Nucleic Acids Research, 1997, Vol. 25, No. 1 **181–187**

The human type I collagen mutation database

Raymond Dalgleish*

Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, UK

Received August 30, 1996; Accepted October 8, 1996

DDBJ/EMBL/GenBank accession nos Z74615 and Z74616

ABSTRACT

Type I collagen is the most abundant and ubiquitously distributed of the collagen family of proteins. It is a heterotrimer comprising two α1(I) chains and one α2(I) chain which are encoded by the unlinked loci *COL1A1* and *COL1A2* respectively. Mutations at these loci result primarily in the connective tissue disorders osteogenesis imperfecta and Ehlers–Danlos syndrome types VIIA and VIIB. Two instances of osteoporosis and a single instance of Marfan syndrome are also the result of mutations at these loci. The mutation data are accessible on the world wide web at http://www.le.ac.uk/depts/ge/collagen/collagen.html

INTRODUCTION

The collagens are primary components of the extracellular matrix. They are the most abundant proteins in the human body and are members of a complex superfamily (1). Some collagen types are ubiquitously expressed, while others have a more limited distribution. Each has a specific function, or set of functions, and there are extensive interactions with other connective tissue components. These interactions combined with the complex nature of collagen biosynthesis (2,3) result in an exquisitely mutation-sensitive biological system. The clinical phenotypes resulting from collagen mutations are wide-ranging in their manifestations and severity (4). Of the 19 known vertebrate collagen types, type I is the most abundant and widely expressed collagen in humans. Consequently, it is the best studied and more is known about mutations of type I collagen than any other type.

The basic subunits of collagen are the α-chains which consist predominantly of repeating Gly-Xaa-Yaa tri-peptide motifs. The presence of glycine at every third amino acid is essential to allow the α-chains to adopt the characteristic collagen triple helix. These chains are initially synthesised with N- and C-terminal propeptides which are later enzymatically cleaved. A signal peptide (preceding the N-terminal propeptide) directs the nascent protein to endoplasmic reticulum upon which it is enzymatically removed (Table 1). The Xaa and Yaa positions of the tri-peptide repeats are frequently occupied by the imino acids proline and hydroxyproline respectively and there are 338 such tri-peptide repeats in the mature α-chains of type I collagen. The combination of glycines at every third amino acid and the presence of the imino acids allows three collagen α-chains to self assemble into a right-handed triple helix which is stabilised by hydrogen bonding and other charge interactions. The winding of the triple helix initiates by way of interaction of the C-terminal propeptides of the three participating chains and proceeds in a C- to N-terminal direction (2,3). Type I collagen is a heterotrimer comprising two α1(I) chains and one α2(I) chain.

THE GENES ENCODING TYPE I COLLAGEN

The α1(I) and α2(I) chains of type I collagen are encoded at the unlinked loci *COL1A1* and *COL1A2* respectively. *COL1A1* is ~18 kb in size and is located at 17q21.3–q22. *COL1A2* is ~38 kb and is located at 7q21.3–q22.1 (5,6).

The most striking feature of these genes is that the exons encoding the triple-helical portions of the collagen chains are mostly 54 bp in length. Those that are not are either twice 54, three times 54, or combinations of 45 and 54 bp exons (Table 2). In every instance the exon size is an exact multiple of 9 bp which can encode a Gly–Xaa–Yaa triplet and exons begin with a Gly codon and end with a Yaa position codon. The consequence of this organisation is that exon-skipping mutations maintain the repeating Gly–Xaa–Yaa triplet collagen structure.

THE NUMBERING OF AMINO ACIDS, DNA BASES AND EXONS

For historical reasons the numbering systems used to define the amino acids of the α1(I) and α2(I) chains, and the exons of the genes which encode them, are complicated. This has been a constant source of confusion, especially for those new to the field, and so some discussion of the numbering systems in past and present use is warranted.

Amino acid numbers

Collagen α-chains can be considered to comprise three domains; the N-terminal domain, the collagen domain and the C-terminal domain (Table 1). These domains are functional as well as descriptive. Strictly speaking, the α1(I) and α2(I) collagen chains are initially synthesised with signal peptides of 22 amino acids (Table 1) and are known as preproα-chains. Cleavage of the signal peptides produces molecules known as proα-chains which have both N- and C-terminal propeptides (Table 1). Once the proα-chains have assembled into the characteristic collagen triple helix, they are secreted into the extracellular space where the propeptides are cleaved off by specific proteases. The cleavage process leaves telopeptides flanking the extended triple-helical collagen region at each end of both chains (Table 1). The individual chains of fully processed collagen molecules are known as α-chains.

*Tel: +44 116 252 3425; Fax: +44 116 252 3378; Email: ray@le.ac.uk

Table 1. The domains of type I collagen and the number of amino acids comprising them

Domain	proα1(I)	proα2(I)
N-terminal domain	161	79
signal peptide	22	22
N-propeptide	139	57
Collagen domain	1057	1040
N-telopeptide	17	11
Triple helix	1014	1014
C-telopeptide	26	15
C-terminal domain		
C-propeptide	246	247

By convention, amino acid 1 is the first glycine of the first Gly–X–Y repeat of the α-chains rather than of the first amino acid of the primary translation products, of the proα-chains, or of the mature α-chains. In practice, this has rarely led to any confusion in spite of there being no formalised numbering system for the signal peptides, N-propeptides or the N-telopeptides. However, a potential source of confusion lies in the numbering of the amino acids of the C-propeptides. Formally, the conventional numbering system ends at 1014, the last amino acid of the triple-helical region. In practice, it is convenient to directly continue the numbering such that the C-telopeptide comprises amino acids 1015–1040 in the α1(I) chain and 1015–1029 in the α2(I) chain. Similarly, the C-propeptides comprise amino acids 1041–1286 and 1030–1276 respectively for the two chains. It should be noted that this numbering practice is by no means universal and many reports of C-propeptide mutations use a numbering system that starts at the first amino acid of the C-propeptide.

