Biophysical Journal
Article

Biophysical Society

Sequence-dependent mechanics of collagen reflect its structural and functional organization

Alaa Al-Shaer,[1] Aaron Lyons,[2] Yoshihiro Ishikawa,[5] Billy G. Hudson,[6,7,8,9,10,11,12] Sergei P. Boudko,[6,7,8] and Nancy R. Forde[1,2,3,4,*]

[1]Department of Molecular Biology and Biochemistry, [2]Department of Physics, [3]Department of Chemistry, and [4]Centre for Cell Biology, Development and Disease, Simon Fraser University, Burnaby, British Columbia, Canada; [5]Department of Ophthalmology, University of California San Francisco, School of Medicine, San Francisco, California; and [6]Department of Medicine, Division of Nephrology and Hypertension and [7]Vanderbilt Center for Matrix Biology, Vanderbilt University Medical Center, Nashville, Tennessee; and [8]Department of Biochemistry, [9]Department of Pathology, Microbiology, and Immunology, [10]Department of Cell and Developmental Biology, [11]Vanderbilt-Ingram Cancer Center, and [12]Vanderbilt Institute of Chemical Biology, Vanderbilt University, Nashville, Tennessee

ABSTRACT Extracellular matrix mechanics influence diverse cellular functions, yet surprisingly little is known about the mechanical properties of their constituent collagen proteins. In particular, network-forming collagen IV, an integral component of basement membranes, has been far less studied than fibril-forming collagens. A key feature of collagen IV is the presence of interruptions in the triple-helix-defining (Gly-X-Y) sequence along its collagenous domain. Here, we used atomic force microscopy to determine the impact of sequence heterogeneity on the local flexibility of collagen IV and of the fibril-forming collagen III. Our extracted flexibility profile of collagen IV reveals that it possesses highly heterogeneous mechanics, ranging from semi-flexible regions as found for fibril-forming collagens to a lengthy region of high flexibility toward its N-terminus. A simple model in which flexibility is dictated only by the presence of interruptions fit the extracted profile reasonably well, providing insight into the alignment of chains and demonstrating that interruptions, particularly when coinciding in multiple chains, significantly enhance local flexibility. To a lesser extent, sequence variations within the triple helix lead to variable flexibility, as seen along the continuously triple-helical collagen III. We found this fibril-forming collagen to possess a high-flexibility region around its matrix-metalloprotease binding site, suggesting a unique mechanical fingerprint of this region that is key for matrix remodeling. Surprisingly, proline content did not correlate with local flexibility in either collagen type. We also found that physiologically relevant changes in pH and chloride concentration did not alter the flexibility of collagen IV, indicating such environmental changes are unlikely to control its compaction during secretion. Although extracellular chloride ions play a role in triggering collagen IV network formation, they do not appear to modulate the structure of its collagenous domain.

SIGNIFICANCE Collagens are the predominant proteins in vertebrates, forming diverse hierarchical structures to support cells and form connective tissues. Despite their mechanical importance, surprisingly little is established about the molecular encoding of mechanics. Here, we image single collagen proteins and find that they exhibit variable flexibility along their backbones. By comparing collagens with continuous and discontinuous triple-helix-forming sequences, we find that the type of helix interruption correlates with local flexibility, providing the first steps toward a much-needed map between sequence, structure, and mechanics in these large proteins. Our results inform our understanding of collagen's ability to adopt compact conformations during cellular secretion and suggest a physical mechanism by which higher-order structure may be regulated by the distinct molecular properties of different collagens.

INTRODUCTION

Collagen is the most abundant protein in the animal kingdom and represents one third of the total protein in the human body (1,2). It is a major structural component of the extracellular matrix, contributing to the mechanical stability, organization, and shape of a wide variety of tissues. Twenty-eight distinct collagen types have been reported in humans, with different higher-order organizational structures (1–3). The most prevalent are fibril-forming collagens; these have a unique hierarchical structure whereby collagen molecules assemble in a parallel, staggered fashion into long fibrillar nanostructures that form higher-order structures (Fig. 1). These fibers are the predominant load- and

Submitted December 18, 2020, and accepted for publication August 6, 2021.

*Correspondence: nforde@sfu.ca

Editor: Markus Buehler.

https://doi.org/10.1016/j.bpj.2021.08.013

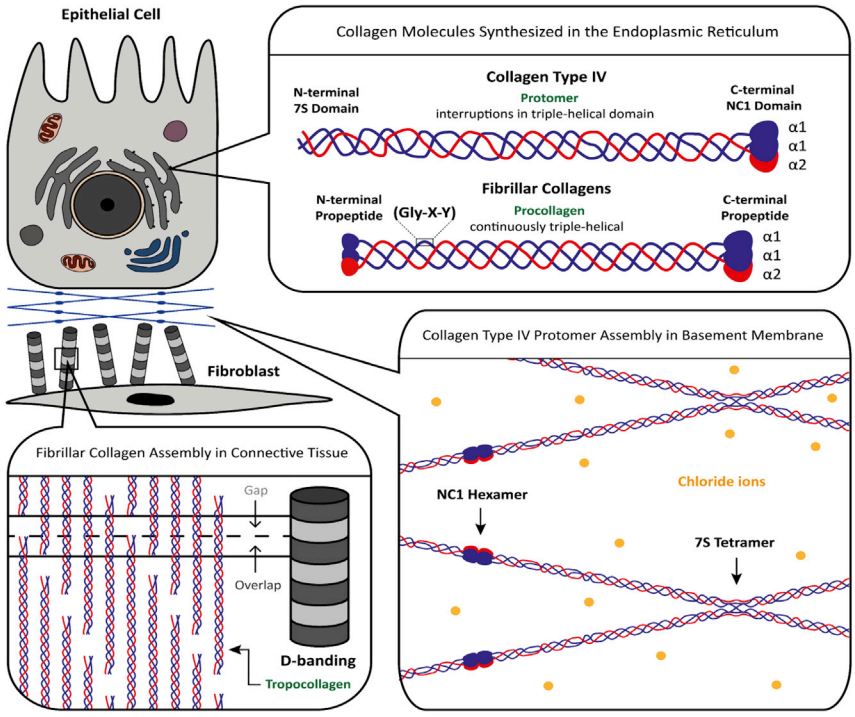

FIGURE 1 Localization and supramolecular structures of fibril-forming and network-forming collagens. All collagen chains are synthesized in the ER. The folded proteins are secreted from the cell, where they can form a variety of supramolecular structures in different tissues. The fibril-forming collagen molecule, procollagen, is post-translationally processed, whereby its propeptides are cleaved off to yield tropocollagen. Tropocollagens align laterally with an offset that gives rise to the characteristic D-banding pattern of fibrils. On the other hand, the globular domain of the network-forming collagen IVis not post-translationally cleaved. Its end domains serve an important role in network formation. The N-terminal end aligns four molecules and forms a 7S tetramer. On the other end, two molecules form head-to-head assemblies forming an NC1 hexamer. Recent studies show that chloride ions are essential in forming an organized network. Lateral interactions between collagens may also contribute to network assembly (not indicated in this schematic). Schematics are not to scale. To see this figure in color, go online.

tension-bearing structures in connective tissues. Network-forming collagens such as collagen IV associate end on to form sheet-like networks rather than fibrils (Fig. 1) (4,5). These collagen IV networks played an integral role in the evolution of multicellular life (3,6). Collagen IV networks are predominantly found in basement membranes (BMs), where they provide mechanical support and anchorage for cells and tissues and serve as a filtration barrier to macromolecules in organs such as the kidney (5,7). The mechanics of fibril-forming collagens have been the subject of numerous studies at the fibrillar and molecular levels (8–10 and many references therein). By contrast, the mechanics of network-forming collagens are far less studied (11). Similarly, knowledge of the structure of collagen IV lags significantly behind fibril-forming collagens at both the supramolecular and molecular scales.

All collagens share a characteristic triple-helical structure, formed by a $(\text{Gly-X-Y})_n$ repeating sequence in each of the three composite α-chains. However, two key molecular features distinguish collagen IV and fibril-forming collagens that have been shown to impact higher-order assembly. One is the globular C-terminal domain, whose presence inhibits fibril formation and is proteolytically removed from fibril-forming collagens before assembly (12). By contrast, the C-terminal noncollagenous domain of collagen IV (NC1) is not removed and plays a central, chloride-directed role in forming networks (13,14). The second distinguishing feature of collagen IV is the presence of natural discontinuities (interruptions) in the $(\text{Gly-X-Y})_n$ repeating unit within its collagenous domain. The absence or replacement of glycine every third residue prevents continuously triple-helical collagenous domains and promotes local destabilization (15,16). Interruptions are also present in other nonfibrillar collagens such as the FACITs, MACITs, and MULTIPLEXINs (1). Enhanced flexibility attributed to the presence of interruptions in collagens has been seen using rotary shadowing electron microscopy (EM) (17–20); however, a detailed study on characterizing the functionality of distinct interruptions is lacking. Because triple-helix interruptions can result from disease-associated mutations in fibril-forming collagens and because for collagen IV, they may play a biological role in providing recognition sites for other macromolecular components in BMs (21–23), the ability to interrogate the physical properties of collagen in a sequence-specific manner is needed.

The size of the most abundant mammalian collagen proteins ($\gtrsim$300 kDa; $\gtrsim$300 nm length) has imposed practical limitations on the structural insight available from conventional approaches. Diffraction-based analysis of molecular structure has been limited to studies of periodic structure within collagen fibrils (24,25) and to investigations of short triple-helical peptides (~30 amino acids or 10 nm in length) that are amenable to crystallization (23,26). These and other approaches have provided insight into the molecular determinants of triple-helix structure and stability, suggesting, for example, that imino acids (proline and hydroxyproline) are important for structural stability and that local interruptions in the $(\text{Gly-X-Y})_n$ repeating sequence associated with disease in fibril-forming collagens can disrupt the local structure (16,23). NMR studies on peptides have also found

interruptions to result in increased local chain dynamics (23). Because collagen IV exists as heterotrimers, an additional structural question that arises—beyond whether or not a local sequence possesses a triple-helical structure—is the relative arrangement or stagger among its composite chains (27). Alternative methods for investigating the sequence dependence of collagen structure in the context of the full-length protein are needed.

Here, we use atomic force microscopy (AFM) imaging of individual full-length collagen proteins to provide insight into the local determinants of structure. Image analysis provides a map of local flexibility as a function of position along the collagen chain. By applying this technique to collagen IV and to collagen III, we contrast the mechanics of network-forming and fibril-forming collagens at the molecular scale. By mapping the position-dependent flexibility, we relate the local sequence within full-length collagen proteins to the local bending mechanics. Because of the role of chemical environment in regulating collagen supramolecular assembly, we apply this AFM-based flexibility mapping to investigate the effects of distinct chemical environments on flexibility of collagens, globally and in a sequence-specific manner. We interpret these results in light of outstanding physiological questions regarding control of collagen conformations during secretion and its supramolecular assembly in the extracellular space.

MATERIALS AND METHODS

Collagen sources

Heterotrimeric [α1(IV)]$_2$-α2(IV)] collagen IV (113 μg/mL in 0.5 M acetic acid) was purified from Engelbreth-Holm-Swarn (EHS) tumor in lathrytic mouse and was a gift of Albert Ries (28,29). Bovine pro-N collagen III (pN-III) was extracted from fetal bovine skin following a previously published protocol (30). Reducing SDS-PAGE analysis showed both samples to be predominantly free of intramolecular cross-links. Rat tail tendon-derived collagen I was purchased from Trevigen (Cultrex 3440-100-01; Gaithersburg, MD) and is pepsin-treated collagen with a stock concentration of 5 mg/mL in 20 mM acetic acid.

Sample preparation for AFM

Collagen flexibility was mapped after deposition from three different solution conditions: 1 mM HCl + 100 mM KCl (pH ~3) (10), sodium acetate buffer (150 mM NaOAc, 25 mM Tris-OAc (pH 5.5)), and Tris-buffered saline (TBS) (150 mM NaCl, 25 mM Tris-Cl (pH 7.4)); the pH of these solutions were confirmed to remain stable over months to years. Collagen was diluted into the desired solution conditions at a final concentration of 0.2 μg/mL, for which 50 μL of the diluted sample was deposited and allowed to incubate for 20 s on freshly cleaved mica (Highest Grade V1 AFM Mica Discs, 10 mm; Ted Pella, Redding, CA). The excess unbound proteins were removed by rinsing with ultrapure water, and the mica was then dried using filtered air. All proteins were imaged under dry conditions, and the solution conditions of the samples refer to the conditions in which they were deposited onto mica. Previous work suggests that collagen conformations on the mica surface reflect their deposition conditions (10). Imaging was done with an Asylum Research MFP-3D atomic force microscope (Goleta, CA) using tapping mode in air. AFM tips with a 160-kHz resonance frequency and 5 N/m force constant (MikroMasch, HQ: NSC14/AL BS; Sofia, Bulgaria) were used.

Chain tracing and analysis

SmarTrace, a custom-built MATLAB code (The MathWorks, Natick, MA), was used to determine the bending flexibility of collagen chains (10). Persistence length determination by SmarTrace has been extensively validated (10,31). Chains were traced to subpixel resolution starting from the C-terminus (NC1-chain boundary for collagen IV; nonglobular end for collagen III). Homogeneous chain analysis assumes chains to have uniform flexibility (persistence length). The persistence length was found from the dependence of mean-square end-to-end distance, $\langle R^2(\Delta s)\rangle$ (Eq. 1), and tangent vector correlation, $\langle \cos\theta(\Delta s)\rangle$ (Eq. 2), on the segment length Δs. This model treats the chains as worm-like chains (WLCs) equilibrated in two dimensions and has previously been shown to describe fibril-forming collagens well in 100 mM KCl, 1 mM HCl (10).

To investigate the sequence dependence of collagen's flexibility, the local effective persistence length was determined using Eq. 4 (described in more detail in Supporting materials and methods, Text S1; Fig. S1). 95% confidence intervals for the estimates of $p(s)$ were determined using the cumulative probability function of the scaled inverse χ^2 distribution (Fig. S2), as described in Supporting materials and methods, Text S1. The dependence of this estimate on the number of chains is provided in Fig. S3. Analysis of simulated chain images showed that a window of length $\Delta s = 30$ nm was the minimal Δs that reliably produced the expected angular distributions for a wide range of persistence lengths tested (Fig. S4). Thus, for a chain of contour length L, persistence lengths could be extracted for segments centered from $s = 15$ nm through $s = L - 15$ nm.

Simulated images of chains with inhomogeneous flexibility were generated in MATLAB (32) and analyzed as described for the experimental images. Chains were generated as described in (10), but with a bending stiffness that varies along the chain and incorporating a "knob" at one end as a directional reference point. Specifically, the chains we simulated here had a total contour length of $L = 400$ nm, interspersing long, stiff regions with short flexible regions (Fig. 3 *B*): $p = 85$ nm ($\Delta s = 79$ nm), $p = 5$ nm ($\Delta s = 1$ nm), $p = 85$ nm ($\Delta s = 78$ nm), $p = 5$ nm ($\Delta s = 2$ nm), $p = 85$ nm ($\Delta s = 77$ nm), $p = 5$ nm ($\Delta s = 3$ nm), $p = 85$ nm ($\Delta s = 76$ nm), $p = 5$ nm ($\Delta s = 4$ nm), and $p = 85$ nm ($\Delta s = 80$ nm). The knob was introduced as a uniform-intensity disk of radius 7 nm, centered at the starting point of the simulated chain. Experimentally realistic background noise was included in the images, as described previously (10).

Sequence alignment

The α1 (P02463) and α2 (P08122) amino acid sequences were obtained from UniProt (33). All chain alignments were initiated at the edge of NC1 domain, starting at the first Gly-X-Y unit of collagenous domain of both α chains. Specifics of the alignments are provided in a Supporting material. Graphical alignment of the amino acid representation and persistence length profile assumes a length of 0.29 nm/aa in the collagenous domain.

Variable flexibility model

Each amino acid of an α-chain was assigned a value of "0" if it was within a (Gly-X-Y) sequence or "1" otherwise (within an interruption). Distinct alignments of the three α-chain sequences were tested (supplied as a Supporting material), including linear alignment, loop(s) within α2 to better align its (Gly-X-Y) sequence blocks with α1 (including a disulfide-bridged loop (34)), and sequence-similarity alignments (based on a Clustal alignment of the sequences (35), adapted to ensure chain continuity). A flexibility class was assigned at each amino acid step along the aligned chains based on whether it was a 0-, 1-, 2-, or 3-chain interruption in the triple helix

(Table 1). Each of these classes was assigned a local bending stiffness (via a local persistence length), and thus, the model assigned a local flexibility to each position along the contour. This model was then fitted to the determined effective persistence length profile $p^*(s)$ to determine the persistence lengths p_0, p_1, p_2, and p_3 that best describe each flexibility class. Conversion factors were included in the fitting routine, to convert between primary (aligned) sequence, measured in amino acids, and position along the contour from image analysis, measured in nanometers. Because of variability in identifying the starting position of the chains (found also when tracing simulated chains), a chain-stagger parameter was included, which represented the standard deviation of model chain starting positions. Finally, because we do not know where the collagenous domains start (for example, their starting position may be obscured by the NC1 domain), an offset parameter was included to linearly shift the modeled chains with respect to the traced chains. The fitting procedure is described in detail in the Supporting materials and methods, Text S2.

Modeling the simulated chains (Fig. 3) required a simpler physical model containing only two classes: rigid p_r and flexible p_f, located along the chains as defined by the inputs to the simulations. No length conversion factors were necessary because the simulated chain profile positions were defined in nanometers. Chain-stagger and offset parameters were necessary to align the model results with the traced and analyzed simulated chain $p^*(s)$ profile.

Data availability

Primary and analyzed data are available from the corresponding author upon request. Sequence alignments used in this work are provided in a Supporting material.

RESULTS AND DISCUSSION

Most studies of collagen mechanics have focused on fibril-forming collagens such as collagen I, which associate laterally to form higher-order structures as shown in Fig. 1 (8). Considerably less is known about the mechanics of collagen IV, which forms the scaffold of basement membranes. We first quantified flexibility differences between these types of collagens by analyzing images of many individual collagen molecules.

AFM images of collagen

We imaged mouse heterotrimeric $[\alpha1(IV)]_2$-$\alpha2(IV)]$ collagen IV molecules (also referred to as protomers) derived from the matrix produced from EHS tumor using AFM. Multiple forms of heterotrimeric collagen IV exist ($\alpha112$, $\alpha345$, and $\alpha556$), coming from six different gene products (7); because our study here is limited to the $\alpha112$ form, we henceforth refer to it simply as "collagen IV." AFM directly images molecules deposited on a surface, so is free from potential metal replication artifacts of rotary shadowing EM, which has previously been used for imaging conformations and binding interactions of collagen IV (4,17,18,36–42). With appropriate imaging conditions, AFM offers superior spatial resolution than the platinum nanocrystallite size of 1–2 nm in EM (17), and our studies avoid the use of glycerol, which has been shown to affect collagen IV assembly (36). Although imaging in liquid conditions is possible with AFM, here we deposited collagens from the specified solution conditions and dried them before imaging. This follows previous work that used AFM to image and quantify the flexibility of collagen types I, II, and III (10,31,43), to observe interactions between the NC1 domains of collagen IV (44,45), and to observe binding sites of laminin on collagen IV (46). For the conditions used in this study (>100 mM ionic strength deposition before drying), numerous statistical measures demonstrate the collagen to be equilibrated on the mica surface (10).

