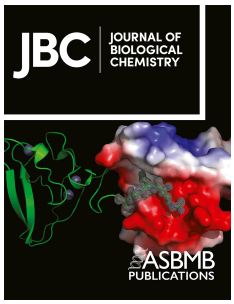


A recombinant technique for mapping functional sites of hetero-trimeric collagen helices: collagen IV CB3 fragment as a prototype for integrin binding

Sergei P. Boudko, Elizabeth H. Konopka, Woojin Kim, Yuki Taga, Kazunori Mizuno, Timothy A. Springer, Billy G. Hudson, Terence I. Moy, Fu-Yang Lin

PII: S0021-9258(23)01929-4

DOI: https://doi.org/10.1016/j.jbc.2023.104901

Reference: JBC 104901

To appear in: *Journal of Biological Chemistry*

Received Date: 8 May 2023

Revised Date: 1 June 2023

Accepted Date: 5 June 2023

Please cite this article as: Boudko SP, Konopka EH, Kim W, Taga Y, Mizuno K, Springer TA, Hudson BG, Moy TI, Lin FY, A recombinant technique for mapping functional sites of hetero-trimeric collagen helices: collagen IV CB3 fragment as a prototype for integrin binding, *Journal of Biological Chemistry* (2023), doi: https://doi.org/10.1016/j.jbc.2023.104901.

Title

A recombinant technique for mapping functional sites of hetero-trimeric collagen helices: collagen IV CB3 fragment as a prototype for integrin binding

Running title

Recombinant production of functional collagen fragments

Authors

Sergei P. Boudko[1,2,3]*, Elizabeth H. Konopka[4], Woojin Kim[4], Yuki Taga[5], Kazunori Mizuno[5], Timothy A. Springer[6], Billy G. Hudson[1,2,3,7-10], Terence I. Moy[4], Fu-Yang Lin[4]*

[1]Department of Medicine, Division of Nephrology and Hypertension, [2]Center for Matrix Biology, Vanderbilt University Medical Center, Nashville, TN 37232; [3]Department of Biochemistry, Vanderbilt University, Nashville, TN 37232; [4]Morphic Therapeutics, Inc. 35 Gatehouse Drive, A2 Waltham, MA 02451; [5]Nippi Research Institute of Biomatrix, 520-11 Kuwabara, Toride, Ibaraki 302 0017, Japan; [6]Department of Biological Chemistry and Molecular Pharmacology, Program in Cellular and Molecular Medicine, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, [7]Department of Pathology, Microbiology, and Immunology, [8]Department of Cell and Developmental Biology, Vanderbilt University Medical Center, Nashville, TN 37232; [9]Vanderbilt-Ingram Cancer Center, [10]Vanderbilt Institute of Chemical Biology, Vanderbilt University, Nashville, TN 37232

*To whom correspondence should be addressed:

Sergei P. Boudko: Division of Nephrology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232;
sergey.budko@vumc.org; Tel. (615) 936-3251; Fax. (615) 322-7381

or

Fu-Yang Lin: Morphic Therapeutic, Inc. 35 Gatehouse Drive, A2 Waltham, MA 02451;
albert.lin@morphictx.com; Tel. (781) 996-0955

Keywords

collagen IV; integrin; basement membrane; receptor; triple helix; extracellular matrix; heterotrimer; CB3 fragment; cystine knot; protein self-assembly; recombinant protein expression; protein folding; atomic force microscopy; circular dichroism spectroscopy (5 to 15)

Abstract

Collagen superfamily of proteins is a major component of the extracellular matrix. Defects in collagens underlie the cause of nearly 40 human genetic diseases in millions of people worldwide. Pathogenesis typically involves genetic alterations of the triple helix, a hallmark structural feature that bestows exceptional mechanical resistance to tensile forces and a capacity to bind a plethora of macromolecules. Yet, there is a paramount knowledge gap in understanding the functionality of distinct sites along the triple helix. Here, we present a recombinant technique to produce triple helical fragments for functional studies. The experimental strategy utilizes the unique capacity of the NC2 hetero-trimerization domain of collagen IX to drive three α-chain selection and registering the triple helix stagger. For proof of principle, we produced and characterized long triple helical fragments of collagen IV that were expressed in a mammalian system. The hetero-trimeric fragments encompassed the CB3 trimeric peptide of collagen IV which harbors the binding motifs for α1β1 and α2β1 integrins. Fragments were characterized and shown to have a stable triple helix, post-translational modifications, and high affinity and specific binding of integrins. The NC2 technique is a universal tool for the high yield production of hetero-trimeric fragments of collagens. Fragments are suitable for mapping functional sites, determining coding sequences of binding sites, elucidating

pathogenicity and pathogenic mechanisms of genetic mutations, and production of fragments for protein replacement therapy.

Introduction

Collagens are a major component of the extracellular matrix. They are comprised of 28 types in humans (I-XXVIII) encoded by over 40 different genes, forming a diversity of triple helical protomers of varying α-chain compositions (1,2). Protomers assemble into diverse super-structures, ranging from networks to fibrils and broadly function in structural, mechanical, and organizational roles that define tissue architecture and influence cellular behavior. Defects in collagens underlie the cause of nearly 40 human genetic diseases, affecting numerous organs and tissues in millions of people worldwide (3). Pathogenesis typically involves genetic alterations of the triple helix, a hallmark structural feature that bestows exceptional mechanical resistance to tensile forces and a capacity to bind a plethora of macromolecules. Such macromolecules include but are not limited to: integrins, DDR1 and 2, fibronectin, nidogen, perlecan, heparin, von Willebrand factor (VWF), decorin, bone morphogenetic proteins (BMPs), and glycoprotein VI (2,4,5). Yet, there is a paramount knowledge gap in understanding the functionality of distinct sites along the triple helix. This lack of knowledge impedes the development of precision therapies aimed at restoring/repairing function of collagen super-structures.

A unique feature of the collagen triple helix is a register (or stagger) of chains (6). There are always leading, middle, and trailing chains shifted by one residue. Even in homo-trimeric types of collagens, identical residues within the triple helix are not structurally equivalent. At least seven types of human collagens, *i.e.* I, IV, V, VI, VIII, IX, XI exist as heterotrimers of either AAB or ABC forms (2), which represent an additional challenge in generating biologically relevant collagen fragments. Examples include collagen I, the most abundant fibrillar type with an α1α1α2 composition and collagen IV of basement membrane with three compositions in mammals - α1α1α2, α3α4α5, and α5α5α6 (7,8).

Generation of short fragments of homo-trimeric collagens using chemical synthesis became a general method and great progress was achieved using this approach. Synthetic peptides were used to solve the first crystal structures of a collagen triple helix (9), the collagen triple helix with a mutation (10), and a collagen triple helix complex with integrin α2β1 (11). Peptide toolkits covering the entire triple-helical domains of homo-trimeric collagens II and III were successfully implemented and used to precisely map and study binding to integrin α2β1 (12), von Willebrand factor (13), DDR1 (14), DDR2 (15), dermatopontin (16), HSP47 (17), gpVI (18), LAIR-1 (19), SPARC (20), MMP-3 (21), multimerin 1 (22), and other proteins (reviewed by Farndale (23)). Several strategies were developed to address production of hetero-trimeric collagen fragments using regio-selective chemical synthesis (24-26) or hosting hetero-trimers with artificial sequence designs (27,28) (reviewed by Xu and Kirchner (29)). However, these approaches are limited to relatively short fragments (up to 12 residues) and require sophisticated approaches of chemical synthesis. So far, no collagen peptide toolkits have been reported for any hetero-trimeric type of collagen.

Here, we present a recombinant technique to produce hetero-trimeric (triple helical) fragments for mapping functional sites, determining coding sequences of binding sites, elucidating pathogenicity and pathogenic mechanisms of genetic mutations, and production of fragments for protein replacement therapy. The experimental strategy utilizes the unique property of the noncollagenous (NC) 2 hetero-trimerization domain of collagen IX. We previously demonstrated that the NC2 domain is sufficient to drive three α-chain selection and register the triple helix stagger using a bacterial system (30-32). For proof of principle, we adapted the NC2 domain for studies of collagen IV (Fig. 1) which harbors binding sites for numerous macromolecules (3,4). Long triple helical fragments which encompass the CB3 site that harbors the binding motifs for α1β1 and α2β1 integrins (Fig. 1) were expressed and characterized. Fragments were shown to have a stable triple helix, post-translational modifications, and specific binding to integrins.

Results

Design and production

Expression of artificial or fragmentary collagen triple helices that exhibit native structure and function has proven difficult as they are composed of two or three distinct α-chains. It has proven difficult to control the stoichiometry of the assembled collagen fragment. Moreover, collagen sequences require post-translational modifications, *i.e.,* hydroxylation and glycosylation, in order to form a stable and functional triple helical structure. We thus sought an expression system in mammalian cells that would facilitate expression of collagen domains whose collagen subunit composition and helix register could be controlled. Previously we demonstrated that three chains of the collagen IX NC2 domain one for each subunit, are sufficient to drive subunit-selection, *i.e.,* appropriate subunit composition and to control the register of the triple helical stagger (30-32). Here, we adapt this system to produce a long fragment of collagen IV (~200 residues) in a mammalian expression system. We selected the CB3 region of human collagen IV as its most well characterized fragment (33-41). The CB3 region of collagen IV was initially identified as a fragment chemically excised with cyanogen bromide that could be prepared in relatively high yield for structural and functional studies (33). Typical of collagen IV domains, CB3 has interruptions and a cystine knot as well as containing other known protein binding sites (4). The cystine knot seems to be an evolutionary conserved structural element in collagen IV also found in such distant organisms as fruit flies, roundworms and even in Cnidaria (Fig. S1). We generated expression constructs which extended the sequences of CB3 to a gelatinase A cleavage site at the N-terminus (38) and completed the last GXY repeat in the α2 chain to ensure the continuity and stability of the triple helices (Figs. 1, 2, 3, S2). We refer to this genetically encoded CB3 domain as extended CB3 (eCB3).

As in the bacterial expression system (30,32), for each chain (α1 and α2) of eCB3 we generated three constructs each containing one of the three non-identical NC2 trimerization domains of collagen IX, *i.e.,* α1 NC2, α2 NC2, and α3 NC2. These domains pack in a specific orientation around the three-fold pseudo-symmetry axis of collagen and define the stagger of the collagen triple helices (31). Each of the three eCB3 domain sequences was cloned upstream of its specific NC2 domain and flanked by GPP repeats to avoid destabilizing effects at the N and C-termini (Figs. 2, S2). These extensions may also contribute to more robust post-translational modifications (PTMs) within the CB3 sequences. To ensure straightforward affinity purifications of hetero-trimeric assemblies each chain contained a specific tag placed at the N-terminus following the signal peptide (Figs. 2, 3, S2).

We selected the expiCHO transient expression system (Thermo Fisher), which is suitable for our need of co-expression of three different chains (Table S3). The transient transfection system we used allows for rapid production as it eliminates three cycles of time-consuming selection for each vector to generate stable clones and the necessity of each vector having a unique selectable marker. To facilitate collagen specific PTMs, the expression cell media was supplemented with fresh ascorbate on a daily basis (42,43).

Assemblies of eCB3

The chain register of the collagen triple helix has long been considered critical for structure, folding, and function. For collagen IV the register was explored only for helix #9 (39) and no rules were reported on how it can be translated through the interruptions to other helices. However, in collagen IX, the register of the collagen triple helix N-terminal to the NC2 domain has been defined and shown to be determined by linkage to the NC2 domain (30,31). This known register allows matching to the correct register of the collagen triple helix if known. As the register of the C-terminal helix (#11, Fig. 1) of eCB3 fragment is not known, we generated plasmids to match each possible register. We refer to an assembly nomenclature in which the His-tagged NC2-A subunit (IX α1) is listed first, the Flag-

tagged NC2-B subunit (IX α2) is listed second, and the Twin-Strep-tagged NC2-C subunit (IX α3) is listed third (Fig. 3). We thus generated 112, 121 and 211 collagen IV eCB3 expressing cells. As it was also not known whether homopolymers assembled and were stable, we also expressed 111 and 222 eCB3 fragments (Fig. 3).

Purification

Secreted proteins were purified from conditioned cell culture media using three sequential rounds of chain-specific affinity purifications (Figs. S3, 4A). This purification scheme ensures the hetero-trimeric composition and high purity (Fig. 4B). We achieved yields of pure proteins ranging from 1.5-2.4 mg from 1L of culture medium for 111, 112, 121, and 211 eCB3 assemblies and ~12 mg for the 222 eCB3 protein.

To determine whether the purified proteins are stabilized by inter-chain disulfide bonds (Figs. 1B and 3) the samples were electrophoresed under both non-reducing and reducing conditions. A single band corresponding to disulfide cross-linked trimer for each of the 111, 112, 121, and 211 eCB3 assemblies was observed at approximately 120 kDa (Fig. 4B). The same samples run under reducing conditions resolved in monomeric bands of approximately 40 kDa. This result suggests that register of the last eCB3 helix (#11, Fig. 1), which is deliberately controlled by the NC2 domain, is not critical for assembly. Another surprising observation of assembly of homotrimeric 111 raises a question whether a new isoform of collagen IV can exist. In contrast, the 222 expressing cells produced heterogeneously assembled material with both trimeric, dimeric, and a minority of monomeric-like material under non-reducing conditions. Under reducing conditions, a dominant band at the expected molecular weight for the monomers is observed along with bands of unknown compositions. Because the 222 eCB3 is likely poorly or inappropriately folded, we hypothesized it would be sensitive to proteolysis. Indeed, after 6 weeks of storage of purified material at 4 °C , only the 222 eCB3 demonstrated significant degradation (Fig. S4).

The affinity purified proteins were further analyzed by size-exclusion chromatography (Fig. 4C). The 111, 112, 121, and 211 eCB3 assemblies demonstrated a single major peak that corresponds to the expected hydrodynamic radius, while the major fraction of 222 was shifted to a smaller apparent size suggesting a more globular structure.

Collagen-specific PTMs

Post-translational modification of collagenous sequence is critical for mediating the stability of the collagen helix and collagen function. Our expression system, while not derived from epithelialized cells, should perform the post-translational modifications observed *in vivo* in collagen proteins. We noticed that the 111, 112, 121 and 211 assembly containing proteins ran as relatively diffused bands on the SDS PAGE gels. Moreover, the 222-eCB3 chains migrated faster than the other chains under non-reducing and reducing conditions (Fig. 4B) suggesting that this protein may not be efficiently post-translationally modified. Consistent with this hypothesis, the amount of secreted and purified 222 was 5-8 times higher, as judged from the SDS-PAGE (Fig. 4B) and gel filtration chromatography (Fig. 4C), possibly because of escaping the PTM machinery by yet unknown mechanism. The 112, 121, and 211 proteins differed from one another in reducing SDS-PAGE (Fig. 4B), suggesting that registration differences affected post-translational modifications at chain level. Collectively, these results suggest that the α1-containing eCB3 fragments are post-translationally modified.

To quantitate PTMs of the expressed eCB3 proteins, the amino acid hydrolysates were subjected to LC-MS analysis using stable isotope-labeled collagen as a reference (44). Quantities of proline (Pro), 4-hydroxyproline (4-Hyp), 3-hydroxyproline (3-Hyp), lysine (Lys), hydroxylysine (Hyl), galactosyl-hydroxylysine (GHL), and glucosyl-galactosyl-hydroxylysine (GGHL) were calculated as fractions of modified and non-modified residues (Fig. 5). The 222 assembly stands out as least modified for both proline and lysine residues, which should destabilize the triple helix and results in a more compact or

less elongated structure as shown by elution in gel filtration (Fig. 4C). The α1 chain-containing eCB3 assemblies demonstrate approximately equal levels of PTMs. Given that 4-hydroxylation of Pro and hydroxylation and sugar modifications of Lys are restricted to Y position of GXY-tripeptide units, the fraction of modification of these residues reaches ~50% and ~35% respectively.