Exon numbers

When human collagen cDNAs were first isolated, they represented sequences predominantly from the 3′-ends of the mRNAs. When the first genomic clones were isolated, using these cDNAs as probes, they consequently represented the 3′-end of the gene. Not knowing the full exon/intron structure of the genes, investigators decided initially to number the exons from the 3′-end of the genes. It was only when the entire gene structures were elucidated that the exons were then re-numbered in the conventional 5′ to 3′ direction. This lead to instances of mutations being reported using the reverse-order system (7,8). Eventually, it emerged that *COL1A1* consisted of 51 exons and *COL1A2* of 52 and, for a short time, this lead to further confusion. Amino acids 568–603 are encoded by a single exon of 108 bp in *COL1A1*, but by two exons each of 54 bp in *COL1A2*. These two exons of *COL1A2* are exons 33 and 34 and initially the single analogous exon in *COL1A1* was numbered exon 33. However, it was soon realised that such a numbering system would lead to confusion when discussing analogous exons encoding amino acids C-terminal to position 603. For this reason, the 108 bp exon of *COL1A1* encoding amino acids 568–603 of the α1(I) chain is now designated exon ‘33/34’ (Table 2). An example of this exon-numbering confusion is illustrated in the reporting of a 9 bp deletion in *COL1A1* (9) which is described as being in exon 43 when, in fact, it is in exon 44 using the present numbering system.

Table 2. The exon structure of the type I collagen triple-helical region

Exon number	Size (bp)	Amino acids encoded	α1(I) cDNA base numbers	α2(I) cDNA base numbers
6	72[a]	1–3[a]	654–662	410–418
7	45	4–18	663–707	419–463
8	54	19–36	708–761	464–517
9	54	37–54	762–815	518–571
10	54	55–72	816–869	572–625
11	54	73–90	870–923	626–679
12	54	91–108	924–977	680–733
13	45	109–123	978–1022	734–778
14	54	124–141	1023–1076	779–832
15	45	142–156	1077–1121	833–877
16	54	157–174	1122–1175	878–931
17	99	175–207	1176–1274	932–1030
18	45	208–222	1174–1319	1031–1075
19	99	223–255	1320–1418	1076–1174
20	54	256–273	1419–1472	1175–1228
21	108	274–309	1473–1580	1229–1336
22	54	310–327	1581–1634	1337–1390
23	99	328–360	1635–1733	1391–1489
24	54	361–378	1734–1787	1490–1543
25	99	379–411	1788–1886	1544–1642
26	54	412–429	1887–1940	1643–1696
27	54	430–447	1941–1994	1697–1750
28	54	448–465	1995–2048	1751–1804
29	54	466–483	2049–2102	1805–1858
30	45	484–498	2103–2147	1859–1903
31	99	499–532	2148–2246	1904–2002
32	108	533–567	2247–2354	2003–2110
33	54[b]	568–585	2355–2408	2111–2164
34	54[b]	586–603	2409–2462	2165–2218
35	54	604–621	2463–2516	2219–2272
36	54	622–639	2517–2570	2273–2326
37	108	640–675	2571–2678	2327–2434
38	54	676–693	2679–2732	2435–2488
39	54	694–711	2733–2786	2489–2542
40	162	712–765	2787–2948	2543–2704
41	108	766–801	2949–3056	2705–2812
42	108	802–837	3057–3164	2813–2920
43	54	838–855	3165–3218	2921–2974
44	108	856–891	3219–3326	2975–3082
45	54	892–909	3327–3380	3083–3136
46	108	910–945	3381–3488	3137–3244
47	54	946–963	3489–3542	3245–3298
48	108	964–999	3543–3650	3299–3406
49	283[c]	1000–1014[c]	3651–3695	3407–3451

[a]Exon 6 encodes part of the N-propeptide, the entire N-telopeptide and the first three amino acids of the triple-helical region. The cDNA base numbers are for triple-helical region amino acids only.
[b]In COL1A1 there is a single 108 bp exon designated 33/34.
[c]Exon 49 encodes the last 15 amino acids of the triple-helical region, the entire C-telopeptide and part of the C-propeptide. The cDNA base numbers are for triple-helical region amino acids only.

