Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 2555–2559, May 1985
Biochemistry

Identification and characterization of the human type II collagen gene (*COL2A1*)

(cartilage collagen gene/mRNA/DNA sequence)

KATHRYN S. E. CHEAH*†, NEIL G. STOKER*, JANE R. GRIFFIN*, FRANK G. GROSVELD‡, AND ELLEN SOLOMON*

*Somatic Cell Genetics Laboratory, Imperial Cancer Research Fund Laboratories, P.O. Box 123, Lincoln's Inn Fields, London WC2A 3PX, United Kingdom; and ‡Medical Research Council, National Institute for Medical Research, The Ridgeway, London NW7 1AA, United Kingdom

Communicated by Walter Bodmer, December 3, 1984

ABSTRACT **The gene contained in the human cosmid clone CosHcol1, previously designated an α1(I) collagen-like gene, has now been identified. CosHcol1 hybridizes strongly to a single 5.9-kilobase mRNA species present only in tissue in which type II collagen is expressed. DNA sequence analysis shows that this clone is highly homologous to the chicken α1(II) collagen gene. These data together suggest that CosHcol1 contains the human α1(II) collagen gene *COL2A1*. The clone appears to contain the whole gene (30 kilobases in length) and will be extremely useful in the study of cartilage development and for identifying those inherited chondrodystrophies in which defects occur in this gene.**

Collagens are major structural components of the extracellular matrix. In vertebrates they form a large family of proteins represented by at least nine distinct types for which a minimum of 17 genes exist to code for their constituent α chains (1–5). Different tissues are characterized by the types and quantity of collagen expressed. The coordinated expression of these different collagen genes is believed to be important in vertebrate development (6), and collagen abnormalities may be involved in a wide range of inherited connective tissue disorders in man (7, 8). To approach these questions, a variety of cDNA and genomic collagen clones from a number of species have been isolated, including the human α1(I) (9, 10) and α2(I) genes (11, 12).

We previously reported the isolation of the genomic clone CosHcol1 from a human placental cosmid library (13), using the chicken α1(I) cDNA clone pCg54 as a probe (14). This cosmid clone contains a 36-kilobase (kb) insert and cross-hybridizes with collagen α1(I) mRNA. However, the amino acid sequence derived from 1 kb of the clone showed only 60–70% homology to chicken and bovine α1(I) and α2(I) collagen, and it did not match the human α1(III) amino acid sequence. Since interspecies protein sequence homologies between collagens of the same type are usually greater than 80%, we concluded that CosHcol1 did not code for any of these chains. In the absence of positive identification, we labeled the clone an α1(I) collagen-like gene.

To establish the identity of this collagen gene, its homologous mRNA was sought and a more extensive nucleotide sequence was obtained. We report here that CosHcol1 hybridizes strongly with human fetal cartilage mRNA but not to mRNA from a large number of other sources, suggesting that its expression is cartilage specific. Analysis of the DNA sequence obtained shows that CosHcol1 is highly homologous to chicken α1(II) collagen, which is the major hyaline cartilage collagen. We therefore concluded that CosHcol1 probably contains the human α1(II) collagen gene.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

MATERIALS AND METHODS

Enzymes. Restriction endonucleases and DNA-modifying enzymes were purchased from Bethesda Research Laboratories, Boehringer Mannheim, or New England Biolabs.

DNA Preparation, Manipulation, and Sequencing. Standard DNA manipulations were performed as described by Maniatis *et al.* (15). DNA sequencing was carried out as described by Bankier and Barrell (16).

Preparation of Poly(A)$^+$ RNA. mRNA was prepared from 10^8 to 5×10^8 cultured cells and from human fetal sterna (16 and 22 weeks) and calvaria (10, 12, and 14 weeks). Cells were homogenized in a solution containing proteinase K (Boehringer Mannheim) at 200 μg/ml, 20 mM Tris·HCl at pH 7.6, 1 mM EDTA, and 2% sodium dodecyl sulfate. Fetal tissues were frozen in liquid N_2 and ground to a fine powder with a mortar and pestle. The pulverized tissue was then homogenized in the sodium dodecyl sulfate/proteinase K solution described above. Poly(A)$^+$ RNA was isolated as described by Cheah *et al.* (17). Usually 1–5% of total RNA was recovered as poly(A)$^+$ RNA.

RNA Blot Analyses. mRNAs [1–2 μg of poly(A)$^+$ RNA per gel slot] were denatured with glyoxal, electrophoresed in 0.8% agarose gels, and transferred to filters as described by Thomas (18) except that Pall Biodyne (Santa Monica, CA) A nylon membranes were used. Hybridization of the blots was performed as described (13).

Isolation of Overlapping Cosmid Clones. Overlapping clones from three human cosmid libraries were isolated (19), using the *Eco*RI fragments at the ends of the insert in CosHcol1 as probes.

RESULTS

Identification of a Homologous mRNA Species. To establish the identity of CosHcol1, it was necessary to find a homologous mRNA species. mRNAs were prepared from different tissues and cultured cell lines that synthesize characteristic collagen types. These included collagen type I (human fetal calvaria, fibroblast lines) (1), type II (human fetal sterna, rat chondrosarcoma, chicken sterna) (1, 20), type III (human fibroblast lines) (1), type IV (mouse parietal endoderm, human fibrosarcoma line HT1080) (21, 37), type V (human placenta, rhabdomyosarcoma line A204) (1, 22), and type VI (human placenta) (23). The ability of CosHcol1 to hybridize to these mRNAs was tested by blot hybridization.

CosHcol1 hybridized strongly to mRNA in preparations from only three of the tissues tested: a rat chondrosarcoma, chicken sternal cartilage, and human fetal sterna (Fig. 1). The Swarm rat chondrosarcoma is a transplantable tumor of cartilage origin and has been shown to synthesize pre-

Abbreviations: bp, base pair(s); kb, kilobase(s).

†Present address: Department of Biochemistry, Hong Kong University, Hong Kong.