TABLE 1 Physical classes of interruption

Flexibility class	$\alpha1$ interruption?	$\alpha2$ interruption?
Triple helix (p_0)	0	0
One-chain interruption (p_1)	0	1
Two-chain interruption (p_2)	1	0
Three-chain interruption (p_3)	1	1

An example AFM image of collagen IV deposited from acidic solution (100 mM KCl, 1 mM HCl) is shown in Fig. 2 *A*, along with an image of collagen I for comparison. The acidic conditions preclude lateral assembly of both fibril-forming and collagen IV molecules (10,36). Both proteins possess long chains, representing the collagenous domains. The contour length for EHS-derived collagen IV protomers is 360 ± 20 nm ($N = 262$), in agreement with previous reports (36) and, as expected, longer than the ~300 nm contour length of fibril-forming collagens (17). Similar to previous images recorded using rotary shadowing EM (17,36), we observe by AFM that the collagen IV protomer is capped by a globular domain, known as the C-terminal NC1 domain. Evident in the AFM image are two kinds of molecules: single protomers and dimers of protomers linked end-on by NC1 hexamers.

Visual comparison of collagen I and collagen IV images suggests that the triple helix of collagen IV is more flexible; collagen I bends on a long length scale, whereas collagen IV exhibits more frequent and shorter-range bending fluctuations.

Conformational analysis of collagen IV as a homogenous polymer

To quantify the flexibility of the collagenous domain of collagen IV, we applied our previously developed chain-tracing and analysis algorithm, SmarTrace (10). SmarTrace traces and provides conformational analysis of imaged chains, allowing for the implementation of polymer physics tools to determine their mechanical properties. To analyze collagen as a homogeneous polymer, we traced chains collected from AFM images and randomly segmented the contours into nonoverlapping pieces of different segment lengths. We calculated the mean-square end-to-end distance $\langle R^2(\Delta s)\rangle$ and the mean correlation of the beginning and ending unit tangent vectors (given by their dot product) $\langle \hat{t}\,(s) \cdot \hat{t}\,(s+\Delta s)\rangle$ for all segments of length Δs within the

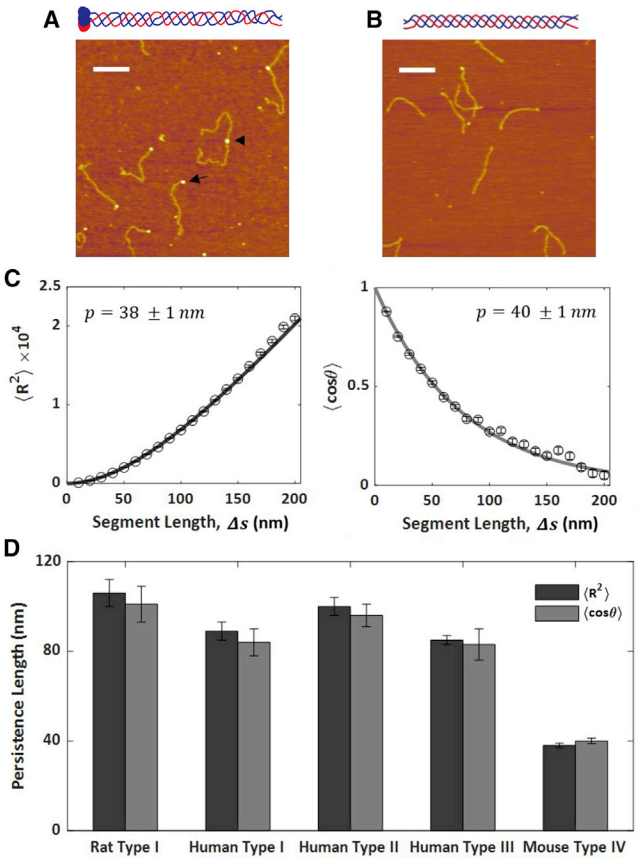

FIGURE 2 Homogeneous WLC analysis comparing flexibility of collagen IV to fibril-forming collagens. AFM images of (*A*) mouse collagen IV and (*B*) rat collagen I are shown, with schematics of their corresponding molecular structures placed above the AFM images. The black arrowhead and arrow point at an NC1 hexamer and an NC1 trimer of a molecule, respectively. Scale bars, 200 nm. All collagen types were deposited from room temperature onto mica from a solution of 100 mM KCl and 1 mM HCl. (*C*) The persistence lengths were obtained using $\langle R^2(\Delta s)\rangle$ (*dark gray*; Eq. 1) and $\langle \cos\theta(\Delta s)\rangle$ (*light gray*; Eq. 2) analyses, as shown for the collagen type IV data here. These data are fitted well by the predictions of the WLC model. (*D*) Bar plot comparing the persistence lengths of fibril-forming collages to collagen IV. Values of p were obtained using the $\langle R^2(\Delta s)\rangle$ (*dark gray*) and $\langle \cos\theta(\Delta s)\rangle$ (*light gray*) WLC fits. Errors represent 95% confidence intervals on p. The fibril-forming collagens possess similar persistence lengths of $p \approx 90$ nm (data reproduced from (10)), whereas collagen IV exhibits a substantially lower effective persistence length of $p = 39$ nm. $n = 262$ collagen IV chains. To see this figure in color, go online.

pool of collected chains. These were fitted with the predictions of the inextensible worm-like chain (WLC) model to estimate persistence length (10,47):

$$\langle R^2(\Delta s)\rangle = 4p\Delta s\left[1-\frac{2p}{\Delta s}\left(1-e^{-\frac{\Delta s}{2p}}\right)\right], \qquad (1)$$

$$\langle \hat{t}\,(s)\cdot\hat{t}\,(s+\Delta s)\rangle = \langle \cos\theta(\Delta s)\rangle = e^{-\frac{\Delta s}{2p}}. \qquad (2)$$

These equations assume the polymers to be equilibrated in two dimensions, which has been shown to be the case for collagens deposited under the solution conditions used here (10). Although we cannot test this explicitly, we assume that the persistence lengths determined from these two-dimensional images represent the bending flexibility of collagen in three dimensions, an assumption that holds reasonable agreement with previous collagen flexibility measurements (see discussion in (10)) and that has been clearly demonstrated for other systems like DNA (47,48). The collagen IV results for both $\langle R^2(\Delta s)\rangle$ and $\langle \cos\theta(\Delta s)\rangle$ appear to be well described by the WLC model (Fig. 2 *C*), from which we find a persistence length of $p = 39 \pm 2$ nm. To our knowledge, this is the first report of a net persistence length of collagen IV.

The persistence length of collagen IV is significantly less than that of fibril-forming collagens deposited under the same solution conditions (Fig. 2 *D*). Our earlier work revealed that the persistence lengths of different fibril-forming collagens (I, II, and III) are very similar, all falling within ~10% of $p = 90$ nm under these solution conditions and being well described as semiflexible polymers (10). By contrast, the persistence length of collagen IV is less than half of this value, reflecting a more flexible protein. This stark difference in bending flexibility between collagen IV and fibril-forming collagens contrasts with their very similar thermal stabilities (49). (This is achieved in part via more extensive proline hydroxylation in collagen IV, which compensates for the presence of interruptions (49).)

This striking difference in flexibility likely relates to the interruptions that characterize collagen IV molecules. The triple helix of collagens is defined by a repetitive $(\text{Gly-X-Y})_n$ sequence, in which the Gly is obligatory to form a stable triple-helical structure. The collagenous domain of collagen I is 96% triple helical (with nonhelical regions confined to its telopeptide ends (50)), whereas in collagen IV it is ~80% triple helical, with interruptions to the $(\text{Gly-X-Y})_n$ sequence occurring throughout the chain (Fig. S5). This distinction is shown schematically above the AFM images in Fig. 2. Because unstructured polypeptide chains are extremely flexible with $p < 1$ nm (51), we expect interruptions in the triple-helix-defining sequence to significantly enhance flexibility. This expectation is consistent with our finding of a lower persistence length of collagen IV and with the visual comparison of images of collagen IV and collagen I chains (Fig. 2, *A and B*).

From the example chain images in Fig. 2 *A*, it appears that the flexibility of the collagen IV molecule varies along its contour and is more flexible away from the NC1 domain, i.e., toward the N-terminus. In a study using rotary shadowing EM to examine collagen IV, Hofmann et al. found that it exhibits regions of variable flexibility along its contour (17). Thus, although the homogeneous WLC model provides a good fit to our experimental data (Fig. 2 *C*), we sought to assess the heterogeneity of flexibility along the collagen IV triple helix.

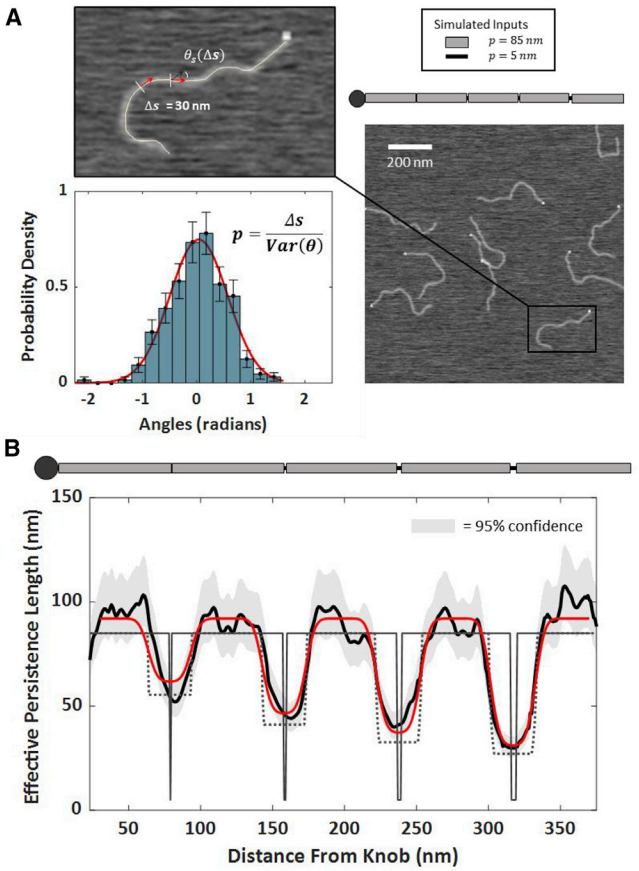

FIGURE 3 Extracted persistence length profiles from simulated inhomogeneous chains. (*A*) Images of WLCs with inhomogeneous flexibility profiles were simulated to validate position-dependent flexibility analysis. For segments of length $\Delta s = 30$ nm centered at position s along the backbone, the effective persistence length is determined from the variance of angular changes along these segments $\theta_s(\Delta s)$ (Eq. 4). Error bars reflect a $\sqrt{n}$ counting error. (*B*) The extracted persistence length profile $p^*(s)$ obtained by tracing and analyzing $n = 297$ simulated chains (*thick black line*) shows inhomogeneous flexibility along the simulated chains. The shaded regions represent 95% confidence intervals on the estimates of persistence length. $p^*(s)$ attains the expected persistence lengths in plateaus longer than $\Delta s = 30$ nm and exhibits minima at locations of enhanced flexibility. Minima are broadened with respect to the input $p(s)$ profile (*thin black line*), in agreement with the expected effective persistence length profile obtained by convolving a $\Delta s = 30$ nm filter with the input $p(s)$ (result: *dotted line*). The two-class model fit (*red*) recovers distinct persistence lengths of the rigid and flexible regions, returning $p_r = 92$ nm and $p_f = 6$ nm, respectively ($\chi^2_r = 1.06$). $n = 297$ simulated chains. To see this figure in color, go online.

Position-dependent flexibility analysis

To address structural variability and to evaluate the contributions of interrupted regions to the overall flexibility of collagen IV, we extended the SmarTrace analysis to include determination of the sequence-dependent chain flexibility. To do so, we use the variance of tangent angles around each position s along the chain backbone to quantify the bending stiffness at that location. This approach was developed by Hofmann et al. to analyze the position-dependent flexibility of collagens (17). This method complements other AFM image analysis methods developed to quantify discontinuous mechanical properties of DNA (52) or to map sequence-dependent properties of individual proteoglycans (53). It is expected to be more appropriate for determining the flexibility of collagen molecules than an alternative position-dependent stiffness analysis developed for more rigid biological filaments (54).

The flexibility of a filament at a position s along its contour is described by its local bending rigidity $\alpha(s)$. The bending rigidity is related to the persistence length via the thermal energy, k_BT:

$$p(s) = \frac{\alpha(s)}{k_BT}. \tag{3}$$

The persistence length thus describes the flexibility of a chain with a given bending stiffness at a given temperature. Practically, we are limited to examining the flexibility of a chain over a segment of finite length Δs, which we define centered at position s; we use the term "effective persistence length" to denote this averaged nature of the response, $p^*(s; \Delta s)$. The effective persistence length of a segment of length Δs is determined from the variance of the tangent angles θ_s between the ends of the segment centered at s:

$$p^*(s; \Delta s) = \frac{\Delta s}{\mathrm{Var}(\theta_s; \Delta s)}. \tag{4}$$

The analysis methodology was validated through tests on simulated chains. We generated AFM images of chains with inhomogeneous flexibility profiles, traced the chains using SmarTrace, and analyzed the resulting contours using Eq. 4. The chains were simulated as previously described (10), here incorporating a position-dependent bending stiffness and a "knob" at one end (which served as a directional reference point, analogous to the NC1 domain of collagen IV). Fig. 3 *A* shows an example of a simulated image, along with a schematic of the inhomogeneous flexibility profile imposed in the simulations (long regions of relatively stiff chains with $p = 85$ nm, interspersed with very short regions with $p = 5$ nm). The short regions of substantially increased flexibility are not visually apparent in the simulated images but are clearly seen in the effective persistence length profile determined from the traced chains (Fig. 3 *B*). A notable difference between the input and extracted persistence length profiles is the apparent broadening of the flexible regions: $p^*(s; \Delta s)$ exhibits much broader wells than the input $p(s)$ profile. This is the expected result of the analysis: the $\Delta s = 30$ nm segment length is substantially longer than the regions of flexibility in our simulated chains and acts to some degree as a low-pass filter of flexibility along the chain (though we stress that $p^*(s)$ is not simply the mean of $p(s)$ over a 30-nm window; see Supporting materials and methods, Text S1; Fig. S1). Fig. 3 *B* includes the

expected effective persistence length profile, which considers this convolution between the ideal chain and the 30-nm filter. This expected profile agrees well with $p^*(s; \Delta s)$ obtained from analyzing the simulated images. Although a shorter segment length should provide more localized information on sequence-dependent flexibility, our validation tests found $\Delta s = 30$ nm to be the minimal length that returned reliable estimates of $p^*(s)$ from simulated chain images (Fig. S4). Thus, $\Delta s = 30$ nm was used for the analysis presented herein.

Having validated the persistence length mapping algorithm, we applied it to experimental images of collagen IV. The effective persistence length profile of collagen IV is shown in Fig. 4. From this result, it is obvious that the flexibility of collagen IV varies markedly along its contour, with the N-terminal half of the molecule being significantly more flexible than the region closer to the NC1 domain. Similar flexibility trends are observed whether collagen chains are traced starting at the NC1 domain (Fig. 4) or in the reverse direction from the 7S domain (Fig. S6). Our flexibility profile agrees well with that of Hofmann et al., who also observed increased flexibility toward the N-terminus of collagen IV (17).

To gain sequence-dependent insight into the flexibility profile requires aligning this map with the collagen IV sequence, which was unavailable to Hofmann et al. (17). An amino acid representation of the full sequence of mouse $\alpha1\alpha1\alpha2$ collagen IV is shown in Fig. 4 *A*. The α-chain polypeptide sequences have been aligned from the beginning of the collagenous domain beside the NC1 domain (from the first Gly-X-Y). The blue and red boxes correspond to segments of $\alpha1$(IV) and $\alpha2$(IV), respectively, that have the $(\text{Gly-X-Y})_n$ sequence required to form a triple helix. In this representation, we have assumed a minimal length of $(\text{Gly-X-Y})_4\text{G}$ to be sufficient to form a triple helix. This assumption of $n = 4$ does not strongly affect the representation (Fig. S7). The thinner lines in Fig. 4 *A* represent interruptions in the triple-helix-competent $(\text{Gly-X-Y})_n$ repetitive sequence.

The position-dependent flexibility profile of collagen IV aligned with the amino acid sequences is shown in Fig. 4, *A and B*. An examination of the sequences of the α-chains reveals that most of the interruptions lie in the N-terminal half of the protein, where we also see an increase in flexibility (lower persistence length). Maxima in local persistence length align well with extended triple-helical stretches in all three chains, found ~30, 90, and 160 nm from the NC1 domain. Within these plateau regions, collagen IV's effective persistence length is $p^* \approx 95$ nm, the persistence length found for continuously triple-helical collagens (see Fig. 2) (10). Between these maxima, minima in persistence length are found ~60 and 130 nm from the NC1 domain. These two locations align well with regions of the sequence in which both $\alpha1$ and $\alpha2$ chains possess interruptions in the $(\text{Gly-X-Y})_n$ sequence. This suggests that interruptions of the triple-helix-forming sequence in all

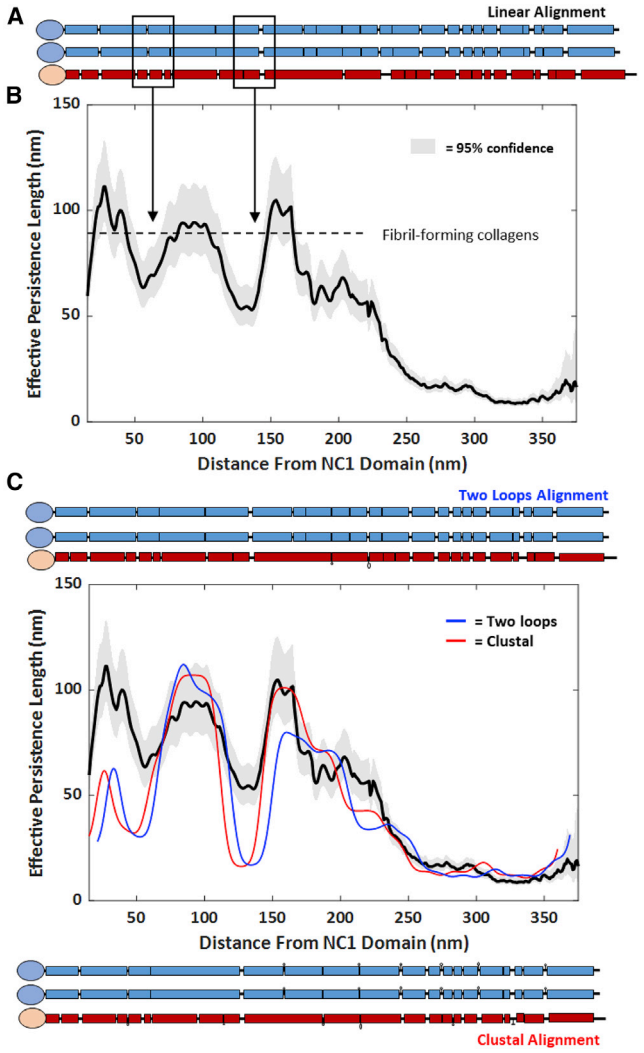

FIGURE 4 Position-dependent flexibility profile of collagen IV. (*A*) Schematic representation of the $\alpha1$ and $\alpha2$ amino acid sequences from mouse collagen type IV. Rectangles indicate regions of the sequence containing triple-helix-competent sequences with $(\text{Gly-X-Y})_n\text{G}$, $n \geq 4$. Interruptions of this repetitive sequence are indicated by thinner lines and occur more frequently toward the N-terminus of the chains (*right side* of the schematic). (*B*) Position-dependent persistence length map of collagen type IV deposited from 100 mM KCl, 1 mM HCl. The dashed line indicates the persistence length of continuously triple-helical fibril-forming collagen molecules (10). Shaded curves represent 95% confidence intervals on the effective persistence length estimate $p^*(s; \Delta s)$. The profile was calculated from $n = 262$ chains. The effective persistence length map begins ~15 nm into the collagen sequence because of the use of centered 30-nm windows for determination of p^*. (*C*) Four-class flexibility model using two-loop (*blue*) and Clustal (*red*) alignments. Alignments differ in the number of amino acids that loop out and thus do not participate in the main backbone, as shown schematically above and below the plot. The persistence length profile is aligned with the amino acid sequence representations using model outputs for best-fit offset (nm) and nanometer/amino acid conversion. The two-loop (Clustal) alignment is best fitted with persistence lengths of $p_0 = 105$ (109) nm, $p_1 = 83$ (86) nm, $p_2 = 21$ (42) nm, and $p_3 = 1.9$ (2.0) nm, with $\chi^2_r = 12.6$ (12.8). Best-fit persistence lengths of these different structural classes, particularly p_1 and p_2, differ for other chain alignments (Table S1). To see this figure in color, go online.

three α-chains of $\alpha1(IV)_2$–$\alpha2(IV)$ strongly impact the flexibility of the molecule, apparently more so than having interruptions in only one or two of the constituent chains.