Table 3. Amino acid substitutions in COL1A1

Mutation	Phenotype	Reference(s)
Gly19Cys GGT→TGT nt708	Osteoporosis	Nicholls *et al.* IV International Conference on OI 48 1990
Gly43Cys GGT→TGT nt780	I	Shapiro *et al.* J Clin Invest 89:567-573 1992
Gly46Cys GGT→TGT nt789	I	Byers *et al.* J Med Genet 28:433-442 1991
Gly79Arg GGA→AGA nt 888	I	Redford-Badwal *et al.* J Clin Invest 97:1035-1040 1996
Gly85Arg GGA→AGA nt906	I	Deak *et al.* J Biol Chem 266:21827-21832 1991
Gly85Val GGA→GTA nt907	I	Valli *et al.* Eur J Biochem 217:77-82 1993
Gly94Cys GGT→TGT nt933	I	Starman *et al.* J Clin Invest 84:1206-1214 1989
Gly97Asp GGT→GAT nt943	II	Lightfoot *et al.* J Biol Chem 267:25521-25528 1992
Gly154Arg GGG→AGG nt1113	III III I	Pruchno *et al.* Hum Genet 87:33-40 1991 Pruchno *et al.* Hum Genet 87:33-40 1991 Zhuang *et al.* Am J Med Genet 61:111-116 1996
Gly172Arg GGT→CGT nt1167	III	Mackay *et al.* Hum Mutat 3:324-326 1994
Gly175Cys GGT→TGT nt1176	III/IV	Wirtz *et al.* Connect Tissue Res 29:1-11 1993
Gly178Cys GGT→TGT nt1185	IV	Valli *et al.* J Biol Chem 266:1872-1878 1991
Gly205Cys GGC→TGC nt1266	I	Byers. Trends Genet 6:293-300 1990
Gly220Asp GGT→GAT nt1312	II	Culbert *et al.* Biochem J 311:815-820 1995
Gly223Cys GGT→TGT nt1320	I	Byers. Trends Genet 6:293-300 1990
Gly226Cys GGT→TGT nt1329	III/IV(?)	Wilcox *et al.* Am J Hum Genet 55:A367 (abstract) 1994
Arg237Stop CGA→TGA nt 1362	I	Redford-Badwal *et al.* J Clin Invest 97:1035-1040 1996
Gly244Cys GGC→TGC nt1383	II	Fertala *et al.* Biochem J 289:195-199 1993
Gly247Ser GGC→AGC nt1392	II/III	Mackay *et al.* Hum Mol Genet 2:1155-1160 1993
Gly256Val GGT→GTT nt1420	II	Patterson *et al.* J Biol Chem 264:10083-10087 1989
Gly298Arg GGA→CGA nt1545	II	Byers *et al.* J Med Genet 28:433-442 1991
Gly352Ser GGT→AGT nt1707	IVB IV II	Bateman *et al.* Biochem J 288:131-135 1992 Marini *et al.* J Biol Chem 268:2667-2673 1993 Mackay *et al.* Hum Mol Genet 2:1155-1160 1993
Gly382Ser GGT→AGT nt1797	IV	Mackay *et al.* Hum Mol Genet 2:1155-1160 1993
Gly382Cys GGT→TGT nt1797	IV	Byers Trends Genet 6:293-300 1990
Gly391Arg GGC→CGC nt1824	II	Bateman *et al.* J Biol Chem 262:7021-7027 1987
Gly415Cys GGC→TGC nt1896	III/IV	Nicholls *et al.* J Med Genet 28:757-764 1991
Gly415Ser GGC→AGC nt1896	III/IV II/III	Bateman *et al.* Biochem J 288:131-135 1992 Mottes *et al.* Hum Mutat 2:196-204 1993
Gly526Cys GGC→TGC nt2229	III	Starman *et al.* J Clin Invest 84:1206-1214 1989
Gly541Asp GGT→GAT nt2275	II	Zhuang *et al.* Am J Hum Genet 48:1186-1191 1991
Gly541Ser GGT→AGT nt2274	III	Mackay *et al.* Hum Mol Genet 2:1155-1160 1993
Gly550Arg GGA→AGA nt2301	II	Wallis *et al.* Am J Hum Genet 46:1034-1040 1990
Gly559Asp GGT→GAT nt2329	II	Cohn *et al.* IV International Conference on OI 47 1990
Gly565Ser GGT→AGT nt2346	II	Bateman *et al.* Biochem J 288:131-135 1992
Gly565Val GGT→GTT nt2347	II	Mackay *et al.* Hum Genet 91:439-444 1993
Gly589Ser GGC→AGC nt2418	III IV(?)	Forlino *et al.* Hum Mol Genet 3: 2201-2206 1995 Zhuang *et al.* Hum Mutat 7:89-99 1996
Gly598Ser GGT→AGT nt2445	II	Westerhausen *et al.* J Biol Chem 265:13995-14000 1990
Arg618His CGT→CAT nt2506	Connective tissue weakness	Superti-Furga *et al.* Matrix Biology 14:385 (abstract) 1994
Gly631Ser GGC→AGC nt2544	II	Westerhausen *et al.* J Biol Chem 265:13995-14000 1990
Gly637Val GGC→GTC nt2563	II	Tsuneyoshi *et al.* J Biol Chem 266:15608-15613 1991
Gly664Ser GGG→AGG nt2643	II	Culbert *et al.* Biochem J 311, 815-820 1995
Gly667Arg GGA→AGA nt2652	II	Bateman *et al.* J Biol Chem 263:11627-11630 1988
Gly673Asp GGC→GAC nt2671	II	Cohn et al. IV International Conference on OI 47 1990
Gly691Cys GGT→TGT nt2724	II	Steinmann *et al.* Biochem J 279:747-752 1991
Gly718Cys GGC→TGC nt2805	II	Starman *et al.* J Clin Invest 84:1206-1214 1989
Gly748Cys GGT→TGT nt2895	II	Vogel *et al.* J Biol Chem 262:14737-14744 1987
Gly802Val GGT→GTT nt3058	II	Bonaventure *et al.* Hum Genet 89:640-646 1992
Gly832Ser GGT→AGT nt3147	IV	Marini *et al.* J Biol Chem 264:11893-11900 1989
Gly844Ser GGT→AGT nt3183	III	Pack *et al.* J Biol Chem 264:19694-19699 1989
Gly847Arg GGA→AGA nt3192	II	Wallis *et al.* J Biol Chem 265:18628-18633 1990
Gly862Ser GGC→AGC nt3237	IIa III III	Virdi *et al.* Hum Genet 93:287-290 1994 Namikawa *et al.* Hum Genet 95: 666-670 1995 Zhuang *et al.* Hum Mutat 7:89-99 1996
Gly883Ser GGC→AGC nt3300	IV	Lightfoot *et al.* J Biol Chem 269: 30352-30357 1994
Gly883Asp GGC→GAC nt3301	II	Cohn *et al.* Am J Hum Genet 46:591-601 1990
Gly901Ser GGC→AGC nt3354	I	Mottes *et al.* Hum Genet 89:480-484 1992
Gly904Cys GGC→TGC nt3363	II	Constantinou *et al.* J Clin Invest 83:574-584 1989
Gly910Ala GGA→GCA nt3382	II	Valli *et al.* Eur J Biochem 211:415-419 1993
Gly913Ser GGC→AGC nt3390	II	Cohn *et al.* Matrix 10:236 (abstract) 1990
Gly928Ala GGC→GCC nt3436	II	Lamande *et al.* J Biol Chem 264:15809-15812 1989
Gly946Cys GGC→TGC nt3489	II	Kurosaka *et al.* J Biochem 115:853-857 1994
Arg963Stop CGA→TGA nt3540	I	Willing *et al.* Am J Hum Genet 55:638-647 1994
Gly964Ser GGT→AGT nt3543	II	Wallis *et al.* Am J Hum Genet 45:A228 (abstract) 1989
Gly973Ser GGC→AGC nt3570	III	Gomez-Lira *et al.* V International Conference on OI 120 1993
Gly973Val GGC→GTC nt3571	II	Lamande *et al.* J Biol Chem 264:15809-15812 1989
Gly976Arg GGA→CGA nt3579	II	Lamande *et al.* J Biol Chem 264:15809-15812 1989
Gly988Cys GGT→TGT nt3615	II	Cohn *et al.* Proc Natl Acad Sci USA 83:6045-6047 1986
Gly1003Ser GGC→AGC nt3660	II	Pruchno *et al.* Hum Genet 87:33-40 1991
Gly1006Val GGA→GTA nt3670	II	Lamande *et al.* J Biol Chem 264:15809-15812 1989
Gly1009Ser GGT→AGT nt3678	III	Cohn *et al.* Matrix 10:236 (abstract) 1990
Gly1009Val GGT→GTT nt3679	II	Cohn *et al.* Matrix 10:236 (abstract) 1990
Gly1017Cys C-telopeptide GGT→TGT nt3702	I	Labhard *et al.* Mol Biol Med 5:197-207 1988; Cohn *et al.* J Biol Chem 263:14605-14607 1988 Note: These papers refer to the same patient
Asp1099His C-propeptide GAC→CAC nt3948	II	Chessler *et al.* J Biol Chem 268:18218-18225 1993
Trp1134Cys C-propeptide TGG→TGT nt4055	II	Bateman *et al.* Am J Med Genet 45:233-240 1993
Leu1210Arg C-propeptide CTG→CGG nt4282	II	Chessler *et al.* J Biol Chem 268:18218-18225 1993
Leu1286Pro C-propeptide CTG→CCG nt4510 with variant Thr1058Pro C-propeptide ACC→CCC nt3582 in α2(I)	III	Oliver *et al.* Hum Mutat 7:318-326 1996