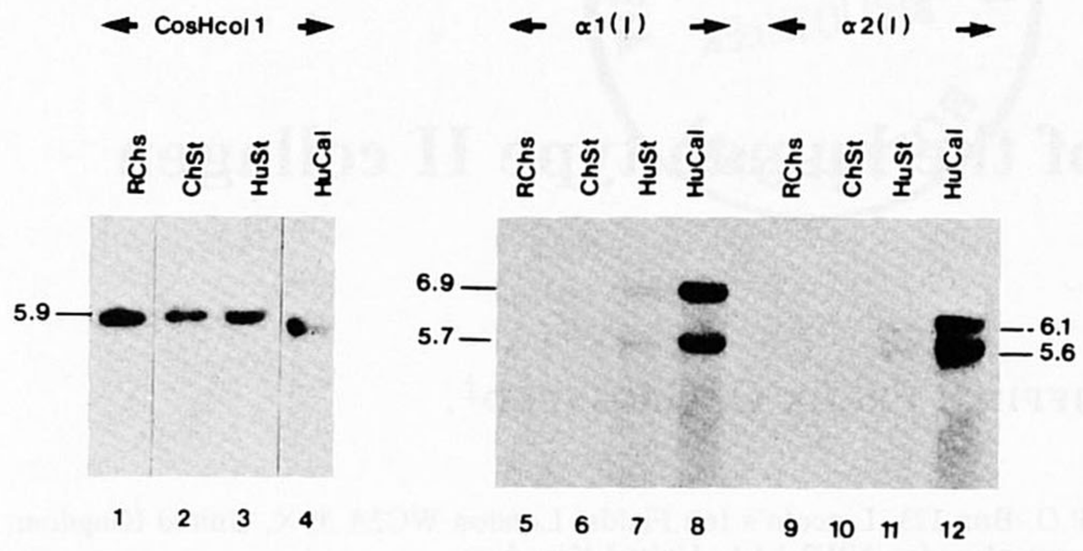

FIG. 1. Hybridization of CosHcol1 to mRNA preparations. Poly(A)$^+$ RNA preparations (1 μg) from rat chondrosarcoma (RChs), chicken sterna (ChSt), human sterna (and rib ends) (HuSt), and human calvaria (HuCal) were blotted and hybridized with ^{32}P-labeled nick-translated CosHcol1, or human α1(I) or α2(I) genomic probes. All tracks shown represent overnight exposures except for the hybridization of CosHcol1 to human calvaria mRNA, which was a 2-day exposure. Sizes are given in kb.

dominantly type II collagen (20), which is the major collagen synthesized by chondrocytes. Chicken sternal cartilage synthesizes mainly type II collagen, as well as other minor types (5, 24, 25). The human fetal sterna used consisted mainly of sternal cartilage with the ends of rib bones attached and would be expected to be synthesizing collagens typical of bone and of cartilage—i.e., types I and II, respectively. CosHcol1 therefore appeared to hybridize to a cartilage-specific mRNA, possibly type II collagen mRNA.

Fig. 1 shows that CosHcol1 hybridizes strongly to a distinct 5.9-kb band in mRNA from the cartilaginous tissues (tracks 1–3). The same probe hybridizes less strongly to two different bands of 5.7 and 6.9 kb in mRNA from human fetal calvaria (skull bones) (track 4), which synthesize type I but not type II collagen. These two bands correspond to the sizes of α1(I) mRNAs of type I collagen. To show that the calvaria were indeed synthesizing type I collagen, and to determine whether type I was also present in the other three tissues, the mRNA preparations were hybridized with human α1(I) and α2(I) probes (tracks 5–8 and 9–12, respectively). The human calvaria can be seen to contain large amounts of type I mRNAs (tracks 8 and 12). The human sterna produced small amounts of type I mRNAs (tracks 7 and 11), as was expected from the presence of the ends of the rib bones, but the rat chondrosarcoma and chicken sterna produced virtually no type I collagen. Cross-hybridization of CosHcol1 to the α1(I) mRNA was only seen where type I was present in very large amounts, as in human calvaria and chicken tendon (latter not shown). Cross-hybridization to α2(I) mRNA was not seen at all.

CosHcol1 did not hybridize with mRNA from other tissues tested, suggesting that it did not code for collagen types III, IV, V, or VI (data not shown).

Nucleotide Sequence Determination. The 3.8-kb *Eco*RI fragment and part of the adjacent 4.3-kb *Eco*RI fragment (see Fig. 4) were sequenced (Fig. 2). Exons were located by comparison with other collagen gene sequences and by following the A-G–G-T splicing rule (38). The nucleotide sequence was combined with that already published (ref. 13; Fig. 2).

The sequenced fragments encode from amino acid 832 to the end of the triple-helical region and the entire C propeptide and extend into the 3′ untranslated region. Comparison with other collagen genes shows the CosHcol1 sequence to be most similar to the chick α1(II) gene (Table 1), and these are shown aligned in Fig. 2. Where there are differences between the published genomic and cDNA chick α1(II) sequences (26, 31), the genomic sequence has been used in preference. The derived amino acid sequence from CosHcol1 is shown in Fig. 3, aligned with the chicken α1(II) and human α1(I) and α2(I) amino acid sequences. Although the DNA sequence of the chick α1(II) gene extends only up to exon 4, direct amino acid sequence analyses for exons 5, 6, and 7 show that the high homology continues further (Table 1 and Fig. 3; W. Butler, personal communication). As can be seen from Table 1, the amino acid homologies between CosHcol1 and the chick α1(II) gene in exons 1–7 range from 83% to 94% (89% overall), whereas the same exons show only 61–83% (71% overall) and 61–72% (65% overall) homology for the human α1(I) and α2(I) chains, respectively. Other published sequences—e.g., chick α1(III) collagen (30)—all show much lower homology than the chick α1(II) gene to CosHcol1 (data not shown). The exon–intron organization of the sequenced region of CosHcol1 is shown in Fig. 4. The sizes of exons 1–4 are conserved between CosHcol1 and the chick α1(II) gene. Intron sizes are different and no significant homology was detected. However, in other collagen genes, both intron and exon sizes have diverged from CosHcol1, although the locations of introns within the coding sequence have been conserved. We conclude from these results, and from the specific hybridization to human sternal mRNA, that CosHcol1 codes for α1(II) collagen.

