

Type I Collagen Molecular Map Lends Insights into the Domain Structure of the Fibril and the Genotype-Phenotype Relationship for Some Collagen Mutations

James San Antonio¹, Anton Persikov², Antonella Forlino³, Joan Marini⁴, Peter Byers⁵, Anne De Paepe⁶, Francis Glorieux⁷, Allan Lund⁸, Gerard Pals⁹, Monica Mottes¹⁰, Osten Ljunggren¹¹, Anne-sophie Lebre¹², Federica Sgariglia¹³, and Olena Jacenko¹⁴

1. Global Quality and Operations, Stryker, Inc., 45 Great Valley Parkway, Malvern, PA 19355, USA, 484-323-8802; 610-640-1714; james.sanantonio@stryker.com
2. Lewis-Sigler Institute for Integrative Genomics, Princeton University, Carl Icahn Lab, Princeton, NJ 08544, USA, 609-258-7195; 609-258-8004; persikov@princeton.edu
3. Dept Molecular Medicine, University of Pavia, Via Taramelli 3/B, 27100 Pavia, Italy; phone number: +39-0382-987235; +39-0382-423108; aforlino@unipv.it
4. Bone & Extracellular Matrix Branch, NICHD, NIH, Bldg 10, Rm 10D39, 9000 Rockville Pike, Bethesda, MD 20892, 301-594-3418; 301-480-3188; oidoc@helix.nih.gov
5. Department of Pathology, University of Washington, D-518 Health Sci. Cntr, Box 357470, Seattle, WA, USA, 98195, 206-543-4206, 206-616-1899, pbyers@u.washington.edu
6. Center for Medical Genetics, Ghent University Hospital, De Pinteleaan 185, 9000 Ghent, Belgium, +32-9332-4979, +32-9332-4970, anne.depaepe@ugent.be
7. Shriner's Hospital for Children, 1529 Cedar Ave., Montreal, Quebec, Canada H3G 1A6 (514) 842-5964, 514-843-5581, glorieux@shriners.mcgill.ca
8. Dept. Clinical Genetics, Copenhagen University Hospital, Juliane Marie Centre 4062, Copenhagen, Denmark, +45 3545 3887 Fax: +45 3545 4072 e-mail: alund@rh.regionh.dk
9. Cntr Connective Tissue Res., Vrije University, P.O. Box 7057, 1007 MB, Amsterdam, Netherlands, +31-20 444 82 78 Fax: + 31 20 444 82 93, g.pals@vumc.nl
10. Dept. Life and Reproductive Sciences, University of Verona, strada le Grazie, 8 37134, Verona, Italy; 39-045-8027184, 39-045-8027180, monica.mottes@univr.it
11. Dept. Med. Sci., Uppsala Univ., P.O. Box 256, SE-751,05, Uppsala, Sweden, +46-186114906, +46-18553601, Osten.Ljunggren@medsci.uu.se
12. Dept. Genetics, Hospital for Sick Children, Batiment Lavoisier 3e etage, 149 rue de Sevres, 75015, Paris, Fr., +33-1444-95164, +33-1711-96420, anne-sophie.lebre@nck.aphp.fr
13. Dept. of Surgery, Children's Hospital of Philadelphia, Philadelphia, PA 19104, 3615 Civic Center Boulevard, ARC 902, Philadelphia, PA 19104, 267-425-2077, 267-426-2215, sgarigliaf@email.chop.edu
14. Department of Animal Biology, School of Veterinary Medicine, University of Pennsylvania, 3800 Spruce Street, Philadelphia, PA 19104, USA, 215-573-9447; 215-573-5188; jacenko@vet.upenn.edu

Abstract

Our molecular map of type I collagen was previously correlated with the Orgel et al., 2006 x-ray fibril diffraction model to identify cell and matrix interaction domains. Here we used two strategies to analyze mutation patterns to pinpoint functionally significant regions. First,

regions of the $\alpha 1(I)$ chains were identified having three or more consecutive glycines either associated with lethal or silent phenotypes. Many of these regions co-localized with sites for interactions with mineralization proteins such as phosphophoryn, cell surface receptors, and matrix metalloproteinases, or for intermolecular crosslinking. Five of the larger runs of silent glycines, although each on separate monomers in the D-period, clustered vertically within a narrow fibril region- herein called the major silent zone (MSZ). Second, the distribution of OI substitution mutations on the COL1A1 and COL1A2 genes were examined and found to be statistically different from that expected on the basis of base pair mutation rates, suggesting differential phenotypic consequences of mutations occurring on different collagen regions. For example, some glycines were predicted to have high mutation rates yet did not; notably, most localized within or near the MSZ or other runs of silent glycines. Together, these results pinpointed several regions of the collagen triple helix- most notably within the cell interaction domain, and a narrow cross-fibril zone just N-terminal to the major cell surface integrin binding site GFOGER- as being particularly sensitive to glycine mutations and likely having highly crucial biological functions. Thus for some collagen mutations, disease phenotypes may result, at least in part, from disruption of crucial protein functions such as mineralization or cell-fibril interactions.