Table 4. Exon skipping mutations in *COL1A1*

Mutation	Phenotype	Reference(s)
Skipping of exon 6 G→A 3′ end of exon 6 nt622	EDS VIIA EDS VIIA	Weil *et al.* EMBO J 8:1705-1710 1989 D'Alessio *et al.* Am J Hum Genet 49:400-406 1991
Skipping of exon 8 G^{+1}→C intron 8 resulting in skipping of exon 8 and insertion of part of exon 7 in case 2, not defined for case 1	III/IV	Bateman *et al.* Biochem J 302: 729-735 1994
Skipping of exon 14 G^{+1}→A intron 14	II	Bonadio *et al.* J Biol Chem 265:2262-2268 1990
Skipping of exon 17 A^{-2}→G intron 16	I	Willing *et al.* Am J Med Genet 45:223-227 1993
Skipping of exon 18 G^{+1}→A intron 18 plus use of cryptic splice site at G^{+4} exon 18 (nt1312) causing frameshift and termination at 32 aa	I I	Willing *et al.* Am J Hum Genet 55:638-647 1994 Willing *et al.* Am J Hum Genet 55:638-647 1994
Skipping of exon 21 not defined	II	Byers *et al.* V International Conference on OI 1993
Skipping of exon 22 not defined	III	Wallis *et al.* Am J Hum Genet 45:A228 (abstract) 1989
Inclusion of intron 26 G^{+1}→A intron 26	I	Stover *et al.* J Clin Invest 92:1994-2002 1993
Skipping of exon 27 A^{-2}→C intron 26	II	Byers. Trends Genet 6:293-300 1990
Skipping of exon 43 not defined	II	Byers. Trends Genet 6:293-300 1990
Skipping of exon 44 C^{-3}→? intron 43	II	Byers. Trends Genet 6:293-300 1990
Skipping of exon 47 G^{+1}→A	II	Wallis *et al.* IV International Conference on OI 55 1990
Skipping of exon 48 G^{+1}→A	I	Willing *et al.* Am J Hum Genet 55:A249 (abstract) 1994

Table 5. Deletions, insertions, duplications and frameshifts in *COL1A1*

Mutation	Phenotype	Reference(s)
Duplication involving exons 14 to 17 causing insertion of 60aa	II	Cohn *et al.* Hum Mutat 2:21-27 1993
Frameshift at Pro318 due to deletion of CC nt 1605-1606; termination 14 aa downstream	I	Willing *et al.* Am J Hum Genet 55:638-647 1994
84 aa deletion (328 to 411) (exons 23→25). Intron mediated recombination	II	Chu *et al.* J Biol Chem 260:691-694 1985 Barsh *et al.* Proc Natl Acad Sci USA 82:2870-2874 1985 Note: These papers refer to the same patient
562 nt deletion from 3′ end of exon 34 and ending in exon 36	III	Wang and Marini Am J Hum Genet 57:A253 (abstract) 1995
Frameshift at Pro444 due to deletion of T nt 1985; termination 142 aa downstream	I	Redford-Badwal *et al.* J Clin Invest 97:1035-1040 1996
75bp insertion derived from intron 35	II	Genovese *et al.* J Biol Chem 264:9632-9637 1989
3aa deletion (GlyProArg) 730→732	II	Byers *et al.* V International Conference on OI 1993
3aa deletion (GlyAlaPro) in region of 868→876	II	Hawkins *et al.* J Biol Chem 266:22370-22374 1991
3aa deletion (GlyAlaHyp) 874→876	II	Wallis *et al.* J Biol Chem 267:25529-25534 1992
Frameshift at Lys918 due to insertion of C before nt 3520; termination 2 aa downstream	I	Willing *et al.* Am J Hum Genet 55:638-647 1994
Frameshift at Ala956 due to deletion of C nt 3520 or 3521; termination 105 aa downstream	I	Willing *et al.* Am J Hum Genet 55:638-647 1994
Frameshift at Asp1019 C-telopeptide due to deletion of GA nt 3708-3709; termination at 22 aa	I	Willing *et al.* Am J Hum Genet 55:638-647 1994
Val1146Cys C-propeptide GTC→TGTC nt4089 Frameshift and truncation	II	Bateman *et al.* J Biol Chem 264:10960-10964 1989
2aa deletion (Glu Tyr) 1159→1160. C-propeptide and Phe1158 TTC→TTT	II	Chessler *et al.* J Biol Chem 268:18218-18225 1993
Deletion of 5bp from1st base of Glu1275 nt4476→4480 C-propeptide Frameshift and 84aa elongation	I	Willing *et al.* J Clin Invest 85:282-290 1990

DNA sequence numbers

Although the entire genomic sequence of *COL1A1* is known (10,11) the same is not true for *COL1A2*. For this reason, a numbering system based on cDNA sequences has been adopted. 'Contig' cDNA sequences were assembled from the complete, but fragmented, data available in the EMBL DNA sequence database. The contigs for the α1(I) and α2(I) cDNAs have been assigned the accession numbers Z74615 and Z74616 respectively. The reporting of all mutation data is based on these sequences. Mutations in intron donor or acceptor sequences, leading to exon skipping, are reported relative to the start or end of the intron (e.g. G^{+1}→A or A^{-2}→T).