The 3′ Untranslated Region. The sequence of the first 229 bp of the 3′ untranslated region of the human α1(II) collagen gene is shown in Fig. 2. A canonical polyadenylylation signal (A-A-T-A-A-A) is present 189 bp downstream from the stop codon. This, or a similar sequence, is necessary but not sufficient for polyadenylylation (36).

Boundaries of the Gene. To determine the extent of the type II gene in CosHcol1, *Eco*RI fragments from the 5′ and 3′ regions of the clone were used to screen other human cosmid libraries. Five overlapping clones were isolated, covering a total of 75 kb, of which 12.5 kb was 5′ and 25.7 kb was 3′ to CosHcol1. Fragments extending 5′ and 3′ to CosHcol1 and fragments from within the clone were tested for hybridization to rat chondrosarcoma mRNA on blots. The results are shown diagramatically in Fig. 4.

At the 5′ end, no hybridization to mRNA was detected with the 9.8-kb *Eco*RI fragment or the 5.9-kb *Eco*RI fragment (which extends 3.2 kb into CosHcol1). The adjacent 4.8-kb *Eco*RI fragment hybridized to mRNA, as did all the other *Eco*RI fragments in CosHcol1. The next 3′ fragment did not hybridize to mRNA. Since the stop codon and a polyadenylylation signal occur within the 3′ terminal *Eco*RI fragment of CosHcol1, and sequences 12.5 kb 5′ and 2.5 kb 3′ to CosHcol1 did not hybridize to mRNA, this clone probably contains the complete type II collagen gene. Hybridization of parts of the 4.8-kb *Eco*RI fragment has located

Table 1. Percent sequence homology between CosHcol1 and other collagen genes (amino acid/nucleotide)

Exon	Chick α1(II)	Human α1(I)	Human α2(I)
10	—/—	67/69	72/63
9	—/—	68/70	53/62
8	—/—	67/56	50/44
7	94/—	83/80	69/58
6	89*/—	61/72	67/72
5	91/—	72/74	61/62
4	91/82	68/68	67/66
3	83/83	71/79	63/73
2	85/84	68/74	62/68
1	94/82	73/70	67/70

