Consortium for Osteogenesis Imperfecta Mutations in the Helical Domain of Type I Collagen: Regions Rich in Lethal Mutations Align With Collagen Binding Sites for Integrins and Proteoglycans

Joan C. Marini[1,*], **Antonella Forlino**[1,2], **Wayne A. Cabral**[1], **Aileen M. Barnes**[1], **James D. San Antonio**[3], **Sarah Milgrom**[1], **James C. Hyland**[4], **Jarmo Körkkö**[4], **Darwin J. Prockop**[4], **Anne De Paepe**[5], **Paul Coucke**[5], **Sofie Symoens**[5], **Francis H. Glorieux**[6], **Peter J. Roughley**[6], **Alan M. Lund**[7], **Kaija Kuurila-Svahn**[8,9], **Heini Hartikka**[10], **Daniel H. Cohn**[11], **Deborah Krakow**[11], **Monica Mottes**[12], **Ulrike Schwarze**[13], **Diana Chen**[13], **Kathleen Yang**[13], **Christine Kuslich**[13], **James Troendle**[14], **Raymond Dalgleish**[15], and **Peter H. Byers**[13]

[1]Bone and Extracellular Matrix Branch, National Institute of Child Health and Human Development (NICHD), NIH, Bethesda, Maryland [2]Department of Biochemistry A. Castellani, University of Pavia, Pavia, Italy [3]Department of Medicine, Hematology/Cardeza Foundation, Thomas Jefferson University, Philadelphia, Pennsylvania [4]Center for Gene Therapy, Tulane University, New Orleans, Louisiana [5]Centre for Medical Genetics, University Hospital Ghent, Ghent, Belgium [6]Genetics Unit, Shriners Hospital for Children, Montreal, Quebec, Canada [7]Department of Clinical Genetics, Rigshospitalet, Copenhagen, Denmark [8]Department of Otorhinolaryngology–Head and Neck Surgery, Vaasa Central Hospital, Vaasa, Finland [9]Department of Otorhinolaryngology–Head and Neck Surgery, Turku University Hospital, Turku, Finland [10]Collagen Research Unit, Biocenter and Department of Medical Biochemistry and Molecular Biology, University of Oulu, Oulu, Finland [11]Medical Genetics Institute, Steven Spielberg Pediatric Research Center, Cedars-Sinai Research Institute, Los Angeles, California [12]Department of Mother and Child, Biology and Genetics, University of Verona, Verona, Italy [13]Department of Pathology, University of Washington, Seattle, Washington [14]Biometry and Mathematical Statistics Branch, National Institute of Child Health and Human Development (NICHD), NIH, Bethesda, Maryland [15]Department of Genetics, University of Leicester, Leicester, United Kingdom

Abstract

*Correspondence to: Joan C. Marini, MD, PhD, Chief, Bone and Extracellular Matrix Branch, NICHD, NIH, Bldg. 10; Rm 9s239, 9000 Rockville Pike, Bethesda, MD 20892. oidoc@helix.nih.gov.
Sarah Milgrom's current address:Yale University School of Medicine, New Haven, CT 06511.
James C. Hyland's current address: Medical Director, Connective Tissue Gene Tests, Allentown, PA18103.

Osteogenesis imperfecta (OI) is a generalized disorder of connective tissue characterized by fragile bones and easy susceptibility to fracture. Most cases of OI are caused by mutations in type I collagen. We have identified and assembled structural mutations in type I collagen genes (COL1A1 and COL1A2, encoding the proα1(I) and proα2(I) chains, respectively) that result in OI. Quantitative defects causing type I OI were not included. Of these 832 independent mutations, 682 result in substitution for glycine residues in the triple helical domain of the encoded protein and 150 alter splice sites. Distinct genotype–phenotype relationships emerge for each chain. One-third of the mutations that result in glycine substitutions in α1(I) are lethal, especially when the substituting residues are charged or have a branched side chain. Substitutions in the first 200 residues are nonlethal and have variable outcome thereafter, unrelated to folding or helix stability domains. Two exclusively lethal regions (helix positions 691–823 and 910–964) align with major ligand binding regions (MLBRs), suggesting crucial interactions of collagen monomers or fibrils with integrins, matrix metalloproteinases (MMPs), fibronectin, and cartilage oligomeric matrix protein (COMP). Mutations in COL1A2 are predominantly nonlethal (80%). Lethal substitutions are located in eight regularly spaced clusters along the chain, supporting a regional model. The lethal regions align with proteoglycan binding sites along the fibril, suggesting a role in fibril–matrix interactions. Recurrences at the same site in α2(I) are generally concordant for outcome, unlike α1(I). Splice site mutations comprise 20% of helical mutations identified in OI patients, and may lead to exon skipping, intron inclusion, or the activation of cryptic splice sites. Splice site mutations in COL1A1 are rarely lethal; they often lead to frameshifts and the mild type I phenotype. In α2(I), lethal exon skipping events are located in the carboxyl half of the chain. Our data on genotype–phenotype relationships indicate that the two collagen chains play very different roles in matrix integrity and that phenotype depends on intracellular and extracellular events.

Keywords

osteogenesis imperfecta; type I collagen; genotype–phenotype; proteoglycan binding; COL1A1; COL1A2

INTRODUCTION

The most common clinical phenotypes associated with mutations in the two type I collagen genes, COL1A1 (MIM# 120150) and COL1A2 (MIM# 120160), are forms of osteogenesis imperfecta (MIM#s OI Type I, 166200; OI Type II, 166210; OI Type III, 259420; and OI Type IV, 166220), also known as brittle bone disease (Table 1). The severity of outcome ranges from lethality in the perinatal period to a barely detectable connective tissue disorder [Byers and Cole, 2002; Marini, 2004; Sillence et al., 1979a]. The two genes containing these mutations are similar in structure; COL1A1 and COL1A2 have 51 and 52 exons, respectively. Although COL1A1 is about half the size of COL1A2, the two genes transcribe mRNAs of about the same length. Each encodes an alpha chain with a core triple helical domain of 1,014 amino acids, composed of uninterrupted repeats of the Gly-Xaa-Yaa tripeptide, flanked by propeptides at both the amino- and carboxyl-terminal ends. Each of the 43 exons that encodes the triple helical domain begins with a glycine codon and ends with a Y-position codon, so that exon-skipping yields in-frame transcripts [Chu and Prockop, 2002]. Type I procollagen chains form a heterotrimer (proα1(I)₂proα2(I)); the

chains associate using regions in the carboxyl-terminal propeptides and nucleate the triple helix, which is propagated linearly toward the amino terminal ends of the chains. During chain synthesis and helix formation, some Y-position proline and lysine residues are hydroxylated, and some lysine residues may be subsequently glycosylated [van der Rest and Garrone, 1991]. Glycine, with its hydrogen side chain, is essential in every third position of the chain because it is the only residue small enough to be accommodated in the sterically restricted inner aspect of the helix. Residues that substitute for glycine disrupt helix folding and propagation temporarily, exposing all three chains to excess hydroxylation and glycosylation [Engel and Prockop, 1991]. In-frame exon skipping mutations also disrupt helix folding. Once assembled in the rough endoplasmic reticulum, the mature procollagen molecule is transported to the Golgi, packaged into transport vesicles in which lateral aggregation, the initial phase of fibril formation, occurs, and the aggregates are then secreted into the pericellular space [Canty and Kadler, 2005]. Specific propeptidases cleave the amino- and carboxyl-terminal propeptides in secretory vesicles and in the pericellular space, and mature collagen molecules self-assemble into heterotypic fibrils with other matrix proteins [Ayad et al., 1998; Chakravarti et al., 1998; Colige et al., 1997; Font et al., 1998; Kyriakides et al., 1998; Prockop et al., 1993; Scott, 1988; Svensson et al., 1999].

There are two general classes of mutations in type I collagen that result in osteogenesis imperfecta—those that cause a quantitative defect, with synthesis of structurally normal type I procollagen at about half the normal amount, and those that result in synthesis of collagen molecules with structural abnormalities. The first class of mutations usually produce premature termination codons in the coding sequence of one COL1A1 allele; these initiate nonsense mediated decay of the mRNA derived from that allele [Genovese and Rowe, 1987; Körkkö et al., 1998; Redford-Badwal et al., 1996; Slayton et al., 2000; Willing et al., 1996, 1992]. The ensuing matrix insufficiency results in the mild type I OI clinical picture. We have not included this class of mutation in our database.

The class of mutations that changes the protein sequence of the chains in the collagen triple helical domain results in a wide phenotypic range, from lethal OI to mild–moderate type IV OI. The most common mutations that cause the more clinically significant forms of osteogenesis imperfecta result in substitutions for one of the invariant glycine residues in the triple helical domain of the two chains of type I collagen [Kuivaniemi et al., 1991; Prockop and Kivirikko, 1984]. In the coding sequence of the triple helical domain, which can be written $GGNNNNNNN_{338}$, substitutions at either G will cause a substitution for glycine. Thus, because of the crucial role of the glycine in helix formation, substitution at two out of every nine nucleotides will result in a clinically apparent phenotype. The second most common type of mutation is alterations in splice sites [Kuivaniemi et al., 1997]. These may lead to exon skipping, intronic inclusion, or activation of cryptic sites in introns or exons. The consequences for the mRNA and protein depend on whether these alterations are inframe or produce translational frameshifts. Structural mutations of type I collagen result in an abnormal protein that is secreted from the cell into the matrix, where it may interfere with fibrillogenesis, collagen-matrix or collagen-cell interactions, or mineralization in a dominant-negative mechanism. The structural mutations have more severe consequences for the extracellular matrix than does matrix insufficiency [Forlino and Marini, 2000]. Finally, in addition to glycine substitutions and splice site alterations, there is a small group of

deletions, insertions, and duplications in the collagen helix, as well as alterations in the carboxyl-terminal propeptide, that result in an OI phenotype. Rare X- and Y-position substitutions form a special phenotype set, none of which are included in this database.