Model for interpreting position-dependent flexibility

To gain a deeper structural interpretation of the persistence length profile, we implemented a simple physical model to describe the sequence-based flexibility of a collagen containing interruptions (Supporting materials and methods, Text S2). This model assumes that the local flexibility of collagen can be completely described by how many of the three component chains contain a triple-helix-compatible sequence at that location. To implement the model, each amino acid in each of the $\alpha1$ and $\alpha2$ chains was assigned a value of either "0" (contained within triple-helix-compatible sequence $(Gly\text{-}X\text{-}Y)_n$) or "1" (within an interruption). The three α-chain sequences were aligned starting from the edge of the NC1 domain, and different chain alignments were tested. These include a linear, sequence-based alignment (Figs. 4 *A* and S8); nonlinear alignments that include previously proposed intrachain loops within $\alpha2$ (34,55,56) and improve the registration of triple-helix-forming sequences; and alignments arising from a Clustal-based analysis of sequence similarity between $\alpha1$ and $\alpha2$ chains (35), which require the incorporation of copious loops and bulges to confer backbone continuity.

For each chain alignment, a flexibility class was assigned at each amino acid step along the backbone based on whether it was a 0-, 1-, 2-, or 3-chain interruption in the triple helix (Table 1). Each of these classes was assigned a local bending stiffness (via a local persistence length) and thus the model assigned a local flexibility to each position along the contour. This model was then fitted to the determined persistence length profile $p^*(s; \Delta s)$ to determine the persistence lengths p_0, p_1, p_2, and p_3 that best describe each flexibility class. The approach was validated by application to the simulated chains; here, a two-class model (rigid or flexible) recovered the distinct persistence lengths of the rigid and flexible regions and described the data well ($p_r = 92$ nm; $p_f = 6$ nm; $\chi^2_r = 1.06$). A slight overestimate of persistence lengths is expected for finite samples (Supporting materials and methods, Text S1). Agreement between the four-class flexibility model and the measured persistence length profile $p^*(s; \Delta s)$ of collagen IV would indicate that this physical classification is a valid approach to describing the variable flexibility of collagen IV. By contrast, disagreement would indicate that other factors (such as the influence of X and Y identity on flexibility, length of interruption, and location with respect to interruptions (extended triple-helix tracts) may be of key importance for quantifying the structural properties of this protein.

This model of a physically heterogeneous collagen IV captured positional variations in flexibility along the chain, comparing favorably to the measured $p^*(s)$ profile (Fig. 4 *C*). The model returned a value for triple-helix local persistence length of p_0 ~100–160 nm for all distinct α-chain alignments tested (Table S1). This estimate of p_0 is somewhat larger than the net persistence length of continuously triple-helical collagens found by treating the chains as homogeneous (Fig. 2 *D*) but is consistent with local values of p^* determined along their backbones (see below). Also consistent among chain alignments was the finding that $p_2 < p_1$. In other words, having an interruption in two of the chains leads to a more flexible structure than having a one-chain interruption.

Some of the chain alignments resulted in fits of the variable flexibility model that were not minimized within the specified parameter range: for example, a linear chain alignment (aligned based only on primary sequence) produced for an overlapping interaction $p_3 = 200$ nm, the maximal value allowed for fitting, a value that represents an unphysically rigid structure of three polypeptide chains that lack any triple-helix-forming sequence. Alignments that returned this high value, however, were the only ones that captured the rigidity in the collagenous region adjacent to the NC1 domain (e.g., Fig. S8).

Alignments that optimized the fit within their parameter bounds returned an optimal length scaling of 0.29 nm/amino acid, commensurate with triple-helix lengths from collagen diffraction analysis (23,57). These alignments also returned the smallest value for $p_3 \approx 2$ nm, i.e., that an overlapping interruption is the most locally flexible structure within the collagenous domain of collagen IV. These optimal alignments require at least some outward loops in the longer $\alpha2$ chain (e.g., "two-loop" alignments) and perhaps extensive minor loops and bulges along the length of each of the three α-chains ("Clustal" alignments).

In some regions, the variable flexibility model underestimates the rigidity, all located in the C-terminal half of the protein (Fig. 4 *C*). The region adjacent to the NC1 domain exhibits a rigidity similar to the other extended triple-helical regions, yet the model significantly underestimates the persistence length, both of the plateau at $s \approx 30$ nm and of the adjacent minimum at $s \approx 50$ nm. These model results imply that the overlapping interruption closest to the NC1 domain possesses a distinctly enhanced stability not captured by our simple physical model. Enhanced stability in this region could be conferred by post-translational glycosylation of hydroxylysines within this overlapping interruption (58–60). Unfortunately, a glycosylation map of collagen IV lacked sequence coverage in this region (61), so this remains a speculative, yet testable, prediction.

The alignment of the three chains of the $\alpha1\alpha1\alpha2$ mouse collagen IV collagenous domain is unknown. A disulfide-bridged loop within $\alpha2$ has been proposed, which helps better to align triple-helix-competent sequences in the chains (34,55); this is included as the larger, N-terminal loop in our "two-loop" alignment (Fig. 4; Supporting

materials and methods). Most of our chain alignments also agree with the determined alignment of α-chains at an integrin-binding site in human collagen IV (Table S1), the only location for which chain alignment in collagen IV has been determined (27). To validate or further constrain chain alignments will require improvements in spatial resolution from this AFM imaging and chain-tracing approach and/or complementary approaches such as proteolysis to identify unstably structured regions within the collagenous domain (62).

In our alignments, we ignored the register of the three chains, as the register has been identified at only one location and only in human collagen IV (27). We tested a staggered alignment of the three chains that included this identified local alignment, but it failed to produce an optimized fit (Table S1). It is highly likely that some of the interruptions alter the local register of the three chains (such that a leading chain at some locations may become the middle or trailing chain at others) (16) and that one "universal" registration does not apply along the entire collagenous domain.

It is also possible that the variable flexibility model is insufficient to completely describe the measured flexibility profile because of its many simplifying assumptions. Beyond stabilization by post-translational modifications, potential intramolecular cross-links (though these were few in our sample; Materials and methods), and local variations in chain stagger, there are other attributes that may be of direct relevance to local stability and flexibility. For example, we have ignored the length dependence of interruptions when defining flexibility class. Short one-chain interruptions may act as "stammers," serving to alter the local register among chains instead of breaking the triple-helix structure (16). Single glycine substitutions may introduce local kinks in the structure (63). Longer interruptions may destabilize flanking triple-helical regions, leading to increased breathing of the adjoining triple helix and a lower persistence length (64). The local sequence within and flanking the interruption also may alter its stability (23). And finally, this physical model ignores all sequence-dependent attributes of not only the interruptions but also the triple helix itself by classifying each amino acid purely by whether or not it is found within a (Gly-X-Y) sequence. We next sought to test the assumption of mechanical homogeneity within the triple helix by mapping the flexibility profile of a fibril-forming collagen.

Triple-helix mechanical inhomogeneity: collagen III

To probe the variability of bending stiffness along the triple helix, we imaged the pN-III molecular variant of bovine collagen III using AFM (Fig. 5 *A*). Collagen III is a homotrimeric [α1(III)]$_3$] fibril-forming collagen devoid of interruptions in its ~300-nm-long triple-helical domain. The pN-III variant retains its N-terminal propeptide (65), providing us with a directional reference point for position-dependent flexibility analysis.

We found that the flexibility of a triple helix varies significantly along its contour, although less than along the interruption-containing collagen IV (Fig. 5 *B*). This provides direct evidence for distinct mechanical properties of different sequences within the collagen triple helix. Our results show considerably more variation in local flexibility along the backbone of collagen III than previous rotary shadowing EM measurements on this molecule (17), perhaps because of the differences in resolution provided by the two imaging and chain-tracing approaches and/or the choice of segment length used to determine the effective persistence length. Our flexibility map shows that the effective persistence length ranges from ~60 to 180 nm in different regions of collagen III. Notably, the larger values are similar to values returned for p_0 of the triple helix in our model fitting, above, which ranged from 100 to 160 nm, and are longer than the net persistence length of continuously triple-helical collagens (Fig. 2 *D*).

In the context of our physical model for collagen IV, this variability of persistence length within collagen III has

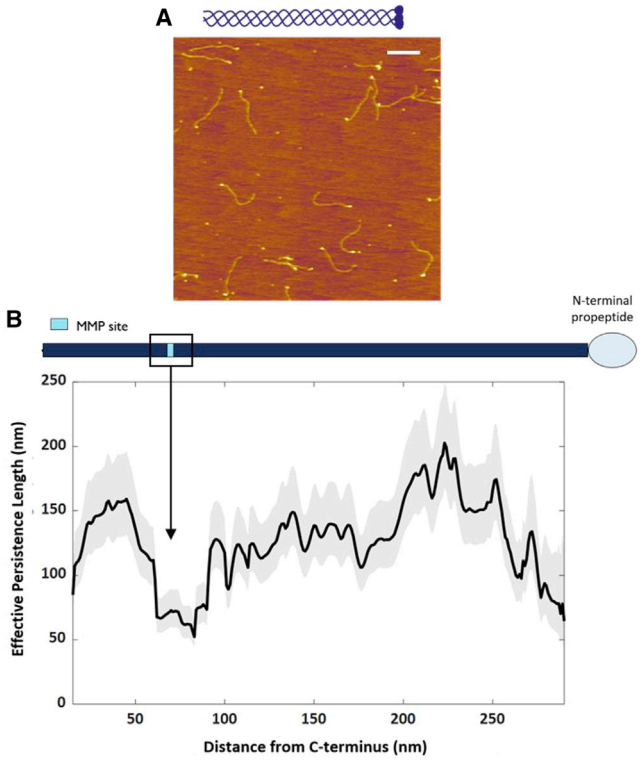

FIGURE 5 Position-dependent flexibility profile of collagen III. (*A*) AFM image of bovine collagen pN-III deposited from room temperature onto mica from a solution of 100 mM KCl and 1 mM HCl. Scale bars, 200 nm. (*B*) Position-dependent persistence length map of collagen III aligned with its amino acid sequence representation. This collagen is continuously triple helical and therefore is represented as a single bar with the MMP site marked. Shaded curves represent 95% confidence intervals on the effective persistence length. The profile was calculated from $n =$ 267 chains. To see this figure in color, go online.

additional implications. The range of p^* found for collagen III suggests that the assumption of a single persistence length p_0 to characterize all triple-helical regions of collagen IV is an oversimplification. However, the model-extracted values of p_2 and p_3 are significantly shorter than the lowest values of p^* along the collagen III chain, with values of p_1 similar to or less than the most flexible regions of collagen III. It is possible that two chains possessing a $(\text{Gly-X-Y})_n$ sequence may be able to induce a triple-helix-compatible structure in the third chain, leading to a value of p_1 similar to regions of the collagen III triple helix. Overall, we can conclude that the enhanced flexibility of collagen IV is dominated by the effects of interruptions, though variability of X, Y residues within a continuous $(\text{Gly-X-Y})_n$ sequence also influences bending flexibility. Previous single-molecule imaging has suggested that sequence may also control the local diameter of collagen (66).

The regions of greatest flexibility along pN-III collagen are found near the N-terminus (which contains conformationally distinct regions (67,68)) and, intriguingly, at the matrix-metalloprotease (MMP) site (Fig. 5 *B*). This site is the target of proteolysis by MMPs and is of key importance for extracellular matrix remodeling, e.g., during embryonic development and cancer metastasis. To our knowledge, ours is the first demonstration of enhanced mechanical flexibility at the MMP site. Our ability to detect this feature when previous studies using EM did not (17) may arise from the higher-resolution chain tracing in our AFM images compared to the particulate nature of rotary shadowed chains (31) and from our directional tracing from the C-terminus toward the N-terminus. When we traced the chains starting at the N-terminus, the contrast in flexibility around the MMP site was diminished (Fig. S9), perhaps resulting from challenges in identifying the start of the triple-helical domain in the presence of the N-propeptide (67).

The MMP site has previously been identified as possessing distinct characteristics within the collagenous domain. Intriguingly, the sequence in this MMP region of collagen I has been found, in peptides, to exhibit structural tautomerism, transitioning between a triple-helix and β1-bend structure (69). It is interesting to speculate that this structural flexibility may underlie the enhanced bending flexibility revealed here through AFM imaging. The MMP region also exhibits enhanced sensitivity to proteolysis by trypsin and has been suggested to have an enhanced propensity to unwind (70–76). We propose that this structural instability is what gives rise to the enhanced bending flexibility in this region; three weakly interacting chains are predicted to bend more easily than a tightly wrapped triple helix. This would produce a lower local persistence length.

The decreased triple-helix stability in the MMP region of fibril-forming collagens has been attributed to a low imino acid content (73,77,78). Thus, we examined how local imino acid content correlates with flexibility (Figs. S10 and S11). We saw no obvious correlations between imino acid content and flexibility along the collagen III contour (Fig. S10 *A*), which was confirmed by a linear correlation coefficient of $R^2 = 0.017$. Although an anticorrelation between imino acid content and flexibility in the C-terminal half of collagen IV might be inferred by visual comparison of these profiles (Fig. S11 *B*), quantitative analysis revealed no statistical correlation between these quantities, neither for the full length of the chain ($R^2 = 0.084$) nor for the first 200 nm from the NC1 domain ($R^2 = 0.15$). Thus, a simple picture of imino acids enhancing rigidity (17,79) does not apply here. Incidentally, recent work has suggested that imino acids may instead enhance the flexibility of triple-helical regions (80). Although the anticorrelation we observe in some regions supports this conclusion, we do not find this to be the case globally along the collagen backbone. Thus, although imino acids may contribute to local flexibility, at present our data suggest that they are not the main driver of bending flexibility of the triple helix.

We also looked for correlations between the locations of the imino acid within the (Gly-X-Y) triplet and the flexibility. This is because most prolines in the Y position are 4-hydroxylated (whereas most in the X position are not, with a small fraction 3-hydroxylated) (61,81–83). (We do not have specific information about the hydroxylation state of each proline in our samples.) Hydroxyproline increases the thermal stability of the triple helix (84,85), and we thus wished to determine whether its presence correlated with increased triple-helix bending rigidity. As seen in Fig. S10 *B*, we find no correlation between flexibility and Y-positioned prolines. X-positioned prolines can enhance conformational flexibility (within Gly-Pro-Hyp triplets) and have recently been correlated with enhanced thermal stability of collagens (80). However, we also find no correlation between local flexibility within collagen III and the local density of X-positioned prolines (Fig. S10 *B*).

The physical model we have developed to assess the structural relationships between interruptions and flexibility could be adapted to study sequence-dependent influences on flexibility in collagens lacking interruptions, such as collagen III. At present, our flexibility analysis is limited by accuracy in chain registry and has an ~100-amino-acid resolution (30-nm filter window) (Supporting materials and methods, Text S2), but improvements in imaging and tracing algorithms should be able to push this to finer length scales. Even with the current resolution, this approach should allow initial comparisons of bending flexibility with measured and predicted variations in local helical symmetry, thermal stability, sequence content, and post-translational modifications. Such comparisons would provide important links among the composition, structure, and mechanics of collagen and would offer a further means to connect high-resolution studies on triple-helical peptide constructs with their in situ properties in the full-length proteins (23,86). Such knowledge of the breadth of properties

exhibited by continuously triple-helical structures would enable elucidation of the effects of glycine substitutions and of sequence interruptions on the molecular structure and mechanics of collagen (63,87,88).

Physiological implications of collagen IV flexibility

Thus far, our analysis has determined collagen IV properties only when deposited from acidic solution conditions. These conditions (pH ~3) are more acidic than the varying physiological conditions during biosynthesis and secretion. Previous studies of collagen IV flexibility also deposited the protein from very nonphysiological conditions (25 mM acetic acid, 50% glycerol) (17). Thus, we sought to investigate its flexibility in more physiologically relevant conditions.

Intracellularly, folded proteins are transported in secretory vesicles with low pH and chloride concentrations relative to the cellular exterior (89,90). Upon secretion, collagen's chemical environment changes to one with a higher chloride concentration and neutral pH. This environmental switch is important for triggering assembly into higher-order structures; NC1 interactions strengthen at neutral pH and higher chloride concentrations (42,91,92), supporting collagen IV assembly into networks, and neutral pH and chloride favor lateral assembly of fibril-forming collagens (93,94). Environmental conditions such as pH and ionic strength have also been found to influence the conformations of fibril-forming collagen molecules (10,43). These observations raise the question of how much the conformations of the collagenous domain of collagen IV are controlled by chloride and pH. We posited that the pH and chloride gradient experienced during secretion might affect the flexibility of collagen IV, conferring a shorter persistence length and enhanced flexibility early in the secretory pathway, which would impart a lower free energy cost for compact configurations.

To test this hypothesis, we analyzed collagen IV under two additional solution conditions of pH and Cl concentrations. To represent vesicular solution conditions, we took an extreme limit by eliminating the chloride ions and used a sodium acetate buffer at a slightly acidic pH (150 mM NaOAc, 25 mM Tris-OAc (pH 5.5)). TBS (150 mM NaCl, 25 mM Tris-Cl (pH 7.4)) served as an extracellular proxy because of its relatively high chloride content and neutral pH.

AFM images of collagen IV molecules deposited from each of these two solution conditions are shown in Fig. 6, *A and B*. Visually, the collagen conformations look similar in both images and appear similar to those found in the previous condition of 100 mM KCl, 1 mM HCl (pH ~3) (Fig. 2 *A*). It is important to note that although collagen was deposited from a chloride-free environment in the AFM image in Fig. 6 *B*, there are still some head-to-head assemblies. This is due to the presence of sulfilimine cross-links in the NC1 hexamer, a post-translational modification that endows mechanical strength and stability to the collagen IV network (95). The presence of such cross-links protects the dimers from dissociation in a chloride-free environment. We do not observe any evidence of higher-order association or network formation in our images in these conditions, likely because of the low solution concentration of collagen IV used for deposition (0.2 μg/mL) (36).

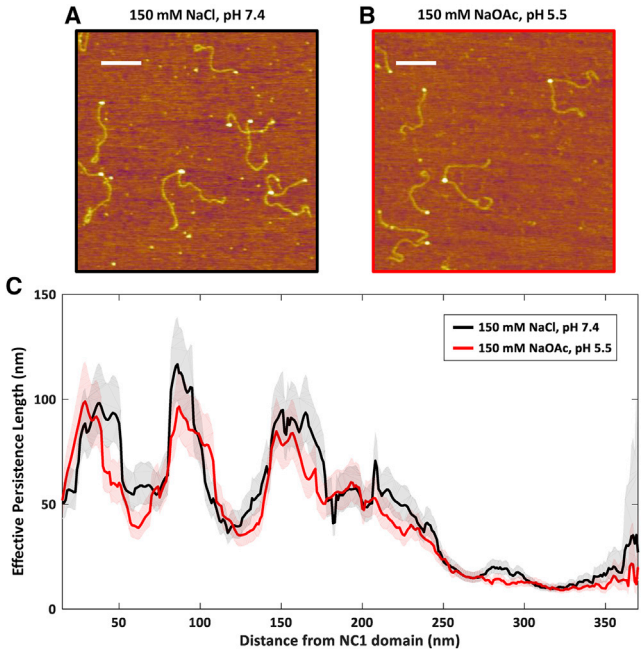

FIGURE 6 Effect of chloride and pH on the flexibility of collagen IV. Inspired by the changes in the chemical environment along collagen's secretion pathway, we imaged collagen IV under two solution conditions, (*A*) TBS (150 mM NaCl, 25 mM Tris-Cl (pH 7.4)) and (*B*) sodium acetate buffer at a slightly acidic pH (150 mM NaOAc, 25 mM Tris-OAc (pH 5.5)). Scale bars, 200 nm. (*C*) Position-dependent flexibility map of collagen type IV. The shaded curves represent 95% confidence intervals on the effective persistence length estimates $p^*(s; \Delta s)$. The profiles were calculated from $n = 272$ (TBS; *black*) and $n = 273$ (NaOAc; *red*) chains. To see this figure in color, go online.