Atomic force microscopy imaging

To determine whether the expressed eCB3 proteins were indeed correctly folded into a collagen helix, atomic force microscopy (AFM) imaging was used to reveal the structural organization of individual molecules. All assemblies except 222 eCB3 demonstrated the presence of ~70-nm-long rod-like structures, which are typical for the triple helical collagenous domain and the α-helical coiled coil domain (NC2 trimer) (Fig. 6). At this resolution we cannot reliably distinguish triple helical domains from the coiled coil. Some of the molecules showed kinks, possibly at sites of interruptions. No filaments were observed in the 222 eCB3 assembly sample indicating the absence of a folded collagen triple helix.

Circular dichroism and thermal transitions

To determine whether the expressed eCB3 assemblies are thermally stable, we analyzed the circular dichroism (CD) spectra and measured the melting temperature of these proteins. The far UV CD spectra of 111, 112, 121, and 211 assemblies are similar to collagen spectra although without a prominent positive peak in the range of 220-230 nm (Fig. 6A). The CD spectrum of the NC2 domain alone was previously found predominantly α-helical with a prominent negative peak at 215-230 nm (32), which compensates the collagen-specific positive peak in the eCB3 assemblies resulting in a "flat" CD spectrum around 225 nm.

The thermal stability of the complexes was studied at pH 4.5 to prevent disulfide bond reshuffling upon denaturation. At least one unpaired cysteine is expected in 111 eCB3 within the interruption between helices #8 and #9 (Fig. 1). Also, there is no data reported on an oxidative state of the cystine knot. The unfolding transitions of the triple helical domain were measured at 225 nm and calculated as a fraction of helix (Fig. 7). The apparent melting temperature was ~37 °C for all the assemblies. When both heating and cooling transitions were recorded they demonstrated a hysteresis phenomenon (Fig. S5) due to slow proline peptide bond cis-trans isomerization (45), which causes slow collagen triple helix denaturation and renaturation (46).

Functional activity of a small molecule β_1 integrin inhibitor

The BOP compound was reported to be an inhibitor of $\alpha_4\beta_1$ and $\alpha_9\beta_1$ integrins (47), and we determined that this compound inhibits β_1 integrins including the collagen-binding integrins ($\alpha_1\beta_1$, $\alpha_2\beta_1$, $\alpha_{10}\beta_1$, and $\alpha_{11}\beta_1$) and $\alpha_3\beta_1$. Purified $\alpha_1\beta_1$, $\alpha_2\beta_1$, $\alpha_3\beta_1$, $\alpha_{10}\beta_1$, and $\alpha_{11}\beta_1$ proteins bind to a fluorescently labeled version of the BOP compound as measured by fluorescence polarization (FP). K_d values between the fluorescent BOP compound and $\alpha_1\beta_1$, $\alpha_2\beta_1$, $\alpha_3\beta_1$, $\alpha_{10}\beta_1$, and $\alpha_{11}\beta_1$ proteins were 13.9 nM, 6.5 nM, 13.4 nM, 16.9 nM, and 30.3 nM, respectively. Using a FP IC_{50} assay with the fluorescent BOP as the FP probe, unlabeled BOP compound has IC_{50} values of 80.7 nM, 16.2 nM, 31.2 nM, 25.3 nM, and 118 nM against $\alpha_1\beta_1$, $\alpha_2\beta_1$, $\alpha_3\beta_1$, $\alpha_{10}\beta_1$, and $\alpha_{11}\beta_1$, respectively. BOP at a concentration of 50 μM in the solid phase assay is expected to result in near complete inhibition of the integrins' ability to bind ligands. Thus, small molecule compound BOP was used as a nonselective integrin inhibitor.

Integrin binding

The solid phase assay was used to measure the binding of the extracellular domains of purified recombinant integrin proteins to collagen IV purified from human tissue and the eCB3 assemblies immobilized onto microtiter plate wells. With immobilized collagen IV purified from human tissue,

integrins α1β1, α2β1, and α11β1 bound with saturating binding (Fig. 8) and binding was inhibited with the non-selective small molecule inhibitor BOP (data not shown). The 111, 112, 121, and 211 eCB3 assemblies bound to integrins α1β1, α2β1, and α11β1 (Fig.8). No significant binding was observed either for full-length collagen IV or the eCB3 assemblies for the α3β1 and α10β1 integrins.

Interestingly, binding to eCB3-111 demonstrated binding to the same integrins as full-length collagen IV suggesting that α2 chain is not absolutely required for integrin interaction.

While integrins were able to bind 111, integrin α1β1 bound to 112, 121, and 211 with higher apparent affinity as judged by EC_{50} values that were lowered by 5.2-fold, 5.7-fold, and 4.4-fold respectively (Table 1). This finding emphasizes a role of α2 chain in modulating binding to integrin α1β1, which was suggested previously as a chain-distributed recognition site within helix #9 (39).

No other significant differences were noticed for binding of α2β1 and α11β1 integrins, suggesting linear nature of binding epitopes within α1 chain of eCB3.

Discussion

The insertion of collagenous sequences into a hetero-trimerization cassette with the NC2 domain of collagen IX (Fig. 9) allowed recombinant production of well-folded and functional homo- and hetero-trimeric collagen fragments. This is the first report of recombinant production of CB3-encompassing region of human collagen IV, which opens the way for investigating the molecular basis of interaction with integrins, heparin, fibronectin, nidogen, perlecan, von Hippel-Lindau protein, HSP47, SPARC (4); collagen IV cleavage by MMP2/9 (4); Hereditary angiopathy, nephropathy, aneurysms, and muscle cramps (HANAC) syndrome mutations (35,48); acute rheumatic fever (ARF) triggered by *Streptococcus pyogenes* (40). It is now possible to assess effects of point mutations on the binding properties of CB3 and refine studies on shorter fragments.

The outcome of this technology is much broader as it allows us to efficiently produce recombinant fragments of any type of collagen with or without interruptions. Even for homotrimeric types of collagens, our technology allows precise mapping of binding to register-specific residues or studying effects of genetic variants in a composition- and register-specific context. Such fragments could be used to determine whether a genetic variant is pathogenic, a characteristic which is often not obvious from genetic studies. Indeed, among thousands of pathogenic variants reported for collagen IV α1 to α6 chains in the two main variant databases LOVD (https://www.lovd.nl/) and Clinvar (https://www.ncbi.nlm.nih.gov/clinvar/) there are 20 to 70% of variants with conflicting interpretations or uncertain significance (after excluding benign and likely benign variants). Our technology can also be used to explore pathogenic mechanisms, such as the impact of a variant on folding, stability, and interactions with receptors and other macromolecules. It can help to explore an Asp326Tyr variant in α3 chain of collagen IV, which was recently found to be protective against several definitions of diabetic kidney disease (49). Furthermore, the technology is applicable for mapping integrin and other receptor binding sites to yet unknown positions or fully unexplored collagen isoforms, like α345 of collagen IV.

Interest in trimerization domains as tools to facilitate folding and stabilization of the collagen triple helix started with the use of a foldon domain of bacteriophage T4 fibritin (50) fused to GPP-repeats mimicking the collagen triple helix (51-53). However, this and other approaches suffered from incompatibility of a three-fold rotation symmetry of the trimerization domain and the staggered structure of collagen causing unfavorable kinking at the site of fusion (54). In addition, assembly of heterotrimeric collagen fragments with register control was a challenge until the discovery of the NC2 domain of collagen IX (30,32). It was successfully used to produce variants of the binding site of collagen I to von Willebrand factor A3 domain in bacteria (30,31) and collagen-like region of complement component C1q in stably transformed human cells (55).

Here we developed a general method for cloning and production of a functional region of collagen IV as a framework applicable basically to any fragment of any type of collagen or collagen-like proteins. It allows to control chain composition, chain register, and is applicable to fragments containing irregularities known as triple helix interruptions. It takes advantage of transient co-expression of three chains, which makes it a robust method for production of multiple fragments and/or testing different compositions and chain registers.

Specifically, we produced the CB3-containing fragment of human collagen IV in non-human cells. Previously, generation of CB3 fragment was only possible by cyanogen bromide (CNBr) cleavage of collagen IV extracted from human tissue. CNBr cleavage is toxic and laborious. Moreover, the sequence analysis revealed that the beginning of α1 chain in CB3 cannot be generated by CNBr as it requires a preceding Met residue. The most frequent natural variant at this position is Pro and other reported variants, Thr and Ser, are extremely rare, less than 0.00002% according to gnomAD database (56). The closest Met residue is located ~100 residues upstream (Fig. S6). Thus, generation of CB3 fragment by CNBr cleavage relies also on the yet unknown cleavage mechanism for α1 chain, possibly relying on proteolysis during preparation. With our NC2 recombinant technology, it is now possible to overcome safety issues and generate CB3 and other collagen fragments of desired length and chain composition.

Originally, the NC2 technology involved individual expression of chains in bacteria, assembly of heterotrimers after combining the cell lysates, and subsequent purification of assembled triple helical fragments (30-32). The approach had several limitations, including the absence of collagen-specific PTMs and solubility issues with certain collagen-derived sequences (unpublished data). The absence of PTMs had several drawbacks including decreased thermal stability of triple helix (lack of stabilizing effect of 4-hydroxylation of prolines) and altered affinity for ligand binding.

These limitations were overcome with the use of a mammalian system and co-expression of all three chains to ensure proper heterotrimer folding inside the cell. Instead of tedious rounds of stable transfections of three compatible plasmids encoding all three chains we selected transient expression system in CHO cells. The ExpiCHO Expression System is a high-yield transient expression system based on suspension-adapted Chinese Hamster Ovary (CHO) cells. It uses serum-free medium and is not of human origin, which makes it advantageous for production of proteins for replacement therapy. To ensure high purity and desired chain stoichiometry in the assembled heterotrimer we used three different chain-specific tags and purified secreted proteins using serially three affinity columns.

The NC2 technology, coupled with expression in CHO cells, allows to combine desired chains in a specific way to obtain different stoichiometries and/or different registers of the chains. In this study we produced and tested three stoichiometries ($\alpha1_3$, $\alpha2_3$, $\alpha1_2\alpha2_1$) and three possible registers for $\alpha1_2\alpha2_1$ composition (α1α1α2, α1α2α1, α2α1α1). For example, α1α1α2 register means α1 in the leading, α1 in the middle, and α2 in the trailing positions. The register of the chains is influenced by the NC2 domain as it has specific register geometry for the adjacent triple helix (31). We could at least expect the designed register for the last triple helical segment of the collagenous sequence, which is placed in front of NC2. Given the flexible nature of collagen IV interruptions (42,57-59), artificial mismatch of the register in the last segment (forced by the NC2 domain) can be tolerated by nearest interruption and correct (native) register can be restored in the following triple helical segments. Recent success in designing self-assembling collagen-like hetero-trimeric peptides with a specific register provides rationale that similar process is possible for native collagen sequences (60-62). Altogether it could be an interplay of collagenous and non-collagenous domains in defining the register for the triple helix.

Collagen triple helix is characterized by the presence of unique PTMs, which were found also in our fragments. Indeed, significant PTMs in 111, 112, 121, and 211 eCB3 fragments demonstrate that mammalian system is a better choice for collagen production. Uniquely, 222 fragment revealed least

amount of such modifications. Is it just a consequence of over-expression of α2 chains (5-8 more than α1 chains), which overloads the cell PTM machinery, or presence of α1 chains somehow coordinates the PTM enzymes in a folding triple helix? Those questions need to be further explored. We did not perform analysis of individual sites of PTMs, including glycosylation sites Hyl393 and Hyl543, which were reported to modulate integrin affinity (63). This would be an interesting and important question in future studies.

AFM imaging revealed worm-like structures for all except 222 eCB3 fragments. Their length (~70 nm) corresponds to the expected lengths of collagenous fragment (66 nm) and NC2 domain (4.4 nm). The collagenous length has been estimated for ~200 residues of collagen IV and 27 residues of flanking GPP repeats and classical triple helix structure, which gives a rise of 2.9 Å per residue (6). AFM imaging in the air is a straightforward tool for validating collagen triple helical structure. Alternatively, rotary shadowing can be applied if available. Unfortunately, length (~200 residues per chain) and presence of multiple interruptions precluded such structural analyses as NMR or X-ray crystallography due to size and increased flexibility, which would otherwise provide atomic resolution details and resolve the chain registers. Nevertheless, this technique has been proven to be successful in solving crystal structures of shorter collagen fragments (31). By generating sub-fragments of collagen IV CB3 (or other fragments of other collagens) it is possible to elucidate specific integrin binding sites (or others) and try to determine high-resolution structures.

Thermal transitions of the 111, 112, 121, and 211 eCB3 assemblies demonstrated melting of the triple helical segments at temperatures higher than 25 °C, which is sufficient for *in vitro* experiments at room temperature. If higher thermal stability would be required, as the case for cell culture or animal experiments, a co-expression of collagen prolyl-4-hydroxylase could be an option. Another solution could be use of “collagen-specialized” cultures of cells. HT1080 (human origin) or PF-HR9 (mouse origin) might be a better choice with respect to collagen-specific post-translational modifications, yet several aspects should be taken into consideration like transfection efficiency, non-human origin, yield, and interaction with deposited matrix. Further studies are required to broaden repertoire of cell lines suitable for recombinant production of collagen fragments.

Finally, integrin binding assays confirmed the biological functionality of the eCB3 assemblies. Out of four collagen-specific integrin receptors (64), we observed binding to three, *i.e.* α1β1, α2β1, and α11β1 integrins demonstrated affinity to collagen IV and the eCB3 assemblies. Under the experimental conditions used we were not able to confirm binding of α3β1 and α10β1 integrins to collagen IV or to the eCB3 assemblies. Here, we re-confirmed the presence of a unique spatially organized binding site to α1β1 integrin, which requires presence of α2 chain of collagen IV. Interestingly, α1-homotrimeric variant reveals similar binding efficiency to other two collagen IV binding integrins, α2β1 and α11β1. Taken together that 111 does form a stable collagen triple helix and binds to integrins, a question is raised whether this combination exists in living tissues, under what conditions and to what extent.

In summary, we developed a system, using collagen IV as a prototype, for recombinant production of triple helical fragments with different combinations of chains and stagger control. We confirmed the structural integrity and stability of the collagenous domain, the collagen specific PTMs, and functionality at the level of integrin binding. This method opens the door for systematic exploring the collagen molecules using routine molecular biology techniques and instrumentation.

Experimental procedures

DNA constructs

The synthetic genes encoding guest inserts of the eCB3 regions of α1 and α2 chains of human collagen IV and the host frameworks, bearing signal peptides, affinity tags, flanking GPP repeats and

NC2 domain chains of collagen IX were ordered from Genewiz (USA). The genes were cloned into the pUC-GW-Amp plasmid vector by the supplier (Table S1).

The coding sequences of the host frameworks were excised with HindIII and BspDI restriction enzymes from the cloning plasmids pUC_His-GPP5-2xBsmBI-GPP4-NC2a1, pUC_Flag-GPP5-2xBsmBI-GPP4-NC2a2, and pUC_TwinStrep-GPP5-2xBsmBI-GPP4-NC2a3 (Table S1) and inserted into the backbone of the expression plasmid pRcX (65) using the same restriction sites.