Given the repetitive sequence of collagen chains and the limitation of most functional mutations to either substitutions for glycine residues in Gly-X-Y triplets or exon skipping, straightforward genotype–phenotype relationships might have been expected. Understanding the rules for structural mutations will enhance genetic counseling, and should facilitate our understanding of normal collagen function and the pathogenesis of bone disease in OI. However, determining the genotype–phenotype relationships of the structural mutations in type I collagen has proven to be difficult. Although more than two decades have passed since the identification of the first COL1A1 mutation in an infant with lethal OI [Chu et al., 1983], a clear relationship of mutation and phenotype has been elusive. The earliest models focused on the chain in which the mutations occurred. Phenotypes resulting from mutations in COL1A1 were thought to be more severe than those from COL1A2 mutations, because type I collagen heterotrimers contain two copies of the $\alpha 1$(I) chain, so that random assortment of mutant chains causes up to 75% of heterotrimers to contain one or more abnormal $\alpha 1$(I) chains vs. up to 50% of trimers with an abnormal $\alpha 2$(I) chain [Prockop et al., 1989; Sykes, 1985]. Later, the focus shifted to the delay in helix folding caused by the glycine substitutions. When it was recognized that mutations in the triple helical domain resulted in additional hydroxylation and glycosylation of residues amino-terminal to the site of the altered sequence as a consequence of delay in helix folding, it was proposed that clinical severity correlated with the position of the alteration; those closest to the carboxyl-terminal end of the $\alpha 1$(I) chains would be lethal, those midchain would be severe nonlethal cases, and those closest to the amino-terminal end would be the moderate and milder cases [Byers et al., 1991]. Another model focused on the stability of the type I collagen helix; mutations disrupting highly stabilizing regions of the helix were predicted to be more severe than those in lower melting regions [Bächinger and Davis, 1991; Bächinger et al., 1993]. The $\alpha 2$(I) chain has a greater proportion of nonlethal mutations than $\alpha 1$(I); the lethal mutations in $\alpha 2$(I) have been noted to occur in regularly spaced regions or clusters along the chain, but their function was not clear [Marini et al., 1993; Wang et al., 1993].

With PCR and automated sequencing, mutation detection became more rapid and now includes mutations that are not biochemically detectable, providing hope that the extended database will facilitate greater insight into genotype–phenotype relationships. The combination of published and unpublished mutations assembled here approximately triples the number of mutations previously available and provides a more detailed context for genotype–phenotype modeling.

ASSEMBLY OF THE DATABASE

Mutations

The published mutations listed in Supplementary Tables S1–S4 (available online at http://www.interscience.wiley.com/jpages/1059-7794/suppmat) are derived from the Database of Collagen Mutations (www.le.ac.uk/genetics/collagen) [Dalgleish, 1997, 1998] at the University of Leicester, supplemented by the addition of recently published cases.

Unpublished mutations come from the laboratories of the members of the consortium. All published and unpublished mutations were verified for accuracy of nucleotide composition of codons and exon/intron junctions and their location within the gene, using the GenBank sequences of the COL1A1 (AF017178, for genomic and NM_000088.3 for cDNA) and COL1A2 (AF004877.1 for genomic and NM_000089.3 for cDNA) genes. The database listed in Supplementary Tables S1–S4 complies with the nomenclature guidelines set by the Human Genome Variation Society (HSVG) [den Dunnen and Antonarakis, 2001; den Dunnen and Paalman, 2003]. After publication, the database may be accessed at www.oiprogram.nichd.nih.gov.

We have included only the two most abundant types of mutations that alter the sequence of the helical region of the type I procollagen chains: point mutations that cause the substitution of a triple helical glycine residue by another amino acid and mutations that alter splice sites. These comprise the overwhelming majority of mutations that result in OI types II, III, and IV. These mutation groups were chosen to facilitate modeling of genotype–phenotype relationships.

Many splice site mutations have been identified by sequence changes found in genomic DNA, but the resulting transcripts have not been studied in cell culture. Information on resulting mRNA structures has been noted when available.

Clinical Data

We relied on the diagnostic category according to the Sillence classification (Table 1) [Sillence et al., 1979b], provided by the referring physicians or the laboratory in which the mutation was identified. The phenotype of cases identified prenatally was determined from ultrasound, radiological, or pathological exam. The recently identified OI types V, VI, and VII [Rauch and Glorieux, 2004] are based on distinct bone histology and some clinical features and were often considered to have OI type IV. The new types do not have mutations in type I collagen and so were not included in this review.

ANALYSIS OF THE DATABASE

Modeling was done separately for the α1(I) and α2(I) chains. Also, to facilitate the emergence of the strongest features of the genotype–phenotype relationship and to avoid confusion from the variable clinical information, as noted above, modeling was done by dividing cases into lethal (type II) and nonlethal (types III and IV) forms of OI. In cases where OI was detected in utero and terminated prenatally, we relied on the contributor's estimation of severity.

The current database almost triples the number of mutations previously available. The complete database contains 832 independent mutations that affect the sequence of one of the chains of type I collagen. All mutations are heterozygous, with one mutant and one normal allele, and dominantly inherited or resulting from new dominant mutations. There are 682 independent mutations that result in substitution of glycine residues; 391 are in COL1A1 and affect the α1(I) chains (Fig. 1) and 291 are in COL1A2 and alter α2(I) chains (Fig. 2). Of these mutations, 458 were previously unpublished. Independent splice site mutations now comprise about 20% of the total structural mutations identified in the collagen helix; they

were about 26% of the total in the previous database. There are now 102 independent splice site mutations in COL1A1 and 48 in COL1A2, for a total of 150 independent splice site mutations.

Reliable modeling requires a reasonable distribution of mutations along the length of the two chains. The relatively recent implementation of primary mutation identification by DNA and cDNA sequencing has resulted in improved detection of mutations in exons coding for the amino end of the helical region, where biochemical abnormalities do not result in overmodified collagen and steady-state collagen protein is usually electrophoretically normal on SDS-Urea-PAGE. There is at least one known substitution for 44.4% of all triple-helical glycine residues in α1(I) and for 42.9% in α2(I), with a good distribution along the length of the chain (Table 2). Since the same glycine can be substituted by four to six different aa residues, depending on the sequence of the glycine codon involved, full mutation saturation of the helical regions would require 1,895 distinct α1(I) and 1,870 distinct α2(I) substitutions. At this point, 211 distinct α1(I) substitutions (11.1% of all possible substitutions) and 184 distinct α2(I) substitutions (9.8% of all possible substitutions) have been identified, leaving considerable room for further insights as mutation identification continues.

For splice-site mutations, there is at least one mutation identified for 35 out of the 42 helical exons in COL1A1 (83.3%), with alterations in 21 acceptor sites and 24 donor sites; 16 exons have mutations in both sites. For COL1A2, there is at least one mutation for 23 out of the 44 (53.4%) helical exons, including nine acceptor sites and 16 donor sites and three exons with mutations at both sites.

MUTATIONS IN COL1A1 THAT RESULT IN SUBSTITUTIONS FOR GLYCINE IN α1(I)

Prevalence of Substituting Residues

The prevalence of most substituting residues in the database does not match the theoretical expectation (Table 3). Substitution of glycine by serine, cysteine, and arginine result from mutations in the first position of glycine codons and account for 78.6% of all α1(I) substitutions. The predicted proportion of these three residues is 47.8% if mutations occurred randomly in the first and second positions of COL1A1 glycine codons. There is an upward distortion of the proportion of cysteine residues because of ease of biochemical detection in the era preceding direct mutation identification. Cysteine is absent from the triple-helical region of type I collagen and its presence in α1(I) chains can be easily detected during protein screening by formation of α1(I) dimers. On the basis of codon distribution, arginine, together with valine and alanine, which result from second position changes in glycine codons, should be the most abundant substitutions. Although substitution by arginine occurred at about the expected rate, substitution by valine and alanine occurred less than one-third as often as expected.

CpG residues appear to have a substantial effect in the discrepancy between observed and theoretical prevalence of α1(I) glycine substitutions [Schorderet and Gartler, 1992]. The majority of sites with multiple independent recurrences of point mutations occurred at CpG-

associated codons (Fig. 3A). Of the COL1A1 codons at which first-position substitutions have been identified, the one-fifth preceded by a C nucleotide (thus generating a CpG dinucleotide) account for almost half of all first-position substitutions. Only one COL1A1 glycine codon not involved in a CpG sequence has more than five independent recurrences of a first-position change (Gly 448, p.Gly626, which is preceded by a G from the intron acceptor site) and no COL1A1 codons with second-position mutations have more than five recurrences. About 9% of glycine substitutions in COL1A1 result in the mild type I OI phenotype that is more commonly the result of nonsense mutations and matrix deficiency.

Clinical Consequences of Substituting Residue

About one-third (35.6%) of all independent glycine substitutions in α1(I) collagen result in perinatal lethal type II OI (Table 3; Fig. 1). The lethal substitutions are not evenly distributed among the substituting residues (Table 3). Valine, with a branched nonpolar side chain, and the charged amino acids aspartic acid, glutamic acid, and arginine are associated with a lethal outcome more often than the average for α1(I). This is especially striking in the case of valine, which has a lethal outcome in about 73% of occurrences. Substitutions by the polar residues cysteine and serine have a lethal outcome less often than average for the α1(I) chain. The large numbers of serine and cysteine substitutions, however, resulted in their contributing the greatest number of lethal α1(I) substitutions overall. Finally, alanine, with a methyl group side chain, is lethal in less than 15% of occurrences.

Collagen Stability and Clinical Outcome of Substitutions

To determine whether loss of local helix stability by the substitution contributed to differences in clinical outcome, we asked if clinical outcome was more severe when the substituted glycines were in high-stability triplets (so-called "helix forming triplets" such as Gly-Pro-Hyp or Gly-Ala-Pro) than when lowerstability triplets were disrupted [Bächinger and Davis, 1991; Bächinger et al., 1993; Persikov et al., 2004, 2005]. If a triplet had both a lethal and nonlethal substitution, it was counted once in each category. The distribution of stability scores of triplets with lethal or nonlethal mutations was analyzed for goodness of fit using a chi-squared test. There was no statistically significant skewing of the stability score distributions of triplets with lethal or nonlethal substitutions, compared to either the distribution of stability scores of all α1(I) triplets (P = 0.59 to 0.72) or to the distribution of all α1(I) triplets in which mutations have been identified (P = 0.68–0.94) (Table 4).

Locations and Concordance of Clinical Outcome of Recurrent Mutations

About one-quarter of all glycines at which substitutions have been identified now have two or more independent occurrences of substitutions; 17 glycines have more than five independent substitutions. Glycine residues with multiple substitution recurrences are scattered along the length of the α1(I) chain, and are often associated with CpG dinucleotides (Fig. 3A) [Schorderet and Gartler, 1992].

Recurrences at the same α1(I) glycine residue frequently have different clinical outcomes. In many instances of recurrences at a single glycine by different substituting amino acids, more severe outcomes resulted from valine, aspartic acid, and arginine substitutions, consistent with the greater frequency of lethal outcomes for residues with charged or branched side

chains noted above. For example, at position 382 (p.Gly560), substitution by arginine produces lethal OI type II, while serine substitutions produce OI types III and IV, and a cysteine substitution resulted in OI type IV. In addition, it is notable that substitutions of a single glycine by the same residue in different individuals can lead to both lethal and nonlethal forms of OI. The six glycines with 10 or more substitutions illustrate typical findings. Four of the five instances in which arginine substituted for glycine 79 (p.Gly257) resulted in mild to moderate type I or IV OI, but in one instance the result was type III OI. At glycine 352 (p.Gly530), substitutions of cysteine or serine usually caused OI types III, III/IV, and IV, but in two instances substitution by serine at this site was lethal. All 16 serine substitutions at glycine residue 589 (p.Gly767) are nonlethal, with severity ranging from types III to IV. At glycine residue 688 (p.Gly866), substitution by cysteine resulted in type III OI, while substitutions by serine caused multiple cases of types III and IV OI, plus a single lethal outcome. At glycine residue 862 (p.Gly1040), serine substitutions usually resulted in OI types III, III/IV, but there was one lethal outcome.