Keywords: type I collagen, interactome, osteogenesis imperfecta, mutations, integrins, mineralization, protein structure, cell interactions

Introduction: The type I collagen interactome.

Simple maps of proteins, showing their dimensions, shapes, domain features, and positions of mutations have been commonplace in the scientific literature. One of the earliest maps of type I collagen included its primary protein sequence and presented a mechanism of charge-based monomer alignment consistent with the “quarter stagger” fibril structure (Chapman, 1974). Others showed the positions of chemically-reactive amino acids and GPP motifs on the type I collagen D-period (Dölz and Heidemann, 1986), or the sites of heparin binding to collagen fibrils, and their spatial relationships to several cell interaction sequences (San Antonio et al., 1994). In the late 1990’s, with reports of more than a hundred cell and ligand-binding sites and substitution mutations mapping to type I collagen, we began constructing a “road map” or “interactome” of the molecule (DiLullo et al., 2002; Sweeney et al., 2008). The unique collagen molecular structure allowed the construction of a map wherein triple helical domains are represented as 2D linear arrays of three polypeptide chains, which are also shown in the quarter stagger, D-period arrangement, i.e., as they are proposed to occur *in vivo*. The map was annotated with positions of functional domains including post-translational modifications, sites mediating cell or ligand binding, proteolysis, or amino acids associated with mutations. The intent of this initiative was to archive data on type I collagen, but unexpectedly, it has provided deeper insights into the structure-function relationships of this ubiquitous macromolecule. This manuscript summarizes recent updates to the type I collagen interactome and how it led to the creation of the domain model of fibril function, and reports new insights on the type I collagen structure-function relationship gleaned from examining the distributions of several classes of mutations relative to structural landmarks on the collagen fibril.

Materials and Methods.

Interactome construction. The interactome includes the primary sequence of human type I collagen $\alpha 1$ and $\alpha 2$ chains, arranged according to the Chapman D-period overlap model of the fibril (Chapman, 1974).

Ligand Binding Sites and Functional Domains. Positions of binding sites and functional domains were obtained from the literature and are indicated by labeled boxes placed next to relevant sequences. Primary literature references for sites indicated on the map, as well as a detailed legend appear elsewhere (Sweeney et al., 2008). Zones of interactions between ligands and broad regions of collagen fibrils, as observed by electron microscopy, are indicated by shaded overlays. Many binding site locations were approximated based on low resolution approaches, such as electron microscopy of ligand-collagen complexes. Other functional sites were mapped at high resolution by various research groups using collagen model triple helical peptides (THPs). *In vivo*, type I collagen-ligand interactions may depend upon the tissue source. Moreover, collagen fibrils may be heterotypic, i.e. contain other collagens such as types III and V, whose arrangements in the fibril are not understood, and that could affect fibril-ligand binding (Sweeney et al., 2008).

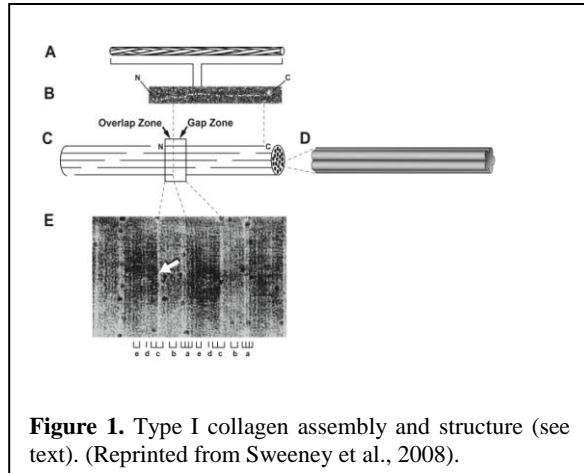
Identifying Interrelationships between Sites. Relationships between binding sites and functional domains were identified in three ways. First, as sequences on the collagen monomer to which two or more ligands have been shown to bind, or that are near neighbors; second, as sequences falling within the borders of a fibril region shown to bind a particular ligand; and third, as neighboring binding sites on adjacent monomers within the D-periodic fibril packing scheme. It was assumed that interactions between ligands on adjacent monomers may occur if their binding sites align vertically within the D-period, and if their molecular reach allows them to simultaneously bind more than one triple helix and contact another ligand or ligand-binding site on neighboring monomer(s). Sites on multiple monomers were considered to be vertically aligned if they overlap with or fall close to an axis drawn perpendicular to the long axes of the monomers, and joining monomers 1 and 5.

Human Mutations. OI-causing mutations leading to OI were obtained from the Database of Human Type I and Type III Collagen Mutations (www.le.ac.uk/genetics/collagen/) and the OI consortium mutation database (Marini et al., 2007 for review). Unpublished mutations were detected by DNA sequencing and were identified in patients referred to a clinical DNA diagnostics laboratory (Reading, PA) for mutation analysis of the COL1A1 and COL1A2 genes. Clinical diagnosis of OI was made by the referring medical personnel.

Results and Discussion.

