Matrix Biology 125 (2024) 73–87

Contents lists available at ScienceDirect

Matrix Biology

journal homepage: www.elsevier.com/locate/matbio

Collagen prolyl 4-hydroxylase isoenzymes I and II have sequence specificity towards different X-Pro-Gly triplets

Antti M. Salo [a,b,*], Pekka Rappu [c], M . Kristian Koski [a,b], Emma Karjalainen [a,b], Valerio Izzi [a,d], Kati Drushinin [a,b], Ilkka Miinalainen [b], Jarmo Käpylä [c], Jyrki Heino [c], Johanna Myllyharju [a,b]

[a] Faculty of Biochemistry and Molecular Medicine, University of Oulu, Oulu, Finland
[b] Biocenter Oulu, University of Oulu, Oulu, Finland
[c] Department of Life Technologies, University of Turku, Turku, Finland
[d] Faculty of Medicine, BioIM Research Unit, University of Oulu, Oulu, Finland

ARTICLE INFO

Keywords:
Collagen
Post-translational
Modification
Hydroxylation
Hydroxyproline
Proline

ABSTRACT

Collagen biosynthesis requires several co- and post-translational modifications of lysine and proline residues to form structurally and functionally competent collagen molecules. Formation of 4-hydroxyproline (4Hyp) in Y-position prolines of the repetitive -X-Y-Gly- sequences provides thermal stability for the triple-helical collagen molecules. 4Hyp formation is catalyzed by a collagen prolyl 4-hydroxylase (C-P4H) family consisting of three isoenzymes. Here we identify specific roles for the two main C-P4H isoenzymes in collagen hydroxylation by a detailed 4Hyp analysis of type I and IV collagens derived from cell and tissue samples. Loss of C-P4H-I results in underhydroxylation of collagen where the affected prolines are not uniformly distributed, but mainly present in sites where the adjacent X-position amino acid has a positively charged or a polar uncharged side chain. In contrast, loss of C-P4H-II results in underhydroxylation of triplets where the X-position is occupied by a negatively charged amino acid glutamate or aspartate. Hydroxylation of these triplets was found to be important as loss of C-P4H-II alone resulted in reduced collagen melting temperature and altered assembly of collagen fibrils and basement membrane. The observed C-P4H isoenzyme differences in substrate specificity were explained by selective binding of the substrate to the active site resulting in distinct differences in Km and Vmax values. Furthermore, our results clearly show that the substrate proline selection is not dependent on the collagen type, but the main determinant is the X-position amino acid of the -X-Pro-Gly- triplet. Although our data clearly shows the necessity of both C-P4H-I and II for normal prolyl 4-hydroxylation and function of collagens, the mRNA expression of the isoenzymes with various procollagens was, surprisingly, not tightly coordinated, suggesting additional levels of control. In conclusion, this study provides a molecular level explanation for the need of multiple C-P4H isoenzymes to generate collagen molecules capable to assemble into intact extracellular matrix structures.

Introduction

Type I collagen is one of the most abundant proteins in the human body. Together with other collagen types, it forms a family of 28 different collagens. Collagens undergo several co- and post-translational modifications of lysine and proline residues to form hydroxylysine, galactosylhydroxylysine, glucosylgalactosylhydroxylysine, 3-hydroxyproline and 4-hydroxyproline (4Hyp). These modifications have important functions in the synthesis of collagen molecules and their assembly into fibrils, networks and other supramolecular collagenous structures [1–4]. The primary sequence of collagen polypeptides consists of repeating -X-Y-Gly- triplets, where X-position is often proline and Y-position is often 4Hyp. Consequently, -Pro-4Hyp-Gly- is the most frequent triplet and other -X-4Hyp-Gly- triplet combinations are present in variable amounts depending on the collagen type [5,6]. Collagens are right-handed triple-helical molecules consisting of three parallel

Abbreviations: 2OG, 2-oxoglutarate; 4Hyp, 4-Hydroxyproline; C-P4H, Collagen prolyl 4-hydroxylase; PSB, peptide substrate-binding domain; CAT, Catalytic domain; ddPCR, Droplet Digital PCR; MEF, Mouse embryonic fibroblast; CrP4H, *Chlamydomonas reinhardtii* P4H; MS, Mass Spectrometry.