There are almost three dozen glycine residues in α1(I), distributed along the length of the chain, in which both lethal and nonlethal substitutions were identified. In 11 of those cases, the same substituting residue (most often a serine) was associated with the divergent phenotype. As noted above, when a glycine residue was substituted by different amino acids, the more severe phenotypes were associated with substitutions by valine, aspartic acid, and arginine.

Genotype–Phenotype Correlations

With only a few exceptions, the substitutions in the amino-terminal fifth of the chain are nonlethal in outcome. A subset of substitutions in the first 90 helical residues at the amino-terminal end of α1(I) result in a combined OI/Ehlers-Danlos syndrome (EDS) phenotype [Cabral et al., 2005]. They occur in a domain rich in highly stable triplets that anchors the folding of the amino-terminal end of the triple-helical region. Disruption of the anchor domain causes the unfolding of the N-propeptide cleavage site and N-propeptide retention or delayed processing. Substitutions carboxyl-terminal to helical residue 200 cause a mixture of lethal and nonlethal outcomes. The overall pattern of severity of phenotypes from glycine substitutions along the α1(I) chain (Fig. 1) is not linear along the chain, with severity decreasing from the carboxyl- to the amino-terminal end, nor are lethal mutations found in clusters along the chain [Byers et al., 1991; Marini et al., 1989; Wang et al., 1993]. Depending on the substituting residue, two patterns emerge. Valine and the charged residues are almost uniformly lethal carboxyl-terminal to helical residue 200. Substitutions by serine and cysteine have interspersed lethal and nonlethal outcomes. This suggests that, in the α1(I) chain, the substituting residue is an important variable in the clinical outcome.

Due to the hierarchical organization of collagen in matrix, structural alterations may exert a detrimental effect after chain assembly and helix folding, such as disruption of fibril formation or of binding to noncollagenous molecules in the extracellular matrix. There are two stretches, helical positions 691–823 and 910–964, in which essentially only lethal substitutions are identified. These long stretches of lethal mutations correlate strongly with

two of the three major ligand binding regions (MLBRs) [Di Lullo et al., 2002] in the type I collagen fibril (Fig. 4).

MLBR 2 extends from helical position 682–830, and includes sites important for collagen self-assembly [Prockop and Fertala, 1998], for cleavage by matrix metalloproteinases (MMPs) 1, 2, and 13 [Lauer-Fields et al., 2000], as well as for binding by $\alpha1\beta1/\alpha2\beta1$ integrins [Xu et al., 2000], fibronectin [Dzamba et al., 1993; Kleinman and McGoodwin, 1976; Kleinman et al., 1978], cartilage oligomeric matrix protein (COMP) [Rosenberg et al., 1998], phosphophoryn [Dahl et al., 1998], and possibly the discoidin domain (DDR2) receptor [Schlessinger, 1997; Shrivastava et al., 1997; Vogel et al., 1997]. MLBR 3 extends from helical residue 920 to the end of the helical region. The gly910–964 stretch of lethal mutations includes one of the two major residues for intermolecular collagen cross-linking (the lysyl residue at helical position 930) [Hanson and Eyre, 1996], as well as sites important for the binding of the decorin proteoglycan core protein [Keene et al., 2000], the glycosaminoglycan heparin [San Antonio et al., 1994; Sweeney et al., 1998], and possibly the DDR2 receptor [Schlessinger, 1997; Shrivastava et al., 1997; Vogel et al., 1997]. MLBR 1 is located toward the amino end of the type I collagen helix, where most glycine substitutions are nonlethal.

Gap regions in the mutation map may also be informative, raising the possibility that some regions have functions that are especially crucial, leading to early demise and missed clinical detection of the case. Currently, the longest stretch of glycine residues in the $\alpha1(I)$ chain without known mutations contains eight glycine residues (helical position 484–505), and includes one of the three sequences in the triple helix proposed to mediate the binding of the $\alpha1\beta1/\alpha2\beta1$ integrins [Xu et al., 2000], the major cell surface receptors for type I collagen. This $\alpha1\beta1/\alpha2\beta1$ integrin binding site has also been shown to be important in type I collagen–induced endothelial cell activation [Baronas-Lowell et al., 2004], osteoblast differentiation [Reyes and Garcia, 2004], and angiogenesis [Sweeney et al., 2003]. There are three stretches of seven glycines without mutations: helical position 328–346 has no special function currently attributed to it, helical position 418–436 overlaps one of the predominant residues for nonenzymatic glycation (although the other three such residues in the triple helix or the cross-fibril glycation zone are not associated with gaps) [Hadley et al., 1998; Reiser et al., 1992], and the region at helical position 805–820 that includes the second of the three putative sites for $\alpha1\beta1/\alpha2\beta1$ integrin binding [Xu et al., 2000]. It is also theoretically possible that gap regions represent nonlethal cases that are undetected because they have a different or very mild phenotype. Delineation of the actual role of the gap regions could take advantage of the identification of SNPs, which are not currently available in these regions.

There are other mutation gaps within the two lethal stretches at 691–823 and 910–964 that coincide with MLBR 2 and 3. The gaps could simply represent incomplete mutation coverage. Alternatively, they may further demonstrate the importance of the MLBRs, since some substitutions in MLBRs may have a more severe embryonic lethal phenotype in addition to the perinatal lethal cases that have already been detected.

MUTATIONS IN THE COL1A2 GENE THAT RESULT IN SUBSTITUTIONS FOR GLYCINE IN α2(I)

Prevalence of Substituting Residues

Substitutions of glycine by serine, aspartic acid, and valine comprise about three-quarters of known independent events in the helical domain of α2(I) (Table 3). If mutations occur randomly in the first and second positions of the COL1A2 glycine codons, substitutions of glycine by arginine, valine, and alanine should comprise the majority of independent events in α2(I). The substitution pattern in α2(I) is also different from the pattern seen in α1(I), in which changes in the first position of glycine codons are most prevalent. Although serines, which comprise almost half (44%) of the substitutions for glycine in α2(I), result from first-position changes in glycine codons, aspartic acid and valine result from second-position codon changes. Substitutions by alanine are markedly underrepresented in α2(I), as they were in α1(I). Given the generally milder outcome of alanine substitutions, perhaps there are individuals with alanine substitutions who have a phenotype at the mildest end of the OI spectrum (i.e., early onset osteoporosis) and elude clinical detection or referral.

One factor that may have reduced the proportion of first-position codon substitutions in the COL1A2 gene, compared to COL1A1, is that there are fewer glycine codons in COL1A2 that are embedded in CpG dinucleotides. COL1A2 codons with first-position mutations that are associated with CpG dinucleotides do, however, account for the majority of independent recurrences (Fig. 3B); no codon that is not part of a CpG dinucleotide has more than four recurrences.

Clinical Consequences of Substituting Residue

Compared to mutations that result in a substitution for glycine in COL1A1, a smaller proportion of mutations in COL1A2 resulted in the lethal type II OI phenotype (less than 20%, compared to 35.6%) (Table 3; Fig. 2). Substitutions by charged residues, glutamic and aspartic acids and arginine, are more likely to lead to a lethal outcome than substitutions by other residues (Table 3), as was found for these residues in α1(I). However, valine, with a branched side chain, which is associated with type II OI in 73% of the cases in which it occurs in α1(I), is lethal in only 17% of occurrences in α2(I). Substitution of glycine by cysteine is associated with a lethal outcome in about 25% of cases in both chains. Similar to the outcomes of glycine substitutions in COL1A1, about 9% of COL1A2 substitutions result in the mild type I OI phenotype that is more commonly the result of nonsense mutation in COL1A1.

The proportion of lethal vs. nonlethal outcomes for each amino acid that substitutes for glycine differs from that in α1(I). The distribution of lethal substitutions along the chain also differs in α1(I) and α2(I). In the amino-terminal 30% of the triple helical region of the α2(I) chain, essentially all substitutions for glycine have a nonlethal outcome (Fig. 2). In the carboxyl-terminal two-thirds of the α2(I) triple helical region, all substituting residues have lethal and nonlethal outcomes along the length of the chain. This is true even for substitutions by the charged amino acids, which are virtually always lethal in α1(I) when they occur carboxyl-terminal to helical residue 200 of the triple helical region.

Collagen Stability and Clinical Outcome of Substitution

We examined the distribution of stability scores for the glycine residues containing lethal and/or nonlethal substitutions and compared those distributions to the stability scores of all α2(I) triplets and to all α2(I) triplets in which substitutions have been identified (Table 4), as described above for α1(I) [Bächinger and Davis, 1991; Bächinger et al., 1993; Persikov et al., 2004, 2005]. A significant skewing of phenotypic outcome with triplet stability scores is not supported by the goodness of fit analysis (P = 0.59–0.97), suggesting that the stability of the triplet disrupted by the glycine substitution is not crucial in the distinction between lethal and nonlethal phenotypes.

Locations and Concordance of Clinical Outcome of Recurrent Mutations

About 20% of the glycine residues in the α2(I) chain have more than one independent substitution; five glycine residues have more than five recurrences. Different substituting residues occur at more than half of the glycines at which recurrences have been identified. In contrast to the situation in the α1(I) chain, substitutions of the same glycine in α2(I), even by different amino acids, generally result in concordant outcomes. Furthermore, there is a narrow range of clinical outcome that results from the same substitution at the five glycine residues in α2(I) at which more than five recurrences have been identified. Gly190Ser (p.Gly280-Ser) results in types I and IV OI. Gly238Ser (p.Gly328Ser) results in types III and IV OI, while Gly238Cys or Gly238Asp result in type III OI. Four different substituting residues have been identified at glycine 247 (p.Gly337): cysteine or proline (requiring a 2-nt change) produce type III OI; arginine produces type III/IV OI; and serine produces types III (one case), IV (nine cases), and I (three cases). Substitutions of serine at glycine 370 (p.Gly460) produce OI type III (six cases), type II/III (two cases), and type IV (one case), while at glycine 922 (p.Gly1012) they produce type IV OI (nine cases), type III/IV (one case), and type III (three cases). Independent mutations in different families usually have similar phenotypic outcomes; even substitution of serine for glycine 859 (p.Gly949), which has lethal and nonlethal mutations, always has a severe phenotype (type III, II/III, or II OI).