Type I collagen structure. Upon procollagen secretion, N- and C-proteinases remove its globular ends, and every 67 nm along the fiber axis, five tropocollagen monomers (M) assemble in a quarter-staggered fashion to form the supramolecular helix, the microfibril (**Fig. 1C and D**). Each microfibril, the subunit of the fibril (**Fig. 1D**) and its neighbors are covalently joined by N- and C-terminal intermolecular cross-links to form fibrils (**Fig. 1E**). The basic repeating structure of the fibril is the D-period, which is 67 nm long and composed of an overlap and gap zone. Each D-period contains the complete monomer sequence derived from overlapping consecutive

elements of five monomers. The type I collagen interactome is an expanded, 2D view of the fibril D-period (**Fig. 2**). For a detailed review, see Piez and Reddi, 1984.



Functional domains. Functional sites recently plotted on the interactome (**Fig. 2**) include a thermally labile/ hydroxyproline-deficient domain (***thermally labile domain***, **Fig. 2**; label designations for functional sites are shown in bold and italicized in the text) near the C-terminus of the triple helix and in the fibril corresponding to the D–C2 bands region, discovered in studies examining the melting profiles of type I collagen (Miles and Bailey, 2001). A ***hydrophobic domain*** was also mapped at the fibril level based on the observation that fibrils fractured in the vicinity of the C bands in response to treatments with surfactants such as

triton-X-100 or non-polar liquids like cedar wood oil; this activity was attributed to the presence of an alanine-rich hydrophobic cross-fibril zone (Hu et al., 1997). Other new sites include two binding sites for pigment epithelium-derived growth factor (**PEDF**), which has anti-angiogenic neurotrophic, and neuroprotective activities (Sekiya et al., 2011). The confluence of three sites involved in angiogenesis regulation: peptide sequences with anti-angiogenic activities, and PEDF- and heparin/HSPG-binding sites, suggests a prominent role for the N- and C-termini of type I collagen in vascular development, as is also seen for other matrix macromolecules including type IV collagen (Parkin et al., 2011). Also plotted are sequences homologous to eleven collagen peptides recovered from the fossilized bones of the dinosaurs *Tyrannosaurus rex* and *Brachylophosaurus canadensis* (**blue or purple boxes**). These peptides co-localize to functionally significant collagen sequences including the integrin binding site and MMP-1 cleavage domain, occupy the more protected regions of the fibril, and exhibit a distinct chemical composition, suggesting mechanisms for their survival through deep geologic time (San Antonio, Schweitzer et al., 2011). Last, the N- and C-telopeptide intermolecular crosslinking sites are now indicated (rectangles labeled **X-link**), as well as the potential glycosylation status of their crosslink partners; hexagons at positions 87 and 930 represent hydroxylysine-O-linked galactose-glucose sugars, labeled *Gal-Glc* (Piez and Reddi, 1984).

Figure 2. Type I collagen interactome. Human collagen primary sequences were obtained from GenBank, accession #s: $\alpha 1(I)$, NP000079; $\alpha 2(I)$, NP000080 and aligned as described (Chapman, 1974). Ligand binding sites are indicated by rectangular boxes adjacent to relevant collagen sequences. Gray boxes denote ligand binding to the monomer. Non-shaded boxes denote ligand binding to one α chain. Major ligand binding regions (MLBR) are designated **1**, **2**, and **3**. Disease-associated mutations are indicated next to affected residues. Broad cross-fibril ligand binding regions are delineated by shaded overlays. See (Sweeney et al., 2008) and text for abbreviations of most mapped sites and associated literature citations. Note: the map may contain minor errors in functional domain and mutation positions as it is currently under construction prior to publication.



Interactome suggests a domain organization of the human collagen fibril.

The most critical functional sites of the fibril including GFPGER₅₀₂₋₅₀₇, the matrix metalloproteinase-1 (MMP-1) cleavage site, and other prominent elements of major ligand binding region (MLBR)2, localize to a small region of M3 and 4 of the fibril that we propose regulates dynamic aspects of collagen biology, thereby comprising the “Cell interaction domain” (**Figs. 2 & 4**)(Sweeney et al., 2008). The remainder of the fibril contains sequences mediating intermolecular cross-links, PG binding, and mineralization, and is proposed to assume structural duties, thereby comprising the “Matrix interaction domain” (**Figs. 2 & 4**). We speculate that similar domains are common among fibrillar collagens.

Translating the interactome to the living collagen fibril.

The interactome and domain model of fibril function suggests how cells may interact with collagen (Sweeney et al., 2008). Thus, the overlap zone of the fibril contains $\alpha 2\beta 1$ integrin binding sequences within each D-period as “landing zones” for cells. Based on the dimensions of integrins and collagen fibrils (Emsley et al., 2000; Piez and Reddi, 1984), the fibril is an optimal substrate for integrin receptor clustering, activation, and signaling. The proximity of the integrin binding site to the MMP-1 cleavage site and predominant ligand-binding regions MLBR1 and MLBR2 (≈ 10 nm apart) suggests that collagen assembly, function, and remodeling may be achieved in a cooperative fashion. The remainder of the fibril is densely decorated by PGs with anionic GAG chains; these are proposed to be constrained from sterically interfering with collagen-ligand interactions via binding to the electropositive charge density stripes of the fibril surface.