HOW TYPE I COLLAGEN MUTATIONS LEAD TO DEFECTIVE COLLAGEN AND PRODUCE DISEASE PHENOTYPES

The vast majority of mutations of type I collagen result in the connective tissue disorder osteogenesis imperfecta (OI) (12,13) which is also known as brittle bone disease. OI may be subdivided into four types that are defined according to clinical phenotype (14). Type I is the mildest and is inherited in an autosomal dominant manner. Types II, III and IV are more severe and generally arise as new dominant mutations. As a general principle, in type I OI the type I collagen is normal but is produced in reduced amounts. OI types II, III and IV result from the production of abnormal type I collagen due to the incorporation of one or more individually abnormal α-chains. Such abnormalities can include the substitution of amino acids or the shortening or lengthening of α-chains due to exon-skipping mutations or more complex gene rearrangements. Mutations that result in premature chain termination, such that no C-terminal propeptide is produced, result in type I OI as the shortened α-chains do not participate in triple helix formation. The mechanisms of phenotype production have been well reviewed in recent years (4,13,15,16) though a truly satisfactory unifying model of the genotype/phenotype relationship is yet to be proposed.

Although type I collagen gene mutations are overwhelmingly dominant in their action, there is a single example of a recessively-inherited case of OI (7). However, evidence suggests that the most cases of recessively-inherited OI type III do not directly involve type I collagen and map to loci other than *COL1A1* and *COL1A2* (17).

Apart from OI, there is one other common connective tissue disorder that results from type I collagen gene mutations. Ehlers–Danlos syndrome types VIIA and VIIB are members of a diverse group of connective tissue disorders (18) and result from mutations in *COL1A1* and *COL1A2* respectively. They share a common basis in that in both instances the cause is the skipping of exon 6 due to mutations in the splice sites at one end or other of the exon. The consequence of the skipping of exon 6 is loss of the site for the cleavage of the N-terminal propeptide which is hence retained. Mutations resulting in the skipping of exon 6 appear to be a much more frequent in *COL1A2* than in *COL1A1*.

MUTATIONS OF *COL1A1* AND *COL1A2*

The first account of a type I collagen gene mutation was of a 0.5 kb deletion in *COL1A1* leading to osteogenesis imperfecta type

Table 6. Polymorphisms in *COl1A1* cDNA

Polymorphism	Reference
Arg59Arg N-propeptide CGG↔CGT nt296	Mackay *et al.* Hum Mol Genet 2:1155-1160 1993
Pro27Ala CCT↔GCT nt732	Spotila *et al.* J Bone Mineral Res 9:923-932 1994
Pro338Pro CCC↔CCT nt1667	Marini *et al.* J Biol Chem 268:2667-2673 1993
Arg386His CGC↔CAC nt1810	Pruchno *et al.* Hum Genet 87:33-40 1991
Ala410Ala GCT↔GCC nt1883	Mackay *et al.* Hum Mol Genet 2:1155-1160 1993
Gly517Gly GGA↔GGT nt2204	Nicholls *et al.* J Med Genet 28:757-764 1991
Pro645Ala CCT↔GCT nt2586	Mackay *et al.* Hum Mol Genet 2:1155-1160 1993
Ala897Thr GCC↔ACC nt3342	Sokolov *et al.* Nucl Acids Res 19:4302 1991
Pro899Pro CCT↔CCC nt3350	Lamande *et al.* J Biol Chem 264:15809-15812 1989
Pro902Pro CCT↔CCC nt3359	Lamande *et al.* J Biol Chem 264:15809-15812 1989
Val903Val GTC↔GTT nt3362	Bateman *et al.* Am J Med Genet 45:233-240 1993
Asp975Asp GAT↔GAC nt3578	Zhuang *et al.* Hum Mutat 7:89-99 1996
Ser1215Ser TCC↔TCT nt4298	Zhuang *et al.* Hum Mutat 7:89-99 1996
Ser1256Thr C-propeptide TCC↔ACC nt4419	Makela *et al.* Nucl Acids Res 16:349 1988
3′ untranslated TCA↔CCA nt4602	Zhuang *et al.* Hum Mutat 7:89-99 1996

II (19). This was later characterised more precisely as a deletion of three exons (8,20).

Subsequently, however, it has emerged that deletions are a relatively infrequent type of mutation in type I collagen genes. By far the most common type of mutation in *COL1A1* and *COL1A2* are single base changes causing substitutions of glycines which are essential for correct folding of the collagen triple helix. Such single base substitutions can result in a glycine being replaced by alanine, arginine, aspartic acid, cysteine, glutamic acid, serine or valine and each has been recognised in both type I collagen α-chains. It is also possible to mutate glycines encoded by GGA to the TGA stop codon by a single base substitution though no examples have yet been detected. Interestingly, the two known examples of premature stop codons caused by single base substitutions have both been the result of mutations in CGA arginine codons rather than the expected glycine codon. Finally, it is worth noting that amino acid substitutions in type I collagen genes have been noted in two cases of osteoporosis and in a single atypical case of Marfan syndrome (21). The amino acid substitutions in *COL1A1* and *COL1A2* are listed in Tables 3 and 7.

Exon skipping is a fairly common mutation type resulting mainly from alterations to splice donor sites (Tables 4 and 8). Large and small alterations to the gene structure are a final heterogeneous group of mutations (Tables 5 and 9).

In the vast majority of instances, the mutations listed below are ‘private’—they have only been recorded in a single individual or within a single family. Where a mutation has been reported to have occurred more than once, each individual report is listed in the tables. This is of value especially where the resulting phenotype has been different, perhaps due to differences in the genetic background on which the primary mutation is expressed.