Clinical outcomes of the same substitution in different families are more concordant in the α2(I) chain than in α1(I). There are only eight glycines in α2(I) at which both lethal and nonlethal outcomes have been found, compared to 35 such glycines in α1(I). Adjusting for the lower number of glycine substitutions identified in α2(I) results in an expectation of 26 nonconcordant outcomes in α2(I), for an incidence comparable to α1(I). In five of the eight nonconcordant cases in α2(I), a serine substitution is nonlethal and a charged or branched amino acid is lethal, demonstrating some role in α2(I) for the effect of particular substituting amino acids. At two glycines (helical positions 811 and 859), substitution by serine has been identified in both lethal and nonlethal cases. The consistently severe phenotype of Gly859Ser (p.Gly949Ser) has been described above, but Gly811Ser (p.Gly901Ser) has both lethal and type IV OI cases.

Genotype–Phenotype Correlations

Lethal mutations in the α2(I) chain occur in eight clusters, which are regularly spaced along two-thirds of the chain (Fig. 5A). The alignment of the cluster boundaries indicated on the map gives the correct phenotype assignment for 86% of glycine substitutions in α2(I). No

particular amino acid is associated with the glycine substitutions that are exceptions to the cluster boundaries. The clustering of lethal glycine substitutions along the α2(I) chain continues to support a "regional model", in which the location of the mutation, rather than the particular substituting residue, is crucial to the clinical outcome. Because the location of the collagen protein alteration is the most important factor in the regional model, the phenotype of exon skipping mutations was predicted to coincide with the lethal clusters. This is generally true, although currently the number of overlaps is still small. Lethal glycine substitution clusters 2, 6, and 8 each overlap with one or two lethal exon-skipping mutations. Lethal cluster 3 overlaps with lethal skipping of exons 32 and 33, if only simple exon skipping is considered. No exon-skipping mutations have been identified within lethal glycine substitution clusters 1, 5, and 7. However, lethal cluster 4 (helical positions 622–637) aligns almost exactly with exon 36, which has a single exon skipping mutation causing type III OI.

The occurrence of the lethal clusters at regular intervals suggests that they disrupt regularly repeating ligand interactions at the level of the collagen heterotrimer or collagen fibril. The α2(I) lethal clusters do not coincide with MLBRs along the collagen helix. Instead, they appear in regions that are proposed to be crucial domains for interactions of proteoglycans with the collagen fibril (Fig. 5B). Lethal Regions 1, 3, and 6 overlap with cross-fibril binding regions for keratan sulfate and heparan sulfate proteoglycans [San Antonio et al., 1994; Scott, 1991]. Lethal Regions 2, 5, and 8 overlap a fibril binding region for keratan sulfate proteoglycans [Scott, 1991]. Lethal region 8 also encompasses the putative monomer-level binding site for the decorin proteoglycan core protein [Keene et al., 2000]. Lethal clusters 4 and 7 overlap with a fibril binding region for dermatan sulfate/chondroitin sulfate proteoglycans, and lethal region 7 overlaps with a monomer-level binding region for dermatan sulfate proteoglycans [Scott et al., 1997].

Examination of the α2(I) glycine substitution map found several gaps of seven or more glycines. The longest gaps (nine to 10 glycines) are located at glycines helical position 46–73, which has no apparent function, and glycines helical position 130–154, which overlaps with the α1β1/α2β1 integrin binding site in MLBR 1 [Xu et al., 2000]. There are three 7-glycine gaps. The first of these (helical position 475–493) overlaps with the cross-fibril zone of nonenzymatic glycation [Hadley et al., 1998; Reiser et al., 1992], and falls close to, but does not overlap, a second putative α1β1/α2β1 integrin binding site [Xu et al., 2000]. The second 7-glycine gap (helical position 718–736) overlaps with binding sites for COMP [Rosenberg et al., 1998] and phosphophoryn [Dahl et al., 1998], while the third gap (helical position 79–97) includes one of the major residues for intermolecular collagen cross-linking [Pietz, 1984] and a site for heparin binding in MLBR 1 [San Antonio et al., 1994; Sweeney et al., 1998].

CHAIN ALIGNMENT IMPLICATIONS

There are two regions in which neither the α1(I) or α2(I) chains have substitutions for glycine, helical residues 484–493 and 721–736. While this may be a consequence of incomplete coverage of the possible substitutions, these regions may encompass a putative site for α1β1/α2β1 integrin-collagen binding across the fibril [Xu et al., 2000], and the

absence of mutations in these domains could reflect an embryonic lethal outcome that is not detected clinically.

SPLICE SITE MUTATIONS IN COL1A1AND COL1A2

Of the 42 exons in the COL1A1 gene that have triple-helical coding domains, there are mutations in 21 acceptor and 24 donor splice-sites; 16 introns have independent mutations at both sites. Of the 43 exons (exons 33 and 34 are fused in COL1A1 but distinct in COL1A2) in the COL1A2 gene that have triple-helical coding domains, there are mutations in nine acceptor and 16 splice donor sites; three introns have independent mutations in both sites. There are, in addition, mutations that lead to misprocessing of exon 6 (five in COL1A1 and 18 in COL1A2), but these mutations all give rise to a phenotype of EDS type VII A and B, respectively.

The 102 separate splice-site mutations in COL1A1 are evenly divided between alterations at acceptor and donor splice sites (Table 5). More than half of the splice-site mutations, including those in the invariant +1, +2, −1, −2 position of splice sites in the introns, result in the type I OI phenotype (27 at acceptor and 32 at donor sites). Only 10 splice site mutations result in lethal type II OI (four at acceptor and six at donor sites), while the remainder result in type III and IV OI (18 acceptor and 13 donor sites). For the majority of splice-site mutations, the mRNA outcome is not known because they were determined directly in genomic DNA and mRNA processing was not studied. Some preliminary generalizations can be made from the 20% of known COL1A1 splice site mutations in which transcript structure was examined. The six lethal cases with mRNA studies included five cases of simple exon skipping (exons 14, 20, 27, 44, and 47) and one case of in-frame retention of part of intron 36. In the instances in which type III or IV phenotypes were studied, there was generally a mixture of mRNA splicing outcomes, including both in- and out-of-frame transcripts. The general situation in the cases with OI type I was that mutations led to the use of cryptic splice-sites, created frameshifts and premature termination codon (PTCs), or, in a few instances, led to intronic retention with a frameshift, so that very little mutant chain was fully translated [Byers and Cole, 2002].

In COL1A2, there are four times more independent mutations at donor sites than at acceptor sites. The 11 lethal mutations are all located in the carboxyl half of the chain; 10 of the cases involved exon skipping (exons 28, 30, 32, 33, 34, 37, 42, and 47) and one involved an in-frame deletion in exon 40. Splice-site mutations in introns 32 and 33 have both lethal and nonlethal cases. In each intron, the lethal case is associated with simple exon skipping, while the nonlethal cases are associated with partial intronic inclusion (6 or 18 nt). Only six COL1A2 splice-site mutations lead to type I OI; these are associated with simple exon skipping in the 5′-half of the mRNA. Splice-site mutations leading to functionally null COL1A2 alleles have an EDS phenotype in at least some cases when both alleles are affected [De Paepe et al., 2002; Schwarze et al., 2004].

DISCUSSION

The database of type I collagen mutations in OI presented here has 832 independent mutations in COL1A1 and COL1A2 genes that alter the sequences of the triple-helical domains of the proα1(I) and proα2(I) chains, respectively. Of these mutations, 682 resulted in substitutions for glycine residues within the triple-helical domains of the chains and 150 alter splice sites. Over 40% of glycine residues within the triple-helical domains now have at least one known substitution and about 10% of all possible mutational events have been identified. We have not included mutations that result in premature termination codons in COL1A1 or COL1A2. Mutations causing PTCs in COL1A1 result in mRNA instability, reduce the amount of normal type I procollagen produced, and result in the mild type I OI phenotype [Willing et al., 1994]. Similar mutations in the COL1A2 gene have no phenotype in the heterozygote and result in a form of EDS in the homozygote or compound heterozygote.

The substantial expansion of the previous database, which is presented here, provides the material for new insights into mutation outcome and mechanism. Several factors contribute to the difficulty of modeling the clinical outcome of collagen mutations. First, the production of type I collagen is a complex process and the mutations in the triple-helical domain may affect chain synthesis and folding into the helix, the transit from the rough endoplasmic reticulum (RER) through the Golgi into the extracellular space, procollagen processing, and fibril assembly, and, in bone, mineralization [Canty and Kadler, 2005]. Second, a considerable number of mutations are required for even partial saturation of the mutation map. Third, *trans*-acting factors are doubtless important in phenotype expression, including polymorphisms in type I collagen and in the noncollagenous proteins that interact with the collagen monomer and fibril.

Distinct genotype–phenotype relationships for the two chains emerge from this analysis, supporting different roles for each chain in supporting extracellular matrix integrity. In the α1(I) chain, substitutions for glycine in the triple-helical domain are more likely to result in lethal type II OI than substitutions in the α2(I) chain. Amino acid residues with a charged or branched side chain, specifically aspartic acid, arginine and valine, usually result in lethal phenotypes when they occur in the α1(I) chain but are less likely to do so in the α2(I) chain. The proportion of specific substituting residues in α1(I) deviates from those predicted by random mutations in the known glycine codons. Substitutions caused by changes in the first position of the codons (serine, cysteine, and arginine) occur more frequently than predicted by random occurrence. Glycine residues with high numbers of independent recurrences in both chains are associated with CpG sequences [Schorderet and Gartler, 1992]. The pattern of lethal and nonlethal mutations along the chain differs depending on the substituting residue in the α1(I) chain. Substitutions in the first 200 helical residues are almost all nonlethal. In the remainder of the chain, residues with charged and branched side chains are lethal, while serine and cysteine have intermixed lethal and nonlethal outcomes along the length of the chain. The variability of outcome for independent recurrences of the same substitutions at glycine residues in the α1(I) chain is notable; at over 30 sites, the same glycine residue has independent lethal and nonlethal outcomes, 10 of which involve the

same substituting residue. This variability may be ascribed to stochastic effects or to unknown modifiers.

There is no discernable correlation of clinical outcome with extent of overmodification in a linear progression along the α1(I) chain, or with recurrent regional clusters. One notable feature that emerges from the COL1A1 mutation map is the occurrence of two regions, helical residues 691–823 and 910–964, containing almost exclusively lethal mutations and gap zones without mutations, which might represent incomplete mutation coverage of the chain or extremely severe cases with demise early in gestation. These two regions align with MLBRs on the collagen map, suggesting that they have a role in binding of integrins, MMPs, fibronectin, and COMP [Di Lullo et al., 2002].