Human mutations.

Mutations mapped on the interactome are limited to missense mutations, although many stop and frame-shift mutations have been identified in type I collagen (Marini et al., 2007). The most prevalent mutations on the map are glycine mutations associated with brittle bone disease, or Osteogenesis imperfecta (OI). OI is typically classified into four major clinical types: type I (mild), II (lethal), III (severe), and IV (moderately severe)(Marini et al., 2007). Glycine residues are 100% conserved at every third position as a structural requirement in fibrillar collagens including $\alpha 1(I)$ and $\alpha 2(I)$ collagen chains. Therefore, none of the DNA missense mutations in the COL1A1 and COL1A2 genes affecting glycine residues in the corresponding $\alpha 1(I)$ and $\alpha 2(I)$ collagen chains should be considered as “neutral”. Indeed, it is expected that every such mutation should have a significant effect on collagen structure (as confirmed by several studies on model peptides) and lead to phenotypic changes. To date, hundreds of such mutations are known to lead to dramatic forms of OI, or other disorders with 716 mutations observed in the COL1A1 gene and 557 in COL1A2 (OI consortium database). It is thus assumed here that glycines observed to have no associated mutations may likely be “silent” or embryonically lethal when mutated; however, it is possible that mutation of such glycines could produce either no phenotype or such a mild phenotype that they are not reported.

The consequences of substitution mutations in type I collagen are generally thought to arise from their disruption of the folding or stability of the triple helix. The “gradient” model suggests that, because the helix folds from the C- to the N-terminus, more C-terminal mutations affect assembly and post-translational modifications more profoundly, resulting in a more severe

phenotype (Byers et al., 1991). However, while mutations in the amino end of the alpha 1 chain are nearly uniformly non-lethal, the phenotypic outcome for the rest of the chain is strongly dependent on the substituting residue, indicating the importance of $\alpha 1(I)$ for helical stability and potentially for interactions with other molecules (Marini et al., 2007). “Regional models” suggest the distribution of some mutations may correlate with their impact upon various functional landmarks on the protein (Marini et al., 1993; Scott and Tenni, 1997). Thus, on the $\alpha 2(I)$ chain clusters of lethal OI mutations are interspersed with non-lethals, with the former corresponding moderately well with PG-fibril interaction zones (Marini et al., 2007). The interactome also supports a regional model: 1) consecutive runs of glycines associated exclusively with lethal, non-lethal, or no mutations exist throughout the protein and in many cases are not distributed according to a gradient; 2) Non-OI, or “atypical” mutations do not exhibit a gradient of phenotype severity according to their N- to C- terminal position along the protein, but rather, cluster to distinct zones on the monomer and fibril (Marini et al., 2007; Sweeney et al., 2008).

Previously we reported that seven regions of the triple helix with five or more consecutive glycines silent for mutations on both alpha chains existed on the fibril and most localized to the cell interaction domain. However, with the recent addition of several hundred new mutations, most of these silent zones no longer exist. Thus, here we have attempted a new set of complementary strategies to exploit mutation distribution and frequency to pinpoint functionally significant regions of collagen based on the presence of: 1) high concentrations of glycines associated exclusively with lethal mutations and/or silent mutations on the $\alpha 1$ chain; and 2) glycines on the $\alpha 1$ and $\alpha 2$ chains that should theoretically be associated with a high frequency of mutations but in reality are not.

Mapping consecutive runs of lethal and silent glycines. Here our analysis has been limited to mutations on the $\alpha 1$ chain, although a similar analysis is planned for the $\alpha 2$ chain. Visual inspection of the collagen map revealed eight regions of the triple helix containing 3 or more consecutive runs of glycines (i.e. 3 or more gly-X-Y amino acid triplets) associated exclusively with lethal mutations and ranging in size from 3 to 8 triplets, with the average being 4.1. These included positions 250-259; 388-394; 550-556; 697-718; 766-772; 910-916; 943-949; and 1000-1006. There were thirteen regions of the triple helix having 3 or more consecutive runs of glycines silent for mutations and ranging in size from 3 to 6 triplets, with the average being 4. These included positions 1-7; 49-58; 262-274; 358-367; 493-502; 616-628; 721-736; 775-781; 808-820; 853-859; 889-895; 934-940; and 952-961. For sites containing consecutive runs of lethal and silent glycines, sometimes the silent or the lethal glycines form clusters, and sometimes they alternate positions or appear randomly distributed. By far the largest concentration of exclusively lethal and