Surprisingly, we find similar overall flexibility among both buffered solution conditions (net persistence lengths $p = 43 \pm 3$ nm for TBS and $p = 41 \pm 3$ nm for NaOAc; Fig. S12) and the previous acidic solution ($p = 39 \pm 2$ nm for 100 mM KCl + 1 mM HCl). Collagen IV exhibited no global curvature in any of these conditions (data not shown). Therefore, collagen IV flexibility is not tuned by changes in pH and chloride concentrations that mimic those experienced during secretion.

The net persistence length of a collagen molecule can be used to estimate its average size in solution, which we can compare with the size of vesicles that have been proposed to transport collagens from the endoplasmic reticulum (ER) to the Golgi. To do so, we assume that collagen molecules adopt a random coil conformation in solution whose extent is approximated as a sphere with a radius given by the radius of gyration, R_g. We find that collagen IV is

predicted to be considerably more compact ($R_g = \sqrt{\frac{pL}{3}} = 74$ nm, assuming a contour length $L = 410$ nm and $p = 40$ nm) than the interstitial, fibril-forming collagens ($R_g = 95$ nm, with $L = 300$ nm and $p = 90$ nm). Thus, despite its longer length, collagen IV would pose less of a size burden to the secretory machinery of the cell if it were to be transported in small vesicles. The 150–190 nm diameter of such random coils is significantly less than the 300-nm length dimension commonly stated as required for transport, determined by (incorrectly) assuming fibril-forming collagens to act as rigid rods. Nonetheless, both collagen IV and fibril-forming collagens would require additional compaction to fit in the standard COPII vesicles involved in ER-to-Golgi transport (60- to 90-nm diameter) (89,96).

Although it is generally accepted that collagens are not trafficked in such small vesicles, there remains considerable uncertainty about the mechanism used for procollagen trafficking during secretion. To our knowledge, there is no evidence of any collagen-associated proteins such as Hsp47 acting to compact collagen to facilitate loading into ER exit vesicles. In fact, some studies have found collagens IV and VII to be transported from the ER to Golgi in larger vesicles with 400 nm diameter (97,98), into which fibril-forming and collagen IV could easily be accommodated. However, recent work has found that direct trafficking of procollagen I from the ER to Golgi occurs and that these large vesicular carriers are not required (99). In this case, collagen (in)flexibility would not pose a strong energetic cost on its transport to the Golgi, consistent with our finding that collagen flexibility is not tuned by chemical environmental changes experienced during secretion.

As shown above, collagen IV should not be viewed as a uniformly flexible structure, so again, we performed a position-dependent analysis. We sought to determine how both triple-helical and interrupted regions of the collagenous domain of collagen IV are affected by chloride and pH. These regions have been implicated in assembly into higher-order networks (40), and we posited that their stability and flexibility would vary between intracellular (lower Cl^- and pH; assembly-disfavoring) and extracellular (higher Cl^- and pH; assembly-favoring) conditions. Specifically, because the triple helix is thought to loosen and destabilize at lower pH and Cl^- (albeit at significantly lower ionic strengths than used here) (10,100), we anticipated that a similar destabilization would enhance microunfolding of the triple-helical regions of collagen IV. This would be seen by a shortening of the high persistence length plateaus in sodium acetate relative to TBS and a concomitant broadening of the minima.

Surprisingly, we find that the flexibility profile of collagen IV is affected very little by these changes in pH and chloride. Fig. 6 *C* shows the superposition of the flexibility profiles of collagen IV in sodium acetate and TBS buffers, including the respective 95% confidence intervals of the effective persistence lengths. Over the vast majority of the collagenous domain, the effective persistence length is statistically unaffected by this change in solution conditions upon deposition. This contrasts with the strong mechanical response of collagens to decreases in ionic strength (10). Over the range of pH from ~3 to 7.4, the net charge on collagen IV is expected to change significantly, from positively charged to approximately neutral (Fig. S13). Despite this change in charge, the net persistence length of collagen IV remains the same ($p = 40$ nm), and its position-dependent flexibility is essentially invariant (Fig. S14). This finding provides further evidence that the conformations we observe are not governed by electrostatic interactions with the surface. Furthermore, it implies that electrostatic interactions such as salt bridges (101) do not contribute substantially to the net bending stiffness of collagen at physiologically relevant ionic strengths.

There are only a few regions in the effective persistence length profile where the flexibility differs between these two buffered solution conditions. The region ~170 nm from the NC1 domain is of interest as a potential site for NC1 binding that may lead to higher-order assembly (36). The sensitivity of the structure and mechanics in this region of the collagenous domain to solution pH and Cl^- could provide a responsive element that prevents premature assembly inside the cell. AFM-based mapping of ligand binding sites along collagen IV could help to reveal how these interactions are modulated by pH and chemical environment, and to relate the binding to local collagen structure. Sequence-based predictions of binding sites on collagen IV have been made for various binding partners (35,102), which could be compared with AFM-based mapping experiments. Of particular interest would be proteins implicated in chaperoning collagen folding (e.g., Hsp47), secretion (e.g., TANGO1), and assembly (e.g., SPARC), as well as other proteins involved in BM assembly (103,104). For example, EM-based mapping has found laminin, a major BM component, to bind 140 nm from the NC1 domain along the collagen IV molecule (37). Analysis of our persistence length profile shows that this binding site is at a flexible region, which coincides with a large overlapping interruption. On the other hand, the glycoprotein nidogen was found to bind 80 nm from the NC1 domain (41), within a triple-helical region. These results suggest that there may be a broad diversity of binding modes to collagen IV, which can be modulated by its local structure and perhaps further tuned by the chemical environment. Studies of the temperature-dependent flexibility profile, binding interactions, and in vitro assembly morphologies could elucidate the impact of sequence-dependent mechanical and thermal stability on guiding BM assembly.

Our findings on the bending flexibility of collagen can be used to quantify energetic costs of various candidate compartments and constraints along its secretory and assembly pathways. For example, the enhanced flexibility of collagen IV imparted by its interruptions means that it is expected to adopt more compact conformations in solution compared with its shorter fibril-forming collagen counterparts. Thus, its

secretory burden on the cell machinery may be less. However, this flexibility of collagen IV endows a greater entropic cost for lateral ordering into higher-order structures compared with the stiffer collagen III. The greater rigidity of continuously triple-helical collagens facilitates their lateral assembly into highly ordered thick fibrils. We speculate that these mechanical properties contributed to the development of collagen in evolution, in which collagen IV arose first but now is a minority component, at least by mass portion, compared with fibril-forming collagens that form the basis of most connective tissues in higher-level organisms (6).

CONCLUSIONS

AFM-based mapping offers the ability to measure local structural properties such as bending stiffness within the context of full-length collagen proteins and correlate these with sequence-dependent properties elucidated from studies on much shorter collagen-mimicking peptides. Our results reveal a heterogeneous mechanical response along the triple-helical domains of collagens. We found that the presence of interruptions provides significant flexibility to collagen IV and were able to describe the shape of its flexibility profile reasonably well with a simple physical model that classifies sequences solely based on the presence of interruptions in the (Gly-X-Y) pattern. In our complementary imaging and analysis of collagen III molecules, we found significant mechanical heterogeneity along their lengths, indicating that a more refined picture of collagen flexibility should include sequence-dependent flexibility within triple-helical regions. Notably, we found collagen III to possess a unique mechanical signature at the MMP site, a key sequence for collagen's extracellular remodeling. Surprisingly, in light of past suggestions that proline content influences local triple-helix rigidity (17,67,68,70,71), we find no net correlation between local proline abundance and flexibility of collagenous domains, though some tracts in both collagens III and IV exhibit an anticorrelation between flexibility and imino acid content. In this study, our analysis was limited to determining effective persistence lengths over 30-nm segments and deconvolving the effect of local sequence; future improvements in imaging and image analysis that allow more refined backbone tracing and registry of chains should further enhance our understanding of how local sequence influences mechanical properties of collagens.

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2021.08.013.

AUTHOR CONTRIBUTIONS

A.A.-S. performed all experiments and analyzed all experimental data. A.L. devised and validated the position-dependent analysis protocols. A.A.-S. and N.R.F. developed the variable flexibility model. All authors contributed to the project design and to writing the manuscript.

ACKNOWLEDGMENTS

We gratefully acknowledge Albert Ries for the gift of collagen IV from the Klaus Kühn lab and Takako Sasaki for help procuring the sample. We thank Hans Peter Bächinger for providing the pN-III collagen and for supportive feedback. We thank Mike Kirkness, Mathew Schneider, Daniel Sloseris, and other members of the Forde lab for useful suggestions on this work and ChangMin Kim for assistance with AFM instrumentation.

This work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant to N.R.F. and a Postgraduate Scholarship – Doctoral (PGS-D) fellowship to A.A.-S. B.G.H. and S.P.B. acknowledge funding from The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK RO1 18381-49).

SUPPORTING CITATIONS

References (105–107) appear in the Supporting material.

REFERENCES

1. Shoulders, M. D., and R. T. Raines. 2009. Collagen structure and stability. *Annu. Rev. Biochem.* 78:929–958.
2. Sorushanova, A., L. M. Delgado, …, D. I. Zeugolis. 2019. The collagen suprafamily: from biosynthesis to advanced biomaterial development. *Adv. Mater.* 31:e1801651.
3. Fidler, A. L., S. P. Boudko, …, B. G. Hudson. 2018. The triple helix of collagens - an ancient protein structure that enabled animal multicellularity and tissue evolution. *J. Cell Sci.* 131:jcs203950.
4. Timpl, R., H. Wiedemann, …, K. Kühn. 1981. A network model for the organization of type IV collagen molecules in basement membranes. *Eur. J. Biochem.* 120:203–211.
5. Knupp, C., and J. M. Squire. 2005. Molecular packing in network-forming collagens. *Adv. Protein Chem.* 70:375–403.
6. Fidler, A. L., C. E. Darris, …, B. G. Hudson. 2017. Collagen IV and basement membrane at the evolutionary dawn of metazoan tissues. *eLife.* 6:e24176.
7. Hudson, B. G. 2004. The molecular basis of Goodpasture and Alport syndromes: beacons for the discovery of the collagen IV family. *J. Am. Soc. Nephrol.* 15:2514–2527.
8. Kirkness, M. W., K. Lehmann, and N. R. Forde. 2019. Mechanics and structural stability of the collagen triple helix. *Curr. Opin. Chem. Biol.* 53:98–105.
9. Fratzl, P. 2008. Collagen: Structure and Mechanics. Springer, New York.
10. Rezaei, N., A. Lyons, and N. R. Forde. 2018. Environmentally controlled curvature of single collagen proteins. *Biophys. J.* 115:1457–1469.
11. Miller, R. T. 2017. Mechanical properties of basement membrane in health and disease. *Matrix Biol.* 57–58:366–373.
12. Kadler, K. E., Y. Hojima, and D. J. Prockop. 1987. Assembly of collagen fibrils de novo by cleavage of the type I pC-collagen with procollagen C-proteinase. Assay of critical concentration demonstrates that collagen self-assembly is a classical example of an entropy-driven process. *J. Biol. Chem.* 260:15696–15701.
13. Brown, K. L., C. F. Cummings, …, B. G. Hudson. 2017. Building collagen IV smart scaffolds on the outside of cells. *Protein Sci.* 26:2151–2161.
14. Cummings, C. F., V. Pedchenko, …, B. G. Hudson. 2016. Extracellular chloride signals collagen IV network assembly during basement membrane formation. *J. Cell Biol.* 213:479–494.

15. Hwang, E. S., and B. Brodsky. 2012. Folding delay and structural perturbations caused by type IV collagen natural interruptions and nearby Gly missense mutations. *J. Biol. Chem.* 287:4368–4375.
16. Bella, J. 2014. A first census of collagen interruptions: collagen's own stutters and stammers. *J. Struct. Biol.* 186:438–450.
17. Hofmann, H., T. Voss, …, J. Engel. 1984. Localization of flexible sites in thread-like molecules from electron micrographs. Comparison of interstitial, basement membrane and intima collagens. *J. Mol. Biol.* 172:325–343.
18. Lunstrum, G. P., H. P. Bächinger, …, J. H. Fessler. 1988. Drosophila basement membrane procollagen IV. I. Protein characterization and distribution. *J. Biol. Chem.* 263:18318–18327.
19. Bächinger, H. P., K. J. Doege, …, J. H. Fessler. 1982. Structural implications from an electronmicroscopic comparison of procollagen V with procollagen I, pC-collagen I, procollagen IV, and a Drosophila procollagen. *J. Biol. Chem.* 257:14590–14592.
20. Bächinger, H. P., N. P. Morris, …, R. E. Burgeson. 1990. The relationship of the biophysical and biochemical characteristics of type VII collagen to the function of anchoring fibrils. *J. Biol. Chem.* 265:10095–10101.
21. Thiagarajan, G., Y. Li, …, B. Brodsky. 2008. Common interruptions in the repeating tripeptide sequence of non-fibrillar collagens: sequence analysis and structural studies on triple-helix peptide models. *J. Mol. Biol.* 376:736–748.
22. Hwang, E. S., G. Thiagarajan, …, B. Brodsky. 2010. Interruptions in the collagen repeating tripeptide pattern can promote supramolecular association. *Protein Sci.* 19:1053–1064.
23. Bella, J. 2016. Collagen structure: new tricks from a very old dog. *Biochem. J.* 473:1001–1025.
24. Orgel, J. P. R. O., T. C. Irving, …, T. J. Wess. 2006. Microfibrillar structure of type I collagen in situ. *Proc. Natl. Acad. Sci. USA.* 103:9001–9005.
25. Antipova, O., and J. P. R. O. Orgel. 2010. In situ D-periodic molecular structure of type II collagen. *J. Biol. Chem.* 285:7087–7096.
26. Bella, J., M. Eaton, …, H. M. Berman. 1994. Crystal and molecular structure of a collagen-like peptide at 1.9 Å resolution. *Science.* 266:75–81.
27. Golbik, R., J. A. Eble, …, K. Kühn. 2000. The spatial orientation of the essential amino acid residues arginine and aspartate within the $\alpha1\beta1$ integrin recognition site of collagen IV has been resolved using fluorescence resonance energy transfer. *J. Mol. Biol.* 297:501–509.
28. Kleinman, H. K., M. L. McGarvey, …, G. R. Martin. 1982. Isolation and characterization of type IV procollagen, laminin, and heparan sulfate proteoglycan from the EHS sarcoma. *Biochemistry.* 21:6188–6193.
29. Vandenberg, P., A. Kern, …, K. Kühn. 1991. Characterization of a type IV collagen major cell binding site with affinity to the alpha 1 beta 1 and the alpha 2 beta 1 integrins. *J. Cell Biol.* 113:1475–1483.
30. Timpl, R., R. W. Glanville, …, K. Kühn. 1975. Isolation, chemical and electron microscopical characterization of neutral-salt-soluble type III collagen and procollagen from fetal bovine skin. *Hoppe Seylers Z. Physiol. Chem.* 356:1783–1792.
31. Schneider, M., A. Al-Shaer, and N. R. Forde. 2021. AutoSmarTrace: automated chain tracing and flexibility analysis of biological filaments. *Biophys. J.* 120:2599–2608.
32. The MathWorks, Inc. 2018. MATLAB and Statistical Toolbox Release. 2018b. The MathWorks, Inc, Natick, MA.
33. UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47:D506–D515.
34. Kühn, K. 1995. Basement membrane (type IV) collagen. *Matrix Biol.* 14:439–445.
35. Parkin, J. D., J. D. San Antonio, …, J. Savige. 2011. Mapping structural landmarks, ligand binding sites, and missense mutations to the collagen IV heterotrimers predicts major functional domains, novel interactions, and variation in phenotypes in inherited diseases affecting basement membranes. *Hum. Mutat.* 32:127–143.
36. Yurchenco, P. D., and H. Furthmayr. 1984. Self-assembly of basement membrane collagen. *Biochemistry.* 23:1839–1850.
37. Charonis, A. S., E. C. Tsilibary, …, H. Furthmayr. 1985. Binding of laminin to type IV collagen: a morphological study. *J. Cell Biol.* 100:1848–1853.
38. Laurie, G. W., J. T. Bing, …, R. J. Feldmann. 1986. Localization of binding sites for laminin, heparan sulfate proteoglycan and fibronectin on basement membrane (type IV) collagen. *J. Mol. Biol.* 189:205–216.
39. Tsilibary, E. C., and A. S. Charonis. 1986. The role of the main non-collagenous domain (NC1) in type IV collagen self-assembly. *J. Cell Biol.* 103:2467–2473.
40. Yurchenco, P. D., and G. C. Ruben. 1987. Basement membrane structure in situ: evidence for lateral associations in the type IV collagen network. *J. Cell Biol.* 105:2559–2568.
41. Aumailley, M., H. Wiedemann, …, R. Timpl. 1989. Binding of nidogen and the laminin-nidogen complex to basement membrane collagen type IV. *Eur. J. Biochem.* 184:241–248.
42. Fox, J. W., U. Mayer, …, M. L. Chu. 1991. Recombinant nidogen consists of three globular domains and mediates binding of laminin to collagen type IV. *EMBO J.* 10:3137–3146.
43. Lovelady, H. H., S. Shashidhara, and W. G. Matthews. 2014. Solvent specific persistence length of molecular type I collagen. *Biopolymers.* 101:329–335.
44. Pedchenko, V., R. Bauer, …, S. P. Boudko. 2019. A chloride ring is an ancient evolutionary innovation mediating the assembly of the collagen IV scaffold of basement membranes. *J. Biol. Chem.* 294:7968–7981.
45. Pedchenko, V., S. P. Boudko, …, B. G. Hudson. 2021. Collagen IV$^{\alpha345}$ dysfunction in glomerular basement membrane diseases. III. A functional framework for $\alpha345$ hexamer assembly. *J. Biol. Chem.* 296:100592.
46. Chen, C. H., and H. G. Hansma. 2000. Basement membrane macromolecules: insights from atomic force microscopy. *J. Struct. Biol.* 131:44–55.
47. Rivetti, C., M. Guthold, and C. Bustamante. 1996. Scanning force microscopy of DNA deposited onto mica: equilibration versus kinetic trapping studied by statistical polymer chain analysis. *J. Mol. Biol.* 264:919–932.
48. Heenan, P. R., and T. T. Perkins. 2019. Imaging DNA equilibrated onto mica in liquid using biochemically relevant deposition conditions. *ACS Nano.* 13:4220–4229.
49. Davis, J. M., B. A. Boswell, and H. P. Bächinger. 1989. Thermal stability and folding of type IV procollagen and effect of peptidyl-prolyl cis-trans-isomerase on the folding of the triple helix. *J. Biol. Chem.* 264:8956–8962.
50. Shayegan, M., T. Altindal, …, N. R. Forde. 2016. Intact telopeptides enhance interactions between collagens. *Biophys. J.* 111:2404–2416.
51. Li, H., W. A. Linke, …, J. M. Fernandez. 2002. Reverse engineering of the giant muscle protein titin. *Nature.* 418:998–1002.
52. Rivetti, C., C. Walker, and C. Bustamante. 1998. Polymer chain statistics and conformational analysis of DNA molecules with bends or sections of different flexibility. *J. Mol. Biol.* 280:41–59.
53. Todd, B. A., J. Rammohan, and S. J. Eppell. 2003. Connecting nanoscale images of proteins with their genetic sequences. *Biophys. J.* 84:3982–3991.
54. Valdman, D., B. J. Lopez, …, P. J. Atzberger. 2013. Force spectroscopy of complex biopolymers with heterogeneous elasticity. *Soft Matter.* 9:772–778.
55. Brazel, D., R. Pollner, …, K. Kühn. 1988. Human basement membrane collagen (type IV). The amino acid sequence of the alpha 2(IV) chain and its comparison with the alpha 1(IV) chain reveals deletions in the alpha 1(IV) chain. *Eur. J. Biochem.* 172:35–42.
56. Engel, J., and D. J. Prockop. 1991. The zipper-like folding of collagen triple helices and the effects of mutations that disrupt the zipper. *Annu. Rev. Biophys. Biophys. Chem.* 20:137–152.