The eCB3 sequences from pUC_a1CB3IV and pUC_a2CB3IV were seamlessly incorporated into the expression vectors pRc_His-GPP5-2xBsmBI-GPP4-NC2a1, pRc_Flag-GPP5-2xBsmBI-GPP4-NC2a2, and pRc_TwinStrep-GPP5-2xBsmBI-GPP4-NC2a3 using the Golden Gate Assembly BsmBI kit (New England Biolabs Inc.) (Tables S1 and S2).

All plasmids were sequence-verified, and their sequences are available upon request.

Transient expression and purification of proteins

For each assembly, *i.e.* 111, 222, 112, 121, and 211, we transfected three corresponding plasmids (Fig. 2, Table S3) into expiCHO-S cells (Gibco) according to the ExpiCHO Expression System User Guide and followed the Max Titer Protocol with three modifications: 1) on the day of transfection and every day until the final collection we supplemented the medium with fresh solution of ascorbic acid to the final concentration of 50 µg/ml; 2) the second feed was added on day 3; and 3) final culture was collected on day 6. The cells were pelleted by centrifugation at 4,000 g for 15 minutes and media were collected for protein purification.

The media were extensively dialyzed against TBS buffer before three rounds of affinity purifications. First round was purification of His-tagged proteins using NiNTA resin and the manufacturer protocol (Qiagen). For the second round the imidazole eluent from round one was directly run over the ANTI-FLAG® M2 Affinity Gel (Merck KGaA) to capture FLAG-tagged proteins and eluted with FLAG-peptide solution as described (66). Finally, for the third round the FLAG-peptide elution fraction was loaded onto the Strep-Tactin®XT Superflow® (IBA Lifesciences GmbH) resin to bind Twin-Strep-tagged proteins and eluted with Buffer BXT containing 50 mM Biotin as described in the manufacturer protocol for native protein purification.

Finally, we achieved yields of pure proteins ranging from 1.5-2.4 mg of 1L of culture medium for 111, 112, 121, and 211 eCB3 assemblies and ~12mg for 222.

Size-exclusion chromatography

Size-exclusion chromatography was conducted with a Superdex 200 Increase 10/300 GL gel-filtration column (GE Healthcare), using ÄKTA FPLC system (GE Healthcare) at a 0.5 ml/min flow rate. The column was equilibrated with 25 mM Tris-HCl, pH 7.5 supplemented with 150 mM NaCl (TBS). Eluting proteins were monitored by absorbance at 215 nm.

Quantification of post-translational modifications (PTMs)

Collagen PTMs were analyzed by LC–MS with high sensitivity after amino acid hydrolysis and quantified using stable isotope-labeled collagen (SI-collagen) as an internal standard. The detailed method is described in Taga *et al*. (44). The molar amounts of Pro, Lys, and their PTMs were calculated from the peak area ratio of the sample (stable isotopically light) relative to the internal standard (stable isotopically heavy), in which molar amounts of each amino acid were predetermined, according to the following formula: (light/heavy) × mol (SI-collagen).

Atomic force microscopy

The sample preparation for atomic force microscopy was done on mica (Highest Grade V1 AFM mica discs, 9.9 mm, Ted Pella). The samples in TBS buffer were diluted 40 times with TBS containing 2

mM $CaCl_2$ into ~2 µg/ml, and 50 µl was deposited onto freshly cleaved mica. After a 30-s incubation period, the excess unbound proteins were washed with ultrapure water for ~10 s, and the mica was dried immediately under filtered air. All proteins were imaged under dry conditions. AFM imaging was done with a Bruker Dimension Icon atomic force microscope using ScanAsyst/PeakForce mode in air using a SCANASYST-AIR tip.

Circular dichroism spectroscopy and thermal unfolding

Far-UV circular dichroism (CD) spectra were recorded on a Jasco model J-810 spectrometer equipped with Peltier temperature control unit (JASCO Corp.) using a quartz cell of 1-mm path length at 20 °C. The proteins were dialyzed against 0.1 M sodium acetate pH 4.5 buffer and their final protein concentrations were 0.98, 1.42, 0.86, and 0.93 µM of trimer for constructs 111, 112, 121, and 211 respectively. The spectra were normalized for concentration and path length to obtain the mean molar residue ellipticity. The thermal unfolding transitions were monitored at 225 nm and calculated as fraction of triple helix as described by Bächinger et al. (67). The heating rate was 1 °C/min.

Generation of human $\alpha_1\beta_1$, $\alpha_2\beta_1$, $\alpha_3\beta_1$, $\alpha_{10}\beta_1$, $\alpha_{11}\beta_1$ integrin ectodomain proteins

Human α_1 (1-1143), α_2 (1-1129), and β_1 (1-728) subunits were cloned separately into the pcDNA3.4 vector. The expression vectors of $\alpha_{10}\beta_1$ and $\alpha_{11}\beta_1$ were constructed into the pcDNA3.4 backbone as a continuous transcript using the P2A approach. Briefly, the α_{10} (1-1120) or α_{11}(1-1139) subunit was followed by a GGGS linker, and P2A ribosome skipping peptide (ATNFSLLKQAGDVEENPGP), and then the β_1 subunit. The α subunits contained an ACID coil and StrepII tag in the C-terminus; β subunits contained a BASE coil and His tag. All constructs were codon optimized. Expression vectors were transfected to 1-4 liters of Expi293 cells (Thermo Fisher) and cultured for 3-10 days. In most cases, the heterodimeric integrin ectodomain proteins were purified using Ni-NTA affinity chromatography, and when necessary, further purified by the StrepTactin affinity column. Finally, the aggregates or monomers were removed from the heterodimer by running a size exclusion column (Superdex 200 Increase) in the HBS buffer (25 mM HEPES pH 7.5, 150 mM NaCl, 1 mM $MgCl_2$, 1 mM $CaCl_2$). The human $\alpha_3\beta_1$ protein used in the assays was purchased from R&D Systems (catalog # 2840-A3-050).

Fluorescence Polarization (FP) assays

In the FP K_d assay, dilution series of the purified integrin proteins were incubated in 50 mM HEPES pH 7.3, 150 mM NaCl, 2 mM Mn^{2+}, 0.1 mM Ca^{2+}, 0.01% Triton X-100, 1% DMSO, and 3 nM of fluorescent BOP compound (R&D Systems, #6048) (68) in a volume of 20 µL in 384-well plates (PerkinElmer 6007270), at 22 ºC for 1 hr. Fluorescence polarization was measured with an EnVision plate reader (PerkinElmer) with excitation at 531 nm and emission at 595 nm. K_d values were determined using the total, one site model (Graphpad Prism). In FP IC_{50} assays, the integrins at their K_d concentration were incubated in FP assay buffer with a dilution series of the unlabeled BOP compound (R&D Systems, #6047) in a volume of 20 µL in 384-well plates at 22 ºC for 1 hr, FP was measured, and IC_{50} values were determined using 4-parameter regression fitting.

Solid Phase assays

Human collagen IV was purchased from Advanced Biomatrix (#5022-5MG).

The collagen proteins were diluted in 25 mM Tris-HCl pH 7.4, 150 mM NaCl, and 2 mM $CaCl_2$. 20 µL of the dilutions were added to wells of a 384 well plate (Corning 3577) and incubated at 4 °C for 16 hrs. The plates were washed 3 times with wash buffer consisting of PBS (Cytiva) with 0.05% Tween 20 (Sigma-Aldrich) and blocked with 40 µL of 2.5% bovine sera albumin (BSA, Fisher) in PBS at 37°C for 1 hr. The plates were washed 3 times with wash buffer and 20 µL of 5 nM integrin protein in 20

mM HEPES pH 7.3, 150 mM NaCl, 2 mM $MnCl_2$, 0.1 mM $CaCl_2$, 0.5% BSA, 0.01% Triton X-100, 1% DMSO with or without 50 µM BOP inhibitor was added (R&D Systems). The plate was incubated at 22 °C for 1.5 hrs and washed 3 times. 20 µL of the anti-β1 biotinylated antibody (R&D Systems, BAF1778) diluted 1:500 in PBS was added, incubated at 22 °C for 1 hr, and washed 3 times. 20 µL of the streptavidin-HRP solution (R&D Systems DY998) was added, incubated at 22 °C for 40 min, and washed 5 times. 20 µL of the QuantaBlu working solution (Thermo Fisher) was added, incubated at 22 °C for 15 min, and 20 µL of the QuantaBlu stop solution (Thermo Fisher) was added. Fluorescence (Ex 325 nm, Em 420 nm) was measured on a Cytation 5 plate reader (Agilent). Wells that were not coated with the collagen proteins served as the nonspecific binding control, and the fluorescence value of each coated sample well was subtracted by the average nonspecific binding control value. Next, for each collagen concentration, the fluorescence value of each inhibitor-free sample well was subtracted from the fluorescence value of each BOP-treated sample well to measure integrin-specific binding levels, which were analyzed using 4-parameter regression fitting to calculate EC_{50} values.

Computer modeling

The amino acid sequences of 112 eCB3 were modeled as six trimeric segments with overlapping. In particular, segment 1 included sequences beginning from the tags to the end of helix #7 (Fig. 1); segment 2: beginning of helix #7 to the end of helix #8; segment 3: beginning of helix #8 to the end of helix #9; segment 4: beginning of helix #9 to the end of helix #10; segment 5: beginning of helix #10 to the end of helix #11; segment 6: beginning of helix #11 to the end of the molecule (the end of NC2). Each segment was modeled using an open-source code for AlphaFold (69). Resulting six models were serially superimposed within overlapping sequences, trimmed to remove overlapping residues, and connected to mimic the whole molecule. Finally, the NC2 model was replaced with the crystal structure (PDB ID 5ctd). We observed several mismatches of registers of the modeled triple helical fragments, which ultimately resulted in artificial deletions and insertions. The final model can only be used for demonstration purpose and does not represent a meaningful model.

Data presentation and analysis

Data was analyzed using the GraphPad Prism (Dotmatics) and the MS Excel (Microsoft Inc.) software.

Plots were generated and visualized with the MS Excel (Microsoft Inc.) or the Grace program (http://plasma-gate.weizmann.ac.il/Grace/). Protein structure figures were generated using ChimeraX (70,71). Editing and labeling of figures were done using the GIMP (www.gimp.org) and Inkscape (inkscape.org) software packages.

Data availability

The DNA sequences of the constructs are available on request to S.P.B. and F.Y.L. All other data are contained within the manuscript or supporting information.

Supporting information

This article contains supporting information.

Acknowledgements

We are grateful to Patrick Page-McCaw for critical reading of the manuscript.

We thank the Vanderbilt Center for Structural Biology for use of Protein Characterization facilities. We are grateful to the Vanderbilt Institute of Nanoscale Science and Engineering for use of Advance Imaging facilities and Dr. Dmitry Koktysh for technical assistance with the Atomic Force Microscopy.

Author contributions

S.P.B., B.G.H., T.A.S., T.I.M., and F.Y.L. designed, coordinated the project, analyzed the data, and wrote the manuscript.

S.P.B. performed the constructs design, DNA cloning, sequence verification, protein expression, protein purification, gel analysis, protein size-exclusion chromatography, atomic force microscopy imaging, and circular dichroism spectroscopy.

Y.K. and K.M. performed PTM quantitation and analysis.

E.H.K. and W.K. performed the integrin binding assays and analyzed the data.

All authors read and approved the final manuscript.

Funding and additional information

This work was supported by the National Institutes of Health grants (R01DK018381, R56DK131101, and R01DK131101) to B.G.H. and S.P.B. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank the Vanderbilt Center for Matrix Biology for its support.

Molecular graphics and analyses performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.

Conflict of interest

Authors declare no conflict of interest

Figure legends

Figure 1. **Collagen IV and CB3-extended fragment used in this study**. (A) Schematic drawing of primary sequences of collagen IV α1, α1, and α2 chains that assemble into the heterotrimeric protomer. Triple helical segments $(GXY)_nG$ are depicted as colored bars connected with black lines representing interruptions of various lengths. Predicted helical segments are numbered from N to C termini. Cysteines are depicted as red pins. Blue arrows mark the reported CB3 cleavage sites (33). Black arrows label positions of the eCB3 sequences used in this study. (B) The extended CB3 (eCB3) sequences used to generate recombinant fragments in this study. Colored bars highlight triple helical segments. Most helical segments are shown at specific registers. Those registers are suggestive and the one for helix #9 is in agreement with the reported one (39). Black underlying boxes highlight binding sites for α1β1 and α2β1 integrins. GFPGER sequence, the major α2β1 integrin binding site and also the α1β1 integrin binding site is highlighted with cyan background. The suggested chain-distributed recognition site for the α1β1 integrin is additionally highlighted for D (blue background) and R (orange background) residues. Cysteines are shown as red pins. Red box highlights sequences forming the CB3 cystine knot. The Uniprot identifiers for sequences are P02462 for α1 chain (isoform 1) and P08572 for α2 chain.

Figure 2. **A tool for producing heterotrimeric collagen fragments.** (**A**) Collagen IX has three unique chains, α1, α2, and α3. The collagen IX assembly has three collagenous (COL1-COL4) and four non-collagenous (NC1-NC4) regions. The NC2 region is sufficient to select chains, trimerize them, and define the register of an adjacent triple helix (30,31). It can be used to drive the assembly of homo- and heterotrimeric fragments of other collagens. (**B**) Cassette of three chimeric sequences A, B, and C, where collagenous region of interest (sequences 1-3 corresponding to three chains of desired composition and register) flanked by collagen-like repeats ($(GPP)_5$ and $(GPP)_4$) is followed by the NC2 domain (chains α1-α3). Combination of all three chains results in the NC2 heterotrimer assembly, which initiates folding of the collagenous part. To facilitate purification and ensure robust heterotrimeric composition each chain contains a unique tag.

Figure 3. **The chains and eCB3 assemblies**. Schematic presentation of primary sequences of chains co-expressed to assemble 111, 222, 112, 121, and 211 eCB3 fragments. Linear scale corresponds to number of residues. The signal peptide sequences are excluded. Triple helical sequences are depicted as bars, yellow color corresponds to α1 chain of human collagen IV, light purple corresponds to α2, white to artificial GPP repeats. The sequences for NC2 domain are shown as pink, light blue and light green rectangles and labeled A, B, and C, which corresponds to α1, α2, and α3 chains of human collagen IX. Interruptions, linkers, and tags are depicted as black lines. Cysteines are depicted as red pins.

Figure 4. **Production scheme and purification of the eCB3 assemblies**. (**A**) Three corresponding plasmids A, B, and C were co-transfected and transiently expressed in expiCHO suspension culture. Every day a fresh ascorbate solution was added to the media to facilitate collagen-specific hydroxylation of proline and lysine residues. The conditioned medium was cleared from cells by centrifugation and extensively dialyzed before a series of affinity purifications. Three affinity columns ensured purification of three differently tagged chains for each eCB3 assembly, *i.e.,* 111, 222, 112, 121, and 211. (**B**) SDS-PAGE analysis of the purified proteins under non-reduced and reduced conditions. The gel was Coomassie stained. (**C**) Size-exclusion chromatography profiles of the purified proteins.