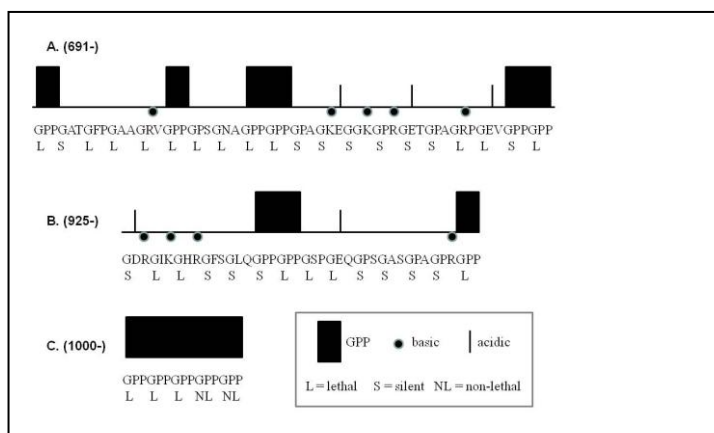


Figure 3. Three $\alpha 1$ chain sequences including consecutive runs of lethal and/or silent glycine residues, shown against a schematic representation of the approximate positions of their GPP triplet and charged residue components drawn according to the convention of Dölz and Heidemann, 1986. *This and other diagrams in Figs. 4 & 5 are not to scale and in the native collagen fibrils, monomers are closely associated in rope-like bundles called microfibrils.*

silent mutations is on M4, beginning at residue 691 and including 13 lethal and 11 silent glycines. Its N-terminal region contains mostly lethal glycines, which are followed by a cluster of silent glycines, and then by a short sequence containing alternating lethal and silent residues- it is partially shown in **Fig. 3A**. This sequence block coincides almost exactly with the reported binding sites for phosphophoryn and cartilage oligomeric matrix protein (COMP). Phosphophoryn is a highly acidic collagen binding protein proposed to nucleate hydroxyapatite formation in dentin, i.e., to mediate tooth mineralization (George and Hao, 2005). COMP is a calcium binding protein abundant in tendons, ligaments, and in the pericellular matrix of chondrocytes; mutations in COMP result in chondrodysplasias (Briggs et al., 1995). It is tempting to speculate that the lethal/silent zone may represent the protein region that regulates the binding of factors necessary for the initiation of bone mineralization. The silent sequence of M4 also includes practically all of the glycines N-terminal to the MMP-1 cleavage site located on the same monomer. Possibly the sequences defined by the lethal/silent zones play a role in accessibility or function of MMP-1, with defects leading to abnormal collagen processing. The M4 sequences may also play a role in fibronectin or secreted protein acidic and rich in cysteine (SPARC) binding, as well as possibly affecting a host of neighboring sites on M3 and M5, including those for integrin binding and C-terminal intermolecular crosslink formation, among others.

The second greatest confluence of lethal and silent mutations begins on M5 at residue 925, includes 6 lethal and 8 silent glycines and is a near neighbor to the M4 sequence just discussed (**Fig. 3B**). This sequence is adjacent to the C-terminal cross-link site, and overlaps with those proposed for the binding of decorin proteoglycan core protein and PEDF. The third largest run of silent/lethal glycines has 5 lethal and 6 silent glycines, begins at residue 316 and maps to M2 to a region distant from the previously discussed sequences. Although there are no major functional domains proposed for this region of M2, notably, across the fibril it is adjacent to the N-terminal crosslink on M1. The fourth largest consecutive run of silent/lethal glycines also resides on M2 beginning at residue 418 and including 5 each of lethal and silent glycines. This sequence roughly co-localizes with a narrow fibril zone proposed crucial for the binding of SPARC, the discoidin domain receptor 2 (DDR2 receptor), and von Willebrand's factor, ligands with proposed functions in collagen-based cell signaling and hemostasis (Farndale et al., 2008). Other shorter runs of lethal/silents will not be discussed here.

We next plotted the positions of just the mutation silent zones on the D-period to see if they occupy interesting patterns relative to structural/functional landmarks on the

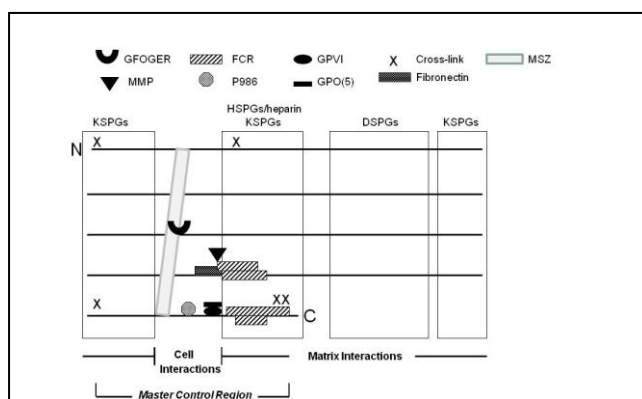


Figure 4. Two-dimensional schematic of the collagen D-period was constructed according to the Chapman overlap model, indicating the positions of the putative cell and matrix interaction domains, and the “master control region” (San Antonio, Parkin et al., 2012) including its constellation of functional sequences. Monomers are shown as unlabeled lines, with M1 through M5 appearing in order from top to bottom. Master control region components include GPO(5), intermolecular cross-links, fibrillogenesis control sequences (FCS), P986, fibronectin, the MMP, 2, 13 cleavage domain, and the GFOGER sequence that mediates $\alpha 1\beta 1/\alpha 2\beta 1/\alpha 11\beta 1$ integrin binding. The approximate position of five large contiguous stretches of glycines with silent phenotypes- the MSZ is indicated by grey diagonal rectangle (see text).