57. Orgel, J. P. R. O., A. V. Persikov, and O. Antipova. 2014. Variation in the helical structure of native collagen. *PLoS One.* 9:e89519.

58. Bann, J. G., D. H. Peyton, and H. P. Bächinger. 2000. Sweet is stable: glycosylation stabilizes collagen. *FEBS Lett.* 473:237–240.

59. Perdivara, I., M. Yamauchi, and K. B. Tomer. 2013. Molecular characterization of collagen hydroxylysine O-glycosylation by mass spectrometry: current status. *Aust. J. Chem.* 66:760–769.

60. Tang, M., X. Wang, …, Y. Gu. 2020. Effect of hydroxylysine-O-glycosylation on the structure of type I collagen molecule: a computational study. *Glycobiology.* 30:830–843.

61. Basak, T., L. Vega-Montoto, …, R. M. Vanacore. 2016. Comprehensive characterization of glycosylation and hydroxylation of basement membrane collagen IV by high-resolution mass spectrometry. *J. Proteome Res.* 15:245–258.

62. Liotta, L. A., S. Abe, …, G. R. Martin. 1979. Preferential digestion of basement membrane collagen by an enzyme derived from a metastatic murine tumor. *Proc. Natl. Acad. Sci. USA.* 76:2268–2272.

63. Srinivasan, M., S. G. M. Uzel, …, M. J. Buehler. 2009. Alport syndrome mutations in type IV tropocollagen alter molecular structure and nanomechanical properties. *J. Struct. Biol.* 168:503–510.

64. Wu, Y.-Y., L. Bao, …, Z.-J. Tan. 2015. Flexibility of short DNA helices with finite-length effect: from base pairs to tens of base pairs. *J. Chem. Phys.* 142:125103.

65. Bächinger, H. P., P. Bruckner, …, J. Engel. 1980. Folding mechanism of the triple helix in type-III collagen and type-III pN-collagen. Role of disulfide bridges and peptide bond isomerization. *Eur. J. Biochem.* 106:619–632.

66. Bhatnagar, R. S., C. A. Gough, …, M. B. Shattuck. 1999. Fine structure of collagen: molecular mechanisms of the interactions of collagen. *Proc. Indian Acad. Sci. Chem. Sci.* 111:301–317.

67. Holmes, D. F., A. P. Mould, and J. A. Chapman. 1991. Morphology of sheet-like assemblies of pN-collagen, pC-collagen and procollagen studied by scanning transmission electron microscopy mass measurements. *J. Mol. Biol.* 220:111–123.

68. Bruckner, P., H. P. Bächinger, …, J. Engel. 1978. Three conformationally distinct domains in the amino-terminal segment of type III procollagen and its rapid triple helix leads to and comes from coil transition. *Eur. J. Biochem.* 90:595–603.

69. Bhatnagar, R. S., J. J. Qian, and C. A. Gough. 1997. The role in cell binding of a β-bend within the triple helical region in collagen α1 (I) chain: structural and biological evidence for conformational tautomerism on fiber surface. *J. Biomol. Struct. Dyn.* 14:547–560.

70. Farndale, R. W., T. Lisman, …, N. Raynal. 2008. Cell-collagen interactions: the use of peptide toolkits to investigate collagen-receptor interactions. *Biochem. Soc. Trans.* 36:241–250.

71. Lisman, T., N. Raynal, …, R. W. Farndale. 2006. A single high-affinity binding site for von Willebrand factor in collagen III, identified using synthetic triple-helical peptides. *Blood.* 108:3753–3756.

72. Lauer-Fields, J. L., D. Juska, and G. B. Fields. 2002. Matrix metalloproteinases and collagen catabolism. *Biopolymers.* 66:19–32.

73. Ravikumar, K. M., and W. Hwang. 2008. Region-specific role of water in collagen unwinding and assembly. *Proteins.* 72:1320–1332.

74. Miller, E. J., J. E. Finch, Jr., …, P. B. Robertson. 1976. Specific cleavage of the native type III collagen molecule with trypsin. Similarity of the cleavage products to collagenase-produced fragments and primary structure at the cleavage site. *Arch. Biochem. Biophys.* 173:631–637.

75. Williams, K. E., and D. R. Olsen. 2009. Matrix metalloproteinase-1 cleavage site recognition and binding in full-length human type III collagen. *Matrix Biol.* 28:373–379.

76. Kirkness, M. W. H., and N. R. Forde. 2018. Single-molecule assay for proteolytic susceptibility: force-induced collagen destabilization. *Biophys. J.* 114:570–576.

77. Fields, G. B. 1991. A model for interstitial collagen catabolism by mammalian collagenases. *J. Theor. Biol.* 153:585–602.

78. Stultz, C. M. 2002. Localized unfolding of collagen explains collagenase cleavage near imino-poor sites. *J. Mol. Biol.* 319:997–1003.

79. Glanville, R. W., T. Voss, and K. Kühn. 1982. A comparison of the flexibility of molecules of basement membrane and interstitial collagens. *In* New Trends in Basement Membrane Research K. Kuehn, H. Schoene, and R. Timpl, eds.. Raven Press, pp. 69–77.

80. Chow, W. Y., C. J. Forman, …, M. J. Duer. 2018. Proline provides site-specific flexibility for in vivo collagen. *Sci. Rep.* 8:13809.

81. Weis, M. A., D. M. Hudson, …, D. R. Eyre. 2010. Location of 3-hydroxyproline residues in collagen types I, II, III, and V/XI implies a role in fibril supramolecular assembly. *J. Biol. Chem.* 285:2580–2590.

82. Montgomery, N. T., K. D. Zientek, …, H. P. Bächinger. 2018. Post-translational modification of type IV collagen with 3-hydroxyproline affects its interactions with glycoprotein VI and nidogens 1 and 2. *J. Biol. Chem.* 293:5987–5999.

83. Taga, Y., K. Tanaka, …, K. Mizuno. 2021. In-depth correlation analysis demonstrates that 4-hydroxyproline at the Yaa position of Gly-Xaa-Yaa repeats dominantly stabilizes collagen triple helix. *Matrix Biol. Plus.* 10:100067.

84. Berg, R. A., and D. J. Prockop. 1973. The thermal transition of a non-hydroxylated form of collagen. Evidence for a role for hydroxyproline in stabilizing the triple-helix of collagen. *Biochem. Biophys. Res. Commun.* 52:115–120.

85. Burjanadze, T. V. 1979. Hydroxyproline content and location in relation to collagen thermal stability. *Biopolymers.* 18:931–938.

86. Uzel, S. G. M., and M. J. Buehler. 2009. Nanomechanical sequencing of collagen: tropocollagen features heterogeneous elastic properties at the nanoscale. *Integr. Biol.* 1:452–459.

87. Gautieri, A., S. Vesentini, …, M. J. Buehler. 2009. Single molecule effects of osteogenesis imperfecta mutations in tropocollagen protein domains. *Protein Sci.* 18:161–168.

88. Yeo, J., Y. Qiu, …, D. L. Kaplan. 2020. Adverse effects of Alport syndrome-related Gly missense mutations on collagen type IV: insights from molecular simulations and experiments. *Biomaterials.* 240:119857.

89. Malhotra, V., and P. Erlmann. 2015. The pathway of collagen secretion. *Annu. Rev. Cell Dev. Biol.* 31:109–124.

90. Stauber, T., and T. J. Jentsch. 2013. Chloride in vesicular trafficking and function. *Annu. Rev. Physiol.* 75:453–477.

91. Dölz, R., J. Engel, and K. Kühn. 1988. Folding of collagen IV. *Eur. J. Biochem.* 178:357–366.

92. Weber, S., J. Engel, …, R. Timpl. 1984. Subunit structure and assembly of the globular domain of basement-membrane collagen type IV. *Eur. J. Biochem.* 139:401–410.

93. Wood, G. C., and M. K. Keech. 1960. The formation of fibrils from collagen solutions. 1. The effect of experimental conditions: kinetic and electron-microscope studies. *Biochem. J.* 75:588–598.

94. Harris, J. R., A. Soliakov, and R. J. Lewis. 2013. In vitro fibrillogenesis of collagen type I in varying ionic and pH conditions. *Micron.* 49:60–68.

95. Vanacore, R., A.-J. L. Ham, …, B. G. Hudson. 2009. A sulfilimine bond identified in collagen IV. *Science.* 325:1230–1234.

96. McCaughey, J., and D. J. Stephens. 2019. ER-to-Golgi transport: a sizeable problem. *Trends Cell Biol.* 29:940–953.

97. Raote, I., M. Ortega-Bellido, …, V. Malhotra. 2018. TANGO1 builds a machine for collagen export by recruiting and spatially organizing COPII, tethers and membranes. *eLife.* 7:e32723.

98. Matsui, Y., Y. Hirata, …, N. Hosokawa. 2020. Visualization of procollagen IV reveals ER-to-Golgi transport by ERGIC-independent carriers. *Cell Struct. Funct.* 45:107–119.

99. McCaughey, J., N. L. Stevenson, …, D. J. Stephens. 2019. ER-to-Golgi trafficking of procollagen in the absence of large carriers. *J. Cell Biol.* 218:929–948.

100. Freudenberg, U., S. H. Behrens, …, C. Werner. 2007. Electrostatic interactions modulate the conformation of collagen I. *Biophys. J.* 92:2108–2119.

101. Keshwani, N., S. Banerjee, …, G. I. Makhatadze. 2013. The role of cross-chain ionic interactions for the stability of collagen model peptides. *Biophys. J.* 105:1681–1688.

102. Hohenester, E., T. Sasaki, …, H. P. Bächinger. 2008. Structural basis of sequence-specific collagen recognition by SPARC. *Proc. Natl. Acad. Sci. USA.* 105:18273–18277.

103. Chioran, A., S. Duncan, …, M. J. Ringuette. 2017. Collagen IV trafficking: the inside-out and beyond story. *Dev. Biol.* 431:124–133.

104. Köhler, A., M. Mörgelin, …, G. Sengle. 2020. New specific HSP47 functions in collagen subfamily chaperoning. *FASEB J.* 34:12040–12052.

105. Landau, L. D., L. P. Pitaevskii, …, E. M. Lifshitz. 1986. Theory of Elasticity. Butterworth-Heinemann, Oxford, UK.

106. Krishnamoorthy, K. 2016. Handbook of Statistical Distribution with Applications, Second Edition. CRC Press, Boca Raton, FL.

107. Gelman, A. 2013. Bayesian Data Analysis, Third Edition. CRC Press, Boca Raton, FL.

Biophysical Journal, Volume *120*

Supplemental information

Sequence-dependent mechanics of collagen reflect its structural and functional organization

Alaa Al-Shaer, Aaron Lyons, Yoshihiro Ishikawa, Billy G. Hudson, Sergei P. Boudko, and Nancy R. Forde

Supporting Information for

Sequence-dependent mechanics of collagen reflect its structural and functional organization

Alaa Al-Shaer, Aaron Lyons, Yoshihiro Ishikawa, Billy G. Hudson, Sergei P. Boudko and Nancy R. Forde

Contents

Supporting Text 1

Inhomogeneous Worm-like Chain Theory and Application

Consider a section of a two-dimensional worm-like chain with length Δs, broken up into n infinitesimal segments of length $\delta s = \frac{\Delta s}{n}$. The energy required to bend segment i into a circular arc is given by

$$E_i = \frac{\alpha_i \delta\theta_i^2}{2\delta s} = \frac{p_i \delta\theta_i^2}{2\delta s} k_\mathrm{B} T. \tag{S1}$$

α_i is the bending rigidity of the segment, related to the persistence length p_i through $p_i = \alpha_i / k_\mathrm{B} T$, where $k_\mathrm{B} T$ is the product of the Boltzmann constant and the absolute temperature (1). The central angle of this arc, $\delta\theta_i$, is equivalent to the angle between the tangents at the beginning and end of the segment. At equilibrium, the distribution of angles $\delta\theta_i$ is given by the Boltzmann distribution, where

$$P(\delta\theta_i) = \sqrt{\frac{p_i}{2\pi\delta s}} \exp\left(-\frac{p_i \delta\theta_i^2}{2\delta s}\right). \tag{S2}$$

This is a normal distribution with $\langle \delta\theta_i \rangle = 0$ and variance $\sigma^2_{\delta\theta_i} = \frac{\delta s}{p_i}$.

Since each $\delta\theta_i$ is a signed angle, the total angle adopted over the total section length Δs is given by $\theta = \sum_{i=1}^{n} \delta\theta_i$. Using the fact that each $\delta\theta_i$ is governed by an independent normal distribution, the sum of each of these angles will also be normally distributed, and thus will have mean $\langle \theta \rangle = \sum_{i=1}^{n} \langle \delta\theta_i \rangle = 0$ and variance $\sigma_\theta^2 = \sum_{i=1}^{n} \sigma^2_{\delta\theta_i} = \sum_{i=1}^{n} \frac{\delta s}{p_i}$. Taking the limit as $n \to \infty$, we therefore find that

$$\langle \theta \rangle = 0 \tag{S3}$$

and

$$\sigma_\theta^2 = \int_0^{\Delta s} \frac{ds\prime}{p(s\prime)}. \tag{S4}$$

Here, $p(s')$ is the persistence length at position $0 \leq s' \leq \Delta s$ along the contour of the chain segment.

Equation (S4) has strong implications for the experimental determination of persistence length profiles: because estimates of persistence length require measurements over a finite segment of length Δs, variations in persistence length within this length cannot be determined. Instead, the angular variance measured across the segment is

$$\sigma_\theta^2 = \frac{\Delta s}{p^*}, \tag{S5}$$

where p^* is the effective persistence length, taken to be uniform along the segment. Thus, for a given segment length Δs, the effective persistence length is calculated as

$$p^* = \frac{\Delta s}{\sigma_\theta^2} = \frac{\Delta s}{\int_0^{\Delta s} \frac{ds\prime}{p(s\prime)}}. \tag{S6}$$

For polymers that have sharp changes in persistence length within their contours (e.g. collagen type IV, DNA with short single stranded regions), this effective persistence length will be biased

towards the smaller values present in the measured segment. Consider the persistence length profile shown in Figure S1: the original persistence length profile $p(s)$ (shown in black) consists of five segments with 85 nm persistence lengths, interspaced with short, variable length sections with 5 nm persistence lengths. The effective persistence length profile (shown in red) calculated using Eq. (S6) is also plotted, using a measurement length of 30 nm. Lastly, for comparison purposes only, a simple average of the persistence length is plotted, calculated by averaging $p(s)$ over the same 30 nm window (shown in blue). We note that this average persistence length is *not* the correct way to view the effects of heterogeneous flexibility; rather, the effective persistence length (red) is expected, based on the angular flexibility of the 30-nm segments. In the main text, it is shown that the effective persistence length profile agrees well with that extracted from tracing images of simulated chains.

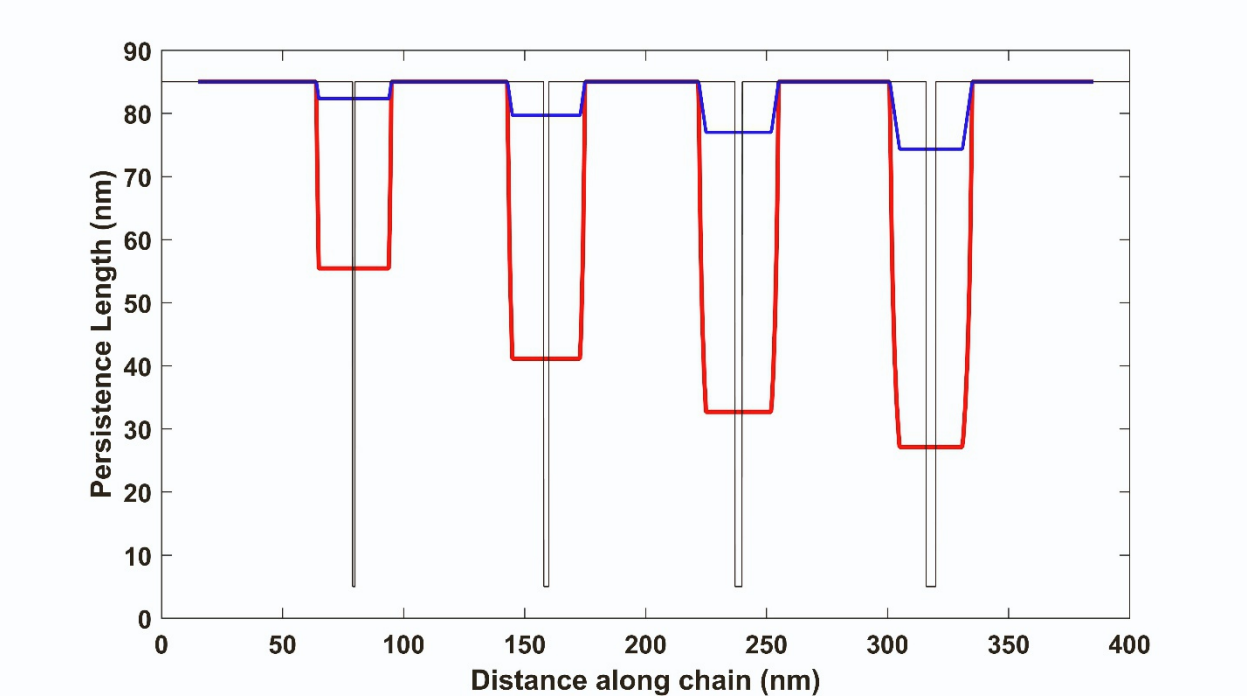

Figure S1. Differences among theoretical persistence length profiles. The original persistence length profile $p(s)$ is shown in black, while the red line shows the effective persistence length profile $p^*(s)$ calculated using Eq. (S6) and centered at the middle of the filter window. The average value of the original persistence length map over this 30 nm window is shown in blue for comparison.

Statistical Properties of Persistence Length Estimates

The persistence length is calculated using the variance of the angular distribution, which we can rewrite as follows, since $\langle\theta\rangle = 0$:

$$p = \frac{\Delta s}{\langle\theta^2\rangle}. \tag{S7}$$

For a finite number of observations, the quantity $\langle\theta^2\rangle$ is an estimate of the true variance, and can be represented by a random variable S^2 such that

$$S^2 = \frac{\Delta s}{p}\sum_{i=1}^{n}\frac{X_i^2}{n}, \tag{S8}$$

where n is the number of observations and each random variable X_i is drawn from a normal distribution with a mean of zero and variance of 1. S^2 can therefore be written in terms of the χ^2 distribution as (2)

$$S^2 = \frac{\Delta s}{np}\chi_n^2. \tag{S9}$$

From this angular variance estimate, an estimate of the true value of the persistence length can be made. This estimate of p is represented by a random variable, P, such that

$$P = \frac{\Delta s}{S^2} = \frac{\Delta s}{\left(\frac{\Delta s}{np}\chi_n^2\right)} = np\chi_n^{-2}. \tag{S10}$$

Here, χ_n^{-2} denotes the inverse chi-squared-distributed random variable. We can incorporate the p and n terms into this random variable, yielding

$$P \sim \text{Scale-Inv-}\chi^2(n,p), \tag{S11}$$

where Scale-Inv-$\chi^2(n,p)$ is a scaled inverse chi-squared-distributed random variable. This particular distribution has the probability density function (PDF) (3)

$$f(x;n,p) = \left(\frac{np}{2}\right)^{n/2}\frac{\exp\left(\frac{-pn}{2x}\right)}{x^{\frac{n}{2}+1}\Gamma\left(\frac{n}{2}\right)}, \tag{S12}$$

where $\Gamma\left(\frac{n}{2}\right) = \int_0^\infty t^{\frac{n}{2}-1}\mathrm{e}^{-t}\mathrm{d}t$ is the standard gamma function and x represents the realization of P. The PDF described by equation (S12) is plotted in Figure S2 (in blue) for the case of $n = 100$ and $p = 85$ nm. The probability density is roughly centered around 85 nm (the actual persistence length), but estimates of the persistence length are more likely to be overestimated than underestimated.