Table 7. Amino acid substitutions in *COl1A2*

Mutation	Phenotype	Reference(s)
Gly121Asp GGT→GAT nt771	I	Zhuang *et al.* Hum Mutat 7:89-99 1996
Gly238Ser GGT→AGT nt1121	III	Rose *et al.* Hum Genet 95:215-218 1995
Gly247Cys GGT→TGT nt1148	III	Marini *et al.* V International Conference on OI 126 1993
Gly247Ser GGT→AGT nt1148	I	Zhuang *et al.* Hum Mutat 7:89-99 1996
Gly259Cys GGT→TGT nt1184	III/IV	Wenstrup *et al.* J of Biol Chem 266:2590-2594 1991
Gly343Glu GGA→GAA nt1437	II	Rose *et al.* Hum Mol Genet 2:2175-2177 1993
Gly370Ser GGC→AGC nt1517	III	Zhuang *et al.* Hum Mutat 7:89-99 1996
Gly457Arg GGT→CGT nt1778	II	Bateman *et al.* Hum Mutat 1:55-62 1992
Gly472Cys GGT→TGT nt1823	II	Edwards *et al.* Hum Mutat 1:47-54 1992
Gly496Arg GGT→CGT nt1895	II	Bateman *et al.* IV International Conference on OI 2 1990
Gly502Ser GGT→AGT nt1913	II	Rose *et al.* Hum Genet 94:497-503 1994
Gly544Val GGT→GTT nt2040	IV	Sztrolovics *et al.* Hum Mol Genet 2:1319-1321 1993
Gly547Asp GGT→GAT nt2049	II	Bonadio *et al.* Collagen & Related Research 8:506-507 (abstract) 1988
Gly580Asp GGC→GAC nt2148	II	Niyibizi *et al.* J Biol Chem 267:23108-112 1992
Gly586Val GGT→GTT nt2166	IV III	Bateman *et al.* Biochem J 276:765-770 1991 Forlino *et al.* Hum Mol Genet 3:2201-2206 1994
Arg618Glu CGG→CAG nt2262	Marfan syndrome	Phillips *et al.* J Clin Invest 86:1723-1728 1990
Gly625Asp GGC→GAC nt2283	II	Byers *et al.* V International Conference on OI 1993
Gly640Cys GGT→TGT nt2327	II/III	Gomez-Lira *et al.* J Med Genet 31:965-968 1994
Gly646Cys GGT→TGT nt2345	I	Wenstrup *et al.* J Biol Chem 266:2590-2594 1991
Gly661Ser GGT→AGT nt2390	Osteoporosis	Spotila *et al.* Proc Natl Acad Sci USA 88:5423-5427 1991
Gly676Val GGT→GTT nt 2436	IV	Wang *et al.* J Biol Chem 268:25162-25167 1993
Gly688Ser GGT→AGT nt2471	III/IV	Raghunath *et al.* Eur J Pediatr 154:123-129 1995
Gly694Arg GGT→CGT nt2489	II	Tsuneyoshi *et al.* J Biol Chem 266:15608-15613 1991
Gly700Asp GGT→GAT nt2508	II	Cohen-Solal *et al.* J Biol Chem 269:14751-14758 1994
Gly706Ser GGT→AGT nt2525	II	Wang *et al.* J Biol Chem 268:25162-67 1993
Gly745Ser GGT→AGT nt2642	I(?)	Zhuang *et al.* Hum Mutat 7:89-99 1996
Gly787Cys GGC→TGC nt2768	II	Fertala *et al.* Biochem J 289:195-199 1993
Gly802Asp GGT→GAT nt 2814	III/IV	Lund *et al.* Eur J Hum Genet 4: 39-45 1996
Gly805Asp GGT→GAT nt2823	II	Grange *et al.* Nucl Acids Res 18:4227-4236 1990
Gly859Ser GGT→AGT nt2984	III	Rose *et al.* Hum Mutat 3:391-394 1994
Gly865Ser GGT→AGT nt3002	II	Lamande *et al.* J Biol Chem 264:15809-15812 1989
Gly907Asp GGT→GAT nt3129	II	Baldwin *et al.* J Biol Chem 264:3002-3006 1989
Gly922Ser GGT→AGT nt3173	IV IV IV	Marini *et al.* J Biol Chem 268:2667-73 1993 Sztrolovics *et al.* Hum Mol Genet 2:1319-21 1993 D'Amato *et al.* V International Conference on OI 63 1993
Gly976Asp GGT→GAT nt3336	II	Byers. Trends Genet 6:293-300 1990
Gly1006Ala GGC→GCC nt3426	III	Lu *et al.* Hum Mutat 5:175-178 1995
Gly1012Arg GGT→CGT nt3443	IV	Wenstrup *et al.* J Biol Chem 263:7734-7740 1988
Thr1058Pro C-propeptide ACC→CCC nt3582 Leu1286Pro C-propeptide in one α1(I) allele CTG→CCG nt4510	III	Oliver *et al.* Hum Mutat 7:318-326 1996