fibril. Notably, the silent components of the two largest lethal/silent zones on M4 and M5 were found to closely align with each other and with three other silent domains across the fibril on M1, M2 and M3, forming a narrow stripe roughly perpendicular to the long axis of the fibril- herein called the major silent zone (MSZ)(**Fig. 4**). It is tempting to speculate that the region of the protein defined by the MSZ- comprising far less than 5% of fibril structure- may define a structurally and/or functionally crucial fibril region. For example, the MSZ sequences could represent a relatively weak fibril region that is more significantly affected by mutations- indeed it is located roughly between the fibril's crosslink sites and other structural domains such as glycine-proline-hydroxyproline (GPP repeats; discussed below). More provocative is the possibility that these sequences mediate the interaction of calcium-binding proteins required for bone mineralization, or function directly as hydroxyapatite nucleation centers. On the other hand, that the MSZ falls within the fibril's cell interaction domain, and directly aligns with and/or falls just to the N-terminal site of the major integrin binding site could suggest its role in supporting cell-collagen interactions. Other possibilities include, for example, that the silent sequences all influence the availability or function of MMP-1, and thus fibril remodeling. Ongoing work is attempting to shed light on these possibilities by visualizing the position of the MSZ sequences in the 3D type I collagen microfibril model, relative to known dimensions of prominent collagen-binding ligands.

On the fibril, the position of the MSZ corresponds with the B2 to C1 bands cross-fibril region and neighbors/overlaps with the alanine rich/hydrophobic zone. Moreover, only the M5 sequence overlaps with the thermally labile domain, albeit partially. The MSZ occupies the center of the overlap zone that also contains the greatest concentration of single and/or tandem GPP triplets on the fibril; these sequences are thought to confer stability to the triple helix and/or function as nucleation domains for the folding of the molecule. Thus, mutation patterns within the two largest sequences of lethal/silent runs of glycines were examined relative to their component GPP sequences, as well as that associated with the largest tandem GPP repeat (**Fig. 3**). For the largest run of lethal/silent mutations, the contiguous stretch of lethal glycines was largely associated with the GPP triplets, whereas that of the silent glycines localized mostly to the intervening, charged sequences (**Fig. 3A**). For the second largest run of lethal/silent mutations, the largest contiguous stretch of silent glycines also localized with a charged sequence between GPP repeats (**Fig. 3B**). Notably, disease phenotypes were not uniform for all glycines in tandem GPP repeats (**Fig. 3A-C**). For example, in GPP₅ located near the C-terminus of the triple helix, the first three glycines were associated with lethal phenotypes, the fourth with lethal/non-lethals, and the last with a non-lethal (**Fig. 3C**). This phenomenon was seen with five of the eight GPP tandem repeats in the fibril (not shown).

Elsewhere in the D-period, towards the right side of the gap zone, two other clusters of 5 and 3 silent glycines align closely with each other across the fibril, localize near GPP-containing sequences, and occupy zones of the triple helix containing charge residue clusters (not shown). This sequence doublet localizes nearly precisely to several proposed sites for decorin core protein binding on M3 and M4, thought crucial for fibril assembly and integrity. Several other small zones silent for mutations localize to yet other fibril areas (not shown).