Figure 5. **Collagen-specific post-translational modifications**. The eCB3 fragments were hydrolyzed and analyzed by LC-MS with stable isotopically labeled collagen used as a reference to

quantitate proline (Pro), 4-hydroxyproline (4-Hyp), 3-hydroxyproline (3-Hyp), hydroxylysine (Hyl), galactosyl-hydroxylysine (GHL), and glucosyl-galactosyl-hydroxylysine (GGHL). (**A**) Overall content of Pro, 4-Hyp, and 3-Hyp. (**B**) Overall content of Lys and total Hyl after acid hydrolysis, which converted GHL and GGHL into Hyl. (**C**) Overall content of Lys, Hyl, GHL, and GGHL as a result of a combination of acid hydrolysis (Lys and total Hyl) and base hydrolysis (Hyl, GHL, and GGHL), which does not eliminate sugar moieties. The ratio [Lys + Hyl + GHL + GGHL = 100%] were calculated based on the ratios [Lys + total Hyl = 100%] determined in (B) and [total Hyl = Hyl + GHL + GGHL] to compare the two different hydrolysis. (**D**) Overall content of Hyl, GHL, and GGHL.

Figure 6. **AFM imaging of purified constructs**. Analysis of the eCB3 assemblies by atomic force microscopy (AFM) revealed the presence of individual molecules. All the assemblies except 222 showed individual ~70-nm-long worm-like structures each containing a triple helical collagenous domain.

Figure 7. **Circular dichroism spectra and thermal denaturation of the eCB3 assemblies**. (A) CD spectra of the eCB3 assemblies in 0.1 M sodium acetate buffer, pH 4.5 at 15 °C demonstrated similar profiles, which indicates similar content of the secondary structure. (B) Thermal unfolding transitions upon heating of the samples were measured at 225 nm and calculated as a fraction of helix. The apparent melting temperature was ~37 °C for all measured variants. The heating rate was 1 °C/min.

Figure 8. **Binding of integrins α1β1, α2β1, α10β1, α11β1, and α3β1 to full-length collagen IV and the eCB3 assemblies**. Collagen IV and the eCB3 assemblies were coated onto plastic at various concentrations and incubated with various integrins at 5 nM concentration for binding. Bound integrins were detected as fluorescence signal after using the anti-β1 biotinylated antibody, which was recognized by peroxidase-coupled streptavidin and fluorogenic peroxidase substrate. Specific binding was calculated as the difference of total and non-specific binding in the excess of integrin inhibitor BOP.

Figure 9. **An overall model of the 112 construct to demonstrate structural organization.** Surface presentation of the model assembled from short overlapping fragments predicted by AlphaFold (69). The model is for demonstration purpose only. Accuracy of the presented model is beyond the scope of this study. The eCB3 fragment is mapped as follows: the triple helical fragments #7-11 are chain-colored with yellow and magenta for α1 and α2 respectively, whereas the interruptions are colored black. Other cassette elements are mapped as white surface for N- and C-terminal $(GPP)_5$ and $(GPP)_4$ fragments, the chains of the NC2 domain are colored as pink, light blue, and light green, and tags are colored red, blue, and green. Suggested binding sites for integrins are highlighted by cyan and blue boxes. Inserts are demonstrating predicted disulfide bonds within the eCB3 fragment, *i.e.,* a disulfide bridge connecting α1 chains of the interruption between helical segments #8 and #9, and three inter-chain disulfide bonds forming a cystine knot within the interruption between segments #9 and #10.

References

1. Bächinger, H. P., Mizuno, K., Vranka, J. A., and Boudko, S. P. (2010) Collagen formation and structure. in *Comprehensive Natural Products II: Chemistry and Biology*, Elsevier Ltd. pp 469-530
2. Ricard-Blum, S. (2011) The collagen family. *Cold Spring Harb Perspect Biol* **3**, a004978
3. Fidler, A. L., Boudko, S. P., Rokas, A., and Hudson, B. G. (2018) The triple helix of collagens - an ancient protein structure that enabled animal multicellularity and tissue evolution. *J Cell Sci* **131**
4. Parkin, J. D., San Antonio, J. D., Pedchenko, V., Hudson, B., Jensen, S. T., and Savige, J. (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
5. San Antonio, J. D., Jacenko, O., Fertala, A., and Orgel, J. (2020) Collagen Structure-Function Mapping Informs Applications for Regenerative Medicine. *Bioengineering (Basel)* **8**
6. Okuyama, K. (2008) Revisiting the molecular structure of collagen. *Connect Tissue Res* **49**, 299-310
7. Khoshnoodi, J., Pedchenko, V., and Hudson, B. G. (2008) Mammalian collagen IV. *Microsc Res Tech* **71**, 357-370
8. Hudson, B. G., Reeders, S. T., and Tryggvason, K. (1993) Type IV collagen: structure, gene organization, and role in human diseases. Molecular basis of Goodpasture and Alport syndromes and diffuse leiomyomatosis. *J Biol Chem* **268**, 26033-26036
9. Okuyama, K., Okuyama, K., Arnott, S., Takayanagi, M., and Kakudo, M. (1981) Crystal and molecular structure of a collagen-like polypeptide (Pro-Pro-Gly)10. *J Mol Biol* **152**, 427-443
10. Bella, J., Eaton, M., Brodsky, B., and Berman, H. M. (1994) Crystal and molecular structure of a collagen-like peptide at 1.9 A resolution. *Science* **266**, 75-81
11. Emsley, J., Knight, C. G., Farndale, R. W., Barnes, M. J., and Liddington, R. C. (2000) Structural basis of collagen recognition by integrin alpha2beta1. *Cell* **101**, 47-56
12. Raynal, N., Hamaia, S. W., Siljander, P. R., Maddox, B., Peachey, A. R., Fernandez, R., Foley, L. J., Slatter, D. A., Jarvis, G. E., and Farndale, R. W. (2006) Use of synthetic peptides to locate novel integrin alpha2beta1-binding motifs in human collagen III. *J Biol Chem* **281**, 3821-3831
13. Lisman, T., Raynal, N., Groeneveld, D., Maddox, B., Peachey, A. R., Huizinga, E. G., de Groot, P. G., and Farndale, R. W. (2006) A single high-affinity binding site for von Willebrand factor in collagen III, identified using synthetic triple-helical peptides. *Blood* **108**, 3753-3756
14. Xu, H., Raynal, N., Stathopoulos, S., Myllyharju, J., Farndale, R. W., and Leitinger, B. (2011) Collagen binding specificity of the discoidin domain receptors: binding sites on collagens II and III and molecular determinants for collagen IV recognition by DDR1. *Matrix Biol* **30**, 16-26
15. Konitsiotis, A. D., Raynal, N., Bihan, D., Hohenester, E., Farndale, R. W., and Leitinger, B. (2008) Characterization of high affinity binding motifs for the discoidin domain receptor DDR2 in collagen. *J Biol Chem* **283**, 6861-6868
16. Jensen, M. M., Bonna, A., Frederiksen, S. J., Hamaia, S. W., Hojrup, P., Farndale, R. W., and Karring, H. (2022) Tyrosine-sulfated dermatopontin shares multiple binding sites and recognition determinants on triple-helical collagens with proteins implicated in cell adhesion and collagen folding, fibrillogenesis, cross-linking, and degradation. *Biochim Biophys Acta Proteins Proteom* **1870**, 140771
17. Cai, H., Sasikumar, P., Little, G., Bihan, D., Hamaia, S. W., Zhou, A., Gibbins, J. M., and Farndale, R. W. (2021) Identification of HSP47 Binding Site on Native Collagen and Its Implications for the Development of HSP47 Inhibitors. *Biomolecules* **11**
18. Jarvis, G. E., Raynal, N., Langford, J. P., Onley, D. J., Andrews, A., Smethurst, P. A., and Farndale, R. W. (2008) Identification of a major GpVI-binding locus in human type III collagen. *Blood* **111**, 4986-4996
19. Lebbink, R. J., Raynal, N., de Ruiter, T., Bihan, D. G., Farndale, R. W., and Meyaard, L. (2009) Identification of multiple potent binding sites for human leukocyte associated Ig-like receptor LAIR on collagens II and III. *Matrix Biol* **28**, 202-210
20. Giudici, C., Raynal, N., Wiedemann, H., Cabral, W. A., Marini, J. C., Timpl, R., Bachinger, H. P., Farndale, R. W., Sasaki, T., and Tenni, R. (2008) Mapping of SPARC/BM-40/osteonectin-binding sites on fibrillar collagens. *J Biol Chem* **283**, 19551-19560
21. Manka, S. W., Bihan, D., and Farndale, R. W. (2019) Structural studies of the MMP-3 interaction with triple-helical collagen introduce new roles for the enzyme in tissue remodelling. *Sci Rep* **9**, 18785

22. Leatherdale, A., Parker, D., Yasheem, S., Wang, Y., Binan, D., Bonna, A., Hamaia, S. W., Gross, P. L., Ni, H., Doble, B. W., Lillicrap, D., Farndale, R. W., and Hayward, C. P. M. (2021) Multimerin 1 supports platelet function in vivo and binds to specific GPAGPOGPX motifs in fibrillar collagens that enhance platelet adhesion. *J Thromb Haemost* **19**, 547-561
23. Farndale, R. W. (2019) Collagen-binding proteins: insights from the Collagen Toolkits. *Essays Biochem* **63**, 337-348
24. Sacca, B., and Moroder, L. (2002) Synthesis of heterotrimeric collagen peptides containing the alpha1beta1 integrin recognition site of collagen type IV. *J Pept Sci* **8**, 192-204
25. Ottl, J., Battistuta, R., Pieper, M., Tschesche, H., Bode, W., Kuhn, K., and Moroder, L. (1996) Design and synthesis of heterotrimeric collagen peptides with a built-in cystine-knot. Models for collagen catabolism by matrix-metalloproteases. *FEBS Lett* **398**, 31-36
26. Boulegue, C., Musiol, H. J., Gotz, M. G., Renner, C., and Moroder, L. (2008) Natural and artificial cystine knots for assembly of homo- and heterotrimeric collagen models. *Antioxid Redox Signal* **10**, 113-125
27. Jalan, A. A., Sammon, D., Hartgerink, J. D., Brear, P., Stott, K., Hamaia, S. W., Hunter, E. J., Walker, D. R., Leitinger, B., and Farndale, R. W. (2020) Chain alignment of collagen I deciphered using computationally designed heterotrimers. *Nat Chem Biol* **16**, 423-429
28. Acevedo-Jake, A. M., Clements, K. A., and Hartgerink, J. D. (2016) Synthetic, Register-Specific, AAB Heterotrimers to Investigate Single Point Glycine Mutations in Osteogenesis Imperfecta. *Biomacromolecules* **17**, 914-921
29. Xu, Y., and Kirchner, M. (2021) Collagen Mimetic Peptides. *Bioengineering (Basel)* **8**
30. Boudko, S. P., and Bachinger, H. P. (2012) The NC2 domain of type IX collagen determines the chain register of the triple helix. *J Biol Chem* **287**, 44536-44545
31. Boudko, S. P., and Bachinger, H. P. (2016) Structural insight for chain selection and stagger control in collagen. *Sci Rep* **6**, 37831
32. Boudko, S. P., Zientek, K. D., Vance, J., Hacker, J. L., Engel, J., and Bachinger, H. P. (2010) The NC2 domain of collagen IX provides chain selection and heterotrimerization. *J Biol Chem* **285**, 23721-23731
33. Vandenberg, P., Kern, A., Ries, A., Luckenbill-Edds, L., Mann, K., and Kuhn, K. (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
34. Underwood, P. A., Bennett, F. A., Kirkpatrick, A., Bean, P. A., and Moss, B. A. (1995) Evidence for the location of a binding sequence for the alpha 2 beta 1 integrin of endothelial cells, in the beta 1 subunit of laminin. *Biochem J* **309 (Pt 3)**, 765-771
35. Plaisier, E., Chen, Z., Gekeler, F., Benhassine, S., Dahan, K., Marro, B., Alamowitch, S., Paques, M., and Ronco, P. (2010) Novel COL4A1 mutations associated with HANAC syndrome: a role for the triple helical CB3[IV] domain. *Am J Med Genet A* **152A**, 2550-2555
36. Kern, A., Eble, J., Golbik, R., and Kuhn, K. (1993) Interaction of type IV collagen with the isolated integrins alpha 1 beta 1 and alpha 2 beta 1. *Eur J Biochem* **215**, 151-159
37. Fleischmajer, R., Perlish, J. S., MacDonald, E. D., 2nd, Schechter, A., Murdoch, A. D., Iozzo, R. V., and Yamada, Y. (1998) There is binding of collagen IV to beta 1 integrin during early skin basement membrane assembly. *Ann N Y Acad Sci* **857**, 212-227
38. Eble, J. A., Ries, A., Lichy, A., Mann, K., Stanton, H., Gavrilovic, J., Murphy, G., and Kuhn, K. (1996) The recognition sites of the integrins alpha1beta1 and alpha2beta1 within collagen IV are protected against gelatinase A attack in the native protein. *J Biol Chem* **271**, 30964-30970
39. Eble, J. A., Golbik, R., Mann, K., and Kuhn, K. (1993) The alpha 1 beta 1 integrin recognition site of the basement membrane collagen molecule [alpha 1(IV)]2 alpha 2(IV). *EMBO J* **12**, 4795-4802
40. Dinkla, K., Talay, S. R., Morgelin, M., Graham, R. M., Rohde, M., Nitsche-Schmitz, D. P., and Chhatwal, G. S. (2009) Crucial role of the CB3-region of collagen IV in PARF-induced acute rheumatic fever. *PLoS One* **4**, e4666
41. Calderwood, D. A., Tuckwell, D. S., Eble, J., Kuhn, K., and Humphries, M. J. (1997) The integrin alpha1 A-domain is a ligand binding site for collagens and laminin. *J Biol Chem* **272**, 12311-12317
42. Lunstrum, G. P., Bachinger, H. P., Fessler, L. I., Duncan, K. G., Nelson, R. E., and Fessler, J. H. (1988) Drosophila basement membrane procollagen IV. I. Protein characterization and distribution. *J Biol Chem* **263**, 18318-18327