Splice-site mutations comprise 20% of the total COL1A1 mutations identified. Most are nonlethal and many result in a mild phenotype. These mutations are most likely to lead to the use of cryptic splice-sites, which often introduce premature termination codons. In most instances, no studies of the effects of the mutations on mRNA processing and stability were completed.

Substitutions for glycine in the triple-helical domain of the α2(I) chain are less likely to be lethal than those in the α1(I) chain (20% vs. 35.6%). Also, substitutions by aspartic acid, glutamic acid, arginine, and valine, which are usually lethal beyond the first 200 helical residues of α1(I), have variable outcomes along the α2(I) chain. Substitutions in α2(I) do not show the strong association with first-position changes in the glycine codons noted previously for α1(I). Of the most prevalent substituting residues, serine, aspartic acid, and valine, only serine results from a first-position change. Substitutions by alanine are strikingly under-represented in α2(I), as in α1(I). It is possible that this smallest of substituting residues for glycine has a milder phenotypic effect that does not generally result in genetic testing. As in α1(I), glycine residues with a high number of independent recurrences have preceding CpG sequences. Unlike α1(I), recurrent mutations that affect the same glycine residues in α2(I) are most often concordant for outcome, even when different substituting residues are involved. There are only five glycines in α2(I) at which both lethal and nonlethal substitutions are known. These distinctions most often involve nonlethal serine and lethal charged residues at the same location, showing some effect of the more destabilizing charged residues. Splice-site mutations in α2(I) have been found preferentially on splice donor sites.

The distribution of lethal and nonlethal glycine substitutions along the α2(I) chain is different than in α1(I). The lethal mutations are located in eight clusters regularly spaced along the chain, supporting a "regional model." Lethal exon skipping coincides with glycine lethal clusters. The alignment of the lethal regions in the α2(I) chain with sites proposed to be important for binding of matrix proteoglycans suggests that they play a crucial role at the level of the collagen fibril and its interactions with other structural molecules of the extracellular matrix. Disruption of fibril-level interactions by glycine substitutions in the lethal regions of the α2(I) chain requires that collagen molecules containing those structural alterations are secreted from the cell and incorporated into mature matrix. Similarly, the proposed effects of lethal substitutions in the α1(I) chain on collagen triple-helix interaction

with integrins, COMP, and other molecules at the MLBRs [Di Lullo et al., 2002] applies to collagen molecules that are secreted from the cell.

There has been a long-standing hypothesis that the OI phenotype of particular collagen mutations is related to the extent to which they disrupt the stability of the collagen helix. Over a decade ago, Bächinger and Davis [1991] and Bächinger et al. [1993] studied the contribution of helix folding and of particular Gly-X-Y units to helix stability, assigning a score to each tripeptide composition. They noted that severity of OI does not correlate with the decrease in helix melting temperature (T_m) resulting from specific glycine substitutions. We found that glycine substitutions in the α1(I) and α2(I) chains which resulted in lethal phenotypes did not correlate with higher or lower scoring triplets. Bächinger and Davis [1991] and Bächinger et al. [1993] also postulated an essential role for the triplet located amino terminal to the glycine substitution, involving renucleation of folding and maintaining the register of the chains. The correlation of lethal phenotypes with substituting residues in the α1(I) chain with charged or branched side chains is consistent with a disruption of renucleation as an important contributor to clinical outcome; we did not find such an effect for α2(I) chain substitutions. More recently, Persikov et al. [2004, 2005] conducted extensive stability experiments with host–guest peptides. Using these data, they calculated stability along the type I heterotrimer, describing stability at each triplet using a window of two triplets on each side to account for neighboring effects and then using a weighted average for chain composition. The resulting helical stability plot does not correlate with phenotype along either type I collagen chain (data not shown). The best interpretation of the current data is that the effect of a particular substitution on collagen stability is not the primary determinant of clinical outcome but that stability effects do play a modifying role in α1(I) chain substitutions.

For collagen, with its successive levels of assembly and interaction, dominant-negative mutations may disrupt multiple nonexclusive processes. Some mutations have intracellular consequences. Delay in helix folding causes variable alpha chain overmodification of collagen. Exon skipping mutations may lead to multiple in- or out-of-frame transcripts. The association of a high proportion of splice mutations in +1, +2, −1, or −2 intronic positions with the mild type I OI phenotype, most likely reflects fast removal of the adjacent intron containing the closest splice site normally used in procollagen mRNA, with default use of a cryptic splice site, creating a frameshift or PTC. Those containing PTCs will be subject to nonsense mediated decay [Slayton et al., 2000]. Glycine substitutions and exon skipping mutations may result in ER retention of some or all mutant chains; this leads to variable extents of matrix insufficiency as well as more general metabolic consequences for the cells. In osteoblasts, cell differentiation or expression of noncollagenous matrix components may be altered; altered expression or function of cell surface receptors may disrupt cell–cell or cell–matrix interactions.

Some glycine substitutions or splice-site defects will have pericellular consequences, in which the removal of either N- or C-propeptides is affected by adjacent mutations or register shifts. Propeptide retention alters collagen fibril geometry and strength.

Other structural defects of type I collagen can affect the ability to form fibrils, that is, to become stably incorporated into matrix with mature cross-links to other collagen heterotrimers. Fibrils containing collagen with glycine substitutions may have abnormal interactions with other molecules in the extracellular matrix [Byers and Cole, 2002]. The proposed effects of lethal substitutions in the α1(I) chain on collagen triple-helix interaction with integrins, COMP, and other molecules at the MLBRs [Di Lullo et al., 2002], and the proposed disruption by glycine substitutions in the lethal regions of the α2(I) chain of fibril-level interactions with proteoglycans, are examples of potential dominant negative functions of secreted mutant collagen.

The mutation maps in this review support disruption of collagen–matrix interactions as a crucial component in phenotype determination. For the α1(I) chain, disruption of interactions in the MLBRs at the level of the collagen monomer almost always has a lethal outcome. For the α2(I) chain, the crucial regions occur at the regularly repeating sites for interactions of matrix proteoglycans with the collagen fibril. The information presented here provides guidelines for genetic counseling and molecular/biochemical correlations for studies of potential treatments. Further study of these mutations and their role in collagen–matrix interactions should provide insight into the pathophysiology of OI bone and into normal bone functions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Ayad, S.; Boot-Handford, RP.; Humphries, MS.; Kadler, CA. The extracellular matrix facts book. London: Academic Press; 1998.

Bächinger HP, Davis JM. Sequence specific thermal stability of the collagen triple helix. Int J Biol Macromol. 1991; 13:152–156. [PubMed: 1911555]

Bächinger HP, Morris NP, Davis JM. Thermal stability and folding of the collagen triple helix and the effects of mutations in osteogenesis imperfecta on the helix of type I collagen. Am J Med Genet. 1993; 45:152–162. [PubMed: 8456797]

Barnes, AM.; Chang, W.; Morello, R.; Cabral, WA.; Weis, MA.; Eyre, DR.; Makareeva, E.; Kouznetsova, N.; Leikin, S.; Uveges, TE.; Mulvihill, JJ.; Wilson, PL.; Sundaram, UT.; Lee, B.; Marini, JC. Recessive lethal form of osteogenesis imperfecta caused by null mutations in CRTAP. 28th Annual Meeting of American Society of Bone and Mineral Research; Philadelphia, PA. 2006. p. Abstract 922

Baronas-Lowell D, Lauer-Fields JL, Fields GB. Induction of endothelial cell activation by a triple helical α2β1 integrin ligand, derived from type I collagen α1(I)496–507. J Biol Chem. 2004; 279:952–962. [PubMed: 14581484]

Byers, PH.; Cole, WG. Osteogenesis imperfecta. In: Royce, PM.; Steinmann, B., editors. Connective tissue and its heritable disorders: molecular, genetic and medical aspects. New York: Wiley-Liss; 2002. p. 385-430.

Byers PH, Wallis GA, Willing MC. Osteogenesis imperfecta: translation of mutation to phenotype. J Med Genet. 1991; 28:433–442. [PubMed: 1895312]

Cabral WA, Makareeva E, Colige A, Letocha AD, Ty JM, Yeowell HN, Pals G, Leikin S, Marini JC. Mutations near amino end of α1(I) collagen cause combined osteogenesis imperfecta/Ehlers-Danlos syndrome by interference with N-propeptide processing. J Biol Chem. 2005; 280:19259–19269. [PubMed: 15728585]

Canty EG, Kadler KE. Procollagen trafficking, processing and fibrillogenesis. J Cell Sci. 2005; 118(Pt 7):1341–1353. [PubMed: 15788652]

Chakravarti S, Magnuson T, Lass JH, Jepsen KJ, LaMantia C, Carroll H. Lumican regulates collagen fibril assembly: skin fragility and corneal opacity in the absence of lumican. J Cell Biol. 1998; 141:1277–1286. [PubMed: 9606218]

Chapman JA. The staining pattern of collagen fibrils. I. An analysis of electron micrographs. Connect Tissue Res. 1974; 2:137–150. [PubMed: 4138005]

Chu ML, Williams CJ, Pepe G, Hirsch JL, Prockop DJ, Ramirez F. Internal deletion in a collagen gene in a perinatal lethal form of osteogenesis imperfecta. Nature. 1983; 304:78–80. [PubMed: 6191221]

Chu, ML.; Prockop, DJ. Collagen: gene structure. In: Royce, PM.; Steinmann, B., editors. Connective tissue and its heritable disorders: molecular, genetic and medical aspects. New York: Wiley-Liss; 2002. p. 223-248.