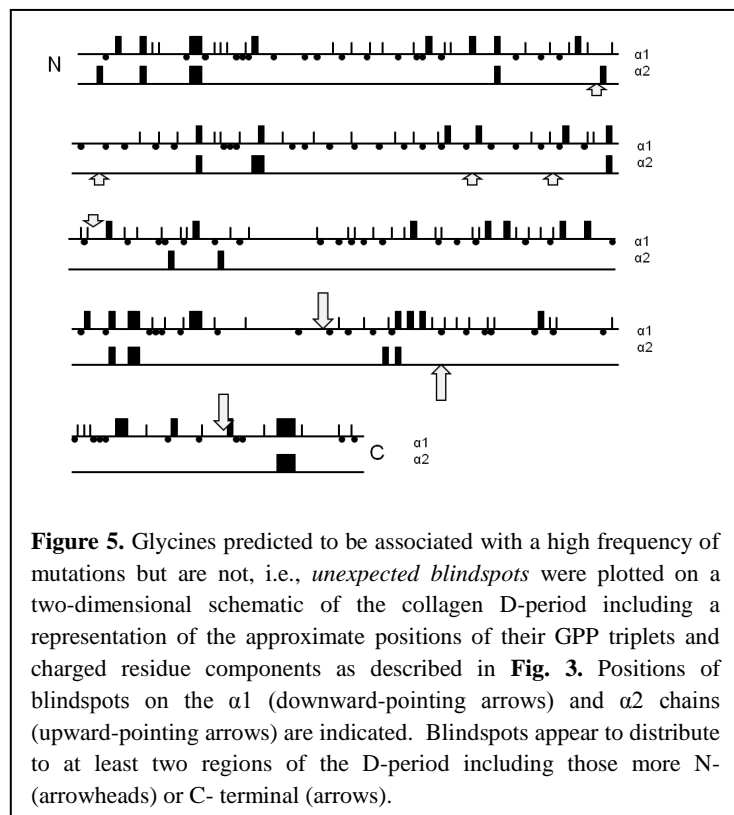


Figure 5. Glycines predicted to be associated with a high frequency of mutations but are not, i.e., *unexpected blindspots* were plotted on a two-dimensional schematic of the collagen D-period including a representation of the approximate positions of their GPP triplets and charged residue components as described in Fig. 3. Positions of blindspots on the $\alpha 1$ (downward-pointing arrows) and $\alpha 2$ chains (upward-pointing arrows) are indicated. Blindspots appear to distribute to at least two regions of the D-period including those more N- (arrowheads) or C- terminal (arrows).

Mapping glycines with unexpected frequencies of associated mutations.

Despite the increasing number of mutations in OI databases, not every position of the 338 glycines of the triple helix has yet to be associated with an OI-causing missense mutation. Therefore, it is not possible to statistically validate the existence of silent glycines. However, some of the glycine positions are subject to an increased mutation probability. Thus the context of the DNA sequence influences the mutation rate (Coulandre et al., 1978), with all those regions containing pairs of the bases C and G (CpG) being subject to elevated mutation rates. For example, CpG dinucleotides are known to be mutational hotspots in mammals, which was explained by spontaneous

oxidative deamidation of methylated cytosines (Coulandre et al., 1978; Nachman and Crowell, 2000). The rate of transition at CpG sites was estimated to be 18x higher than that at non-CpG sites (Kong et al., 2012). An analysis of type I collagen genes shows that a limited number of glycine positions could potentially be mutation “hotspots” due to the CpG effect. More precisely, the most probable G>A transitions on the CpG site could hit the first nucleotide position of the GGN Glycine codon when it is preceded by cytosine (from the previous codon, or from the intron). Practically, such mutations may result in two types of missense mutations with high probability: Gly>Ser, or Gly>Arg.

In an ongoing study (Marini et al., *In Preparation*), all glycine positions were divided into several categories; e.g.: *predicted hotspots* expected to display high numbers of mutations and did; *unexpected blindspots* predicted to display high numbers of mutations but failed to show any; *partial blindspots* predicted to display high numbers of mutations but only associated with ≤ 5 mutations; and *unexpected hotspots* not predicted to display many mutations but did. Only one aspect of these results will be addressed here. Namely, when the locations of *unexpected blindspots* were examined on the D-period model, they clustered to at least two fibril regions

(Fig. 5). The first includes residues beginning at the C-terminus of M1 and ending at the N-terminus of M3, or to approximately twenty five percent of the collagen molecule. On the D-period these sites occupy the gap zone and gap/overlap border, approximately within the E1-C2 bands region. The second cluster of residues maps more C-terminally to M4 and M5, approximately within the B1-A3 bands. Examining the distribution of these residues within the D-period suggests they localize adjacent to GPP repeats, although further analysis is needed to confirm this relationship (Fig. 5). Moreover, several confluences of both *blindspots* and *partial blindspots* localized on or near to some of the MSZ sequences or to other contiguous runs of silent and/or lethal glycine residues identified earlier in this manuscript. These findings further emphasize the crucial roles played by select regions of the collagen fibril.

Conclusion.

Our analysis of mutation distribution on the type I collagen interactome used two different approaches to yield largely complementary findings, pinpointing several regions of the collagen triple helix- namely, segments of M3 and M4 overlapping with the cell interaction domain and MLBRs, and a narrow cross-fibril zone just N-terminal to the major integrin binding site GFOGER- as being particularly sensitive to glycine mutations and likely having crucial biological functions. Further research examining how mutations in those regions might affect collagen conformation and interactions with select ligands may reveal further fundamental insights into the biology of type I collagen.

