter observation and in addition we have found that the Type V-related clone, OK 4, is missing the third cysteine. In the three genes, these variations are the result of single nucleotide changes within the highly conserved stretches of surrounding sequences. Because of the structural homologies and the aforementioned similarities of codon usage between the Types III and V genes, it is most unlikely that the duplication of the pro α1(III) gene followed that of the pro α2(V) with the selective acquisition of a novel cysteine in the same position of the Type V-related gene. In order to reconcile these observations, we therefore suggest that the pro α2(V), and probably the Type V-related gene, arose from a gene containing eight cysteinyl residues, an arrangement ultimately maintained by the pro α1(III) gene and possibly by other members of the same evolutionary branch. The problem of defining the details of these genealogical interrelationships will be solved with the cloning and analysis of other Group 1 collagen genes.

SUMMARY

The isolation of cDNA and genomic clones has enabled us to extend the original analysis of the primary structure of the human α2(V) chain (12,13) and to confirm the structural and evolutionary role of the Type V collagen as a fibrillar-forming molecule. In line with this, we have presented some evidence supporting a close evolutionary kinship between the pro α2(V) and the pro α1(III) genes. However, in doing so we have not used the argument of the cytological linkage between these two loci, because recently we have excluded physical linkage by segregation analyses of specific restriction fragment length polymorphisms (RFLPs) (P. Tsipouras et al., manuscript in preparation). In other words, we believe that following its duplication, the transposition of the pro α2(V) gene to the same segment of chromosome 2 where the pro α1(III) locus resides may have been a stochastic event.

In conclusion, our data strongly weigh in favor of the notion that Type V belongs to the Group 1 collagens, along with the Types I-III, rather than to a distinct subfamily of genes, as Myers et al. (12,13) have suggested. It could be argued that other regions of the pro α2(V) gene may exhibit a different exon pattern. This is obviously a possibility that we can not presently disregard. Experiments are currently in progress to test the validity of this and other hypotheses, and to further and more firmly define the functional and evolutionary interrelationships of the Group 1 collagens.

ACKNOWLEDGEMENTS

The authors wish to thank Drs. P. Berg, M.L. Chu, H. Okayama and G. Todaro for providing us with some of the materials used in these studies and Drs. B. Brodsky, W. deWet, P. Little and M. van der Rest for many helpful discussions and suggestions. We are also very much in debt to V. Benson-Chanda, D. Leviant, M. Leong and S. Salkeld for their help during the course of this work and Ms. M. Schneider for typing the manuscript. This work was supported by grants from the National Institute of Health (AR 38648) and the March of Dimes, Birth Defect Foundation (1-1042).

[2]Unite 12-INSERM, Hôpital des Enfantes Malades, Paris, France and [3]International Institute of Genetics and Biophysics, Naples, Italy

REFERENCES

1. Miller, E.J. (1985) In "Biology, chemistry and pathology of collagen", Ann. N.Y. Acad. Sci. V. 460 pp 1-13, R. Fleischmajer, B.R. Olsen, and K. Kuhn, Eds.
2. Little, C.D. and Church, R. (1977) Cell 10: 287-295.
3. Boedtker, H., Finer, M. and Aho, S. (1985) In "Biology, chemistry and pathology of collagen", Ann. N.Y. Acad. Sci. V. 460 pp 85-116, R. Fleischmajer, B.R. Olsen, and K. Kuhn, Eds.
4. Ramirez, F., Bernard, M., Chu, M.L., Dickson, L., Sangiorgi, F.O., Weil, D., deWet, W., Junien, C., and Sobel, M. (1985) In "Biology, chemistry and pathology of collagen", Ann. N.Y. Acad. Sci. V. 460 pp 117-129, R. Fleischmajer, B.R. Olsen, and K. Kuhn, Eds.
5. Upholt, W.B., Strom C.M., and Sandell, L.J. (1985) In "Biology, chemistry and pathology of collagen", Ann. N.Y. Acad. Sci. V. 460 pp 130-140, R. Fleischmajer, B.R. Olsen, and K. Kuhn, Eds.