Colige A, Li SW, Sieron AL, Nusgens BV, Prockop DJ, Lapiere CM. cDNA cloning and expression of bovine procollagen I N-proteinase: a new member of the superfamily of zinc-metalloproteinases with binding sites for cells and other matrix components. Proc Natl Acad Sci USA. 1997; 94:2374–2379. [PubMed: 9122202]

Dahl T, Sabsay B, Veis A. Type I collagen-phosphophoryn interactions: specificity of the monomer-monomer binding. J Struct Biol. 1998; 123:162–168. [PubMed: 9843670]

Dalgleish R. The human type I collagen mutation database. Nucleic Acids Res. 1997; 25:181–187. [PubMed: 9016532]

Dalgleish R. The Human Collagen Mutation Database 1998. Nucleic Acids Res. 1998; 26:253–255. [PubMed: 9399846]

den Dunnen JT, Antonarakis SE. Nomenclature for the description of human sequence variations. Hum Genet. 2001; 109:121–124. [PubMed: 11479744]

den Dunnen JT, Paalman MH. Standardizing mutation nomenclature: why bother? Hum Mutat. 2003; 22:181–182. [PubMed: 12938082]

Di Lullo GA, Sweeney SM, Körkkö J, Ala-Kokko L, San Antonio JD. Mapping the ligand-binding sites and disease-associated mutations on the most abundant protein in the human, type I collagen. J Biol Chem. 2002; 277:4223–4231. [PubMed: 11704682]

Dzamba BJ, Wu H, Jaenisch R, Peters DM. Fibronectin binding site in type I collagen regulates fibronectin fibril formation. J Cell Biol. 1993; 121:1165–1172. [PubMed: 8501121]

Engel J, Prockop DJ. The zipper-like folding of collagen triple helices and the effects of mutations that disrupt the zipper. Annu Rev Biophys Biophys Chem. 1991; 20:137–152. [PubMed: 1867713]

Font B, Eichenberger D, Goldschmidt D, Boutillon MM, Hulmes DJ. Structural requirements for fibromodulin binding to collagen and the control of type I collagen fibrillogenesis—critical roles for disulphide bonding and the C-terminal region. Eur J Biochem. 1998; 254:580–587. [PubMed: 9688269]

Forlino A, Marini JC. Osteogenesis imperfecta: prospects for molecular therapeutics. Mol Genet Metab. 2000; 71:225–232. [PubMed: 11001814]

Genovese C, Rowe D. Analysis of cytoplasmic and nuclear messenger RNA in fibroblasts from patients with type I osteogenesis imperfecta. Methods Enzymol. 1987; 145:223–235. [PubMed: 3474490]

Hadley JC, Meek KM, Malik NS. Glycation changes the charge distribution of type I collagen fibrils. Glycoconj J. 1998; 15:835–840. [PubMed: 9870360]

Hanson DA, Eyre DR. Molecular site specificity of pyridinoline and pyrrole cross-links in type I collagen of human bone. J Biol Chem. 1996; 271:26508–26516. [PubMed: 8900119]

Keene DR, San Antonio JD, Mayne R, McQuillan DJ, Sarris G, Santoro SA, Iozzo RV. Decorin binds near the C terminus of type I collagen. J Biol Chem. 2000; 275:21801–21804. [PubMed: 10823816]

Kleinman HK, McGoodwin EB. Localization of the cell attachment region in types I and II collagens. Biochem Biophys Res Commun. 1976; 72:426–432. [PubMed: 985489]

Kleinman HK, McGoodwin EB, Martin GR, Klebe RJ, Fietzek PP, Woolley DE. Localization of the binding site for cell attachment in the α1(I) chain of collagen. J Biol Chem. 1978; 253:5642–5646. [PubMed: 209046]

Körkkö J, Ala-Kokko L, De Paepe A, Nuytinck L, Earley J, Prockop DJ. Analysis of the COL1A1 and COL1A2 genes by PCR amplification and scanning by conformation-sensitive gel electrophoresis identifies only COL1A1 mutations in 15 patients with osteogenesis imperfecta type I: identification of common sequences of null-allele mutations. Am J Hum Genet. 1998; 62:98–110. [PubMed: 9443882]

Kuivaniemi H, Tromp G, Prockop DJ. Mutations in collagen genes: causes of rare and some common diseases in humans. FASEB J. 1991; 5:2052–2060. [PubMed: 2010058]

Kuivaniemi H, Tromp G, Prockop DJ. Mutations in fibrillar collagens (types I, II, III, and XI), fibril-associated collagen (type IX), and network-forming collagen (type X) cause a spectrum of diseases of bone, cartilage, and blood vessels. Hum Mutat. 1997; 9:300–315. [PubMed: 9101290]

Kyriakides TR, Zhu YH, Smith LT, Bain SD, Yang Z, Lin MT, Danielson KG, Iozzo RV, LaMarca M, McKinney CE, Ginns EI, Bornstein P. Mice that lack thrombospondin 2 display connective tissue abnormalities that are associated with disordered collagen fibrillogenesis, an increased vascular density, and a bleeding diathesis. J Cell Biol. 1998; 140:419–430. [PubMed: 9442117]

Lauer-Fields JL, Tuzinski KA, Shimokawa K, Nagase H, Fields GB. Hydrolysis of triple-helical collagen peptide models by matrix metalloproteinases. J Biol Chem. 2000; 275:13282–13290. [PubMed: 10788434]

Malfait F, Symoens S, Coucke P, Nunes L, De Almeida S, De Paepe A. Total absence of the α2(I) chain of collagen type I causes a rare form of Ehlers-Danlos syndrome with hypermobility and propensity to cardiac valvular problems. J Med Genet. 2006; 43:e36. [PubMed: 16816023]

Marini JC, Grange DK, Gottesman GS, Lewis MB, Koeplin DA. Osteogenesis imperfecta type IV. Detection of a point mutation in one α1(I) collagen allele (COL1A1) by RNA/RNA hybrid analysis. J Biol Chem. 1989; 264:11893–11900. [PubMed: 2745420]

Marini JC, Lewis MB, Wang Q, Chen KJ, Orrison BM. Serine for glycine substitutions in type I collagen in two cases of type IV osteogenesis imperfecta (OI). Additional evidence for a regional model of OI pathophysiology. J Biol Chem. 1993; 268:2667–2673. [PubMed: 8094076]

Marini, JC. Osteogenesis imperfecta. In: Behrman, RE.; Kliegman, RM.; Jensen, HB., editors. Nelson's textbook of pediatrics. Philadelphia: Saunders; 2004. p. 2336-2338.

Persikov AV, Pillitteri RJ, Amin P, Schwarze U, Byers PH, Brodsky B. Stability related bias in residues replacing glycines within the collagen triple helix (Gly-Xaa-Yaa) in inherited connective tissue disorders. Hum Mutat. 2004; 24:330–337. [PubMed: 15365990]

Persikov AV, Ramshaw JA, Brodsky B. Prediction of collagen stability from amino acid sequence. J Biol Chem. 2005; 280:19343–19349. [PubMed: 15753081]

Pietz, KA. Extracellular matrix biochemistry. New York: Elsevier; 1984.

Prockop DJ, Kivirikko KI. Heritable diseases of collagen. N Engl J Med. 1984; 311:376–386. [PubMed: 6146097]

Prockop DJ, Constantinou CD, Dombrowski KE, Hojima Y, Kadler KE, Kuivaniemi H, Tromp G, Vogel BE. Type I procollagen: the gene-protein system that harbors most of the mutations causing osteogenesis imperfecta and probably more common heritable disorders of connective tissue. Am J Med Genet. 1989; 34:60–67. [PubMed: 2683782]

Prockop DJ, Colige A, Helminen H, Khillan JS, Pereira R, Vandenberg P. Mutations in type I procollagen that cause osteogenesis imperfecta: effects of the mutations on the assembly of collagen into fibrils, the basis of phenotypic variations, and potential antisense therapies. J Bone Miner Res. 1993; 8(Suppl 2):S489–S492. [PubMed: 8122517]

Prockop DJ, Fertala A. Inhibition of the self-assembly of collagen I into fibrils with synthetic peptides. Demonstration that assembly is driven by specific binding sites on the monomers. J Biol Chem. 1998; 273:15598–15604. [PubMed: 9624151]

Rauch F, Glorieux FH. Osteogenesis imperfecta. Lancet. 2004; 363:1377–1385. [PubMed: 15110498]

Redford-Badwal DA, Stover ML, Valli M, McKinstry MB, Rowe DW. Nuclear retention of COL1A1 messenger RNA identifies null alleles causing mild osteogenesis imperfecta. J Clin Invest. 1996; 97:1035–1040. [PubMed: 8613526]

Reiser KM, Amigable MA, Last JA. Nonenzymatic glycation of type I collagen. The effects of aging on preferential glycation sites. J Biol Chem. 1992; 267:24207–24216. [PubMed: 1447170]

Reyes CD, Garcia AJ. α2β1 integrin-specific collagen-mimetic surfaces supporting osteoblastic differentiation. J Biomed Mater Res A. 2004; 69:591–600. [PubMed: 15162400]

Rosenberg K, Olsson H, Morgelin M, Heinegard D. Cartilage oligomeric matrix protein shows high affinity zinc-dependent interaction with triple helical collagen. J Biol Chem. 1998; 273:20397–20403. [PubMed: 9685393]

San Antonio JD, Lander AD, Karnovsky MJ, Slayter HS. Mapping the heparin-binding sites on type I collagen monomers and fibrils. J Cell Biol. 1994; 125:1179–1188. [PubMed: 8195298]

Schlessinger J. Direct binding and activation of receptor tyrosine kinases by collagen. Cell. 1997; 91:869–872. [PubMed: 9428509]

Schorderet DF, Gartler SM. Analysis of CpG suppression in methylated and nonmethylated species. Proc Natl Acad Sci USA. 1992; 89:957–961. [PubMed: 1736311]

Schwarze U, Hata R, McKusick VA, Shinkai H, Hoyme HE, Pyeritz RE, Byers PH. Rare autosomal recessive cardiac valvular form of Ehlers-Danlos syndrome results from mutations in the COL1A2 gene that activate the nonsense-mediated RNA decay pathway. Am J Hum Genet. 2004; 74:917–930. [PubMed: 15077201]

Scott JE. Proteoglycan-fibrillar collagen interactions. Biochem J. 1988; 252:313–323. [PubMed: 3046606]

Scott JE. Proteoglycan: collagen interactions in connective tissues. Ultrastructural, biochemical, functional and evolutionary aspects. Int J Biol Macromol. 1991; 13:157–161. [PubMed: 1911556]

Scott JE, Ritchie M, Glanville RW, Cronshaw AD. Peptide sequences in glutaraldehyde-linked proteodermatan sulphate: collagen fragments from rat tail tendon locate the proteoglycan binding sites. Biochem Soc Trans. 1997; 25:S663. [PubMed: 9450091]

Shrivastava A, Radziejewski C, Campbell E, Kovac L, McGlynn M, Ryan TE, Davis S, Goldfarb MP, Glass DJ, Lemke G, Yancopoulos GD. An orphan receptor tyrosine kinase family whose members serve as nonintegrin collagen receptors. Mol Cell. 1997; 1:25–34. [PubMed: 9659900]

Sillence DO, Rimoin DL, Danks DM. Clinical variability in osteogenesis imperfecta-variable expressivity or genetic heterogeneity. Birth Defects Orig Artic Ser. 1979a; 15:113–129. [PubMed: 393318]

Sillence DO, Senn A, Danks DM. Genetic heterogeneity in osteogenesis imperfecta. J Med Genet. 1979b; 16:101–116. [PubMed: 458828]

Slayton RL, Deschenes SP, Willing MC. Nonsense mutations in the COL1A1 gene preferentially reduce nuclear levels of mRNA but not hnRNA in osteogenesis imperfecta type I cell strains. Matrix Biol. 2000; 19:1–9. [PubMed: 10686420]