* Corresponding author.
E-mail address: antti.salo@oulu.fi (A.M. Salo).

https://doi.org/10.1016/j.matbio.2023.12.001
Received 28 July 2023; Received in revised form 29 November 2023; Accepted 5 December 2023
Available online 9 December 2023

collagen polypeptides, each with a left-handed poly-L-proline type II helix conformation [7]. The 4Hyp residues have an essential role to provide the triple-helical collagen molecules the thermal stability required at body temperature [7,8].

4Hyp formation in procollagen polypeptides is catalyzed by collagen prolyl 4-hydroxylase (C-P4H, EC 1.14.11.2) within the lumen of the endoplasmic reticulum [4]. It is a 2-oxoglutarate (2OG) -dependent dioxygenase requiring 2OG, Fe^{2+}, ascorbate and O_2 in the reaction. There are three C-P4H isoenzymes that are all $\alpha_2\beta_2$ tetramers where protein disulfide isomerase (PDI) functions as the β subunit. The α subunits are products of three different genes *P4HA1* [9], *P4HA2* [10,11] and *P4HA3* [12,13] encoding α(I), α(II) and α(III) subunits that form the C-P4H-I, C-P4H-II and C-P4H-III tetramers, respectively, with PDI [9–13]. Both α subunits in mammalian C-P4H tetramers are always identical [10].

The C-P4H α subunit consists of an N-terminal dimerization domain followed by a central peptide substrate-binding (PSB) domain and a C-terminal catalytic (CAT) domain. Small-angle X-ray scattering data suggests an elongated symmetric βααβ assembly of the C-P4H tetramer [14]. Very recent crystal structure of a truncated C-P4H-II confirmed that the β/PDI subunit interacts tightly with the CAT domain of the α subunit and the interactions are complemented by two inter domain disulfide bridges [15]. Crystal structures of several other P4H enzymes have also been determined. The structures of PHD2 (a hypoxia-inducible factor P4H) and a *Chlamydomonas reinhardtii* P4H (CrP4H), which is the closest homologue of the C-P4H CAT domain, have been determined also in complex with their peptide substrates [16,17]. In both structures, the peptidyl proline to be hydroxylated sits in the active site and points to the metal ion coordinated by the three catalytic residues of the His-x-Asp —— His-motif, but the conformation of the bound peptides and their mode of binding are strikingly different in these two P4Hs.

Kinetic properties of the C-P4H isoenzymes I and II are largely similar with distinct differences in peptide substrate and inhibitor functions, however. The Km value for a $(Pro\text{-}Pro\text{-}Gly)_{10}$ peptide substrate or a full-length procollagen is 3–6-fold lower for C-P4H-I than for C-P4H-II, whereas Km values for 2OG, Fe^{2+} and ascorbate are quite similar [10,11,18,19]. Distinct differences are also found in the binding and inhibition rates of poly-L-proline, C-P4H-I being very effectively inhibited by it, while C-P4H-II is not [10,11,19]. These differences largely correlate with differences in binding of Pro-Pro-Gly-triplet peptides and poly-L-proline to the PSB domains of the α(I) and α(II) subunits [19–21].

Expression analyses at protein and mRNA level have suggested that C-P4H-I is the main isoenzyme in most cells and tissues, whereas the amount of C-P4H-II is generally lower, although it seems to be abundant in certain tissues and cell types, for example in chondrocytes and endothelial cells [22]. *P4HA3* mRNA is expressed in many tissues but only at low levels [12].