Table 8. Exon skipping mutations in COL1A2

Mutation	Phenotype	Reference(s)
Skipping of exon 6 ATG→ATA 3' end of exon 6	EDS VIIB	Weil *et al.* J Biol Chem 264:16804-16809 1989
Missplicing of exon 6 G^{-1}→C intron 5. Cryptic site used at +14/15 in exon 6	EDS VIIB	Chiodo *et al.* J Biol Chem 267:6361-6369 1992
Skipping of exon 6 G^{+1}→A intron 6	EDS VIIB EDS VIIB EDS VIIB EDS VIIB EDS VIIB	Weil *et al.* J Biol Chem 265:16007-16011 1990 Vasan *et al.* Am J Hum Genet 48:305-317 1991 Nicholls *et al.* Hum Genet 87:193-198 1991 Watson *et al.* J Biol Chem 267:9093-9100 1992 Lehmann *et al.* Arch Dermatol Res 286:425-428 1994
Skipping of exon 6 T^{+2}→C intron 6	EDS VIIB EDS VIIB	Weil *et al.* J Biol Chem 263:8561-8564 1988 Ho *et al.* Hum Mutat 3:358-364 1994
Skipping of exon 9 Deletion of 11bp in intron 9, +3→+13	OI type?	Nicholls *et al.* Hum Genet 88:627-633 1992
Skipping of exon 11 Deletion of 19bp across intron 10/exon 11	Atypical OI	Kuivaniemi *et al.* J Biol Chem 263:11407-11413 1988
Skipping of exon 12 T^{+2}→G intron 12	IV	Chipman *et al.* J Bone Mineral Res 7:793-805 1992
Skipping of exon 13 Deletion of 19bp in intron 13, +4→+22.	I	Zhuang *et al.* Hum Genetics 91:210-216 1993
Skipping of exon 16 G^{+1}→A intron16	IV	Filie *et al.* Hum Mutation 2:380-388 1993
Skipping of exon 16 T^{+2}→C intron 16.	III/IV	Zolezzi *et al.* Hum Mutat 6:268-271 1995
Skipping of exon 20 G^{-1}→C intron 19.	I	Mottes *et al.* Hum Genet 93:681-687 1994
Skipping of exon 21 Deletion of 39bp in intron 21, +2→+40	I	Superti-Furga *et al.* Connect Tissue Res 29:31-40 1993
Skipping of exon 21 G^{+5}→A intron 21	I(?) or dentinogenesis imperfecta(?)	Nicholls *et al.* Hum Mutat 7:219-227 1996
Skipping of exon 26 not defined	IV	Wenstrup *et al.* Annal NY Acad Sci 580:546-548 1990
Skipping of exon 28 A^{-2}→G intron 27	II	Tromp and Prockop. Proc Natl Acad Sci USA 85:5254-5258 1988
Missplicing of intron 33 G^{+4}→A intron 33. Alternative site at +19. Inclusion of 6aa	IV	Wenstrup *et al.* Am J Med Genet 45:228-232 1993
Skipping of exon 33 G^{+1}→A intron 33	II	Ganguly *et al.* J Biol Chem 266:12035-12040 1991

Table 9. Deletions, insertions, duplications and frameshifts in COL1A2

Mutation	Phenotype	Reference(s)
Deletion of Val255 nts1172→1174	III	Molyneux *et al.* Hum Genet 90:621-628 1993
180aa deletion, 586→765 (exons 34→40). Intron mediated recombination	II	Willing *et al.* J Biol Chem 263:8398-8404 1988
3aa deletion (GlyProPro) 1003→1006 nts 3418-3426	IV	Lund *et al.* Hum Genet 97: 287-290 1996
Deletion of 4bp, Asn1244. Frameshift but no chain length change. C-propeptide	III	Pihlajaniemi *et al.* J Biol Chem 259:12941-12944 1984

POLYMORPHISMS OF *COL1A1* AND *COL1A2* CODING REGIONS

There are several reported polymorphisms in the coding regions of *COL1A1* (Table 6) and *COL1A2* (Table 10) though few are either well characterised or frequent enough to be useful as genetic markers. Such markers might be useful in the analysis of the expression of individual alleles. In *COL1A1* the potentially useful markers include a sequence polymorphism in the 3′ untranslated region (22) and amino acid 897 of the triple-helical domain can be either alanine or threonine (23). In *COL1A2* there

Table 10. Polymorphisms in *COl1A2* cDNA

Polymorphism	Reference
Thr29Thr N-propeptide ACT↔ACC nt226	Zhuang *et al.* Hum Mutat 7:89-99 1996
Pro59Thr N-propeptide CCA↔ACA nt314	Kuivaniemi *et al.* Biochem J 252:633-640 1988
Asp82Asp N-telopeptide GAT↔GAC nt385	Strobel *et al.* Matrix 12 87-91 1992
Gly127Gly GGG↔GGT nt790	Filie *et al.* Hum Mutat 2:380-388 1993
Gly139Gly GGT↔GGC nt826	Filie *et al.* Hum Mutat 2:380-388 1993
Gly145Gly GGA↔GGC nt844	Filie *et al.* Hum Mutat 2:380-388 1993
Val153Val GTA↔GTG nt868	Filie *et al.* Hum Mutat 2:380-388 1993
Pro158Pro CCT↔CCC nt883	Filie *et al.* Hum Mutat 2:380-388 1993
Asn159Ile AAT↔ATT nt885	Filie *et al.* Hum Mutat 2:380-388 1993
Gly166Gly GGT↔GGC nt907	Filie *et al.* Hum Mutat 2:380-388 1993
Gly172Gly GGT↔GGC nt925	Filie *et al.* Hum Mutat 2:380-388 1993
Thr186Ala ACT↔GCT nt965	Filie *et al.* Hum Mutat 2:380-388 1993
Gly187Gly GGA↔GGT nt970	Filie *et al.* Hum Mutat 2:380-388 1993
Ser275Ser TCT↔TCC nt1234	Filie *et al.* Hum Mutat 2:380-388 1993
Gly277Gly GGT↔GGG nt1240	Filie *et al.* Hum Mutat 2:380-388 1993
Pro392Pro CCA↔CCC nt1585	Constantinou *et al.* Nucl Acids Res 18:5577 1990
Val420Ala GTT↔GCT nt1668	Wenstrup *et al.* J Biol Chem 266:2590-2594 1991
Ala459Pro GCT↔CCT nt1784	Bateman *et al.* Hum Mutat 1:55-62 1992
Val536Val GTG↔GTT nt2017	Bateman *et al.* Am J Med Genet 45:233-240 1993
T↔G +661bp within IVS 33	Strobel *et al.* Matrix 12:87-91 1992
Ala653Gly GCC↔GGC nt2367	Kuivaniemi *et al.* Biochem J 252:633-640 1988
Arg732His CGT↔CAT nt2604	Zhuang *et al.* Hum Mutat 7:89-99 1996
Pro795Pro CCT↔CCA nt2794	Baldwin *et al.* J Biol Chem 264:3002-3006 1989
Gly862Gly GGC↔GGT nt2995	Baldwin *et al.* J Biol Chem 264:3002-3006 1989
Phe932Leu TTC↔TTA nt3205	Baldwin *et al.* J Biol Chem 264:3002-3006 1989
Gly955Gly GGC↔GGT nt3274	Strobel *et al.* Matrix 12:87-91 1992
Thr983Thr ACG↔ACA nt3358	Baldwin *et al.* J Biol Chem 264:3002-3006 1989
Leu1011Pro CTA↔CCA nt3441	Baldwin *et al.* J Biol Chem 264:3002-3006 1989
Thr1058Pro ACC↔CCC nt3581	Oliver *et al.* Hum Mutat 7:318-326 1996
Glu1099Asp C-propeptide GAA↔GAT nt3706	Marini *et al.* J Biol Chem 268:2667-2673 1993
Ala1100Ala C-propeptide GCC↔GCT nt3709	Marini *et al.* J Biol Chem 268:2667-2673 1993
Cys1105Cys C-propeptide TGC↔TGT nt3724	Marini *et al.* J Biol Chem 268:2667-2673 1993
Pro1108Ser C-propeptide CCT↔TCT nt3731	Mäkelä *et al.* Biochim Biophys Acta. 1049:171-176 1990
Val1152Val C-propeptide GTT↔GTA nt3865	Marini *et al.* J Biol Chem 268:2667-2673 1993
Gly1194Gly C-propeptide GGC↔GGA nt3991	Marini *et al.* J Biol Chem 268:2667-2673 1993