43. Murad, S., Grove, D., Lindberg, K. A., Reynolds, G., Sivarajah, A., and Pinnell, S. R. (1981) Regulation of collagen synthesis by ascorbic acid. *Proc Natl Acad Sci U S A* **78**, 2879-2882
44. Taga, Y., Kusubata, M., Ogawa-Goto, K., and Hattori, S. (2014) Stable isotope-labeled collagen: a novel and versatile tool for quantitative collagen analyses using mass spectrometry. *J Proteome Res* **13**, 3671-3678
45. Bruckner, P., Eikenberry, E. F., and Prockop, D. J. (1981) Formation of the triple helix of type I procollagen in cellulo. A kinetic model based on cis-trans isomerization of peptide bonds. *Eur J Biochem* **118**, 607-613
46. Mizuno, K., Boudko, S. P., Engel, J., and Bachinger, H. P. (2010) Kinetic hysteresis in collagen folding. *Biophys J* **98**, 3004-3014
47. Pepinsky, R. B., Mumford, R. A., Chen, L. L., Leone, D., Amo, S. E., Riper, G. V., Whitty, A., Dolinski, B., Lobb, R. R., Dean, D. C., Chang, L. L., Raab, C. E., Si, Q., Hagmann, W. K., and Lingham, R. B. (2002) Comparative assessment of the ligand and metal ion binding properties of integrins alpha9beta1 and alpha4beta1. *Biochemistry* **41**, 7125-7141
48. Chen, Z., Migeon, T., Verpont, M. C., Zaidan, M., Sado, Y., Kerjaschki, D., Ronco, P., and Plaisier, E. (2016) HANAC Syndrome Col4a1 Mutation Causes Neonate Glomerular Hyperpermeability and Adult Glomerulocystic Kidney Disease. *J Am Soc Nephrol* **27**, 1042-1054
49. Salem, R. M., Todd, J. N., Sandholm, N., Cole, J. B., Chen, W. M., Andrews, D., Pezzolesi, M. G., McKeigue, P. M., Hiraki, L. T., Qiu, C., Nair, V., Di Liao, C., Cao, J. J., Valo, E., Onengut-Gumuscu, S., Smiles, A. M., McGurnaghan, S. J., Haukka, J. K., Harjutsalo, V., Brennan, E. P., van Zuydam, N., Ahlqvist, E., Doyle, R., Ahluwalia, T. S., Lajer, M., Hughes, M. F., Park, J., Skupien, J., Spiliopoulou, A., Liu, A., Menon, R., Boustany-Kari, C. M., Kang, H. M., Nelson, R. G., Klein, R., Klein, B. E., Lee, K. E., Gao, X., Mauer, M., Maestroni, S., Caramori, M. L., de Boer, I. H., Miller, R. G., Guo, J., Boright, A. P., Tregouet, D., Gyorgy, B., Snell-Bergeon, J. K., Maahs, D. M., Bull, S. B., Canty, A. J., Palmer, C. N. A., Stechemesser, L., Paulweber, B., Weitgasser, R., Sokolovska, J., Rovite, V., Pirags, V., Prakapiene, E., Radzeviciene, L., Verkauskiene, R., Panduru, N. M., Groop, L. C., McCarthy, M. I., Gu, H. F., Mollsten, A., Falhammar, H., Brismar, K., Martin, F., Rossing, P., Costacou, T., Zerbini, G., Marre, M., Hadjadj, S., McKnight, A. J., Forsblom, C., McKay, G., Godson, C., Maxwell, A. P., Kretzler, M., Susztak, K., Colhoun, H. M., Krolewski, A., Paterson, A. D., Groop, P. H., Rich, S. S., Hirschhorn, J. N., Florez, J. C., and Summit Consortium, D. E. R. G. G. C. (2019) Genome-Wide Association Study of Diabetic Kidney Disease Highlights Biology Involved in Glomerular Basement Membrane Collagen. *J Am Soc Nephrol* **30**, 2000-2016
50. Letarov, A. V., Londer, Y. Y., Boudko, S. P., and Mesyanzhinov, V. V. (1999) The carboxy-terminal domain initiates trimerization of bacteriophage T4 fibritin. *Biochemistry (Mosc)* **64**, 817-823
51. Frank, S., Kammerer, R. A., Mechling, D., Schulthess, T., Landwehr, R., Bann, J., Guo, Y., Lustig, A., Bachinger, H. P., and Engel, J. (2001) Stabilization of short collagen-like triple helices by protein engineering. *J Mol Biol* **308**, 1081-1089
52. Frank, S., Boudko, S., Mizuno, K., Schulthess, T., Engel, J., and Bachinger, H. P. (2003) Collagen triple helix formation can be nucleated at either end. *J Biol Chem* **278**, 7747-7750
53. Boudko, S., Frank, S., Kammerer, R. A., Stetefeld, J., Schulthess, T., Landwehr, R., Lustig, A., Bachinger, H. P., and Engel, J. (2002) Nucleation and propagation of the collagen triple helix in single-chain and trimerized peptides: transition from third to first order kinetics. *J Mol Biol* **317**, 459-470
54. Stetefeld, J., Frank, S., Jenny, M., Schulthess, T., Kammerer, R. A., Boudko, S., Landwehr, R., Okuyama, K., and Engel, J. (2003) Collagen stabilization at atomic level: crystal structure of designed (GlyProPro)10foldon. *Structure* **11**, 339-346
55. Fouet, G., Bally, I., Signor, L., Haussermann, K., Thielens, N. M., Rossi, V., and Gaboriaud, C. (2021) Headless C1q: a new molecular tool to decipher its collagen-like functions. *FEBS J* **288**, 2030-2041
56. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J. X., Samocha, K. E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A. H., Minikel, E. V., Weisburd, B., Lek, M., Ware, J. S., Vittal, C., Armean, I. M., Bergelson, L., Cibulskis, K., Connolly, K. M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R.,

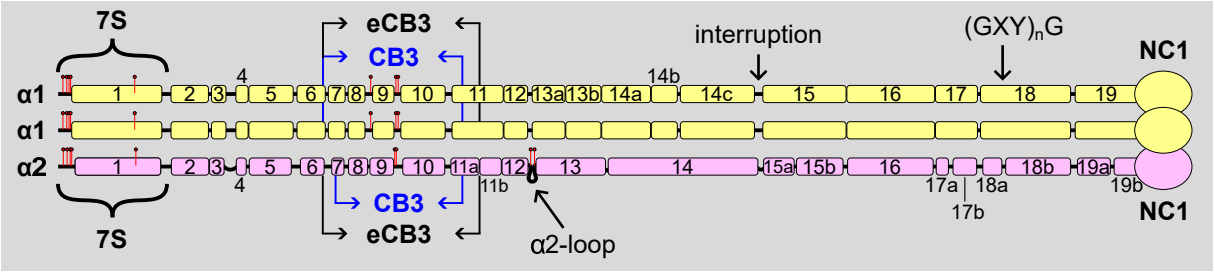
Novod, S., Petrilli, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M. E., Genome Aggregation Database, C., Neale, B. M., Daly, M. J., and MacArthur, D. G. (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443
57. Al-Shaer, A., Lyons, A., Ishikawa, Y., Hudson, B. G., Boudko, S. P., and Forde, N. R. (2021) Sequence-dependent mechanics of collagen reflect its structural and functional organization. *Biophys J* **120**, 4013-4028
58. Bachinger, H. P., Doege, K. J., Petschek, J. P., Fessler, L. I., and Fessler, J. H. (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
59. Hofmann, H., Voss, T., Kuhn, K., and Engel, J. (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
60. Zheng, H., Lu, C., Lan, J., Fan, S., Nanda, V., and Xu, F. (2018) How electrostatic networks modulate specificity and stability of collagen. *Proc Natl Acad Sci U S A* **115**, 6207-6212
61. Jalan, A. A., and Hartgerink, J. D. (2013) Simultaneous control of composition and register of an AAB-type collagen heterotrimer. *Biomacromolecules* **14**, 179-185
62. Fallas, J. A., and Hartgerink, J. D. (2012) Computational design of self-assembling register-specific collagen heterotrimers. *Nat Commun* **3**, 1087
63. Stawikowski, M. J., Aukszi, B., Stawikowska, R., Cudic, M., and Fields, G. B. (2014) Glycosylation modulates melanoma cell alpha2beta1 and alpha3beta1 integrin interactions with type IV collagen. *J Biol Chem* **289**, 21591-21604
64. Hynes, R. O., Ruoslahti, E., and Springer, T. A. (2022) Reflections on Integrins-Past, Present, and Future: The Albert Lasker Basic Medical Research Award. *JAMA* **328**, 1291-1292
65. Netzer, K. O., Leinonen, A., Boutaud, A., Borza, D. B., Todd, P., Gunwar, S., Langeveld, J. P., and Hudson, B. G. (1999) The goodpasture autoantigen. Mapping the major conformational epitope(s) of alpha3(IV) collagen to residues 17-31 and 127-141 of the NC1 domain. *J Biol Chem* **274**, 11267-11274
66. Boudko, S. P., Danylevych, N., Hudson, B. G., and Pedchenko, V. K. (2018) Basement membrane collagen IV: Isolation of functional domains. *Methods Cell Biol* **143**, 171-185
67. Bachinger, H. P., Bruckner, P., Timpl, R., Prockop, D. J., and Engel, J. (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
68. Cao, B., Hutt, O. E., Zhang, Z., Li, S., Heazlewood, S. Y., Williams, B., Smith, J. A., Haylock, D. N., Savage, G. P., and Nilsson, S. K. (2014) Design, synthesis and binding properties of a fluorescent alpha(9)beta(1)/alpha(4)beta(1) integrin antagonist and its application as an in vivo probe for bone marrow haemopoietic stem cells. *Org Biomol Chem* **12**, 965-978
69. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589
70. Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H., and Ferrin, T. E. (2021) UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70-82
71. Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H., and Ferrin, T. E. (2018) UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci* **27**, 14-25

Table 1. EC_{50} values of the collagen proteins in solid phase assays with 5 nM purified integrin proteins. Integrin-specific binding was analyzed using 4-parameter regression fitting using the GraphPad Prism software.

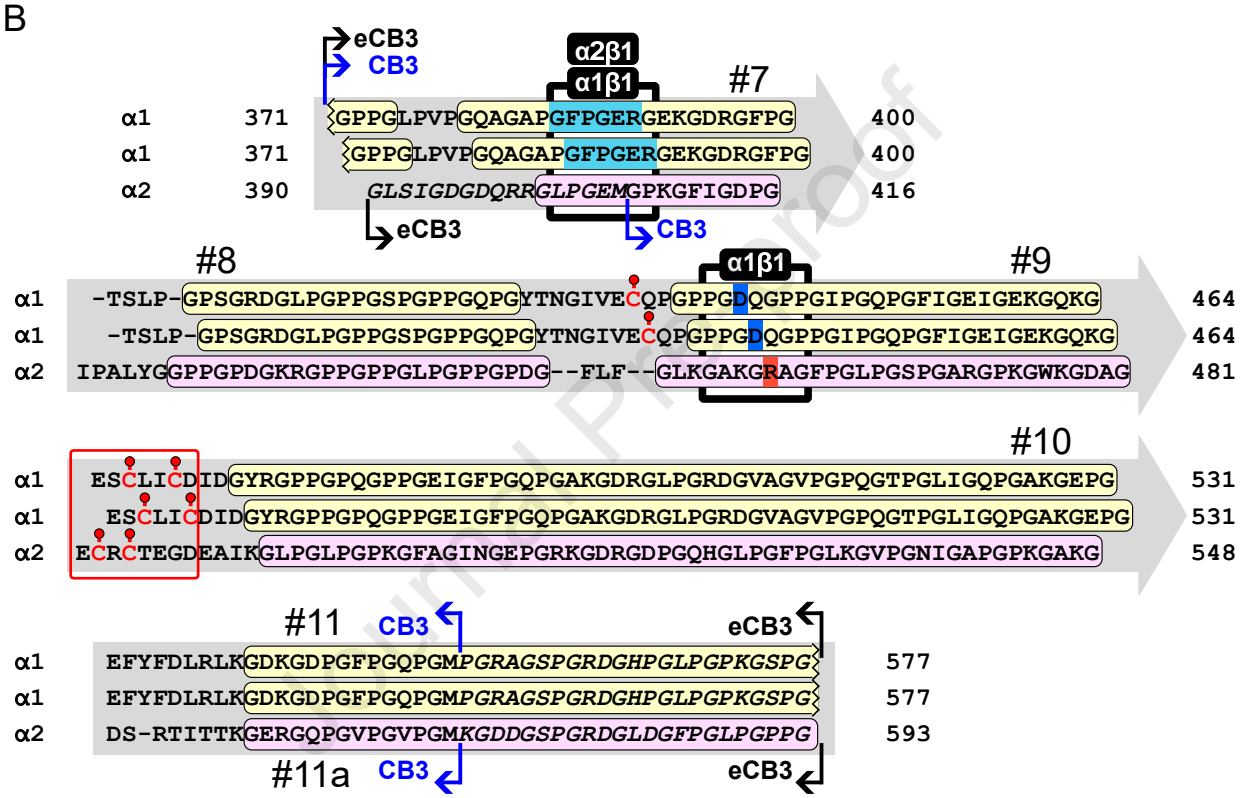
Integrin protein	Collagen IV (µg/mL)	eCB3-111 (µg/mL)	eCB3-112 (µg/mL)	eCB3-121 (µg/mL)	eCB3-211 (µg/mL)
$\alpha_1\beta_1$	0.40	1.33	0.25	0.23	0.29
$\alpha_2\beta_1$	0.20	0.32	0.18	0.21	0.34
$\alpha_3\beta_1$	ND	ND	ND	ND	ND
$\alpha_{10}\beta_1$	ND	ND	ND	ND	ND
$\alpha_{11}\beta_1$	0.93	1.16	1.11	1.07	2.09