The first human patient with heterozygous compound mutations in *P4HA1* leading to reduced total C-P4H activity was identified in 2017 [23]. The disorder manifests as early-onset joint hypermobility, joint contractures, muscle weakness, bone dysplasia and high myopia. Heterozygous human *P4HA2* mutations have been associated with myopia [24]. Risk alleles of *P4HA2* have also been associated with giant cell arteritis [25]. C-P4Hs have also been implicated as potential cancer and fibrosis drug targets [26–28].

Transgenic mouse models have provided valuable information about the roles of C-P4H-I and II isoenzymes. *P4ha1* knockout mice exhibit early embryonal lethality at 10.5 days post coitum (dpc) with severe type IV collagen production and hence basement membrane assembly defect. Total C-P4H activity was reduced to 20 % in the knockout embryos, the remaining activity apparently being derived from the other two isoenzymes [29]. *P4ha2* knockout mice have no overt phenotype and only minor abnormalities in their tissues and extracellular matrix have been described [30,31]. However, combining complete lack of C-P4H-II with reduced C-P4H-I amount (*P4ha1*$^{+/-}$;*P4ha2*$^{-/-}$ mice) leads to chondrodysplasia and extracellular matrix defects in many tissues [30,31].

While the existence of several C-P4H isoenzymes has been identified decades ago, the actual reason why we have multiple C-P4Hs is still largely unknown. In this study, we provide a molecular level explanation to this question. We show that C-P4H-I and C-P4H-II have no overall collagen type specificity but have both distinct and overlapping roles in the hydroxylation of various -X-Pro-Gly- triplets present in collagen molecules, and they cannot fully compensate for each other. This selectivity rises from differences in the respective active sites of C-P4H-I and II, resulting in distinct differences in their catalytic properties.

Results

C-P4H-I loss affects several but not all hydroxylation sites in type I collagen

The minimum sequence requirement for C-P4H for hydroxylation has been identified as XPG (from here on one-letter amino acid codes are used in the results for simplicity) [32]. However, the Km value of, for example, a triplet PPG peptide is very high [32], which means that *in vitro* studies investigating the role of the X-position amino acid using recombinant C-P4H and various synthetic XPG triplet peptides is not feasible. To study the exact role of C-P4H-I in the hydroxylation of various XPG collagen triplets in collagen, we established mouse embryonic fibroblast (MEF) cell lines derived from *P4ha1*$^{-/-}$ embryos. First, we employed Droplet Digital PCR (ddPCR) to measure the absolute C-P4H α subunit mRNA levels in WT MEFs. The data showed that *P4ha1* is the most abundant transcript and thus C-P4H-I is the main isoenzyme in MEFs (Fig. 1A). *P4ha1* accounted for about 65 % of the *P4ha* transcripts, whereas *P4ha2* and *P4ha3* transcripts contributed 15 % and 20 %, respectively. This is in accordance with our previous data showing that *P4ha1*$^{-/-}$ MEFs retain only 20 % of C-P4H activity when compared to WT [29]. Collagen produced by the mutant and WT cells was collected from the culture medium, partially purified and gel bands corresponding to the α1 and α2 chains of type I collagen (Fig. S1A) were subjected to mass spectrometry (MS). Vast majority of the peptides identified were derived from the two collagen I chains, only minor amounts corresponding to other collagen types were detected, and no differences were present between the genotypes in this distribution (Fig. S2A). To analyse the data quality, we calculated the sequence coverage that was reasonably high for type I collagen chains (Fig. S3A). We also show that sampling bias is minimal as the analyzed tryptic peptides are largely the same between the WT and *P4ha1* knockout genotypes (Fig. S4A). Only the spectra matching to XPG peptides identified in both genotypes were used for hydroxylation analysis. We assessed the hydroxylation level for each XPG sequence by counting the number of MS/MS spectra matching to any hydroxylated XPG-containing peptide *vs*. the number of spectra matching to any XPG-containing peptide regardless of its hydroxylation state. Another option for calculating relative hydroxylation would have been to utilize extracted ion intensities of the peptide and calculate the ratio between the sum of extracted ion intensities of the hydroxylated XPG containing peptides and the sum of extracted ion intensities of the hydroxylated or non-hydroxylated XPG containing peptides. However, since modifications can alter peptide ion intensity, the latter approach is likely to be more error-prone than spectral counting used here. Moreover, since the spectra used for calculating relative hydroxylation are counted within a sample and not between the samples, problems arising from technical variation between samples can be largely avoided by using a spectral counting-based method.