are silent base substitutions in the codons for amino acid 82 of the primary translation product (24) and amino acids 392 and 955 of the triple-helical domain (24,25).

ACCESSING THE DATA

The type I collagen mutation data may be accessed on the University of Leicester web server at http://www.le.ac.uk/depts/ge/collagen/collagen.html . At present, the data are in simple static lists but it is hoped that data will be made available at some time in a more comprehensive manner in a relational database that can be queried from a web page. If you make use of the data from the web server, please cite this article in any materials which you prepare for publication.

ACKNOWLEDGEMENTS

I am grateful to Matthew Lewis who helped in the initial compilation of the data presented here. I would also like to thank those members of the OI Mutation Consortium who have pointed out errors and omissions in earlier versions of the data.

REFERENCES

1 Brown,J.C. and Timpl,R. (1995) *Int. Arch. Allergy Immunol.*, **107**, 484–490.
2 Engel,J. and Prockop,D.J. (1991) *Annu. Rev. Biophys. Biophys. Chem.*, **20**, 137–152.
3 Kielty,C.M., Hopkinson,I. and Grant,M.E. (1993) In Royce,P.M. and Steinmann,B. (eds), *Connective Tissue and its Heritable Disorders. Molecular, Genetic, and Medical Aspects.* Wiley-Liss, New York, pp. 103–147.
4 Prockop,D.J. and Kivirikko,K.I. (1995) *Annu. Rev. Biochem.*, **64**, 403–434.
5 Vuorio,E. and de Crombrugghe,B. (1990) *Annu. Rev. Biochem.*, **59**, 837–872.
6 Chu,M.-L. and Prockop,D.J. (1993) In Royce,P.M. and Steinmann,B. (eds), *Connective Tissue and its Heritable Disorders. Molecular, Genetic, and Medical Aspects.* Wiley-Liss, New York, pp. 149–165.
7 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.
8 Chu,M.-L., Gargiulo,V., Williams,C.J. and Ramirez,F. (1985) *J. Biol. Chem.*, **260**, 691–694.
9 Hawkins,J.R., Superti-Furga,A., Steinmann,B. and Dalgleish,R. (1991) *J. Biol. Chem.*, **266**, 22370–22374.
10 D'Allesio,M., Bernard,M., Pretorious,P.J., de Wet,W. and Ramirez,F. (1988) *Gene*, **67**, 105–115.
11 Westerhausen,A., Constantinou,C.D., Pack,M., Peng,M., Hanning,C., Olsen,A.S. and Prockop,D.J. (1991) *Matrix*, **11**, 375–379.
12 Byers,P.H. (1990) *Trends Genet.*, **6**, 293–300.
13 Byers,P.H. (1993) In Royce,P.M. and Steinmann,B. (eds), *Connective Tissue and its Heritable Disorders: Molecular, Genetic and Medical Aspects.* Wiley Liss, New York, pp. 317–350.
14 Sillence,D.O., Senn,A. and Danks,D.M. (1979) *J. Med. Genet.*, **16**, 101–116.
15 Dalgleish,R. (1994) In Humphries,S.E. and Malcolm,S. (eds), *From Genotype to Phenotype.* BIOS Scientific Publishers, Oxford, pp. 49–65.
16 Cole,W.G. and Dalgleish,R. (1995) *J. Med. Genet.*, **32**, 284–289.
17 Wallis,G.A., Sykes,B., Byers,P.H., Mathew,C.G., Viljoen,D. and Beighton,P. (1993) *J. Med. Genet.*, **30**, 492–496.
18 Steinmann,B., Royce,P.M. and Superti-Furga,A. (1993) In Royce,P.M. and Steinmann,B. (eds), *Connective Tissue and its Heritable Disorders: Molecular, Genetic, and Medical Aspects.* Wiley Liss, New York, pp. 351–407.
19 Chu,M.-L., Willaims,C.J., Pepe,G., Hirsch,J.L., Prockop,D.J. and Ramirez,F. (1983) *Nature*, **304**, 78–80.
20 Barsh,G.S., Roush,C.L., Bonadio,J., Byres,P.H. and Gelinas,R.E. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 2870–2874.
21 Phillips,C.L., Shrago-Howe,A.W., Pinnell,S.R. and Wenstrup,R.J. (1990) *J. Clin. Invest.*, **86**, 1723–1728.
22 Westerhausen,A.I., Constantinou,C.D. and Prockop,D.J. (1990) *Nucleic Acids Res.*, **18**, 4968.
23 Sokolov,B.P., Constantinou,C.D., Tsuneyoshi,T., Zhuang,J. and Prockop,D.J. (1991) *Nucleic Acids Res.*, **19**, 4302.
24 Strobel,D., Tsuneyoshi,T., Kuivaniemi,H., Tromp,G., Spotila,L.D., Baldwin,C.T., Constantinou,C.D., Ganguly,A., Sereda,L., Sokolov,B.P. and Prockop,D.J. (1992) *Matrix*, **12**, 87–91.
25 Constantinou,C.D., Spotila,L.D., Zhuang,J., Sereda,L., Hanning,C. and Prockop,D.J. (1990) *Nucleic Acids Res.*, **18**, 5577.