Positions where deletion or insertion events have occurred have not been used in this comparison.
*Part of the sequence of this exon is based on amino acid composition and alignment for maximum homology (see Fig. 3 legend).

```
66TGAACCTG GACGAGAGgt gagcagtgag acccccctggg gtggccctga ttggggagag gggccctgtg agtctctgtg ctgggtcagc aaggacaagc  100
cccagtcagg gcctcggaga agggggcggc agcgctggcc gacaggcgaa agcctaggta caatgggaag gttgtcgggg agagagacgg gcatagagac  200
caagggctgc ttctggaagg aggagggaaa cttggtgagg aaactttggc ttcaaagtgt gagtgagttg ggcagaagag gagaggcctg ggcttctgag  300
aggggctggg ggagcagagg gggaggtgga cagaggacag ctctaggtgc gttcttgttt cactttgtcc ag66AA6CCC C6GT6CT6AT 66CCCCCCT6  400
6CASA6AT66 C6CT6CT66A 6TCAA6gtga gtgtctggtg tctgtgtgtg cagtgggttg gggaggacat tgcctcgggc ctgacaggtc agctgggggt  500
ggcaggttgg aacaagtctc atctcagcct agaaggacct tctgttcctg tctcttctgg aacattcttc tctgagcctg agacctctct cctgacag66  600
TAATC6T66T GAAACC66T6 CTGT666A6C TCCT66AACC CCT666CCCC CT66CTCCCC T66CCCC6CT 66TCCAACT6 6CAA6CAA66 A6ACA6A66A  700
6AA6CTgtaa gtatcctgga attcagtaaa agccgccttc ccctgcgcgg tggggctgag gcagtccctg ggtttccgca gtctctggac taaggagcag  800
tggcctcaga tgcagaggag gcccccacct gtcctggctt ttctctgacg ctgcgctcac tctctcctca g66T6CACAA 66CCCCATG6 GACCCTCA66  900
ACCA6CT66A 6CCC6666AA TCCA6gtgag tatccaagtg tcctgcactg agtccccacc agggataggc tgggagggca gccagcctcc aggtggttcc 1000
tggcctccag ccctgtgttt ccggggattc ctcagcttgg gtgggacagg agggggctcc tgtcctggcc ctgacctgac tcaatcggtg tctgtcttgt 1100
tcccag66TC CTCAA66CCC CA6A66T6AC AAA66A6AG6 CT66AGA6CC TG6C6A6A6A 66CCT6AA66 6ACACC6T6G CTTCACT66T CT6CA666TC 1200
T6CCC66CCC TCCTgtgagt gtcactgcct gcgtgggact tcccgaggcc tcctgccaca cagagcccac ttgagctccc tgtgctgcca ggacagcttg 1300
ggatcaccct aagcagtttc taggatttcc tcagggctgg agggaggagg aagtggaaag ggaatggggc tgggacataa agctgttccc ccagctccca 1400
gaatatagat agatatgtct gtgctgaccg tggccttttg cctcttcctt ctacacag66 TCCTTCT66A 6ACCAA66T6 CTTCT66TCC T6CT66TCCT 1500
TCT66CCCTA 6Agtaagtga catggagttg gaagatggag ggggcccttc agagagtgtg ggcctgtgtt cccatgggga gggaaatgct gctgcttctg 1600
gggaagctgt gggctcaggg gtcctcactc agtaatgggg gcaggactgg ctcatgtgcc tatggccaga aaagcgcctg aggccacaat ggctgtaaga 1700
caaacatgaa tcagcctctc gctgtcagac agaacagcat tttacaaaga ggagcttagg agggtaggca agccatggag ctatcctgct ggttcttggc 1800
caaatagaga ccaacttagg gttccatgac tgagcatgtg aagaactggg ggcggagtgg ctggtgctat caggacagcc acctacccag ccccagcgac 1900
tccccagcct tccctgtggt gaccactctt tcctcacgac ctctctctct tgcag66TCC TCCT66CCCC GTC6GTCCCT CTG6CAAA6A TG6TGCTAAT 2000
66AATCCCTG 6CCCCATTG6 6CCTCCT66T CCCC6TG6AC GATCA66CGA AACCG66CCCT 6CTgtaagtg tcctgactcc ttccctgctg tcgaggtgtc 2100
cctaccatcc gggaggcttg agctcttttt tgctcagggc ctcttttagg gcatcagcct gcagctaaca gtgatggcat cctttatcct gaggtctcct 2200
cagaggtcac agggcccatg atcagtgctg ggaaactgaa gagaagggct aaggaagaaa tagacatggt gctgtggttt ccttggtcct cgcctgctac 2300
acctccgccc cacccatggg gctgggaaga gggacactct agtacattct agcaaatggg gatggacatg gaggggcact ttcacacaat cctggctgat 2400
ctctctgttt cctgctgcag 66TCCTCCTG 6AAATCCT66 ACCCCCT66T CCTCCA66TC CCCCT66CCC T66CATCGAC ATGTCC6CCT TTGCT66CTT 2500
                     ......... ....C..C.. T..T..C... .....T..C. ....C...A. C......... .....T..T. .......AC.
A66CCC6A6A 6A6AA666CC CC6ACCCCCT 6CA6TACAT6 C666CC6ACC A66CA6CC66 T66CCTGA6A CA6CAT6AC6 CC6A66T66A T6CCACACTC 2600
6..T.A..C6 ......... ........A. C.6C...... A....A...6 ....6..... A..6...C.6 .....C.... T6........ C.....C...
AAGTCCCTCA ACAACCA6AT T6A6A6CATC C6CA6CCCC6 A666CTCCC6 CAA6AACCCT 6CTC6CACCT 6CA6A6ACCT 6AAACTCT6C CACCCT6A6T 2700
..A....... ....T..... ......... ......... ........AA 6......... ..CA.6.... ..C.C...A. C......... ..T..C....
66AA6A6T6g taagcttgga gaacaggatc ccctgccccg ggaagcaggg agtcatccct taggcctagc agcaagggag gagatgcccc ctagtacagg 2800
.......C.
gcagagctgg gcctggaagt ttccgccaga gggttcctct cttatttcac agcagagaag ctgcagccct ggcccctgtc ctgccatggc tacctggccg 2900
aggtgacctc agggtggact ccatccacca gctgggcact gcttctgctc tctttgcatg tgttcttcct tagggctgga cttagctcat gcagatctcc 3000
ctgcccctgc atcctcccag gtccccctcc tttcaggcca catgtgaacc tcatcccttg tccctgtagg cctctctgtc tctttcagtc aggcctgggt 3100
ctctcaagct tttgtgtctg tgcctgtctg agcccccatg ggtgctgcct cttccccctg cag6A6ACTA CT66ATT6AC CCCAACCAA6 6CT6CACCTT 3200
                                                                    ....T.. ......... ..6.....6. .........
66AC6CCAT6 AA6GTTTTCT 6CAACAT66A 6ACT66C6A6 ACTT6C6TCT ACCCCAATCC A6CAAACGTT CCCAA6AA6A ACT66T66A6 CA6CAA6A6C 3300
.........C ..A..A.... ......... ...A...... ..C....... ....G.CC.. CA6C.6.A.C ....6..... .........C ........C6
AAG6A6AA6A AACACATCT6 GTTT66A6AA ACCATCAAT6 GT66CTTCCA Tgtgagtacc tgggtgccct agatgatgag cagagatggc tcctcaaact 3400
..A..C.... .6...6.... .....C...6 ........C. .C..T..... C
ctttcttttc tttctccctg gaagctttta gcaccttccc catattttcc tccagttttc tgttgggctt gagaggaggg aaagaggagg aaaagtattt 3500
tttccccacg tggaggtggg aaaagaggtc ctctgagctt gctccactcc tggaagcaaa aatgtccaac tagctccctg ctgccccagt acccttgagg 3600
tccttgaacc atgaactctt ggcagcccct acagcccctg gtcccattga atgccagctc ccaggcctca cactgccgct ctctgcccca acagTTCA6C 3700
                                                                                                      .....
TAT66A6AT6 ACAATCT66C TCCCAACACT 6CCAAC6TCC A6AT6ACCTT CCTAC6CCT6 CT6TCCAC66 AA66CTCCCA 6AACATCACC TACCACT6CA 3800
..C..C.... .6..C...T. C........C ....6.A... .......... ...6.....C ........C. .6........ ....6..... ..........
A6AACA6CAT T6CCTATCT6 6AC6AA6CA6 CT66CAACCT CAA6AA66CC CT6CTCATCC A666CTCCAA TGAC6T66A6 ATCC666CA6 A666CAATA6 3900
.......... C.....CA.. .....6.A6A .6........ 6.....A... A.C....... .A..A..... C......... ...A.A..C. .......C..
CA66TTCAC6 TACACT6CCC T6AA66AT66 CT6CAC6gtg agtggggctg ccagagagaa gagctgcctg tgcccaaact gcctggagca gggctgaggg 4000
.........C ....6C.T.T ..6....C.. .......
ttggcccgcg gcagctgtca ggtcctaaag tgacaggatc atcagaggca tgagtttgag ggtcatgtag agaagatagg ctgagtgaca ggtgagagag 4100
aggcacatat cattccatct tctccattcc cctggctcag gggaacaaaa ccctacctgg aacccagtga ctactgtaga agtgttctcg caatgtgtac 4200
agggtgaaga agcggtcaca ggttgggagc tcactgtggg gagtggggaa ggaggggaag ggcagggtgg agaagggccc tgccgctaag gataggagtt 4300
gaagtggaga ggcctttggc aagccaagaa gaggtctcag gagccccctc agtgtggttc aaccttgtgg gctctgatgc tcgccagttt gttcagtttt 4400
gggcttctgg gcagctggaa ctgggtagca aggcatctac tgaacagagc ctcctccttt tttctccccc agAAACATAC C66TAA6T66 66CAA6ACT6 4500
                                                                    .....C.. T..C..A... ........6.
TTATC6A6TA CC66TCACA6 AA6ACCTCAC GCCTCCCCAT CATT6ACATT 6CACCCAT66 ACATA66A66 6CCC6A6CA6 6AATTC66T6 T66ACATA66 4600
.6........ ......6... ........6. ....6..... T6.A..T... .....T.... ....T..C.. A6....T... ..6..T..C. ....T..T..
6CC66TCT6C TTCTT6TAAA AACCT6AACC CA6AAACAAC ACAATCC6TT GCAAACCCAA A66ACCCAA6 TACTTTCCAA TCTCA6TCAC TCTA66ACTC 4700
C..A...... ......***
T6CACT6AAT 66CT6ACCT6 ACCT6AT6TC CATTCATCCC ACCCTCTCAC A6TTCG6ACT TTTCTCCCCT CTCTTTCTAA 6A6ACCT6AA CT666CA6AC 4800
T6CAAAATAA AATCTC66T6 TTCTATTTAT TTATT6TCTT CCT6T
     *****
     POLY A
```