MS analysis showed that, overall, hydroxylation of the Y-position prolines of secreted collagen that was mainly type I collagen (Fig. S2A) was reduced from 84 % in WT to 71 % in *P4ha1*$^{-/-}$ cells ($P = 0.0004$) (Fig. 1B). We then proceeded to site-specific analysis to identify whether underhydroxylation occurs in any particular XPG sites. Prolines in tryptic peptides identified by MS were categorized according to the X-

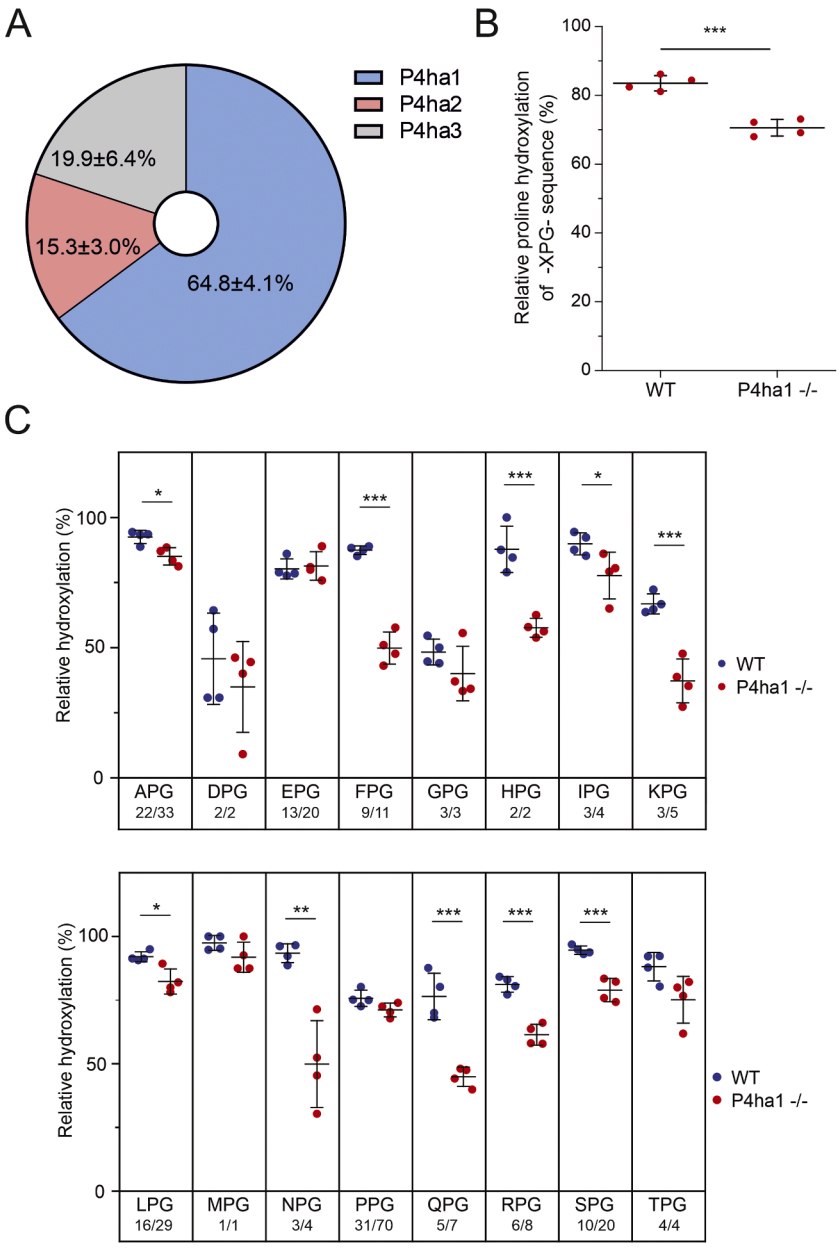

Fig. 1. Hydroxylation of prolines in XPG triplets in collagen fractions that are predominantly type I collagen from WT and *P4ha1*$^{-/-}$ MEFs. (A) ddPCR analysis of the relative abundance of *P4ha1, P4ha2* and *P4ha3* transcripts in WT MEFs. (B) MS analysis of relative hydroxylation of all XPG sequences. (C) MS analysis of relative hydroxylation of different XPG sequences analyzed separately. The number of XPG sites identified in each genotype per the number of all XPG sites present in mouse type I collagen chains is shown below each XPG triplet. Only peptides identified in both genotypes were included in the hydroxylation analysis. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, Student's t-test (Mann-Whitney U test was used for LPG since it did not pass the Shapiro-Wilk test of normality). Data is from 4 WT and 4 *P4ha1*$^{-/-}$ mouse cell lines each derived from different embryo.