ND, Not Defined

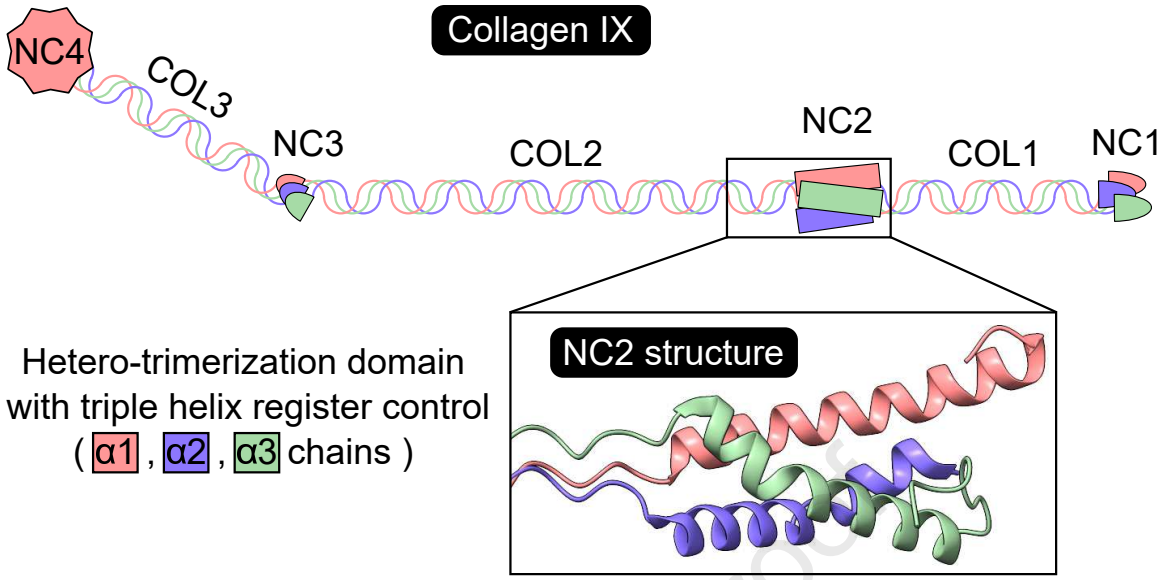
A



B



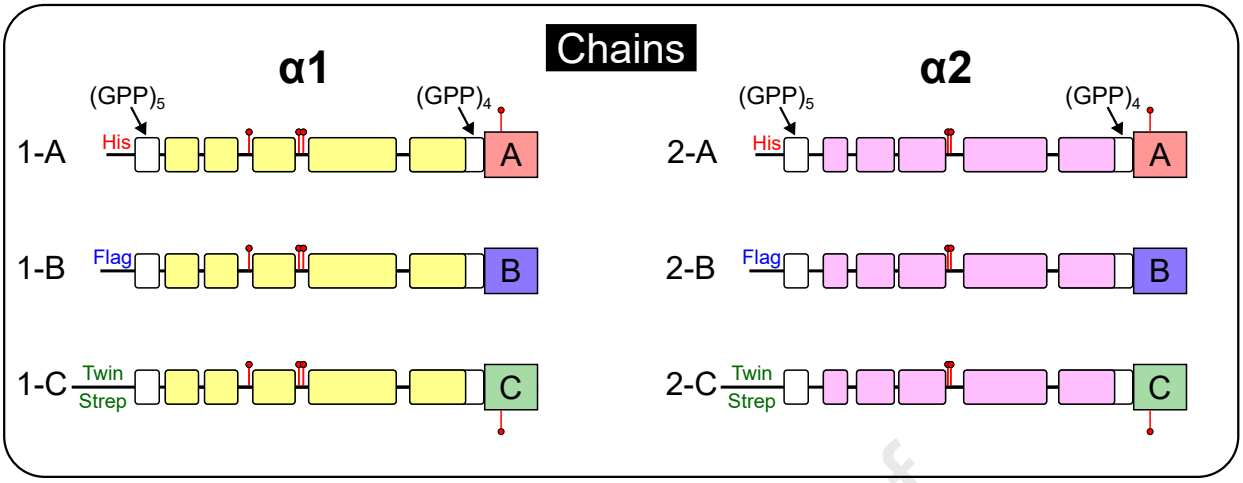
A

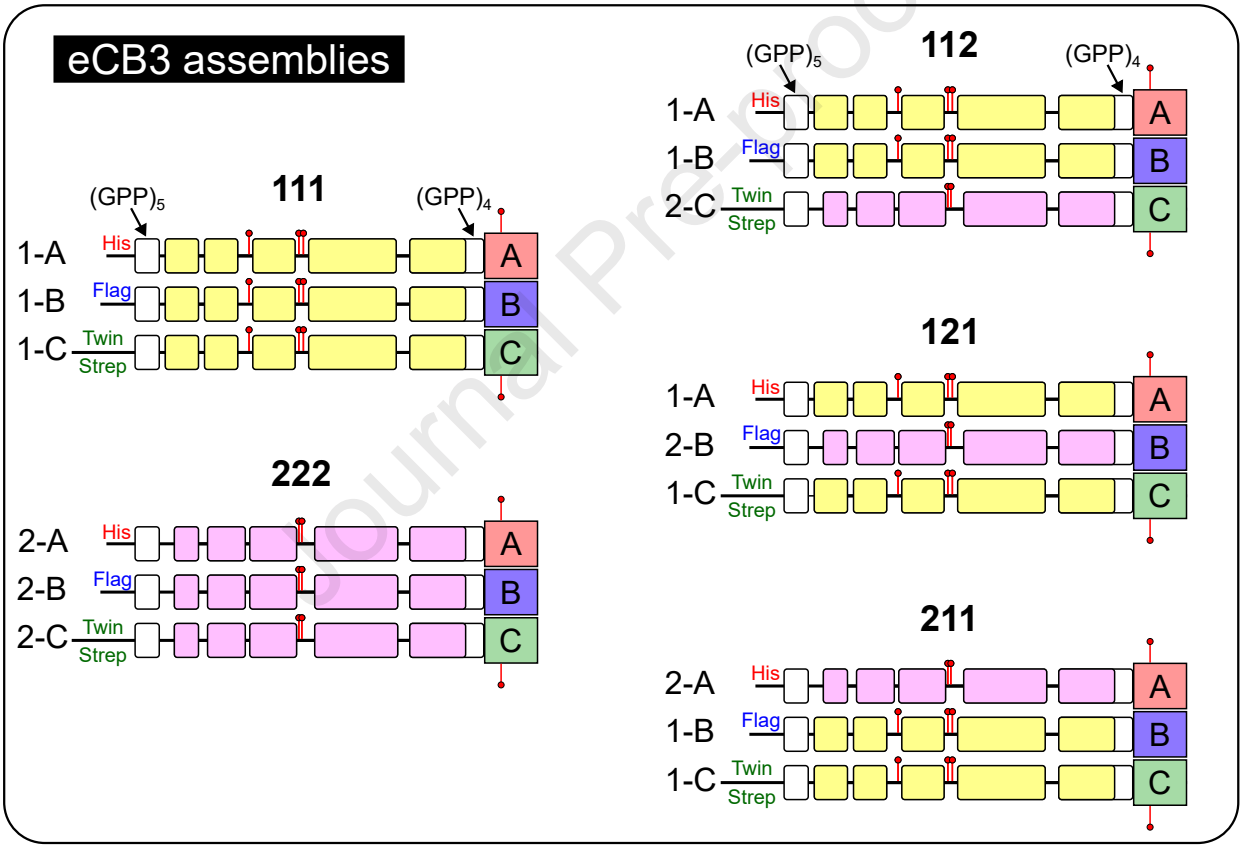


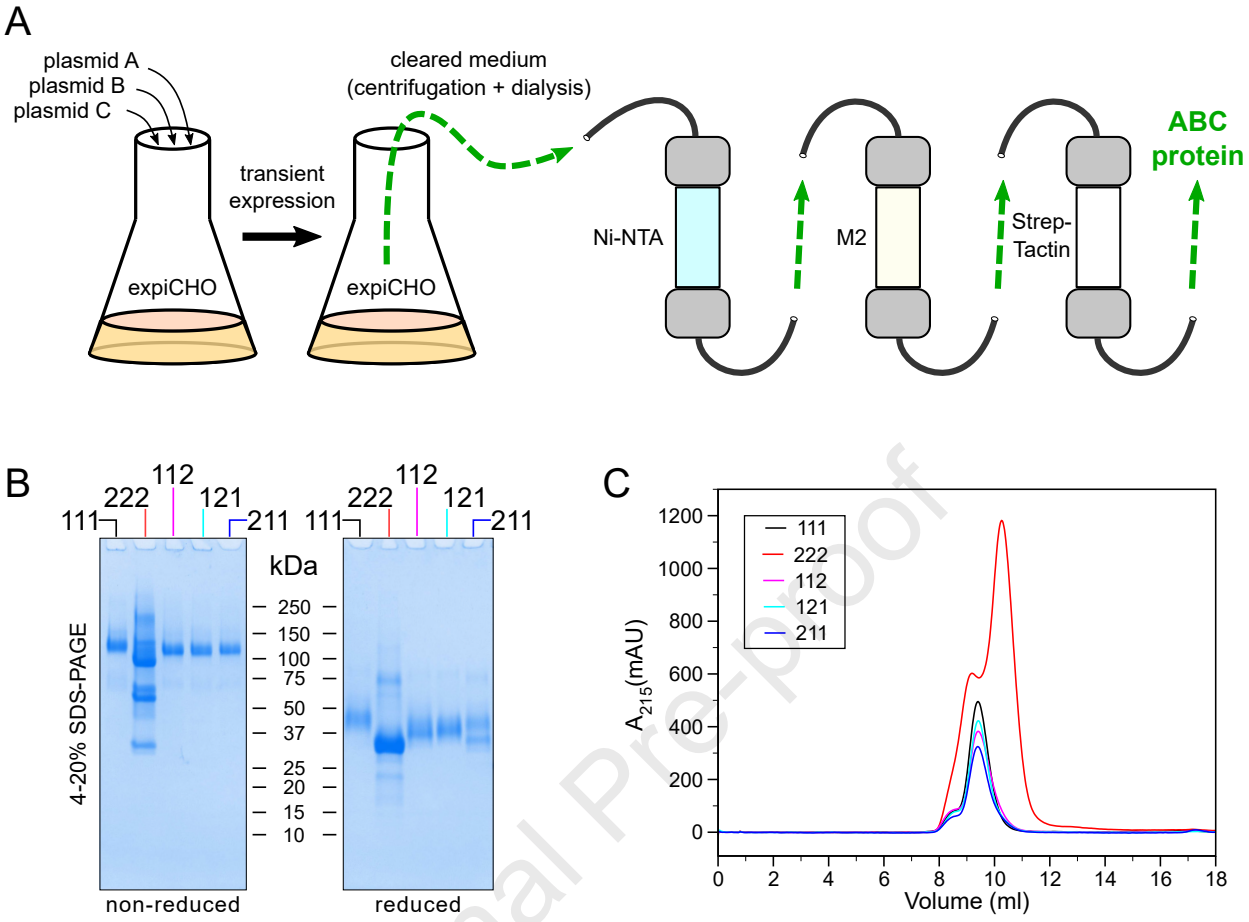
B

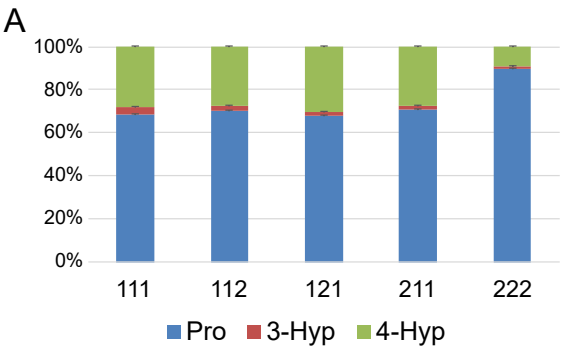
Cassette for heterotrimeric collagen fragment

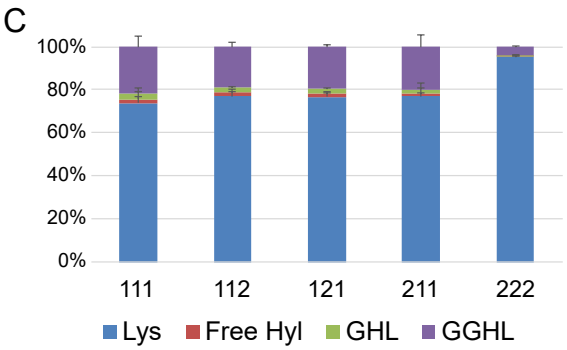
	Tags	$(GPP)_5$	collagen fragment	$(GPP)_4$	NC2
A:	(His-tag)--GPPGPPGPPGPPGPPGPP-		sequence 1	-GPPGPPGPPGPPGPP-	(IX_α1)
B:	(Flag-tag)--GPPGPPGPPGPPGPPGPP-		sequence 2	-GPPGPPGPPGPPGPP-	(IX_α2)
C:	(Twin-Strep)-GPPGPPGPPGPPGPPGPP-		sequence 3	-GPPGPPGPPGPPGPP-	(IX_α3)