FIG. 2. Nucleotide sequence from CosHcol1. The sequence of the 3.8-kb *Eco*RI fragment and part of the 4.3-kb *Eco*RI fragment (see Fig. 4) was determined. This was combined with the sequence previously published (13), which extends 720 base pairs (bp) into the 9.3-kb *Eco*RI fragment. A few errors in the earlier sequence have been corrected. Of the sequence not previously published, 93% of the protein-coding and 3′ untranslated regions and 70% of the intron sequences were determined on both strands. Upper-case letters, exons; lower-case letters, introns. Exons 1–4 are compared with the chicken α1(II) sequence (26). Only bases that differ from the CosHcol1 sequence are shown. The termination codon is marked at position 4617, and a canonical poly(A) addition signal is marked at 4806.

all of the mRNA homology within this fragment to a 1.3-kb region beginning 1 kb from the 5′ end (data not shown). This information, combined with the location of the putative polyadenylylation signal, provides us with an estimate of the gene length of 30 kb.

DISCUSSION

We have identified the human cosmid clone CosHcol1. Strong hybridization to a 5.9-kb cartilage-specific mRNA and comparison with the chick α1(II) collagen gene sequence suggest that CosHcol1 contains the human α1(II) collagen gene (*COL2A1*), of which we have sequenced 10 exons at the 3′ end. This represents approximately 15% of the estimated gene length and over 30% of the protein-encoding sequence. mRNA hybridization and DNA sequence data together provide evidence that CosHcol1 may contain the entire human type II collagen gene and that the gene is approximately 30 kb in length.

The isolation of a genomic human α1(II) collagen clone has recently been reported elsewhere (39), and the published sequence of the 3′ end of exon 4 and of the small fragment of adjacent intron matches exactly with CosHcol1. This suggests that we have cloned the same gene even though no homologous mRNA was described in that report. Furthermore, a 540-bp cDNA clone has recently been isolated from human fetal cartilage (E. Vuorio, personal communication)