position amino acid. The results showed that there was a significant reduction in the amount of 4Hyp in several triplets, whereas a subset of triplets had no changes in hydroxylation (Fig. 1C). Reduction in the hydroxylation was thus not uniform but concentrated on certain sites, whereas some other sites were unaffected. This suggests that C-P4H isoenzymes have site or sequence specific preferences. Our results indicated that the triplets most affected by lack of C-P4H-I had either A, F, H, I, K, L, N, Q, R or S in the X-position (Fig. 1C, Table S1). Of these, the APG and LPG triplets were only marginally affected. The apparent reduction in the hydroxylation of TPG triplets did not quite reach statistical significance ($P = 0.052$). Generally, a positively charged or a polar uncharged X-position amino acid seems unfavorable for Y-position proline 4-hydroxylation in the absence of C-P4H-I (Fig. 1C, Table S1). In addition, hydroxylation of FPG sites was strongly affected by C-P4H-I loss. Altogether, these results show the general importance of C-P4H-I for collagen hydroxylation and that the remaining isoenzymes C-P4H-II and III cannot efficiently hydroxylate a Y-position proline when there is a positively charged or polar uncharged amino acid in the preceding X-position. Interestingly, many triplets (Fig. 1C) remained unaffected by loss of C-P4H-I suggesting that the other isoenzymes, most likely C-P4H-II, can effectively hydroxylate these sites.

C-P4H-II is required for hydroxylation of prolines in EPG and DPG sequences

Next, we studied the role of C-P4H-II in collagen hydroxylation using skin collagen isolated from WT, *P4ha1*$^{+/+}$;*P4ha2*$^{+/-}$, *P4ha1*$^{+/-}$;*P4ha2*$^{+/-}$, *P4ha1*$^{+/+}$;*P4ha2*$^{-/-}$ and *P4ha1*$^{+/-}$;*P4ha2*$^{-/-}$ mice. Similarly to MEFs,