Svensson L, Aszodi A, Reinholt FP, Fassler R, Heinegard D, Oldberg A. Fibromodulin-null mice have abnormal collagen fibrils, tissue organization, and altered lumican deposition in tendon. J Biol Chem. 1999; 274:9636–9647. [PubMed: 10092650]

Sweeney SM, Guy CA, Fields GB, San Antonio JD. Defining the domains of type I collagen involved in heparin-binding and endothelial tube formation. Proc Natl Acad Sci USA. 1998; 95:7275–7280. [PubMed: 9636139]

Sweeney SM, DiLullo G, Slater SJ, Martinez J, Iozzo RV, Lauer-Fields JL, Fields GB, San Antonio JD. Angiogenesis in collagen I requires α2β1 ligation of a GFP*GER sequence and possibly p38 MAPK activation and focal adhesion disassembly. J Biol Chem. 2003; 278:30516–30524. [PubMed: 12788934]

Sykes B. The molecular genetics of collagen. Bioessays. 1985; 3:112–117. [PubMed: 3842588]

van der Rest M, Garrone R. Collagen family of proteins. FASEB J. 1991; 5:2814–2823. [PubMed: 1916105]

Vogel W, Gish GD, Alves F, Pawson T. The discoidin domain receptor tyrosine kinases are activated by collagen. Mol Cell. 1997; 1:13–23. [PubMed: 9659899]

Wang Q, Orrison BM, Marini JC. Two additional cases of osteogenesis imperfecta with substitutions for glycine in the α2(I) collagen chain. A regional model relating mutation location with phenotype. J Biol Chem. 1993; 268:25162–25167. [PubMed: 7693712]

Willing MC, Pruchno CJ, Atkinson M, Byers PH. Osteogenesis imperfecta type I is commonly due to a COL1A1 null allele of type I collagen. Am J Hum Genet. 1992; 51:508–515. [PubMed: 1353940]

Willing MC, Deschenes SP, Scott DA, Byers PH, Slayton RL, Pitts SH, Arikat H, Roberts EJ. Osteogenesis imperfecta type I: molecular heterogeneity for COL1A1 null alleles of type I collagen. Am J Hum Genet. 1994; 55:638–647. [PubMed: 7942841]

Willing MC, Deschenes SP, Slayton RL, Roberts EJ. Premature chain termination is a unifying mechanism for COL1A1 null alleles in osteogenesis imperfecta type I cell strains. Am J Hum Genet. 1996; 59:799–809. [PubMed: 8808594]

Xu Y, Gurusiddappa S, Rich RL, Owens RT, Keene DR, Mayne R, Hook A, Hook M. Multiple binding sites in collagen type I for the integrins $\alpha 1\beta 1$ and $\alpha 2\beta 1$. J Biol Chem. 2000; 275:38981–38989. [PubMed: 10986291]
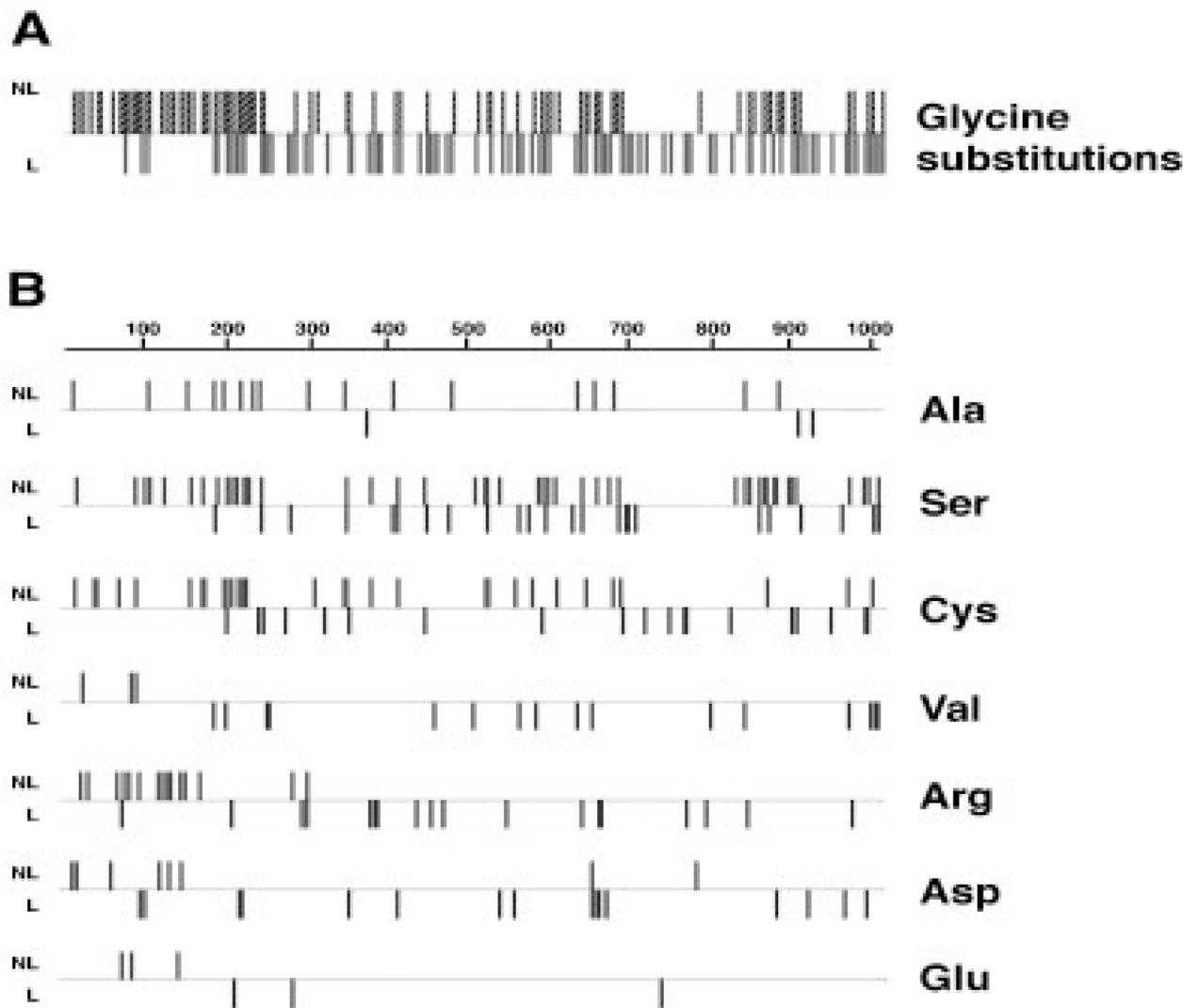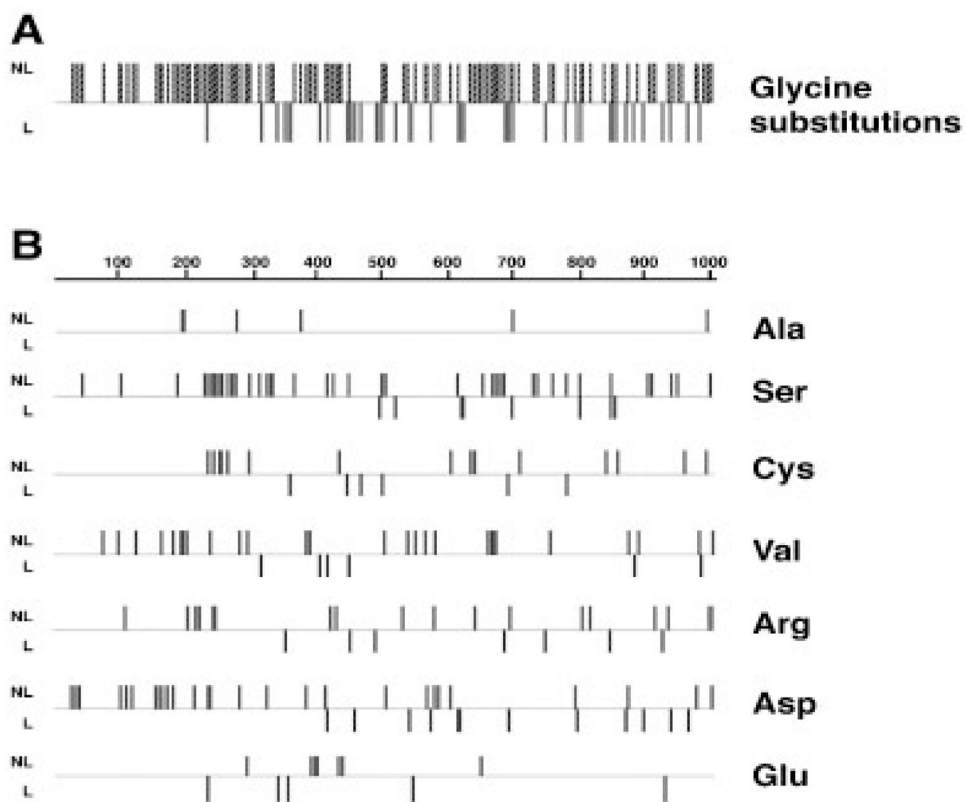
FIGURE 1.
Distribution of the mutations along the α1(I) collagen chain. **A:** Glycine substitutions caused by single-nucleotide changes. **B:** Glycine mutations shown for each substituting amino acid, indicated to the right end of each line. The top line indicates the scale used based on residue in the triple-helical region. The vertical bars indicate the mutations. Each horizontal line represents an α1(I) chain. NL, nonlethal; L, lethal.