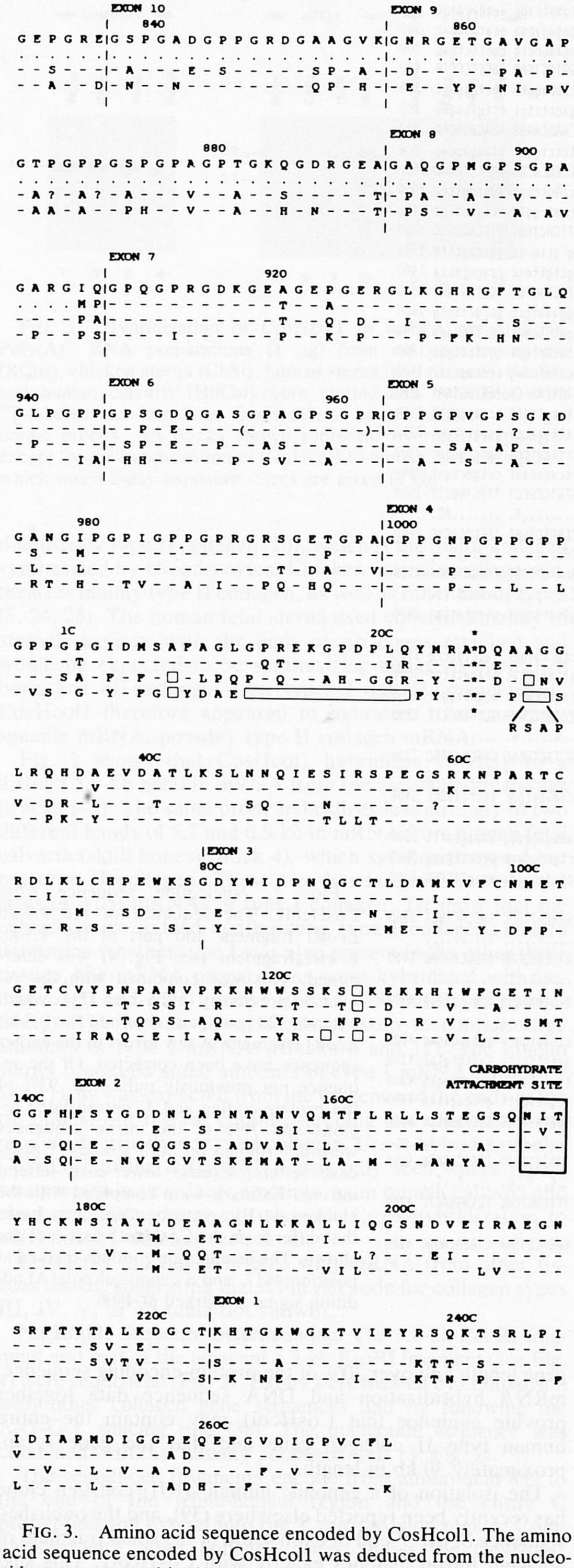

FIG. 3. Amino acid sequence encoded by CosHcol1. The amino acid sequence encoded by CosHcol1 was deduced from the nucleotide sequence, and is shown here, compared with other sequences. The standard one-letter code (27) is used. Line 1, CosHcol1-derived amino acid sequence; line 2, chicken α1(II) amino acid sequence (26); line 3, human α1(I) amino acid sequence (28); line 4, human α2(I) amino acid sequence (29). A dash indicates the presence of the same residue as in the CosHcol1-derived sequence. Numbers refer to the position in the helix, or in the carboxyl nonhelical domain, and are based on the α1(I) sequence; the numbers begin above the residue to which they refer. All sequences were derived from nucleotide sequences except for the chicken α1(II) residues 908–999, which have been determined directly (W. Butler, personal communication). The order of residues 955–962 within this is not known, but the amino acid composition is known, and residues have been aligned for maximum homology. ..., region not sequenced; ?, a residue not confirmed; ★, end of the mature collagen molecule; |, exon boundary; open boxes, deletions/insertions introduced for maximum interchain homology. The carbohydrate attachment site, which lies within a highly conserved region at the nucleotide level (30), is shown.

that is identical in DNA sequence with CosHcol1, extending from exon 1 to exon 4. The poly(A)$^+$ mRNA from which the cDNA clone was derived was shown to program the synthesis of α1(II) collagen *in vitro* (40). This strongly supports the idea that CosHcol1 does not carry a pseudogene.

It remains possible that CosHcol1 carries an α1(II)-related gene. For example, the minor cartilage collagen chain 3α is highly homologous to α1(II) collagen (25, 41, 42) and may or may not be genetically distinct. However, no evidence of other sequences homologous to CosHcol1 has been found in Southern hybridizations, under conditions in which cross-hybridization with the α1(I) gene was visible (43), and copy number estimates are consistent with only one copy of the gene per haploid genome (R. Dalgleish, personal communication). It has been claimed that the 3α collagen chain differs from α1(II) collagen in that it has a much larger peptide in place of the α1(II) cyanogen bromide peptide CB9,7 (42). The sequence presented in this paper covers the whole of the region encoding CB9,7 and agrees with the human α1(II) cyanogen bromide map (44). Nevertheless, absolute proof of this gene's identity will require a comparison of the amino acid sequence of human type II collagen with that derived from the DNA sequence.

The α1(II) gene has been assigned to chromosome 12 (43, 45). The gene is therefore not linked to the α1(I) or α2(I) collagen genes, which map to chromosomes 17 and 7, respectively (46–49), or to the α1(III) or α1(IV) collagen genes, which map to chromosomes 2 and 13, respectively (43).

The isolation of the α1(II) collagen gene is a major step towards the identification of those connective tissue disorders for which an abnormality in this gene is the primary defect. CosHcol1 should prove to be particularly useful in this respect because it appears to carry the entire gene. Several polymorphisms with high allele frequencies have been identified in this gene (50–52) and are being used for linkage analyses in families with some of these disorders.

We thank the following for their generous gifts: David Rowe and Raymond Dalgleish for human α1(I) and α2(I) DNA probes, respectively; Markku Kurkinen for parietal endoderm and HT1080 mRNAs; Claudio Schneider for human placental mRNA; Roger Mason, Lance Liotta, and Bryan Sykes for Swarm rat chondrosarcoma tissue, A204 cell line, and human fetal tissue, respectively; and Linda Sandell, Bill Butler, and Eero Vuorio for chicken α1(II) DNA sequence, chicken α1(II) amino acid sequence, and human α1(II) cDNA sequence, respectively. We are grateful to Frances Benham, Elizabeth Weiss, Richard Flavell, Adrian Kelly, Toby Gibson, and Mike Owen for helpful discussions and technical advice. This work was supported in part by a grant from Hong Kong University and the Croucher Foundation, Hong Kong.

1. Bornstein, P. & Sage, H. (1980) *Annu. Rev. Biochem.* **49,** 957–1103.
2. Furthmayr, H., Wiedemann, H., Timpl, R., Odermatt, E. & Engel, J. (1983) *Biochem. J.* **211,** 303–311.
3. Bentz, H., Morris, N. P., Murray, L. W., Sakai, L. Y., Hollister, D. W. & Burgeson, R. E. (1983) *Proc. Natl. Acad. Sci. USA* **80,**3168–3172.
4. Sage, H., Trueb, B. & Bornstein, P. (1983) *J. Biol. Chem.* **258,** 13391–13401.
5. Ninomiya, Y. & Olsen, B. R. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 3014–3018.