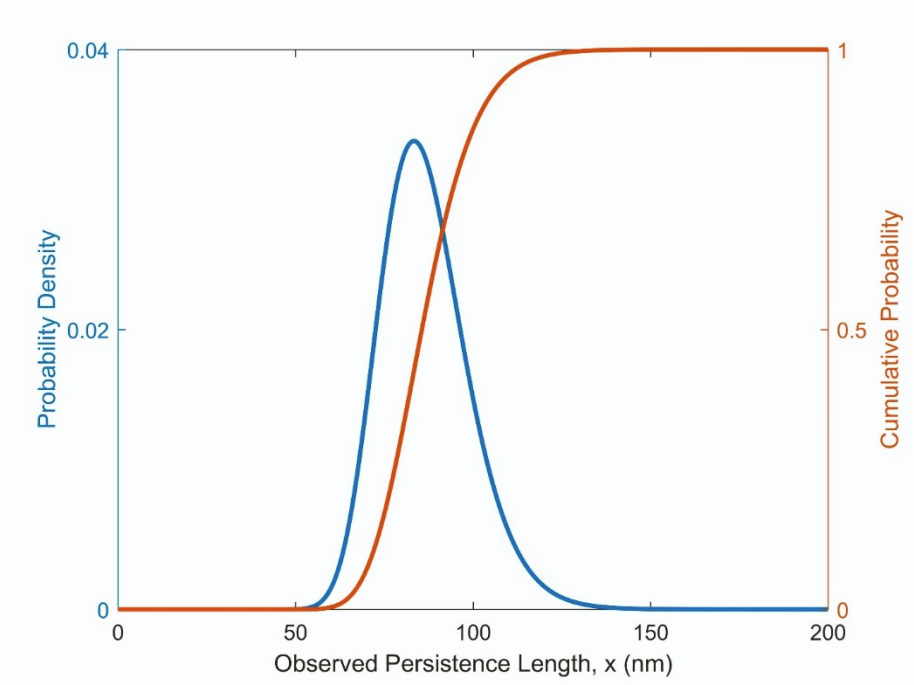

Figure S2. Probability density of persistence length estimates, for p = 85 nm and n = 100 chains. The blue line shows the expected distribution of persistence length estimates, obtained from equation (S12). The red line shows the cumulative probability of these estimates, given by equation (S13).

To calculate the confidence intervals on estimates of persistence length, we use the cumulative density function $F(x;n,p) = \int_0^x f(t;n,p)dt$. This function is given by

$$F(x;n,p) = \frac{1}{\Gamma\left(\frac{n}{2}\right)}\int_0^x \frac{1}{t}\left(\frac{pn}{2t}\right)^{n/2} e^{-\left(\frac{pn}{2t}\right)}dt = \frac{1}{\Gamma\left(\frac{n}{2}\right)}\int_{\frac{pn}{2x}}^{\infty} u^{\frac{n}{2}-1}e^{-u}du = \frac{\Gamma\left(\frac{n}{2},\frac{pn}{2x}\right)}{\Gamma\left(\frac{n}{2}\right)}, \quad \text{(S13)}$$

where the substitution $u = \frac{pn}{2t}$ was used to simplify the integral, and $\Gamma\left(\frac{n}{2},\frac{pn}{2x}\right) = \int_{\frac{pn}{2x}}^{\infty} t^{\frac{n}{2}-1}\mathrm{e}^{-t}\mathrm{d}t$ is the upper incomplete gamma function. This is plotted in red in Figure S2, for the same case of $n = 100$ and $p = 85$ nm.

Finally, we determine how the mean and 95% confidence intervals of the observed persistence length (the realization of P) depend on sample number. The mean value of the persistence length estimate is given by (3)

$$\langle P\rangle = p\left(\frac{n}{n-2}\right). \quad \text{(S14)}$$

This means that the observed persistence length will, on average, be overestimated relative to the true value for small n. To calculate the confidence intervals, the cumulative density function can be numerically inverted to find x at $F(x;n,p) = 0.975$ and $F(x;n,p) = 0.025$ to extract the bounds on the 95% confidence interval. A plot of these values for $p = 85$ nm and $n = 100$ is shown in Figure S3.

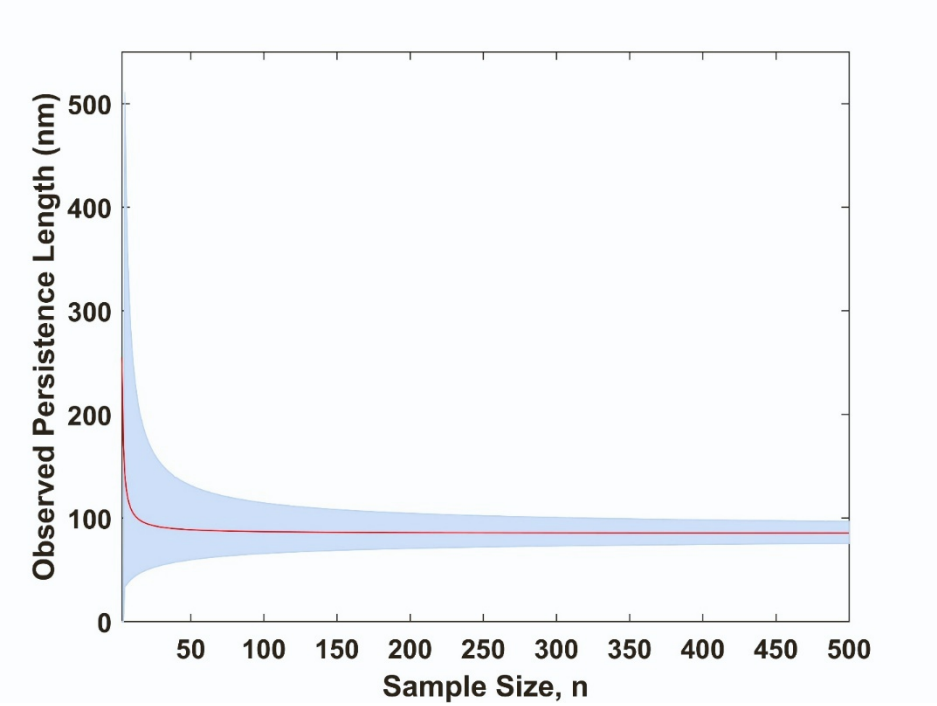

Figure S3. Dependence of persistence length estimate and 95% confidence interval on sample size. These values correspond to chains with persistence length p = 85 nm. The mean value is obtained using equation (S14) and the error bounds are determined from the cumulative probability distribution equation (S13) as described in the text.

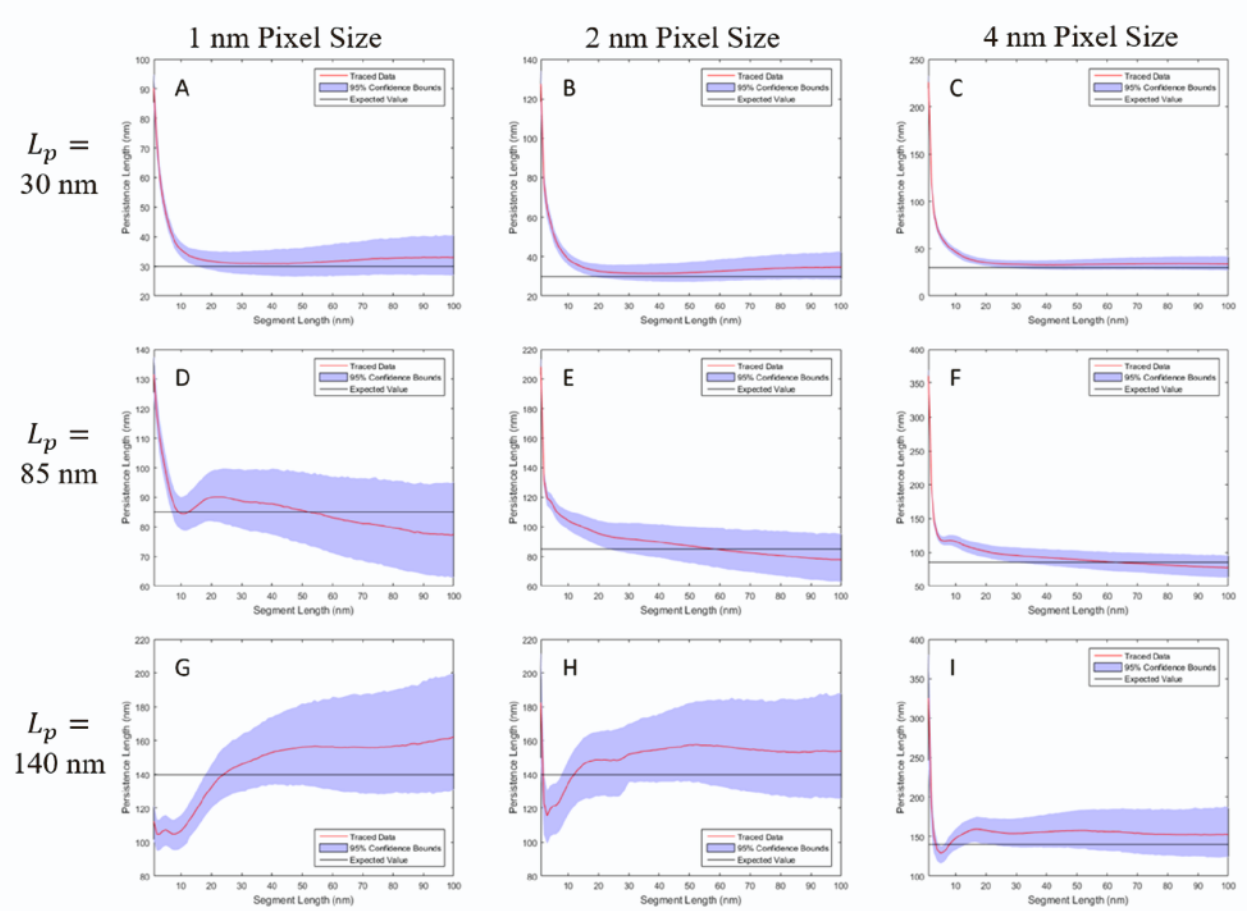

Figure S4. Effective persistence length vs. segment length. The effective persistence length p^* was determined using Equation S5, from traced, simulated chains of contour length L=300 nm and with different, uniform values of bending stiffness, given by persistence lengths of 30 nm (N=73 chains), 85 nm (N=78 chains), and 140 nm (N=76 chains) for the top, middle and bottom rows, respectively. Three different pixel sizes for the simulated images were tested, with the 4 nm pixel size (right column) representing that used in the experimental scans of this work. Segment lengths of $\Delta s \geq 30$ nm recover the input persistence length within error. Error range is shown as 95% confidence interval.

(α1) SP – KGDCGGSGCGKCDCHGV 44
(α2) SP – LLAQSVLGGVKKLDVPCGGRDCSGGCQCYPEKGARGQPGAVGPQGYNGPPGLQGFPGLQ 84

KGQKGERGLPGLQGVIGFPGMQGPEGPHGPPGQKGDAGEPGLPGTKGTRGPPGAAGYPGNPGLPG 109
GRKGDKGERGVPGPTGPKGDVGARGVSGFPGADGIPGHPGQGGPRGRPGYDGCNGTRGDAGPQGP 149

IPGQDGPPGPPGIPGCNGTKGERGPLGPPGLPGFSGNPGPPGLPGMKGDPGEILGHVPGTLLKGE 174
SGSGGFPGLPGPQGPKGQKGEPYALSKEDRDKYRGEPGEPGLVYQGPPGRPGPIGQMGPMGAPGG 214

RGFPGIPGMPGSPGLPGLQGPVGPPGFTGPPGPPGPPGPPGEKGQMGSSFQGPKGDKGEQGVSGP 239
RPGPPGPPGPKGQPGNRGLGFYGQKGEKGDIGQPGPNGIPSDITLVGPTTSTIHPDLYKGEKGDE 279

PGVPGQAQVKEKGDFAPTGEKGQKGEPGFPGVPGYGEKGEPGKQGPRGKPGKDGEKGERGSPGIP 304
GEQGIPGVISKGEEGIMGFPGIRGFPGLDGEKGVVGQKGSRGLDGFQGPSGPRGPKGERGEQGPP 344

GDSGYPGLPGRQGPQGEKGEAGLPGPPGTVIGTMPLGEKGDRGYPGAPGLRGEPGPKGFPGTPGQ 369
GPSVYSPHPSLAKGARGDPGFQGAHGEPGSRGEPGEPGTAGPPGPSVGDEDSMRGLPGEMGPKGF 409

PGPPGFPTPGQAGAPGFPGERGEKGDQGFPGVSLPGPSGRDGAPGPPGPPGPPGQPGHTNGIVEC 434
SGEPGSPARYLGPPGADGRPGPQGVPGPAGPPGPDGFLFGLKGSEGRVGYPGPSGFPGTRGQKGW 474

QPGPPGDQGPPGTPGQPGLTGEVGQKGQKGESCLACDTEGLRGPPGPQGPPGEIGFPGQPGAKGD 499
KGEAGDCQCGQVIGGLPGLPGPKGFPGVNGELGKKGDQGDPGLHGIPGFPGFKGAPGVAGAPGPK 539

RGLPGRDGLEGLPGPQGSPGLIGQPGAKGEPGEIFFDMRLKGDKGDPGFPGQPGMPGRAGTPGRD 564
GIKGDSRTITTKGERGQPGIPGVHGMKGDDGVPGRDGLDGFPGLPGPPGDGIKGPPGDAGLPGVP 604

GHPGLPGPKGSPGSIGLKGERGPPGGVGFPGSRGDIGPPGPPGVGPIGPVGEKGQAGFPGGPGSP 629
GTKGFPGDIGPPGQGLPGPKGERGFPGDAGLPGPPGFPGPPGPPGTPGQRDCDTGVKRPIGGGQQ 669
Loop 1
GLPGPKGEAGKVVPLPGPPGAAGLPGSPGFPGPQGDRGFPGTPGRPGIPGEKGAVGQPGIGFPGL 694
VVVQPGCIEGPTGSPGQPGPPGPTGAKGVRGMPGFPGASGEQGLKGFPGDPGREGFPGPPGFMGP 734

PGPKGVDGLPGEIGRPGSPGRPGFNGLPGNPGPQGQKGEPGIGLPGLKGQPGLPGIPGTPGEKGS 759
RGSKGTTGLPGPDGPPGPIGLPGPAGPPGDRGIPGEVLGAQPGTRGDAGLPGQPGLKGLPGETGA 799
Loop 2
IGGPGVPGEQGLTGPPGLQGIRGDPGPPGVQGPAGPPGVPGIGPPGAMGPPGGQGPPGSSGPPGI 824
PGFRGSQGMPGMPGLKGQPGFPGPSGQPGQSGPPGQHGFPGTPGREGPLGQPGSPGLGGLPGDRG 864

KGEKGFPGFPGLDMPGPKGDKGSQGLPGLTGQSGLPGLPGQQGTPGVPGFPGSKGEMGVMGTPGQ 889
EPGDPGVPGPVGMKGLSGDRGDAGMSGERGHPGSPGFKGMAGMPGIPGQKGDRGSPGMDGFQGML 929

PGSPGPAGTPGLPGEKGDHGLPGSSGPRGDPGFKGDKGDVGLPGMPGSMEHVDMGSMKGQKGDQG 954
GLKGRQGFPGTKGEAGFFGVPGLKGLPGEPGVKGNRGDRGPPGPPPLILPGMKDIKGEKGDEGPM 994

EKGQIGPTGDKGSRGDPGTPGVPGKDGQAGHPGQPGPKGDPGLSGTPGSPGLPGPKGSVGGMGLP 1019
GLKGYLGLKGIQGMPGVPGVSGFPGLPGRPGFIKGVKGDIGVPGTPGLPGFPGVSGPPGITGFPG 1059

GSPGEKGVPGIPGSQGVPGSPGEKGAKGEKGQSGLPGIGIPGRPGDKGDQGLAGFPGSPGEKGEK 1084
FTGSRGEKGTPGVAGVFGETGPTGDFGDIGDTVDLPGSPGLKGERGITGIPGLKGFFGEKGAAGD 1124

Figure S5. Amino acid sequences of the collagen IV collagenous domain. The α1 (P02463, in blue) and α2 (P08122, in red) amino acid sequences were aligned at the first Gly-X-Y overlap from the NC1 domain, as the assembly of collagen is initiated at that end. The signal peptide sequences for both chains are not displayed in this alignment, but are denoted by SP at the start of the sequence. The underlined portions at the beginning of the sequences correspond to the 7S-forming regions of collagen IV. The highlighted segments are interruptions in the triple-helical-defining $(Gly\text{-}X\text{-}Y)_nG$ amino acid sequence. The triple-helical regions have been defined as $(Gly\text{-}X\text{-}Y)_nG$ with $n \geq 4$. The longest interruption, 26 amino acids long in α2, has two cysteines (bolded in black) near its edges, and is treated as loop 1 in the "two loops" sequence alignments. A second interruption in α2 ("loop 2"; centered at LGA) is also removed from the backbone contour in the two loops alignments.

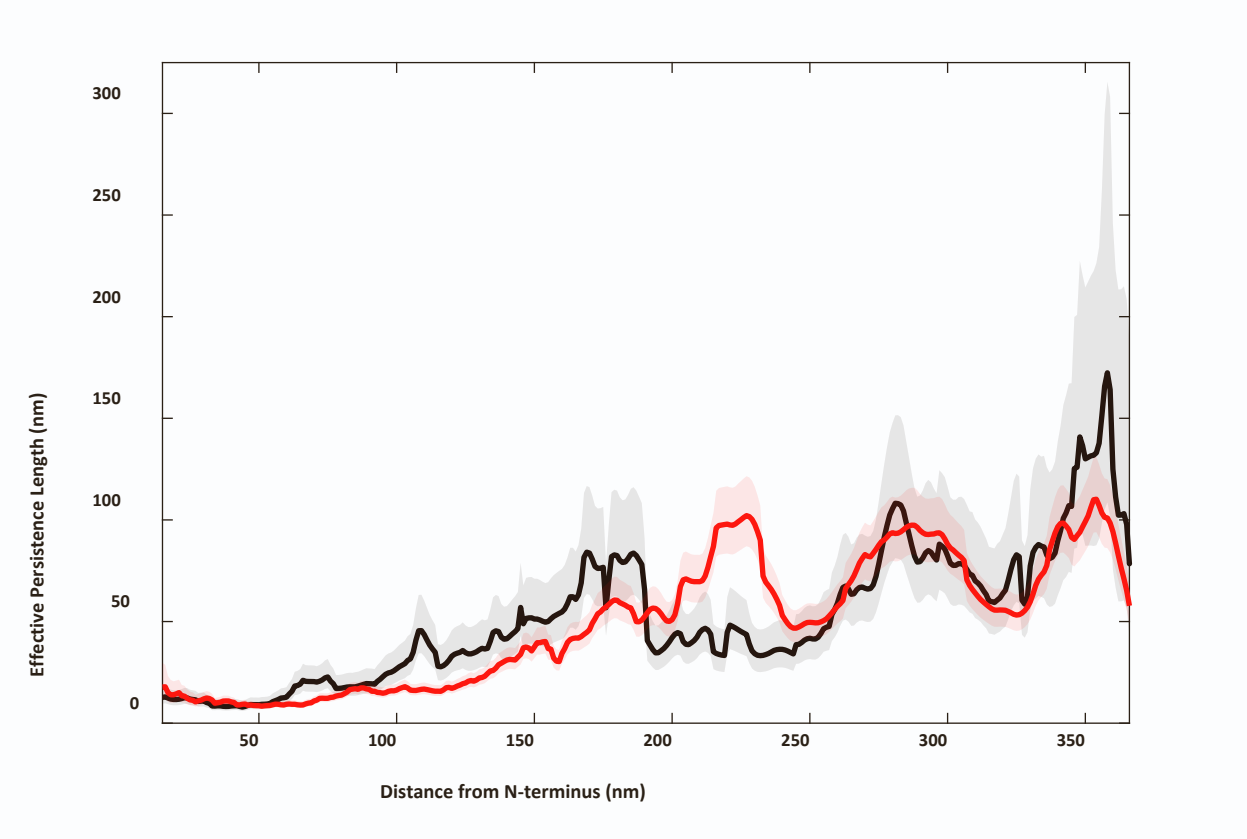

Figure S6. Flexibility profiles of collagen IV traced in both directions. Collagen type IV was deposited from 100 mM KCl, 1 mM HCl and traced in both directions. The shaded curves represent 95% confidence intervals on the effective persistence length estimate $p^*(s; \Delta s)$. The profile in black was traced from the N terminus (7S domain) towards the C terminus (NC1 domain) and was calculated from $N = 84$ chains. The profile in red (reproduced from Fig. 4) was traced from the C terminus (NC1 domain) towards the N terminus (7S domain) and was calculated from $N = 262$ chains. The two profiles displayed are aligned at the N terminus and appear comparable, with the exception of the region around 225 nm from the N terminus. This may be a result of the many fewer chains that were traced from the N terminus.