P4ha1 is the main isoform in mouse skin as relative abundances of the *P4ha1, P4ha2* and *P4ha3* transcripts were 83 %, 15 % and 2 %, respectively (Fig. 2A). The partially purified skin collagen fraction was digested with trypsin and subjected to MS. The analysis showed that, as expected, the sample contained mainly type I and III collagens, with trace amounts of type V collagen (Fig. S2B). Data quality was quite similar to the MEF dataset with even higher sequence coverage (Fig. S3B) and minimal sampling bias (Fig. S4B) and only the spectra matching to XPG peptides identified in all genotypes were used for hydroxylation analysis. The overall hydroxylation of Y-position prolines in the skin collagen was decreased in a *P4ha* allele dose-dependent manner from 77 % in WT to 68 % in *P4ha2*$^{-/-}$ and 66 % in *P4ha1*$^{+/-}$;*P4ha2*$^{-/-}$ mice (Fig. 2B). The tryptic peptides of the samples were then categorized again according to the X-position amino acid. Strikingly, a massive reduction in the hydroxylation of DPG and EPG triplets was observed upon lack of C-P4H-II, while all other triplets were mainly unaffected (Fig. 2C, Tables S1,S2). In *P4ha2*$^{-/-}$ or *P4ha1*$^{+/-}$;*P4ha2*$^{-/-}$ collagen no hydroxylation of DPG sites was detected in contrast to close to 100 % hydroxylation in WT, and only 23 % of the EPG sites were hydroxylated compared to 74 % in WT (Fig. 2C). This clearly suggests that C-P4H-I cannot efficiently hydroxylate prolines that follow an amino acid with a negatively charged side chain. The lower hydroxylation in these two triplets has been independently found also from different starting material by Wilhelm and co-workers [33]. Jointly these findings make a very strong conclusion that C-P4H-II is required for hydroxylation of DPG and EPG sites.

We then wanted to find out if a similar sequence specific hydroxylation is present also in a non-fibrillar collagen and partially purified type IV collagen from the kidneys of *P4ha2*$^{-/-}$ mice (Fig. S1B). The bands cut for MS analysis mainly consisted of the α1(IV) and α2(IV) chains (Fig. S2C). The sequence coverage was lower (Fig. S3C) than in the other datasets, probably due to only partial purification of type IV collagen, but tryptic peptides were mostly the same between the genotypes (Fig. S4C) suggesting only minimal sampling bias. Like in the other analyses, only the spectra matching to XPG peptides identified in both genotypes were used for hydroxylation analysis. In addition, only type