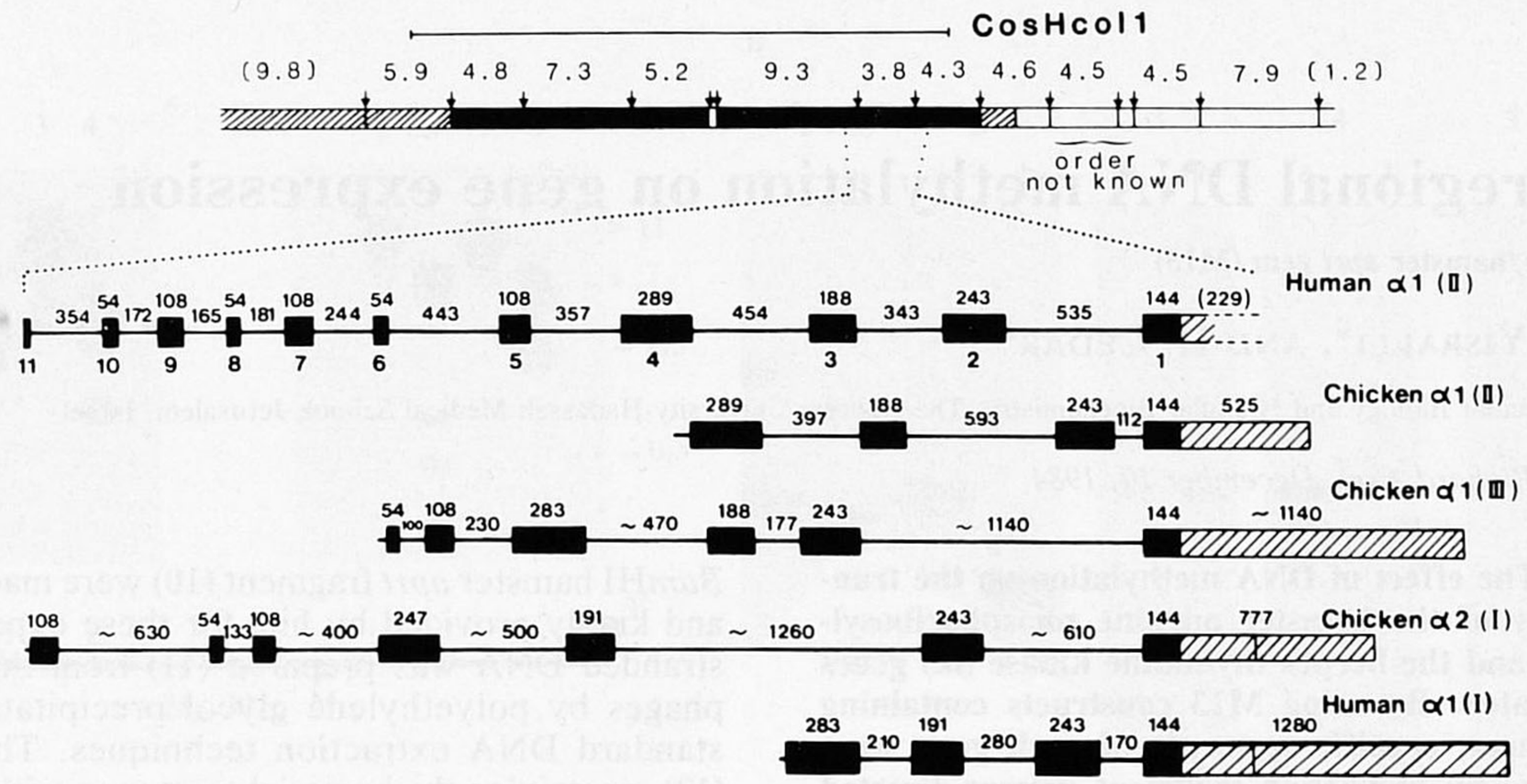

FIG. 4. Organization of the human α1(II) collagen gene. (*Upper*) Restriction map showing CosHcol1 within 75 kb of genomic DNA. Fragment sizes (in kb) and positions are composites from five different cosmid clones overlapping CosHcol1 and extending 5′ and 3′ to it. The positions of *Eco*RI sites are indicated by the arrows. The order of the 4.5- and 1.1-kb fragments (bracketed) was not determined. Filled boxes, *Eco*RI fragments hybridizing to rat chondrosarcoma mRNA (data not shown); hatched boxes, fragments that do not hybridize to mRNA; open boxes, fragments not tested. Only 2.5 kb of the 4.6-kb *Eco*RI fragment was tested. Repetitive sequences were present in all fragments represented by hatched or empty boxes except for the 0.8-kb (the small fragment in CosHcol1) and 1.2-kb fragments, and also in the 5.2-kb fragment. (*Lower*) The region sequenced is expanded to show its organization into exons and introns (sizes given in bp) and is compared with the chicken α1(II) gene (26), the chicken α1(III) gene (30, 32), the chicken α2(I) gene (30, 33–35), and the human α1(I) gene (10). Filled boxes indicate protein-coding regions, while hatched boxes indicate 3′ untranslated regions. The end of this region has not been determined for CosHcol1. Vertical lines within the 3′ untranslated regions indicate the presence of alternative polyadenylylation sites.