Figure S7. Varying assumptions of the minimum number of tripeptide units required to form triple-helical segments. Schematic depiction of sequences when varying the minimal sequence requirements for a triple helix from $(Gly\text{-}X\text{-}Y)_nG$, *n*=2-5. Each of the two bold arrows indicates a triple-helix-forming segment that is lost when increasing *n* (from 2 to 3, and from 4 to 5). All other triple-helical segments are longer than $(GXY)_4G$.

Supporting Text 2

Variable flexibility model fitting

The variable flexibility model was first used to define a position-dependent profile $[p_{\mathrm{in},i}]=[p_{\mathrm{in},1}, p_{\mathrm{in},2}, \ldots, p_{\mathrm{in},N}]$ in which the input flexibility (via local persistence length) is defined at each position $i=(1, 2, \ldots, N)$ along the chain backbone. For the simulated chains, each $p_{\mathrm{in},i}$ can take the value p_{r} or p_{f}, corresponding to a rigid or flexible monomer in the chain, while for collagen IV, each $p_{\mathrm{in},i}$ can take one of four values: p_0, p_1, p_2 and p_3 (Table 1 in main text) The profile $[p_{\mathrm{in},i}]$ was fit to the measured effective persistence length profile $p^*(s)$ as follows, with the aim of determining values for the local flexibilities (p_{r} and p_{f}; or p_0, p_1, p_2 and p_3).

1. A `filter` window of length Δs was used to determine the effective persistence length profile, which depends on all local persistence lengths within this window as given by equation (S6). For simulated chains, whose monomer spacing is defined in nanometers, `filter` = Δs. For collagen IV chains, the monomer spacing is defined in amino acids, and so a scaling factor `nmaa` (nanometers / amino acid) is applied to express the filter length as the number of amino acid steps:
$$\mathrm{filter} = \frac{\Delta s}{\mathrm{nmaa}}. \tag{S15}$$
The effective persistence length profile $[p_{\mathrm{eff},i}]$ is then given by
$$p_{\mathrm{eff},i} = \frac{\mathrm{filter}}{\sum_i^{i+\mathrm{filter}} (p_{\mathrm{in},j}{}^{-1})}, \tag{S16}$$
where the step spacing remains in its original units (e.g. amino acid steps for collagen IV). The p_{eff} profiles are shorter than the initial profiles: $i=(1{:}N$-`filter`$+1)$. In the results reported here, Δs = 30 nm for all measurements and fitting.
2. A `stagger` parameter was used, which represents the standard deviation of chain starting positions in the traced population. This was incorporated to account for variability in identifying the starting position of the chains from images. `stagger` is defined in units of nanometers and converted, for collagen chains, to amino acid steps as for the filter window length (S15). A Gaussian smoothing window was applied to the profile from (S16), with contributions from positions up to ±3 standard deviations included:
$$p_{\mathrm{eff,stag},i} = \sum_i^{i+6\times\mathrm{stagger}} P(j) p_{\mathrm{eff},j}. \tag{S17}$$
Here
$$P(j) = Ae^{-\frac{(j-(i+3\times\mathrm{stagger}))^2}{2(\mathrm{stagger})^2}} \tag{S18}$$
is the Gaussian smoothing function, where A is a normalization constant determined so that $\sum_i^{i+6\times\mathrm{stagger}} P(j) = 1$. The $p_{\mathrm{eff,stag}}$ profiles run from $i=(1{:}N$-`filter`$-6\times$`stagger`$+1)$.
3. Prior to fitting, the collagen chain $p_{\mathrm{eff,stag}}$ profiles in amino acid steps were converted into nanometer integer increments, via linear interpolation. The profiles then run from s = ((`filter`/2 + (6*`stagger`-1)/2) : (`length` - `filter`/2 - 6*`stagger`/2 + 1/2)).

4. An `offset` parameter was added to the contour positions ($s \rightarrow s +$ `offset`), which has the effect of linearly displacing the simulated profile relative to the measured profile. `offset` accounts for systematic errors in determining the start position of the chain (for example, if it is obscured by the NC1 domain).
5. Before fitting, the simulated and traced chain persistence length profiles must be the same length. Non-overlapping beginning and/or end segments of $p_{\mathrm{eff,stag}}(s)$ and/or $p^*(s)$ were removed to generate two linear arrays of identical length and range in s (and which have identical increments along the contour, of 1 nm).
6. The difference between $p_{\mathrm{eff,stag}}(s)$ and $p^*(s)$ was minimized by varying p_0, p_1, p_2 and p_3 [p_r and p_f in the case of simulated chains], and weighting the estimates of each value of $p^*(s)$ in the traced chain profile by its variance (estimated from equation (S13) by assuming normally distributed errors):

$$f(p_0, p_1, p_2, p_3) = \sum_s \frac{[p^*(s) - p_{\mathrm{eff,stag}}(s)]^2}{\mathrm{Var}(s)}. \quad \text{(S19)}$$

   Persistence length values (p_0, p_1, p_2 and p_3) were constrained to lie within the range [0,200] nm. The χ_r^2 value that resulted from minimizing the function f with the best-fit values of p_0, p_1, p_2 and p_3 was recorded:

$$\chi_r{}^2 = \frac{f_{\min}(p_0,p_1,p_2,p_3)}{N_{\mathrm{pts}} - N_{\mathrm{params}}}. \quad \text{(S20)}$$

   N_{pts} is the length of the resulting $p^*(s)$ and $p_{\mathrm{eff,stag}}(s)$ arrays used for minimizing f, and $N_{\mathrm{params}} = 4$ for collagen and 2 for the simulated chains.
7. Steps 2-6 were repeated for different parameters `nmaa`, `stagger` and/or `offset` to determine the values of p_0, p_1, p_2 and p_3 that minimized χ_r^2. In practice, this was implemented in a series of nested loops, with `nmaa` taking possible values of [0.27, 0.29, 0.31, 0.33, 0.35] nm/aa, and `stagger` and `offset` each taking integer values. For simulated chains, an optimal `stagger` = 5 nm was determined. For collagen IV, we found χ_r^2 to decrease as `stagger` was increased, in effect smearing out and making more homogeneous the simulated chain $p_{\mathrm{eff,stag}}$ profile. Thus, we kept `stagger` = 5 nm fixed, while varying `offset` and `nmaa` to determine their optimal values. These parameters, and the resulting best-fit values of p_0, p_1, p_2 and p_3, are presented in Table S1 for some of the tested chain alignments.

All data fitting was implemented within MATLAB (4).

Obtaining a well resolved effective persistence length profile from experimental images relies on accurate contour tracing and chain start-point determination. Experimentally, AFM images of collagen IV were obtained with settings that saturated the intensity in the NC1 domain. This led to a plateau-like intensity profile of the NC1 domains (typical diameter ~12 nm), from which we estimate the error of edge determination to be <2 pixels. This is commensurate with the 5 nm value of `stagger` used in smoothing the model flexibility profiles. Further refining this start-

point of chain tracing would decrease blurring of chain registry and improve the mapping of the underlying flexibility profile.

Table S1: Model outputs of flexibility for different chain alignments. The model was fit to the $p^*(s)$ profile obtained for 100 mM KCl, 1 mM HCl using a Δs = 30 nm filter window. All fits imposed a σ = 5 nm chain stagger and constrained the values of each persistence length to $0 \leq p_i \leq$ 200 nm and offset to $-20 \leq offset \leq 0$. Red values highlight those results that were optimized at the fitting boundary and thus did not minimize within the parameter range. The fits for bolded alignments are shown in the main manuscript. Clustal alignments vary in the loops imposed and assumptions about their contributions to the main contour. Two loops alignment names indicate the number of amino acids looped out from the α2 chain around residues 656-676 (disulfide-bridged loop) and 771-774 (Fig. S6) (5). Detailed sequence alignments are provided in a supporting file. The last column indicates whether α1 441D and α2 456R are aligned, as found for this integrin-binding region in the human homolog (6). Only the final, italicized alignment includes staggered starting positions in the three chains; here, the relative chain offset produces the local register found for the mapped integrin-binding site in human type IV collagen (6).

Chain Alignment	Offset (nm)	nm/aa	p (nm)	p_1 (nm)	p_2 (nm)	p_3 (nm)	χ^2_r	DDR aligned?
Linear	-11	0.33	154	10.8	2.3	200	7.7	No
Clustal A	-17	0.29	146	200	15	1.9	12.2	Yes
Clustal B	-17	0.29	109	86	42	2.0	12.8	Yes
Clustal C	-20	0.31	125	33	11	1.3	14.2	Yes
One loop	-7	0.33	154	9.9	1.8	200	10.6	No
Two loops (19,3)	-11	0.29	163	20	6.6	1.7	14.4	No
Two loops (19,4)	-10	0.29	138	28	8.1	1.8	14.0	No
Two loops (21,4)	-10	0.29	105	83	21	1.9	12.6	Yes
Two loops (21,5)	-9	0.29	102	112	24	2.0	12.8	No
Two loops (21,4) – staggered	*-7*	*0.29*	*129*	*200*	*3.5*	*2.1*	*14.0*	*Yes*

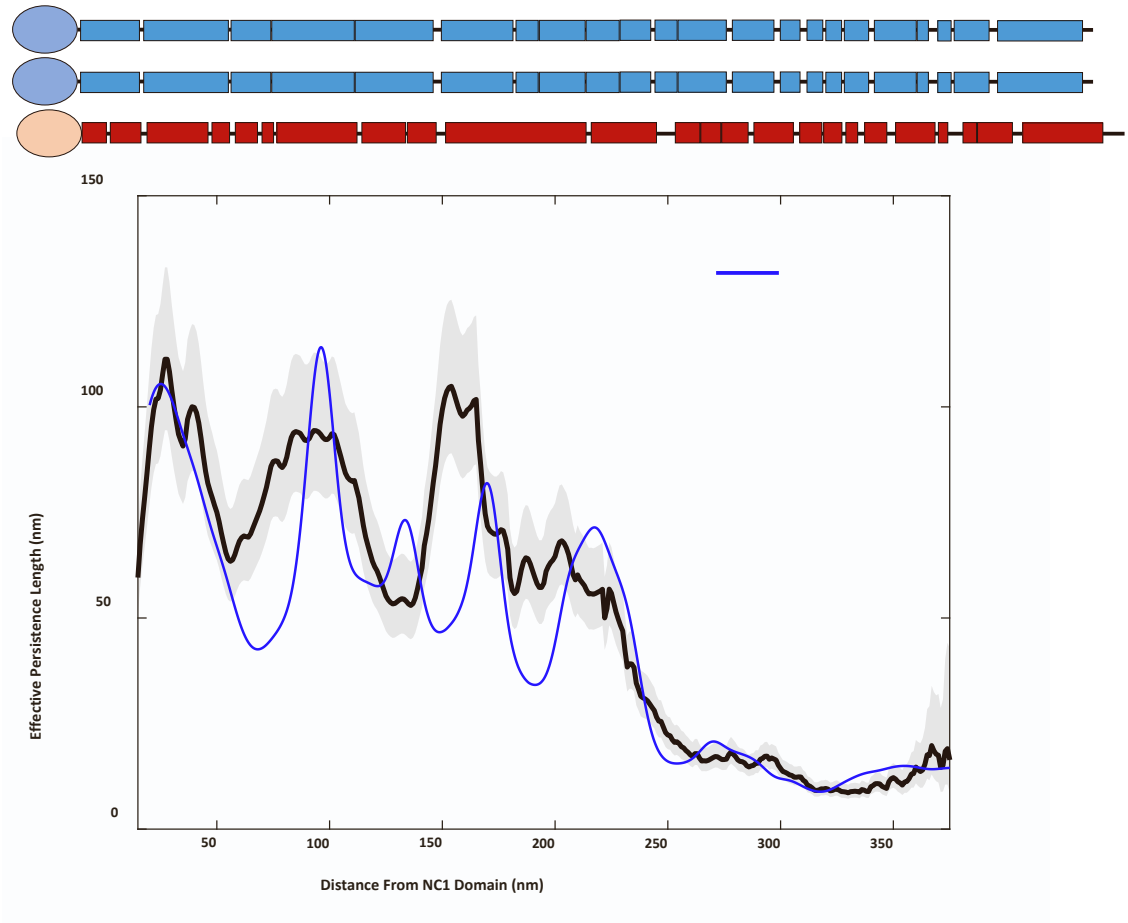

Figure S8. Four-class flexibility model using a linear chain alignment. The effective persistence length profile $p^*(s)$ of collagen IV deposited from 100 mM KCl 1 mM HCl is aligned with the amino acid sequence representations using the offset (-11 nm) and 0.33 nm/aa conversion parameters obtained from the fitting procedure. The model produced for an overlapping interaction p_3 = 200 nm, the maximum value allowed for fitting, an unphysical value. However, fitting $p^*(s)$ by using this linear chain alignment did capture the rigidity in the collagenous region adjacent to the NC1 domain.

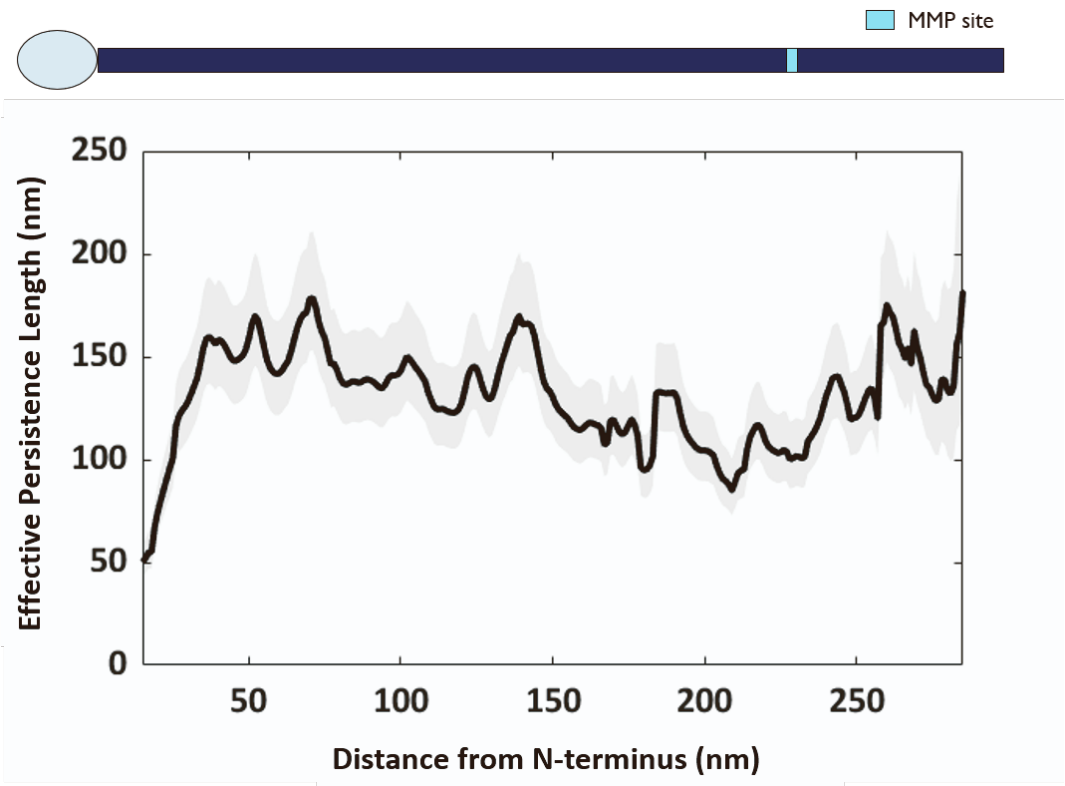

Figure S9. Position-dependent flexibility profile of collagen pN-III. Position-dependent effective persistence length map of collagen pN-III traced from the N-terminus. The profile is aligned with an amino acid sequence representation where the MMP site location is marked. Shaded curves represent 95% confidence intervals on the effective persistence length estimate $p^*(s; \Delta s)$. The profile was calculated from $N = 267$ chains.