References

- Chapman, J.A., 1974, The staining pattern of collagen fibrils. I. An analysis of electron micrographs, *Connect. Tiss. Res.*, 2, 137-150p.
- Coulondre, C., Miller, J.H., Farabaugh, P.J., Gilbert, W., 1978, Molecular basis of base substitution hotspots in *Escherichia coli*, *Nature*, 274, 775-780p.
- DiLullo, G., Sweeney, S.M., Koriko, J., Ala-Kokko, L., and San Antonio, J.D., 2002, Mapping the ligand-binding sites and disease associated mutations on the most abundant protein in the human- type I collagen, *J. Biol. Chem.*, 277, 4223-4231p.
- Dölz, R. and Heidemann, E., 1986, Influence of different tripeptides on the stability of the collagen triple helix. I. Analysis of the collagen sequence and identification of typical tripeptides, *Biopolymers*, 25, 1069–1080p.
- Emsley, J., Knight, C.G., Farndale, R.W., Barnes, M.J., and Liddington, R.C., 2000, Structural basis of collagen recognition by integrin $\alpha 2\beta 1$, *Cell*, 101, 47-56p.
- Farndale, R.W., Lisman, T., Bihan, D., Hamaia, S., Smerling, C.S., Pugh, N., Konitsiotis, A., Leitingner, B., de Groot, P.G., Jarvis, G.E., and Raynal, N., 2008, Cell-collagen interactions: the use of peptide Toolkits to Investigate collagen-receptor interactions, *Biochem. Soc. Trans.*, 36, 241-250p.
- Hao, G.A., and George, A., 2005, Role of phosphophoryn in dentin mineralization, *Cells Tissues Organs*, 181, 232-240p.
- Hu, X.W., Knight, D.P., and Chapman, J.A., 1997, The effect of non-polar liquids and non-ionic detergents on the ultrastructure and assembly of rat tail tendon collagen fibrils in vitro, *Biochim. Biophys. Acta*, 1334, 327-337p.

Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A.Wong, W.S., Sigurdsson, G., Walters, G.B., Steinberg, S., Helagson, H., Thorleifsson, G., Gubdjartsson, D.F., Helgason, A., Magnusson, O.T., Thorsteinsdottir, U., Stefansson, K., 2012, Rate of de novo mutations and the importance of father's age to disease risk, *Nature*, 488, 471-475p.

Marini, J. C., Lewis, M.B., and Chen, K., 1993, 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*, 268, 2667-2673p.

Marini, J.C., Forlino, A., Cabral, W.A., Barnes, A.M. San Antonio, J.D., Milgrom, S., Hyland, J., Korkko, J., Prockop, D., de Paepe, A., Coecke, P., Glorieux, F.H., Roughly, P., Lund, A., Killuria, K., Cohn, D., Krakow, D., Mottes, M., Troendle, J., Dalgleish, R., and Byers, P.H., 2007, Consortium for Osteogenesis Imperfecta Mutations: Database of Glycine Substitutions and Exon Skipping Defects: Lethal Regions in the Helical Portion of Type I Collagen Chains Align with Collagen Binding Sites for Integrins and Proteoglycans, *Human Mutation*, 28, 209-221p.

Miles, C.A., and Bailey, A.J., 2001, Thermally labile domains in the collagen molecule, *Micron*, 32, 325-332p.

Nachman, M.W. and Crowell, S.L., 2000, Estimate of the mutation rate per nucleotide in humans, *Genetics*, 156, 297-304p.

Parkin, J.D., San Antonio, J.D., Pedchenko, V., Hudson, B., Jensen, S.T., and Savage, J., 2011, Mapping structural landmarks, ligand binding sites, and missense mutations to the collagen IV heterotrimers predicts major functional domains, novel interactions, and variations in phenotypes in inherited diseases affecting basement membranes, *Human Mutation*, 32, 127-143p.

Piez, K. A. and A. H. Reddi, 1984, *Extracellular Matrix Biochemistry*, New York, Elsevier.

Orgel J.P., Irving, T.C., Miller, A., and Wess, T.J., 2006, Microfibrillar structure of type I collagen in situ", *PNAS*, 103, 9001-9005p.

San Antonio, J. D., Lander, A.D., Karnovsky, M.J., and Slayter, H.S., 1994, Mapping the heparin-binding sites on type I collagen monomers and fibrils, *J. Cell Biol.*, 125, 1179-1188p.

San Antonio, J.D., Schweitzer, M.H., Jensen, S.T., Kalluri, R., Buckley, M., and Orgel, J.P.R.O., 2011, Dinosaur peptides suggest mechanisms of protein survival, *PLoS One*, 6, e20381.

San Antonio, J.D., Parkin, J.D., Savage, J., Orgel, J.P.R.O., and Jacenko, O., 2012, Collagen interactomes: mapping functional domains and mutations on fibrillar and network-forming collagens, in, *Extracellular Matrix: Pathobiology and Signaling*, N Karamanos, editor, 575-591p.

Scott, J. E. and R. Tenni, R., 1997, Osteogenesis imperfecta mutations may probe vital functional domains (e.g. proteoglycan binding sites) of type I collagen fibrils, *Cell Biochem. Funct.*, 15, 283-286p.

Sekiya, A., Okano-Kosugi, H., Yamazaki, C.M., and Koide, T., 2011, Pigment epithelium-derived factor (PEDF) shares binding sites in collagen with heparin/heparin sulfate proteoglycans. *J. Biol. Chem.*, 286, 26364-26374p.

Sweeney, S.M., Orgel, J.P., Fertala, A., McAuliffe, J.P., Turner, K.R., DiLullo, G.A., Chen, S., Antipova, O., Perumal, A., Ala-Kokko, L., Forlino, A., Cabral, W.A., Barnes, A.M., Marini, J.C., and San Antonio, J.D., 2008, Candidate Cell and Matrix Interaction Domains on the Collagen Fibril, the Predominant Protein of Vertebrates, *J. Biol. Chem.*, 283, 21187–21197p.