6. Adamson, E. D. (1982) in *Collagen in Health and Disease*, eds. Weiss, J. B. & Jayson, M. I. V. (Churchill–Livingstone, London), pp. 218–243.
7. McKusick, V. A. (1972) *Heritable Disorders of Connective Tissue* (Mosby, St. Louis, MO), 4th Ed.
8. Hollister, D. W., Byers, P. H. & Holbrook, K. A. (1982) *Adv. Hum. Genet.* **12,** 1–87.
9. Chu, M.-L., Myers, J. C., Bernard, M. P., Ding, J.-F. & Ramirez, F. (1982) *Nucleic Acids Res.* **10,** 5925–5934.
10. Chu, M.-L., de Wet, W., Bernard, M., Ding, J.-F., Morabito, M., Myers, J., Williams, C. & Ramirez, F. (1984) *Nature (London)* **310,** 337–340.
11. Myers, J. C., Chu, M.-L., Faro, S. H., Clark, W. J., Prockop, D. J. & Ramirez, F. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 3516–3520.
12. Myers, J. C., Dickson, L. A., de Wet, W. J., Bernard, M. P., Chu, M.-L., Di Liberto, M., Pepe, G., Sangiorgi, F. O. & Ramirez, F. (1983) *J. Biol. Chem.* **258,** 10128–10135.
13. Weiss, E. H., Cheah, K. S. E., Grosveld, F. G., Dahl, H. H. M., Solomon, E. & Flavell, R. A. (1982) *Nucleic Acids Res.* **10,** 1981–1994.
14. Lehrach, H., Frischauf, A. M., Hanahan, D., Wozney, J., Fuller, F. & Boedtker, H. (1979) *Biochemistry* **18,** 3146–3152.
15. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
16. Bankier, A. T. & Barrell, B. G. (1983) *Techniques in the Life Sciences* (Elsevier, Limerick, Ireland), Vol. B5, pp. 1–34.
17. Cheah, K. S. E., Grant, M. E. & Jackson, D. S. (1979) *Biochem. Biophys. Res. Commun.* **91,** 1025–1031.
18. Thomas, P. S. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 5201–5205.
19. Grosveld, F. G., Dahl, H. M., de Boer, E. & Flavell, R. A. (1981) *Gene* **13,** 227–237.
20. Smith, B. D., Martin, G. R., Miller, E. J., Dorfman, A. & Swarm, R. (1975) *Arch. Biochem. Biophys.* **166,** 181–186.
21. Kurkinen, M., Barlow, D. P., Helfman, D. M., Williams, J. G. & Hogan, B. L. M. (1983) *Nucleic Acids Res.* **11,** 6199–6209.
22. Alitalo, K., Myllyla, R., Pritzl, P., Vaheri, A. & Bornstein, P. (1982) *J. Biol. Chem.* **257,** 9016–9024.
23. Odermatt, E., Ristell, J., van Delden, V. & Timpl, R. (1983) *Biochem. J.* **211,** 295–302.
24. Miller, E. J. (1971) *Biochemistry* **10,** 1652–1659.
25. Burgeson, R. E. & Hollister, D. W. (1979) *Biochem. Biophys. Res. Commun.* **87,** 1124–1131.
26. Sandell, L. J., Prentice, H. L., Kravis, D. & Upholt, W. B. (1984) *J. Biol. Chem.* **259,** 7826–7834.
27. IUPAC-IUB Commission on Biochemical Nomenclature (1968) *Eur. J. Biochem.* **5,** 151–153.
28. Bernard, M. P., Chu, M.-L., Myers, J. C., Ramirez, F., Eikenberry, E. F. & Prockop, D. J. (1983) *Biochemistry* **22,** 5213–5223.
29. Bernard, M. P., Myers, J. C., Chu, M.-L., Ramirez, F. & Eikenberry, E. F. (1983) *Biochemistry* **22,** 1139–1145.
30. Yamada, Y., Kuhn, K. & de Crombrugghe, B. (1983) *Nucleic Acids Res.* **11,** 2733–2744.
31. Ninomiya, Y., Showalter, A. M., van der Rest, M., Seidah, N. G., Chretien, M. & Olsen, B. R. (1984) *Biochemistry* **23,** 617–624.
32. Yamada, Y., Liau, G., Mudryj, M., Obici, S. & de Crombrugghe, B. (1984) *Nature (London)* **310,** 333–337.
33. Dickson, L. A., Ninomiya, Y., Bernard, M. P., Pesciotta, D. M., Parsons, J., Green, G., Eikenberry, E. F., de Crombrugghe, B., Vogeli, G., Pastan, I., Fietzek, P. P. & Olsen, B. R. (1981) *J. Biol. Chem.* **256,** 8407–8415.
34. Wozney, J., Hanahan, D., Morimoto, R., Boedtker, H. & Doty, P. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 712–716.
35. Aho, S., Tate, V. & Boedtker, H. (1983) *Nucleic Acids Res.* **11,** 5443–5450.
36. Fitzgerald, M. & Shenk, T. (1981) *Cell* **24,** 251–260.
37. Pihlajaniemi, T., Myllylä, R., Alitalo, K., Vaheri, A. & Kivirikko, K. I. (1981) *Biochemistry* **20,** 7409–7415.
38. Breathnach, R. & Chambon, P. (1981) *Annu. Rev. Biochem.* **50,** 349–383.
39. Strom, C. M. & Upholt, W. B. (1984) *Nucleic Acids Res.* **12,** 1025–1038.
40. Vuorio, E., Elima, K., Pulkkinen, J. & Viitanen, A.-M. (1984) *FEBS Lett.* **174,** 238–242.
41. Furuto, D. K. & Miller, E. J. (1983) *Arch. Biochem. Biophys.* **226,** 604–611.
42. Eyre, D. R., Wu, J.-J. & Woolley, D. E. (1984) *Biochem. Biophys. Res. Commun.* **118,** 724–729.
43. Solomon, E., Hiorns, L. R., Spurr, N., Kurkinen, M., Barlow, D., Hogan, B. L. M. & Dalgleish, R. (1985) *Proc. Natl. Acad. Sci. USA* **82,** in press.
44. Miller, E. J. (1984) in *Extracellular Matrix Biochemistry*, eds. Piez, K. A. & Reddi, A. H. (Elsevier, New York), pp. 41–81.
45. Solomon, E., Hiorns, L., Cheah, K. S. E., Parkar, M., Weiss, E. & Flavell, R. A. (1984) *Cytogenet. Cell Genet.* **37,** 588.
46. Huerre, C., Junien, C., Weil, D., Chu, M.-L., Morabito, M., Van Cong, N., Myers, J. C., Foubert, C., Gross, M.-S., Prockop, D. J.,Bone, A., Kaplan, J.-C., De la Chapelle, A. & Ramirez, F. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 6627–6630.
47. Solomon, E., Hiorns, L., Sheer, D. & Rowe, D. (1984b) *Ann. Hum. Genet.* **48,** 39–42.
48. Junien, C., Weil, D., Myers, J. C., Van Cong, N., Chu, M.-L., Foubert, C., Gross, M.-S., Prockop, D. J., Kaplan, J.-C. & Ramirez, F. (1982) *Am. J. Hum. Genet.* **34,** 381–387.
49. Solomon, E., Hiorns, L., Dalgleish, R., Tolstoshev, P., Crystal, R. & Sykes, B. (1983) *Cytogenet. Cell Genet.* **35,** 64–66.
50. Driesel, A. J., Schumacher, A. M. & Flavell, R. A. (1982) *Hum. Genet.* **62,** 175–176.
51. Sykes, B. (1983) *Dis. Markers* **1,** 141–146.
52. Sykes, B., Smith, R., Vipond, S., Paterson, C., Cheah, K. & Solomon, E. (1985) *J. Med. Genet.*, in press.