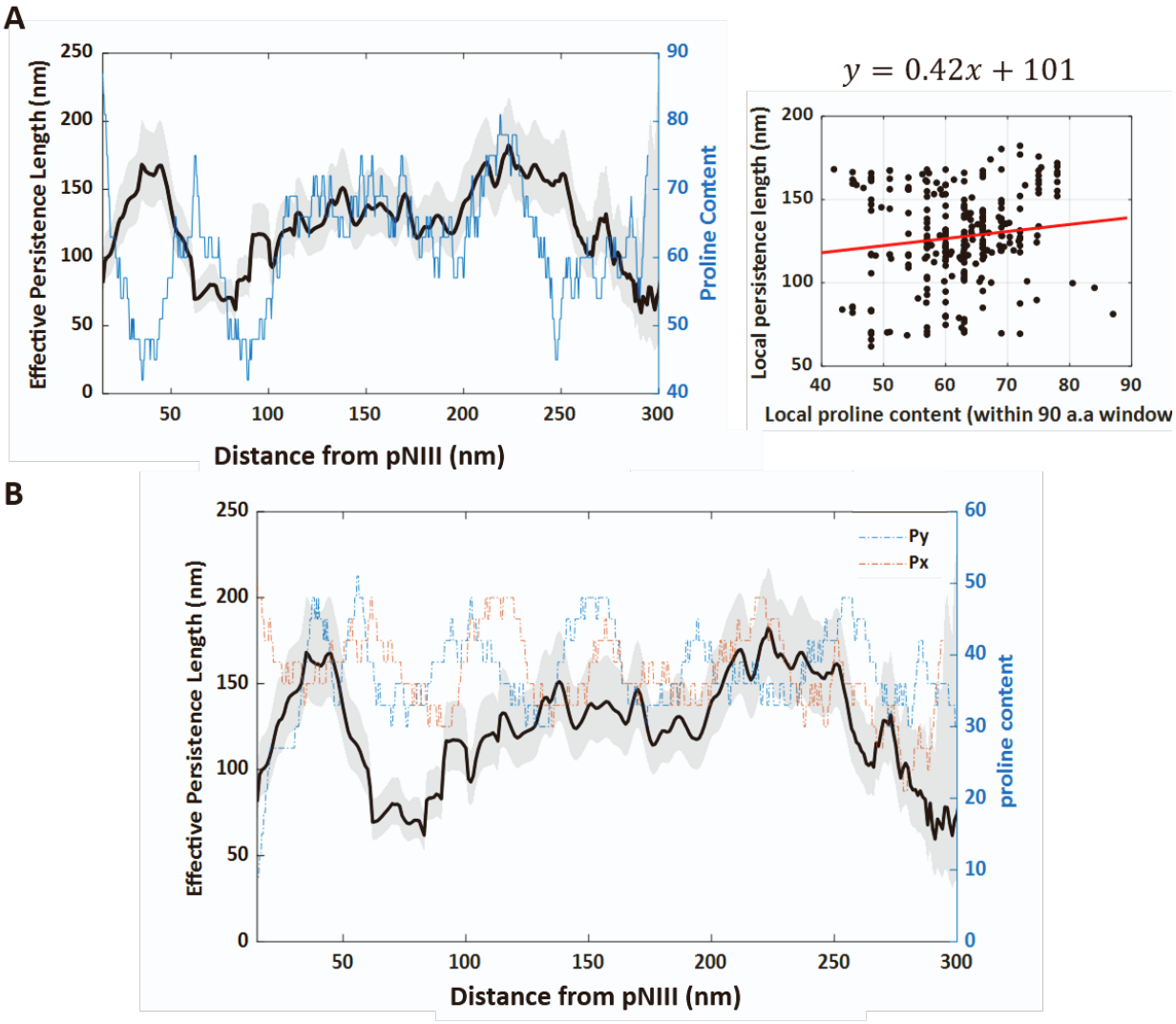

Figure S10. Local imino-acid (proline) content and persistence length profile of pN-III collagen. (A) The local proline content (calculated over a 90-amino-acid sliding window within each of the three chains) and effective persistence length of pN-III collagen (Q08E14) are uncorrelated, shown by a linear correlation coefficient of R^2 = 0.017. (B) There appears to be no strong correlation between either X- or Y-positioned proline content and effective persistence length. X-positioned prolines are expected to be unmodified, while Y-positioned prolines are expected to be 4-hydroxylated. Proline content in bovine pN-III collagen is given by the number of proline residues found in three α1(III) (Q08E14) chains, within a 90-amino-acid sliding window centered at the position noted. Proline content was determined at positions centered every 3 amino acids along the chain and was linearly interpolated to obtain values at nanometer spacing.

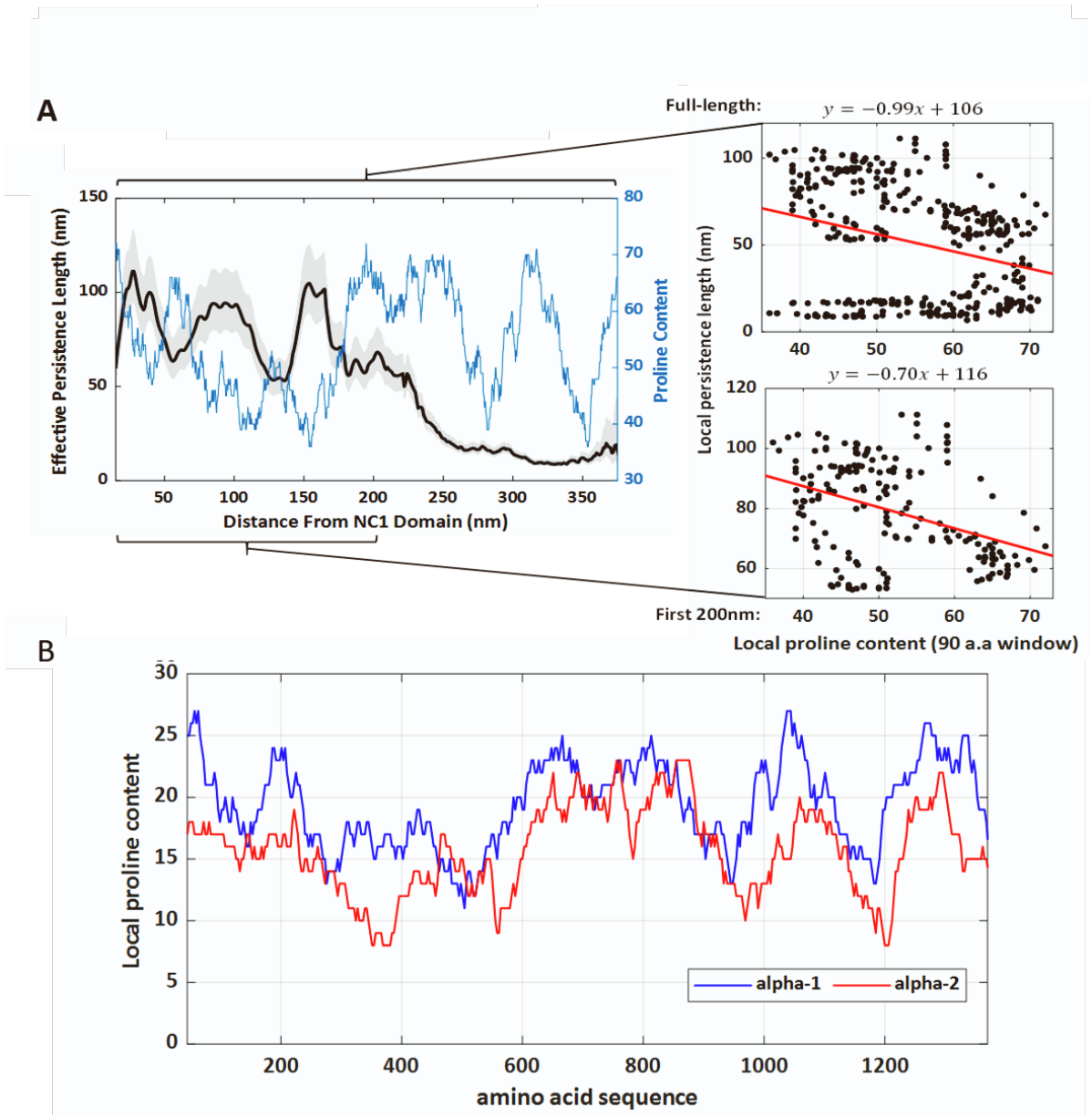

Figure S11. Local imino-acid (proline) content and persistence length profile of collagen IV. (A) The local proline content and effective persistence length appear anti-correlated in the C-terminal half of collagen IV, and otherwise uncorrelated. Quantitative analysis, however, reveals no statistical correlation between proline content and flexibility, neither for the full length of the chain (R^2 = 0.084) nor for the first 200 nm from the NC1 domain (R^2 = 0.15). The analysis assumes the three chains of collagen IV to be linearly aligned (no loops) and uses a conversion of 0.29 nm / amino acid. (C) Proline content in mouse collagen IV is given by the number of proline residues found in α1(IV) (P02463) and α2(IV) (P08122) chains, within a 90-amino-acid sliding window centered at the position noted. Proline content was determined at positions centered every 3 amino acids along the chain and was linearly interpolated to obtain values at nanometer spacing.

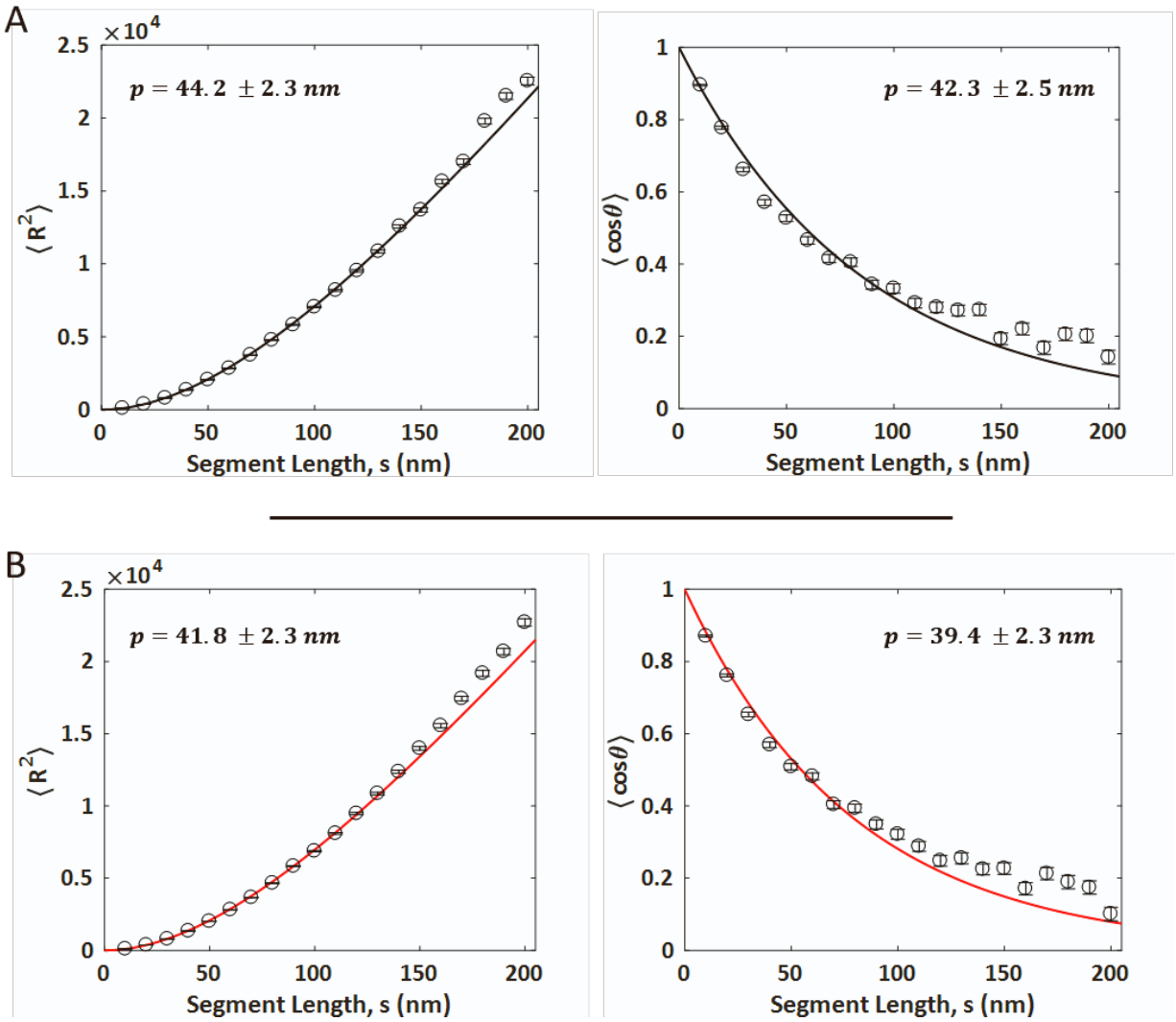

Figure S12. Homogeneous worm-like chain determination of persistence lengths of collagen IV in different solution conditions. The persistence lengths were obtained using $\langle R^2(\Delta s)\rangle$ (plots to the left) and $\langle \cos\theta(\Delta s)\rangle$ (plots to the right) analyses. The data correspond to collagen type IV deposited from A) TBS (chloride-containing), and from B) sodium acetate buffer.

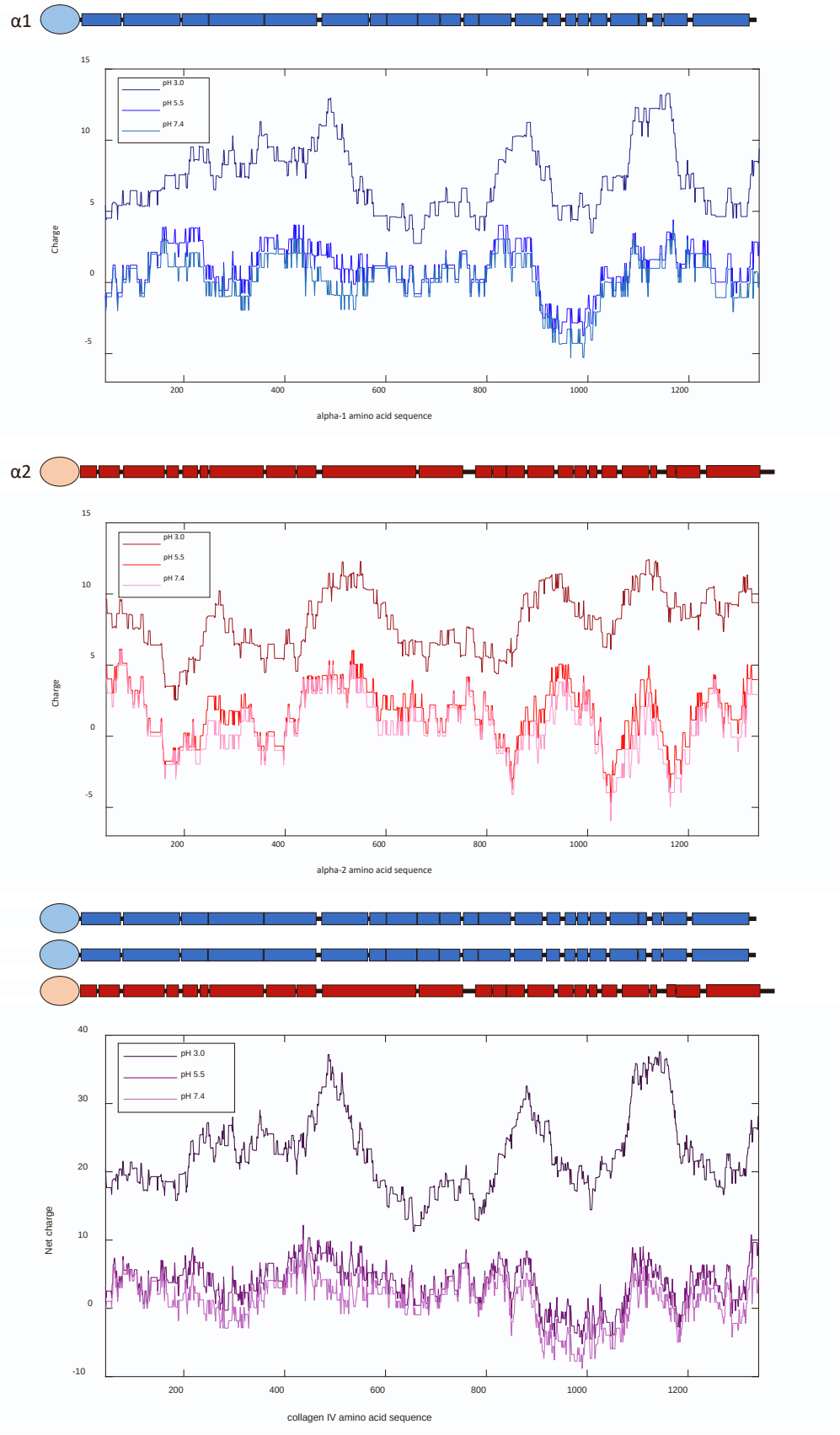
α1
pH 3.0
pH 5.5
pH 7.4
Charge
alpha-1 amino acid sequence
α2
pH 3.0
pH 5.5
pH 7.4
Charge
alpha-2 amino acid sequence
pH 3.0
pH 5.5
pH 7.4
Net charge
collagen IV amino acid sequence

Figure S13. Estimated charge profiles of the α1 and α2 chains of collagen IV at different pH. The charge on each amino acid at each pH was calculated using the Henderson-Hasselbach equation and the pKa of the amino acid side chain. Amino acids considered in the calculation, with their corresponding assumed pKa, are Aspartate, $pKa_D = 3.9$; Glutamate, $pKa_E = 4.3$; Histidine, $pKa_H = 6.1$; Cysteine, $pKa_C = 8.3$; Tryptophan $pKa_Y = 10.1$; Lysine, $pKa_K = 10.5$; Arginine, $pKa_R = 12.0$. The local charge along the sequence was averaged over 90 amino acids centred at each amino acid along the sequence. This estimate assumes that the pKa and charge on each amino acid are unaffected by the surrounding amino acids. **(A)** Charge profile for α1(IV) (P02463). **(B)** Charge profile for α2(IV) (P08122). **(C)** Net charge profile of collagen IV, assuming the three chains of collagen IV to be linearly aligned (no looping).

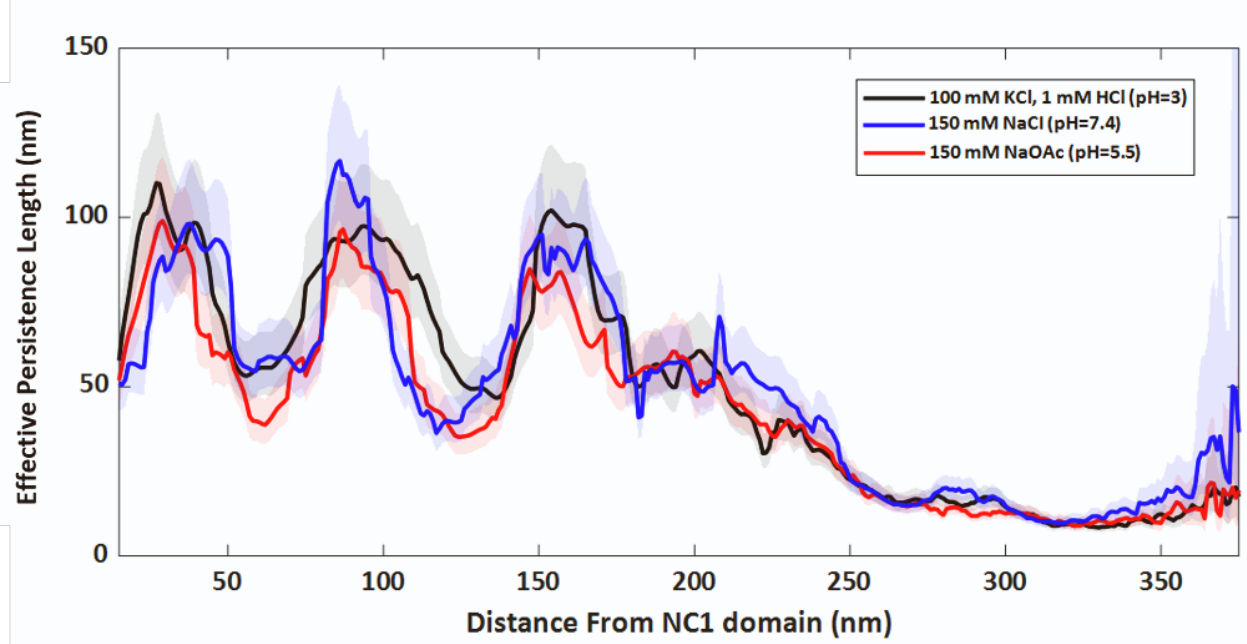

Figure S14. Superposition of experimental persistence length profiles. The position-dependent effective persistence length profile $p^*(s)$ of collagen IV changes remarkably little over a pH range from 3 to 7.4, and in the presence or absence of ~150 mM Cl^- ions.

Supporting References

1. Landau, L. D., L. P. Pitaevskii, A. M. Kosevich, and E.M.Lifshitz (1986) Theory of Elasticity. Butterworth-Heinemann.

2. Krishnamoorthy,K. (2016) Handbook of statistical distribution with applications (2nd edition). CRC Press: Boca Raton, Florida.

3. Gelman,A. (2013) Bayesian Data Analysis (3rd edition). CRC Press: Boca Raton, Florida.

4. MATLAB Release 2020a. The MathWorks, Inc., Natick, Massachusetts, United States.

5. Kühn,K. (1995) Basement membrane (type IV) collagen. *Matrix Biol.*, **14**, 439–445. https://doi.org/10.1016/0945-053X(95)90001-2

6. Golbik,R., Eble,J.A., Ries,A. and Kühn,K. (2000) The spatial orientation of the essential amino acid residues arginine and aspartate within the α1β1 integrin recognition site of collagen IV has been resolved using fluorescence resonance energy transfer. *J. Mol. Biol.*, **297**, 501–509.