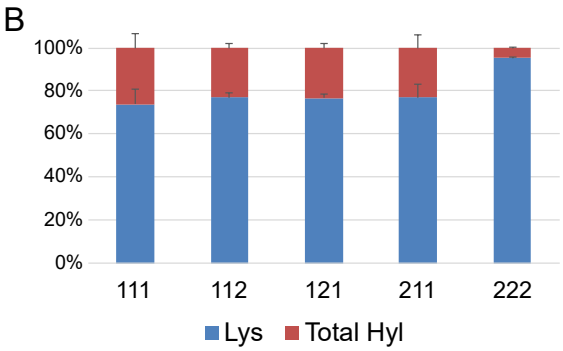


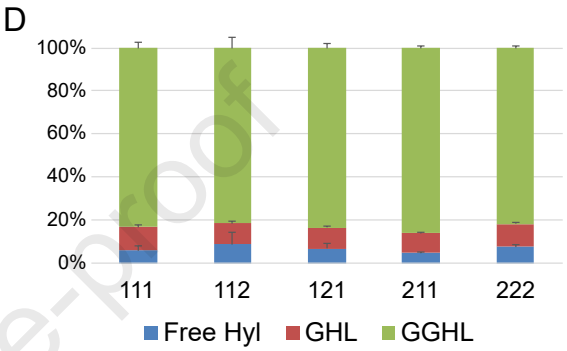


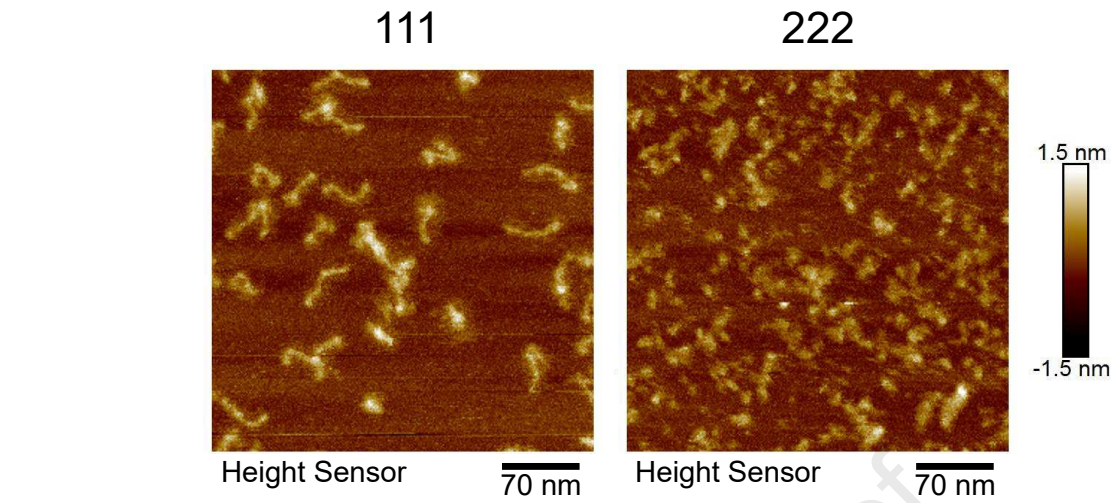


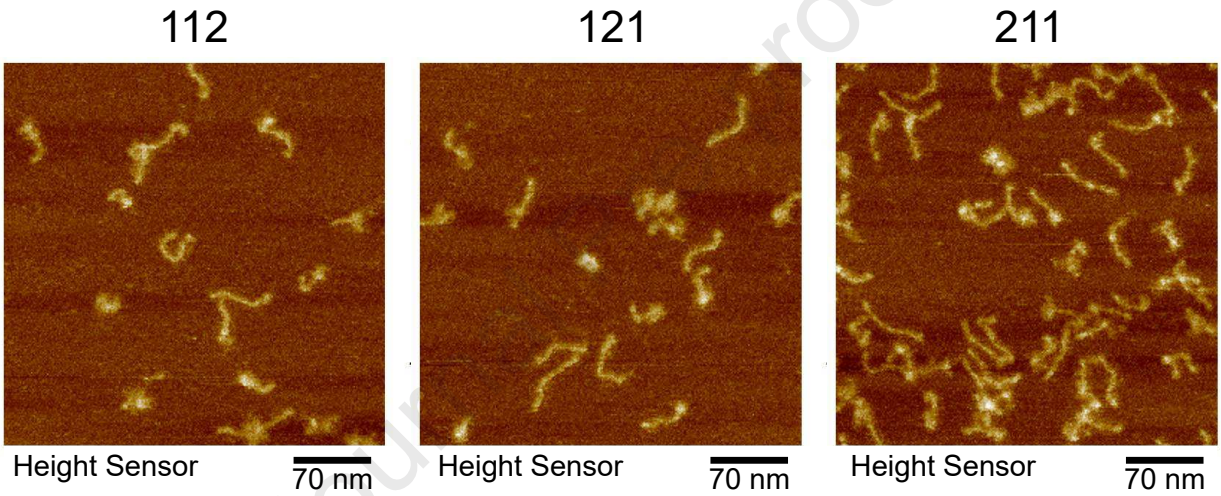


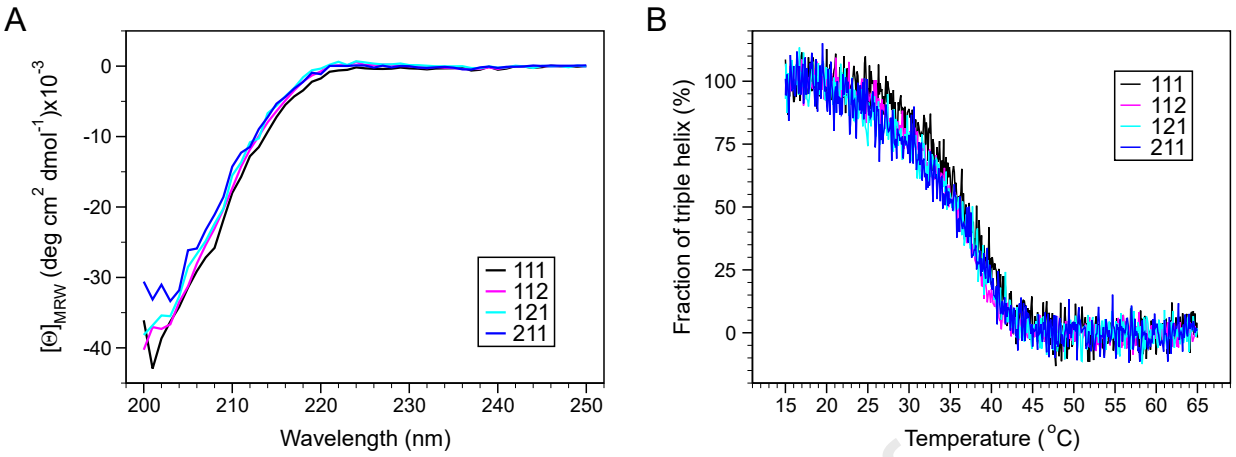


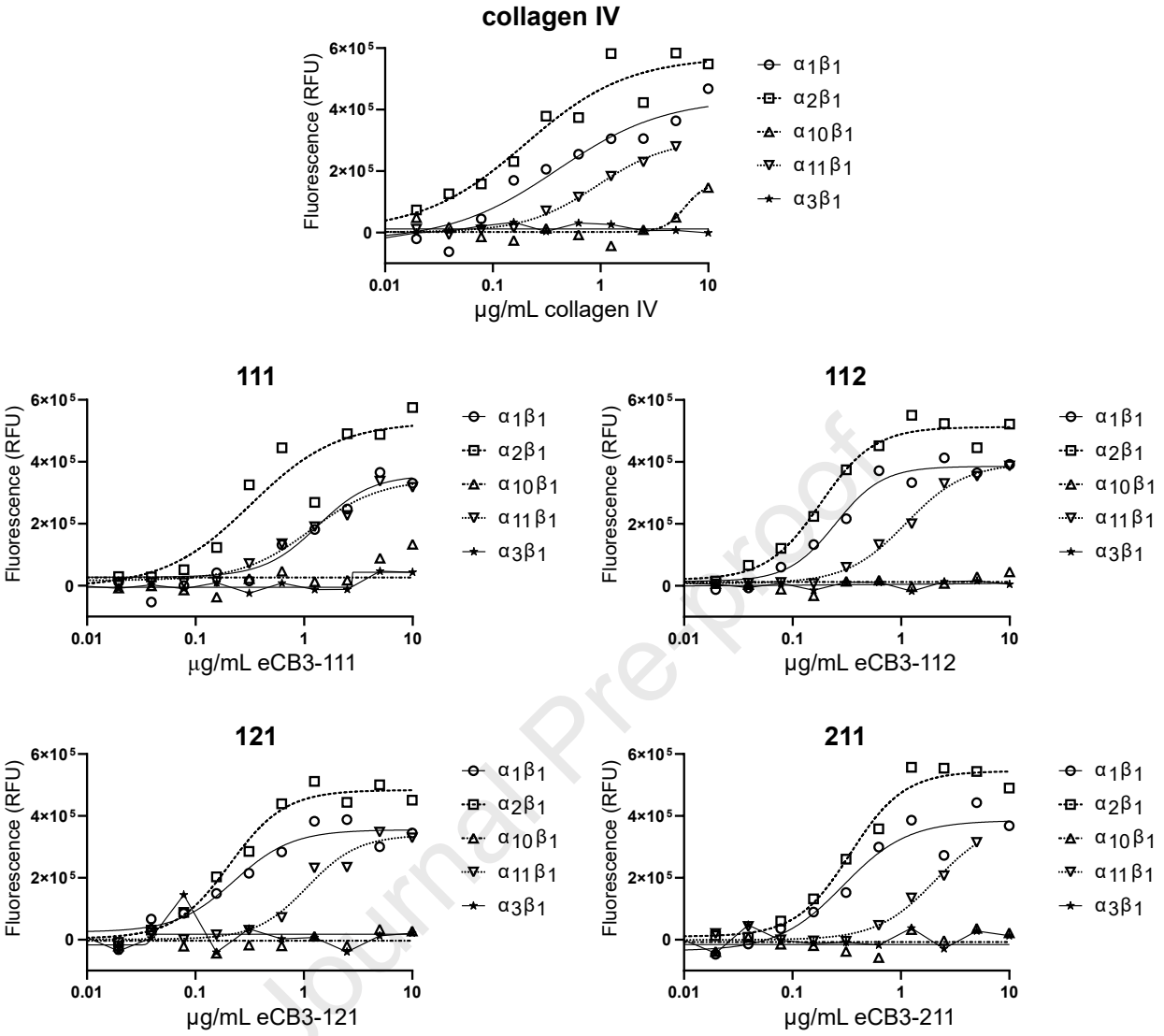


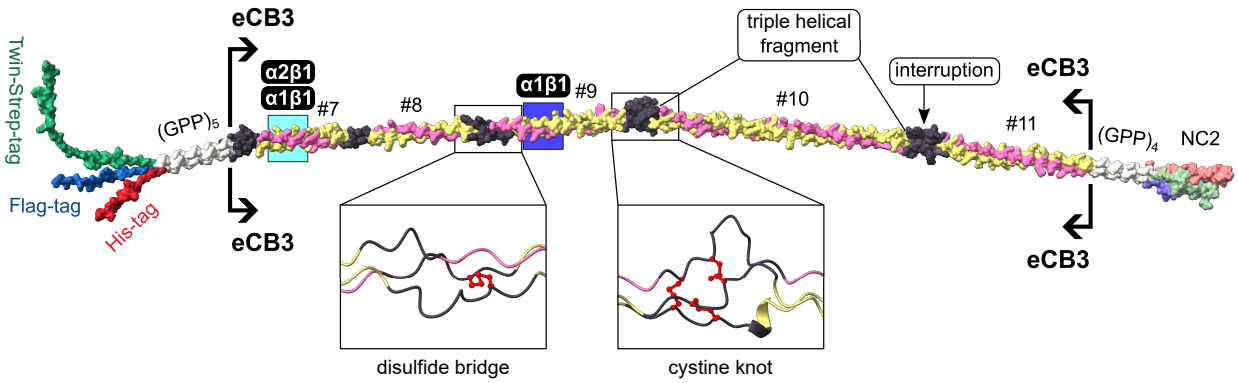












Conflict of interest

Authors declare no conflict of interest

CRediT author statement

Sergei P. Boudko: Conceptualization, Methodology, Investigation, Validation, Formal analysis, Visualization, Writing- Original draft preparation, Writing- Reviewing and Editing, Supervision; **Elizabeth H. Konopka**: Investigation, Validation; **Woojin Kim**: Investigation, Validation; **Yuki Taga**: Investigation, Validation, Formal analysis, Visualization; **Kazunori Mizuno**: Investigation, Validation, Formal analysis; **Timothy A. Springer**: Conceptualization, Writing- Reviewing and Editing; **Billy G. Hudson**: Conceptualization, Resources, Writing- Reviewing and Editing; **Terence I. Moy**: Conceptualization, Methodology, Validation, Formal analysis, Visualization, Writing- Reviewing and Editing; **Fu-Yang Lin**: Conceptualization, Methodology, Resources, Writing- Reviewing and Editing, Supervision.

Supporting Information

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: **A recombinant technique for mapping functional sites of hetero-trimeric collagen helices: collagen IV CB3 fragment as a prototype for integrin binding**

Authors

Sergei P. Boudko, Elizabeth H. Konopka, Woojin Kim, Yuki Taga, Kazunori Mizuno, Timothy A. Springer, Billy G. Hudson, Terence I. Moy, Fu-Yang Lin

Content

Table S1. **Plasmids used for cloning and expression**.

Plasmid name	Content	Type
pUC_His-GPP5-2xBsmBI-GPP4-NC2a1	Host framework with His-tag and α1 chain of NC2 domain of human collagen IX	cloning
pUC_Flag-GPP5-2xBsmBI-GPP4-NC2a2	Host framework with His-tag and α2 chain of NC2 domain of human collagen IX	cloning
pUC_2xStrep-GPP5-2xBsmBI-GPP4-NC2a3	Host framework with His-tag and α3 chain of NC2 domain of human collagen IX	cloning
pRcX	Expression vector with CMV promoter	expression
pRc_His-GPP5-2xBsmBI-GPP4-NC2a1	Host framework with His-tag and α1 chain of NC2 domain of human collagen IX	cloning/expression
pRc_Flag-GPP5-2xBsmBI-GPP4-NC2a2	Host framework with His-tag and α2 chain of NC2 domain of human collagen IX	cloning/expression
pRc_2xStrep-GPP5-2xBsmBI-GPP4-NC2a3	Host framework with His-tag and α3 chain of NC2 domain of human collagen IX	cloning/expression
pUC_a1CB3IV	CB3 extended sequence of α1 chain of human collagen IV	cloning
pUC_a2CB3IV	CB3 extended sequence of α2 chain of human collagen IV	cloning
pRc_His-GPP5-CB3a1-GPP4-NC2a1	CB3 α1 guest sequence in the NC2 α1 host, His-tagged	expression
pRc_His-GPP5-CB3a2-GPP4-NC2a1	CB3 α2 guest sequence in the NC2 α1 host, His-tagged	expression
pRc_Flag-GPP5-CB3a1-GPP4-NC2a2	CB3 α1 guest sequence in the NC2 α2 host, FLAG®-tagged	expression
pRc_Flag-GPP5-CB3a2-GPP4-NC2a2	CB3 α2 guest sequence in the NC2 α2 host, FLAG®-tagged	expression
pRc_2xStrep-GPP5-CB3a1-GPP4-NC2a3	CB3 α1 guest sequence in the NC2 α3host, Twin-Strep®-tagged	expression
pRc_2xStrep-GPP5-CB3a2-GPP4-NC2a3	CB3 α2 guest sequence in the NC2 α3 host, Twin-Strep®-tagged	expression

Table S2. **Alias names for the expression plasmids**.

Plasmid name	Alias
pRc_His-GPP5-CB3a1-GPP4-NC2a1	1-1
pRc_His-GPP5-CB3a2-GPP4-NC2a1	2-1
pRc_Flag-GPP5-CB3a1-GPP4-NC2a2	1-2
pRc_Flag-GPP5-CB3a2-GPP4-NC2a2	2-2
pRc_2xStrep-GPP5-CB3a1-GPP4-NC2a3	1-3
pRc_2xStrep-GPP5-CB3a2-GPP4-NC2a3	2-3

Table S3. **Scheme for co-expression of plasmids to produce eCB3 assemblies**. The register is defined following the BAC-rule, *i.e.* chain B in the leading position, A in the middle, and C in the trailing (1).

eCB3 assembly	Plasmids co-transfected (see Table S2 or Fig. 3 for reference)	Anticipated register of the last helical segment
111	1-1 1-2 1-3	α1α1α1
222	2-1 2-2 2-3	α2α2α2
112	1-1 1-2 2-3	α1α1α2
121	1-1 2-2 1-3	α2α1α1
211	2-1 1-2 1-3	α1α2α1

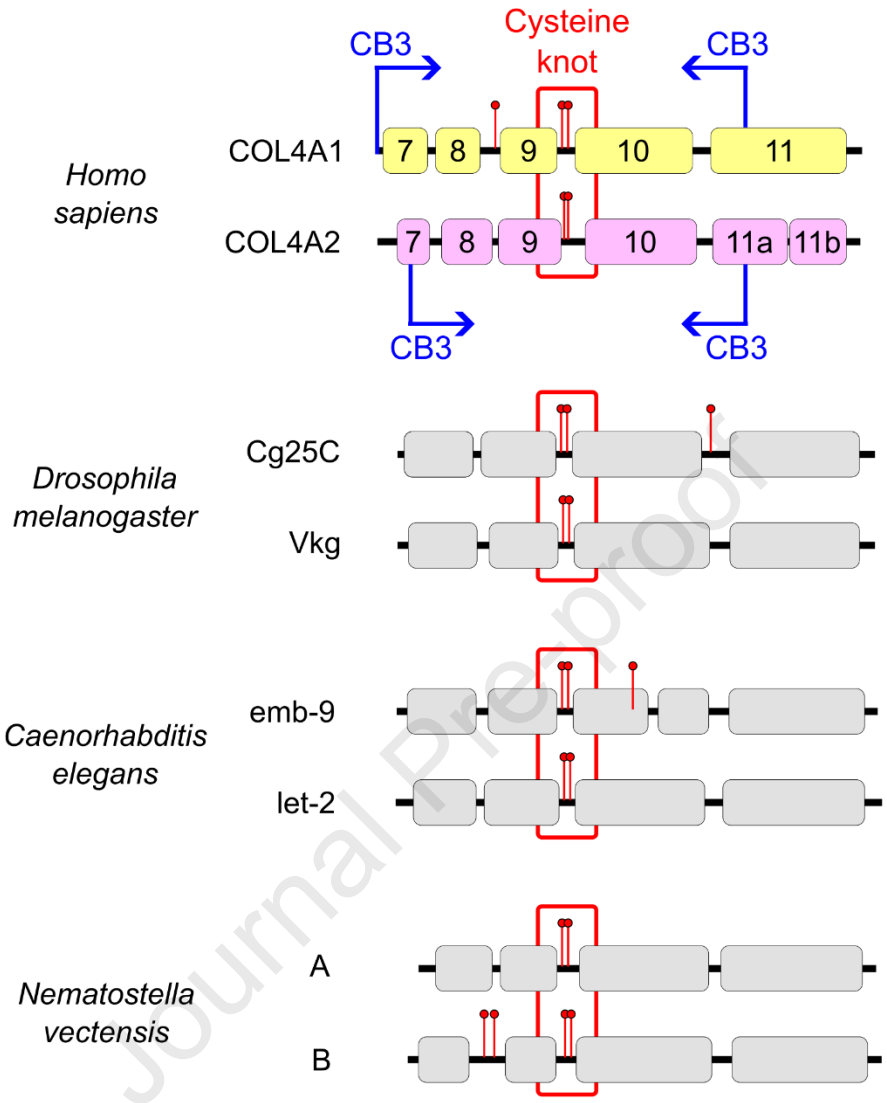
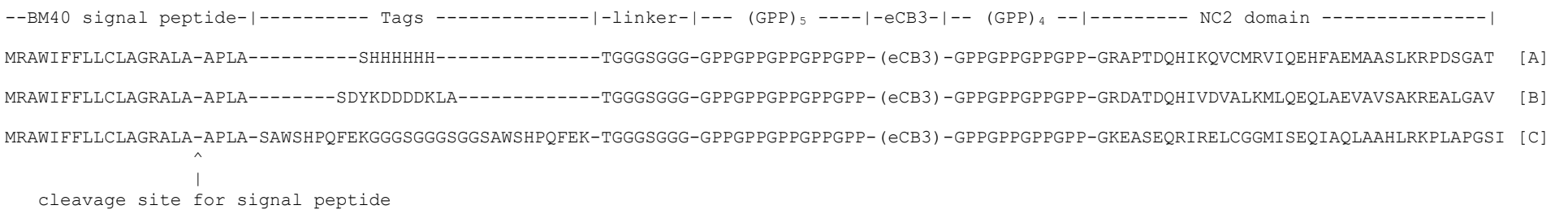


Figure S1. **CB3 cystine knot conservation throughout the animal kingdom**. The cystine knot was recognized within the CB3 fragment of human collagen IV (2). It is presumably formed by interchain disulfides between pairs of cysteines in α1 and α2 chains within the interruption connecting helical segments 9 and 10. Although there is variability of helical segments, interruptions, and cysteines within the CB3-corresponding region across the animal species, pairs of cysteines in α1 and α2 chains from human to cnidarians are conserved, which suggests also evolutionary conservation of the cystine knot. Cysteines are depicted as red pins. Those cysteines forming the CB3 cystine knot are highlighted by red underlying boxes. Uniprot identifiers for sequences are: P02462 for human COL4A1 (isoform 1) and P08572 for human COL4A2; P08120 for Cg25C and Q9VMV5 for Vkg of *D. melanogaster*; P17139 for emb-9 and P17140 for let-2 of *C. elegans*; V9GW22 for chain A of *N. vectensis*. The sequence for chain B of *N. vectensis* was assembled from an open reading frame of the genome (will be published elsewhere).