6. deCrombrugghe, B., Schmidt, A., Liau, G., Seroyam, C., Mudryj, M., Yamada, Y., and McKeon, C. (1985) In "Biology, chemistry and pathology of collagen", Ann. N.Y. Acad. Sci. V. 460 pp 154-162, R. Fleischmajer, B.R. Olsen, and K. Kuhn, Eds.
7. Lozano, G., Ninomiya, Y., Thompson, H., and Olsen, B.R. (1985) Proc. Natl. Acad. Sci. USA 82: 4050-4054.
8. Ninomiya, Y., Gordon, M., van der Rest, M., Schmidt, T., Linsenmayer, T., Olsen, B.R. (1986). J. Biol. Chem. 261: 5041-5050.
9. Kurkinen, M., Bernard, M., Barlow, D., and Chow, L.P. (1985) Nature 317: 177-179.
10. Soininen, R., Tikka, L., Chow, L., Pihlajaniemi, T., Kurkinen, M., Prockop, D.J., Boyd, C.D., and Tryggvason, K. (1986) Proc. Natl. Acad. Sci. USA 83: 1568-1572.
11. Sakurai, Y., Sullivan, M., and Yamada, Y. (1986) J. Biol. Chem. 261: 6654-6657.
12. Myers, J.C., Loidl, H.R., Stolle, C.A., and Seyer, J.M. (1985) J. Biol. Chem. 260: 5533-5541.
13. Myers, J.C., Loidl, H.R., Seyer, J.J., and Dion, A.S. (1985) J. Biol. Chem. 260: 11216-11222.
14. Alitalo, K., Vaheri, A., Krieg, T., and Timpl, R. (1980) Eur. J. Biochem. 109: 247-255.
15. Alitalo, K., Myllyla, R., Sage, H., Pritzl, P., Vaheri, A., and Bornstein, P. (1982) J. Biol. Chem. 257: 9016-9024.
16. Chu, M.L., Weil, D., deWet, W., Bernard, M., Sippola, M., and Ramirez, F. (1985) J. Biol. Chem. 260: 4357-4363.
17. Okayama, H., and Berg, P. (1983) Mol. Cell. Biol. 3: 280-289.
18. Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
19. Chu, M.L., Gargiulo, V., Williams, C.J., and Ramirez, F. (1985) J. Biol. Chem. 260: 691-694.
20. Maxam, A.M., and Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA 74: 560-564.
21. Sanger, R., Nicklen, S., and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 5463-5467.
22. Zagursky, R.N., Baumeister, N., Lomax, N., and Berman, M. (1985) Gene Anal. Tech. 2: 89-94.
23. Fessler, J., Shigaki, N., and Fessler, L.I. (1985) In "Biology, chemistry and pathology of collagen", Ann. N.Y. Acad. Sci. V 460 pp 181-186, R. Fleischmajer, B.R. Olsen, and K. Kuhn, Eds.
24. Sangiorgi, F.D., Benson-Chanda, V., deWet, W., Sobel, M.E., and Ramirez, F. (1985) Nucl. Acids. Res. 13: 2815-2826.
25. Emanuel, B.S., Cannizzaro, L.A., Seyer, J.M., and Myers, J.C. (1985) Proc. Natl. Acad. Sci. USA 82: 3385-3389.
26. Solomon, E., Hiorns, L.R., Spurr, N., Kurkinen, M., Barlow, D., Hogan, B.L.M., and Dalgleish, R. (1985) Proc. Natl. Acad. Sci. USA 82: 3330-3334.
27. Huerre-Jeanpierre, C., Mattei, M.A., Weil, D., Grzeschik, K.H., Chu, M.L., Sangiorgi, F.O., Sobel, M.E., Ramirez, F., and Junien, C. (1985) Am. J. Hum. Genet. 38: 26-37.
28. Huerre-Jeanpierre, C., Henri, I., Bernard, M., Gallano, P., Weil, D., Grzeschik, K., Ramirez, F. and Junien, C. (1986) Human Genet. 73: 64-67.

29. Martinez-Hernandez, A., Gay, S., and Miller, E.J. (1982) J. Cell Biol. 92: 343-349.
30. Fitch, J.M., Gross, J., Mayne, R., Johnson-Wint, B., and Linsenmayer, T.F. (1984) Proc. Natl. Acad. Sci. USA 81: 2791-2795
31. Adochi, E., and Hayashi, T. (1986) Connect. Tiss. Res. 14: 257-266.
32. Eyre, D.R., Paz, M.A., and Gallop, P.M. (1984) Ann. Rev. Biochem. 53: 717-748.
33. Rhodes, R.K., Gibson, K.D., and Miller, E.J. (1981) Biochemistry 20: 3117-3121.
34. Wyman, A.R., Wolfe, L.B., and Botstein, D. (1985) Proc. Natl. Acad. Sci. USA 82: 2880-2884.
35. Liotta, L.A., Abe, S., Robey, P.G., and Martin, G.R. (1979) Proc. Natl. Acad. Sci. USA 76: 2268-2272.
36. Sage, H., and Bornstein, P. (1979) Biochemistry 18: 3815-3822.
37 Fuller, F., and Boedtker, H. (1981) Biochemistry 20: 996-1006.
38. Bernard, M.P., Myers, J.C., Chu, M.L., Ramirez, F., Eikenberry, E., and Prockop, D.J. (1983) Biochemistry 22: 1139-1145.
39. Bernard, M.P., Chu, M.L., Myers, J.C., Ramirez, F., Eikenberry, E., and Prockop, D.J. (1983) Biochemistry 22: 5213-5223.
40. Yamada, Y., Kuhn, K., and deCrombrugghe, B. (1983) Nucl. Acids. Res. 11: 2733-2746.
41. Sandell, L.J., Prentice, H.L., Kravis, D., and Upholt, W.B. (1984) J. Biol. Chem. 259: 7826-7834.
42. Sangiorgi, F.O., Benson-Chanda, V., deWet, W., Sobel, M.E., Tsipouras, P., and Ramirez, F. (1985) Nucl. Acids Res. 13: 2207-2225.
43. Pihlajaniemi, T., Dickson, L.A., Pope, F.M., Korhonen, V.R., Nicholls, A., Prockop, D.J., and Myers, J.C. (1984) J. Biol. Chem. 259: 12941-12944.