FIGURE 2.
Distribution of the mutations along the α1(I) collagen chain. **A:** Glycine substitutions caused by single-nucleotide changes. **B:** Glycine mutations shown for each substituting amino acid, indicated to the right end of each line. The top line indicates the scale used based on residue number in the triple-helical region. The vertical bars indicate the mutations. Each horizontal line represents an α2(I) chain. NL, nonlethal; L, lethal.
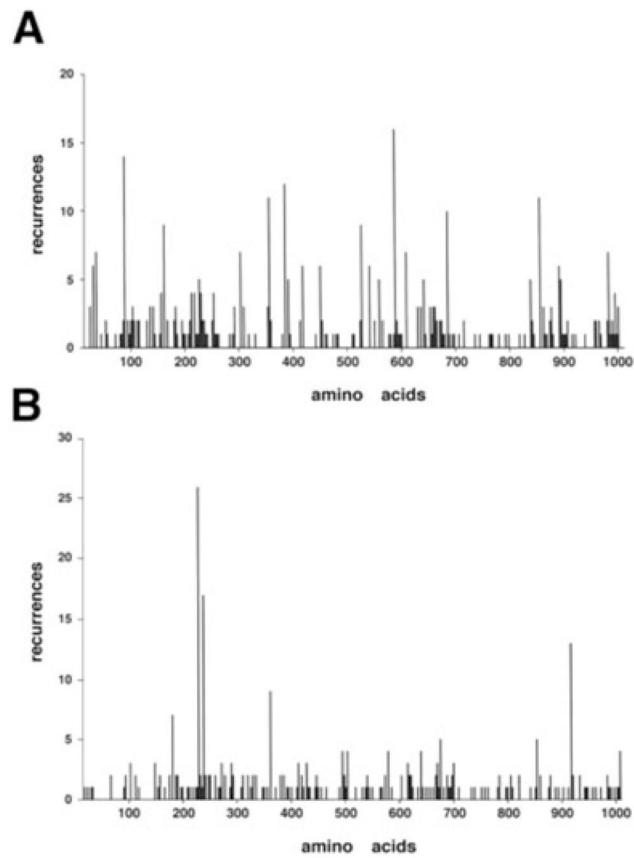
FIGURE 3.
Recurrences of mutations at the same glycine residue in α1(I) (**A**) and α2(I) (**B**) collagen chains.
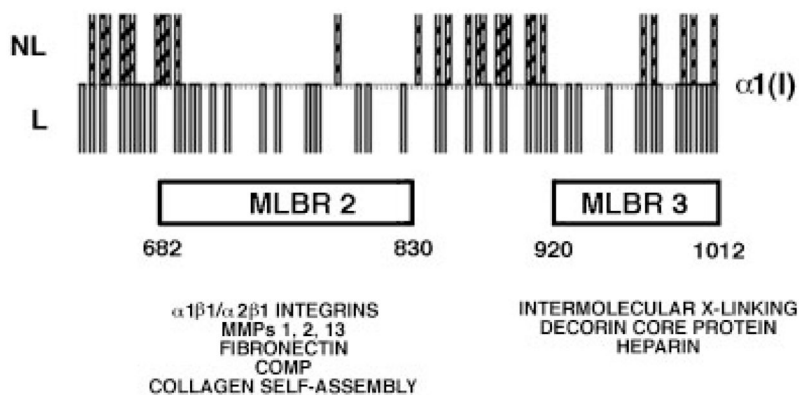
FIGURE 4.
Glycine substitution map of the last 600 C-terminal amino acids of the α1(I) chain. There are three stretches of seven glycines without known substitutions at positions 328–346, 418–436, and 805–820. The last two regions are included within MLBR2 and MLBR3. The vertical bars indicate the glycine substitutions. The bottom boxes represent MLBR2 and MLBR3. The numbers flanking the two boxes denote glycine residue positions in the α1(I) chain.
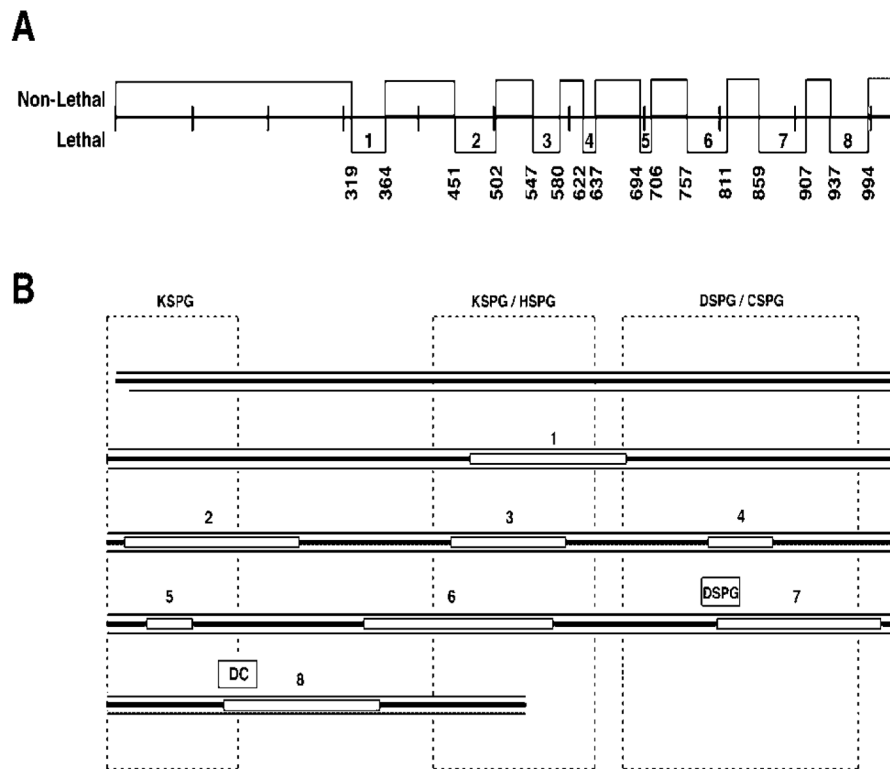
FIGURE 5.
A: Regional model for the distribution of mutations along the α2(I) chain. The lethal mutations are located in eight regularly spaced clusters. Amino acid residues at the boundary of each domain are indicated. Scale bars mark each 100 amino acid along the helix. **B:** Localization of the lethal clusters of the mutations (white boxes) on a collagen fibril D-period model composed of five collagen molecules (horizontal lines) [Chapman, 1974]. The dotted boxes show the binding region on the type I collagen fibril of keratin sulfate (KSPG), keratan sulfate and heparan sulfate (KSPG/HSPG), and dermatan sulfate and chondroitin sulfate proteoglycans (DSPG/CSPG). The monomer-level binding site for decorin core protein (DC) and DSPG are also shown as white boxes overlapping, respectively, lethal cluster number 8 and lethal cluster number 7.

TABLE 1

Clinical Classification of Osteogenesis Imperfecta

Type	Genetics	Clinical findings	Ultrasound findings	First ultrasound detection
Modified Sillence classification				
OI I	Autosomal dominant	Fractures with little or no limb deformity, blue sclerae, normal stature, hearing loss, DI	Rarely, long bone bowing or fracture	>20 weeks but not common
OI II	Autosomal dominant	Lethal perinatal type: undermineralized skull, micromelic bones, "beaded" ribs on x-ray, bone deformity, platyspondyly	Undermineralization, broad, crumpled and shortened limbs, thin beaded ribs, fractures, angulation or bowing of long bones, normal appearing hands, deformable calvarium	14 weeks
OI III	Autosomal dominant	Progressively deforming type: moderate deformity of limbs at birth, scleral hue varies, very short stature, dentinogenesis imperfecta (DI)	Thin ribs, short limbs, fractures, undermineralized skull, long bone length falls away from normal 16–18 weeks	18 weeks
OI IV	Autosomal dominant	Normal sclerae, mild/moderate limb deformity with fracture, variable short stature, DI, some hearing loss	Rarely, long bone bowing and/or fracture	After 20 weeks but not common
New OI types based on bone histology				
OI V	Autosomal dominant	Similar to OI IV plus calcification of interosseous membrane of forearm, radial head dislocation, and hyperplastic callus formation	Unknown	Not described
OI VI	Unknown	More fractures than OI type IV, vertebral compression fractures, no DI	Unknown	Not described
OI VII	Autosomal recessive	Congenital fractures, white sclerae, early deformity of legs, coxa vara, osteopenia	Skeleton poorly mineralized with severe micromelia; Barnes et al. [2006]	Not described

TABLE 2

Mutations by Type and Distribution

Helical residues	Glycine substitutions			Exons	Splice-site defects		
	Lethal	Nonlethal	Total		Lethal	Nonlethal	Total
α1(I) chain							
0–199	5	84	89	7–17	2	36	38
200–399	40	48	88	18–25	1	19	20
400–599	29	44	73	26–34	3	19	22
600–799	30	31	61	34–41	1	7	8
800–1012	35	45	80	42–49	3	11	14
Total	139	252	391		10	92	102
α2(I) chain							
0–199	0	35	35	7–17	0	20	20
200–399	8	93	101	18–25	0	5	5
400–599	20	29	49	26–34	6	5	11
600–799	13	36	49	34–41	2	3	5
800–1012	14	43	57	42–49	3	4	7
Total	55	236	291		11	37	48

TABLE 3

Effect of Specific Amino Acid Residue Substitutions for Glycine

	α1(I) chain			α2(I) chain		
	Mutations identified (%)	Theoretical (%)	Mutations lethal (%)	Mutations identified (%)	Theoretical (%)	Mutations lethal (%)
Arg	71 (18.2)	20.8	32 (45.0)	30 (10.3)	21.7	7 (23.3)
Cys	84 (21.5)	13.5	25 (29.8)	25 (8.6)	12.8	6 (24.0)
Ser	152 (38.9)	13.5	37 (24.3)	128 (44.0)	12.8	14 (10.9)
Ala	22 (5.6)	17.2	3 (13.6)	6 (2.1)	17.3	0
Asp	29 (7.4)	13.5	19 (65.5)	46 (15.8)	12.8	15 (32.6)
Glu	7 (1.8)	3.7	4 (57.1)	14 (4.9)	4.4	6 (42.9)
Val	26 (6.6)	17.2	19 (73.1)	40 (13.9)	17.3	7 (17.5)
Trp	0	0.6		1 (0.4)	0.8	0
Pro						
Lethal substitutions (%)			139 (35.6)			55 (18.9)

TABLE 4

Bächinger Triplet Stability Scores

	Bächinger scores									
	α1(I)					α2(I)				
	0	1	1.5	2	0	1	1.5	2		
Lethal	14 (15.2%)	38 (41.3%)	2 (2.2%)	37 (40.6%)	4 (9.5%)	23 (54.8%)	3 (7.1%)	12 (28.6%)		
Nonlethal	19 (20.7%)	39 (42.4%)	3 (3.3%)	31 (33.7%)	11 (10.0%)	53 (48.2%)	7 (6.4%)	39 (35.5%)		
Triplets with mutation	26 (17.3%)	61 (40.7%)	4 (2.7 %)	59 (39.3%)	15 (10.4%)	71 (49.3%)	10 (6.9%)	48 (33.3%)		
All triplets	56 (16.6%)	136 (40.2%)	15 (4.4%)	131 (38.8%)	34 (10.1%)	183 (54.1%)	17 (5.0%)	104 (30.8%)		

TABLE 5

Splice-Site Mutations and Clinical Outcome

	OI type II	OI type III/IV	OI type I
COL1A1			
All acceptor sites	4	18	27
−1, −2 positions	4	14	22
All donor sites	6	14	32
+1, +2 positions	3	8	21
COL1A2			
All acceptor sites	2	4	3
−1, −2 positions	1	3	2
All donor sites	6	21	8
+1, +2 positions	3	13	2