Extended CB3 (eCB3) sequences for α1 and α2 chains

α1 chain:

GLPVPGQAGAPGFPGERGEKGDRGFPGTSLPGPSGRDGLPGPPGSPGPPGQPGYTNGIVECQPGPPGDQGPPGIPGQPGFIGEIGEKGQKGESCLICDIDGYRGPPGPQGPPGEIGFPGQPGAKGDRGLPGRDGVAGVPGP
QGTPGLIGQPGAKGEPGEFYFDLRLKGDKGDPGFPGQPGM*PGRAGSPGRDGHPGLPGPKGSP*

α2 chain:

*GLSIGDGDQRRGLPGEM***GPKGFIGDPGIPALYGGPPGPDGKRGPPGPPGLPGPPGPDGFLFGLKGAKGRAGFPGLPGSPGARGPKGWKGDAGECRCTEGDEAIKGLPGLPGPKGFAGINGEPGRKGDRGDPGQHGLPGFPG**
LKGVPGNIGAPGPKGAKGDSRTITTKGERGQPGVPGVPGM*KGDDGSPGRDGLDGFPGLPGPP*

Figure S2. **Schemes and amino acid sequences of encoded polypeptides used for production of eCB3 assemblies.** Each sequence begins with the signal peptide followed by a specific tag sequence, i.e. His tag for chain A, Flag tag for chain B, and TwinStrep for chain C. The collagenous sequence (eCB3) is flanked by five and four GPP tripeptide units for avoiding side effects. Each construct ends with a unique chain of the NC2 domain of collagen IX (α1 in chain A, α2 in chain B, and α3 in chain C), which is responsible for chain selection and registering. In cursive are sequences not found in original CB3 fragments (3).

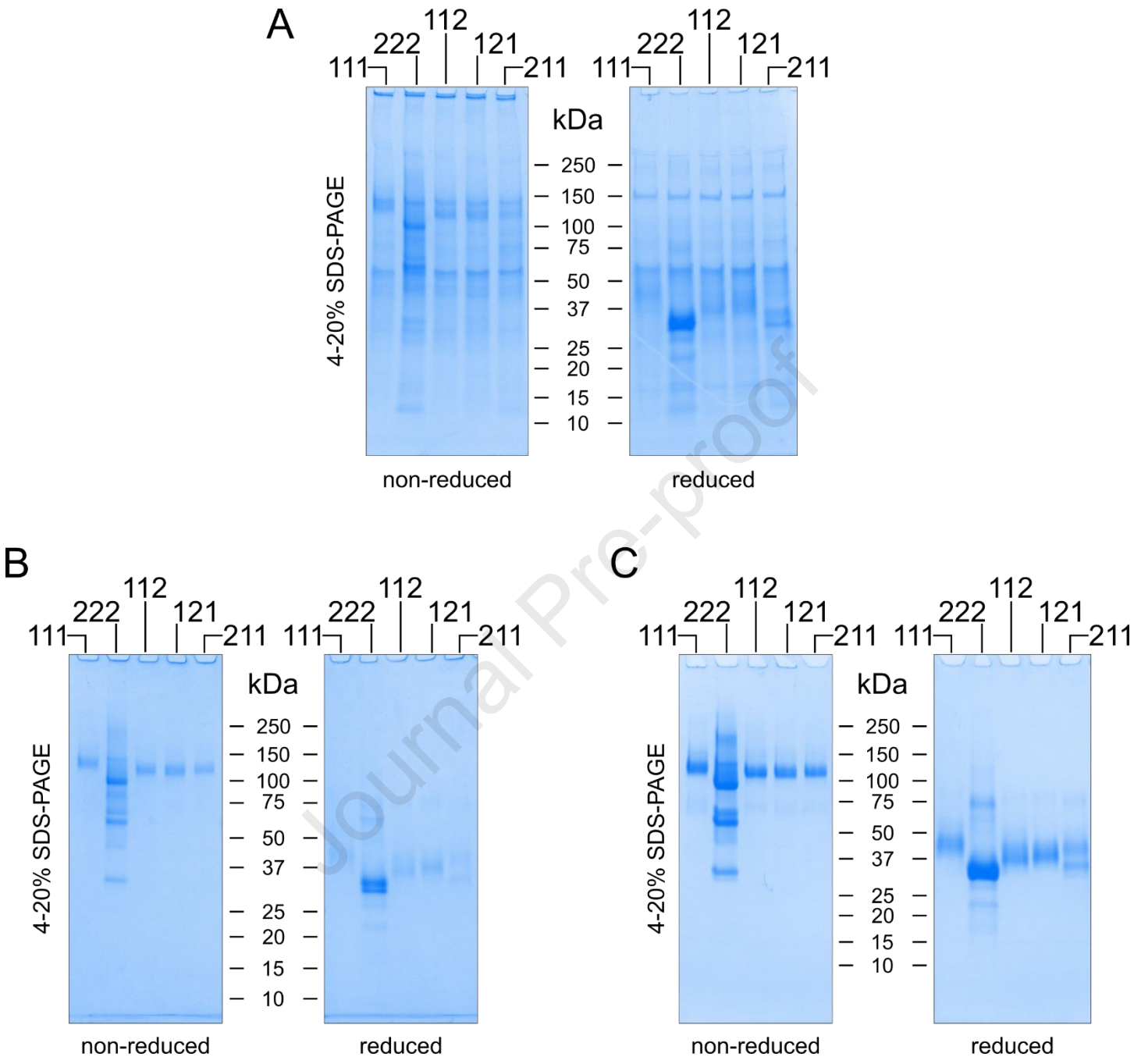


Figure S3. **Purification of eCB3 fragments**. The proteins were purified serially over the three columns. Eluates after each type of column were analyzed on the SDS-PAGE with Coomassie staining. (**A**) Ni-NTA column for His-tagged proteins. (**B**) M2 agarose column for Flag-tagged proteins. (**C**) Strep-Tactin column for Twin-Strep-tagged proteins. Panel (**C**) is reused from Fig. 4B for comparison purpose.

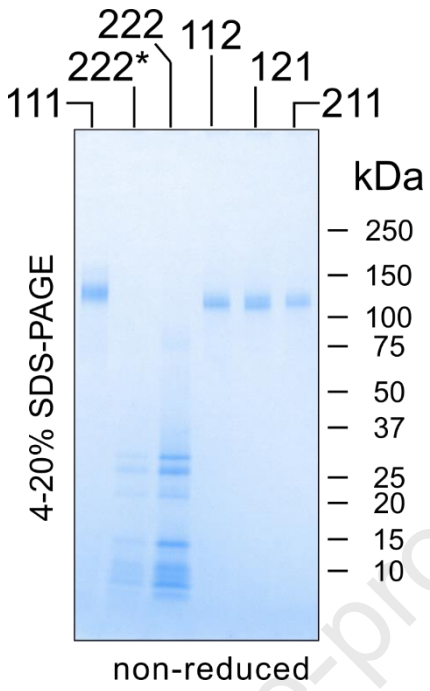


Figure S4. **Long-term stability of eCB3 assemblies.** The major peak for each sample was pooled after the size-exclusion chromatography, stored at 4 °C for 20 weeks, and re-analyzed on the SDS-PAGE with Coomassie staining. 222* - analysis of an additional sample pooled from the left shoulder of the major peak of 222. Only 222 samples demonstrated significant degradation.

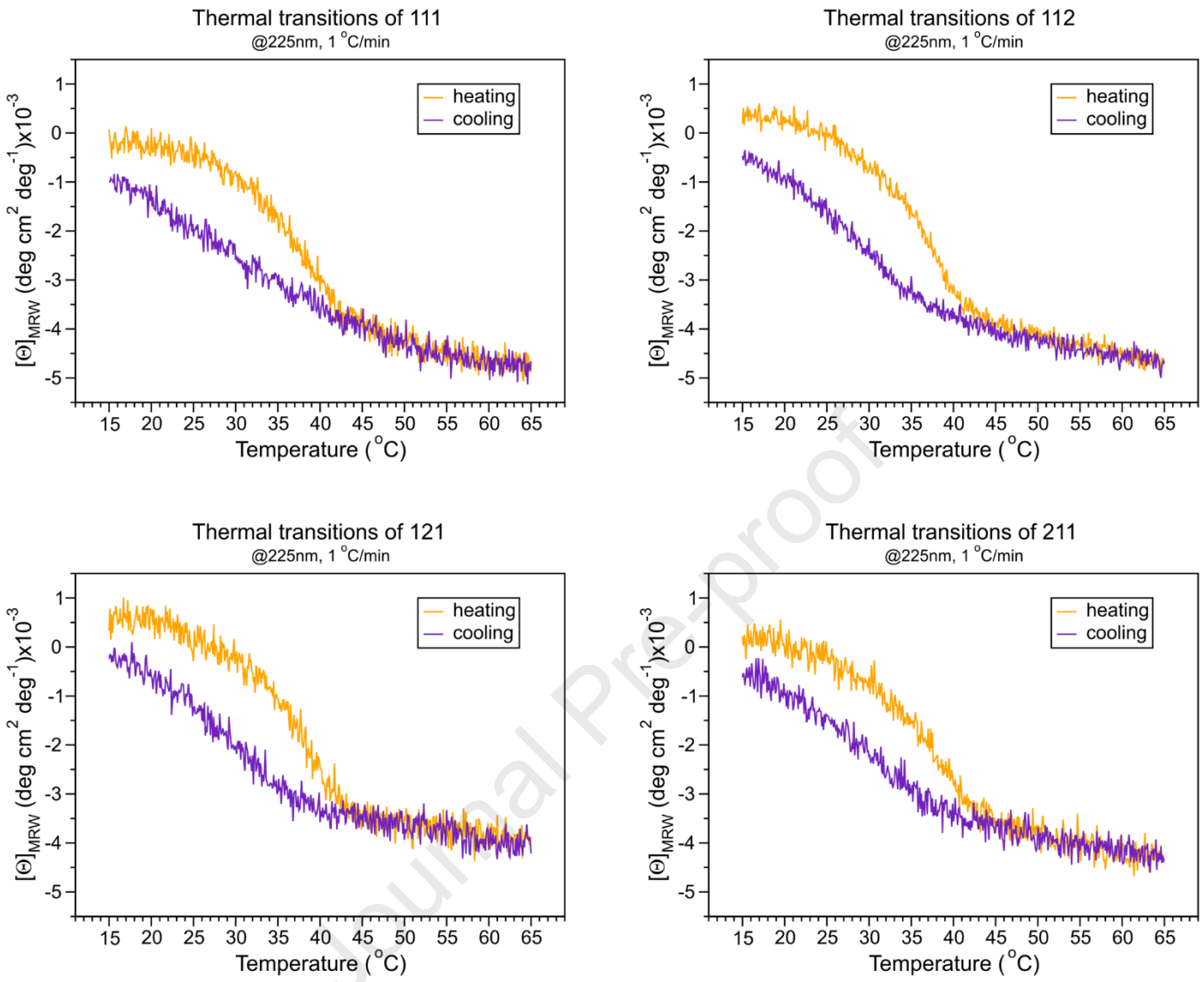


Figure S5. **Thermal transitions of eCB3 assemblies measured by CD.** Mean molar ellipticity is shown at 225 nm wavelength. Heating (orange) and cooling (purple) curves were measured at 1 °C/min rate. All assemblies demonstrated pronounced lag in gaining back the secondary structure signal upon cooling, a well-known phenomenon of collagen hysteresis (4,5).

Human α1:

```
        260         270         280         290         300
KGDFATKGEK GQKGEPGFQG MPGVGEKGEP GKPGPRGKPG KDGDKGEKGS
        310         320         330         340         350
PGFPGEPGYP GLIGRQGPQG EKGEAGPPGP PGIVIGTGPL GEKGERGYPG
        360         370         380         390         400
TPGPRGEPGP KGFPGLPGQP GPPGLPVPGQ AGAPGFPGER GEKGDRGFPG
        410         420         430         440         450
TSLPGPSGRD GLPGPPGSPG PPGQPGYTNG IVECQPGPPG DQGPPGIPGQ
        460         470         480         490         500
PGFIGEIGEK GQKGESCLIC DIDGYRGPPG PQGPPGEIGF PGQPGAKGDR
        510         520         530         540         550
GLPGRDGVAG VPGPQGTPGL IGQPGAKGEP GEFYFDLRLK GDKGDPGFPG
        560
QPGM
```

Human α2:

```
        260         270         280         290         300
NGIPSDTLHP IIAPTGVTFH PDQYKGEKGS EGEPGIRGIS LKGEEGIMGF
        310         320         330         340         350
PGLRGYPGLS GEKGSPGQKG SRGLDGYQGP DGPRGPKGEA GDPGPPGLPA
        360         370         380         390         400
YSPHPSLAKG ARGDPGFPGA QGEPGSQGEP GDPGLPGPPG LSIGDGDQRR
        410         420         430         440         450
GLPGEMGPKG FIGDPGIPAL YGGPPGPDGK RGPPGPPGLP GPPGPDGFLF
        460         470         480         490         500
GLKGAKGRAG FPGLPGSPGA RGPKGWKGDA GECRCTEGDE AIKGLPGLPG
        510         520         530         540         550
PKGFAGINGE PGRKGDRGDP GQHGLPGFPG LKGVPGNIGA PGPKGAKGDS
        560         570
RTITTKGERG QPGVPGVPGM
```

Figure S6. **Putative cyanogen bromide cleavage sites in human collagen IV.** Reported sequences for CB3 fragment span residues G371-M554 for α1 chain and G407-M570 for α2 chain (3). In case of α1 chain the expected residue for specific cyanogen bromide cleavage in position 370 should be M, but it is P (shown in red). The nearest possible cleavage site is M271, which is located ~100 residues upstream. The cleavage at P370 was possibly non-specific during CB3 preparation. The Uniprot identifiers for sequences are: P02462 for α1 (isoform 1) and P08572 for α2.

References

1. Boudko, S. P., and Bachinger, H. P. (2016) Structural insight for chain selection and stagger control in collagen. *Sci Rep* **6**, 37831
2. Vandenberg, P., Kern, A., Ries, A., Luckenbill-Edds, L., Mann, K., and Kuhn, K. (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
3. Eble, J. A., Golbik, R., Mann, K., and Kuhn, K. (1993) The alpha 1 beta 1 integrin recognition site of the basement membrane collagen molecule [alpha 1(IV)]2 alpha 2(IV). *EMBO J* **12**, 4795-4802
4. Engel, J., and Bachinger, H. P. (2000) Cooperative equilibrium transitions coupled with a slow annealing step explain the sharpness and hysteresis of collagen folding. *Matrix Biol* **19**, 235-244
5. Mizuno, K., Boudko, S. P., Engel, J., and Bachinger, H. P. (2010) Kinetic hysteresis in collagen folding. *Biophys J* **98**, 3004-3014