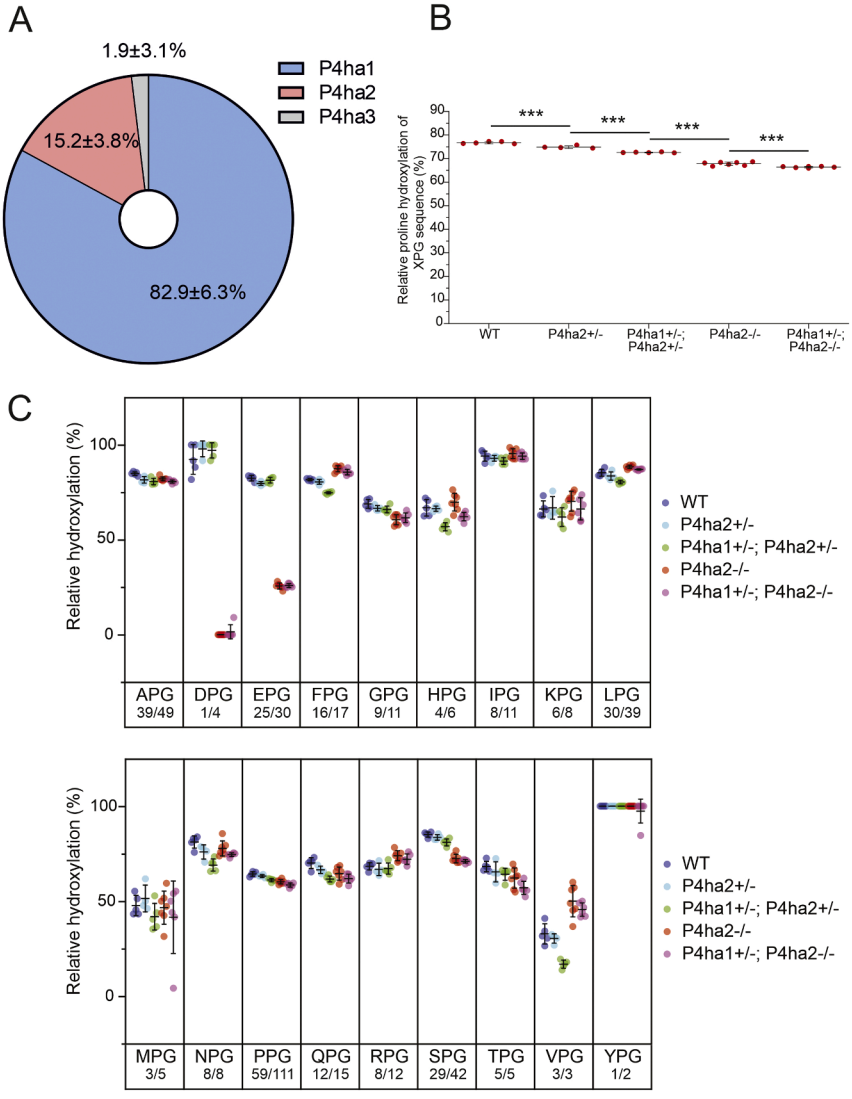

Fig. 2. Hydroxylation of prolines in XPG triplets in fibrillar collagen fractions that mainly consist of type I and type III collagens extracted from WT, *P4ha2*$^{+/-}$, *P4ha1*$^{+/-}$;*P4ha2*$^{+/-}$, *P4ha2*$^{-/-}$ and *P4ha1*$^{+/-}$;*P4ha2*$^{-/-}$ mouse skin. (A) ddPCR analysis of the relative abundance of *P4ha1, P4ha2* and *P4ha3* transcripts in WT mouse skin. Data is from 4 WT mice. (B) MS analysis of relative hydroxylation of all XPG sequences. *$P < 0.05$, ***$P < 0.001$, one-way ANOVA test followed by Tukey HSD test. (C) MS analysis of relative hydroxylation of different XPG sequences analyzed separately. Extracted collagen had been digested with trypsin and analyzed by LC-MS/MS (Sipilä et al., 2018). The number of XPG sites identified in each genotype per the number of all XPG sites in mouse type I and III collagen chains is shown below each XPG triplet. Only peptides identified in both genotypes were included in the hydroxylation analysis. Data is from 5 WT, 4 *P4ha2*$^{+/-}$, 5 *P4ha1*$^{+/-}$;*P4ha2*$^{+/-}$, 7 *P4ha2*$^{-/-}$, 6 *P4ha1*$^{+/-}$;*P4ha2*$^{-/-}$ mice. The statistical significance of differences between genotypes is shown in Table S2.

IV collagen peptides were used in the analysis. ddPCR analysis showed that the relative transcript abundance of *P4ha1* in the kidney was 85.1 %, *P4ha2* abundance was 14.9 % whereas *P4ha3* was not detected (Fig. 3A). The overall hydroxylation of XPG sites in type IV collagen was reduced from 81 % in WT to 76 % in *P4ha2*$^{-/-}$ mice (Fig. 3B) and lack of C-P4H-II markedly affected the hydroxylation of EPG and DPG sites also in the type IV collagen sample (Fig. 3C). The sequence coverage was lower due to only partially purified material. In addition to EPG and DPG, only IPG and PPG sites had a reduced hydroxylation, but the decrease was not as prominent as in the former two sites (Fig. 3C). A slight increase was observed in the hydroxylation of VPG sites in *P4ha2*$^{-/-}$ samples compared to WT (Fig. 3C). In conclusion, a similar sequence specific hydroxylation was observed in both fibrillar and type IV collagens upon lack of C-P4H-II. This suggests that C-P4H-II displays site specificity but not specificity towards an individual collagen type, at least in the case of type I and IV collagens analyzed here.

Next, we wanted to study if the X-position amino acid immediately preceding the proline is solely responsible for determining hydroxylation specificity. We chose to explore the EPG sites of type I collagen from *P4ha2*$^{-/-}$ mice in detail to study if all EPG sites were similarly affected independent of the further sequence context. The data showed that hydroxylation of the EPG sites varied between no hydroxylation to

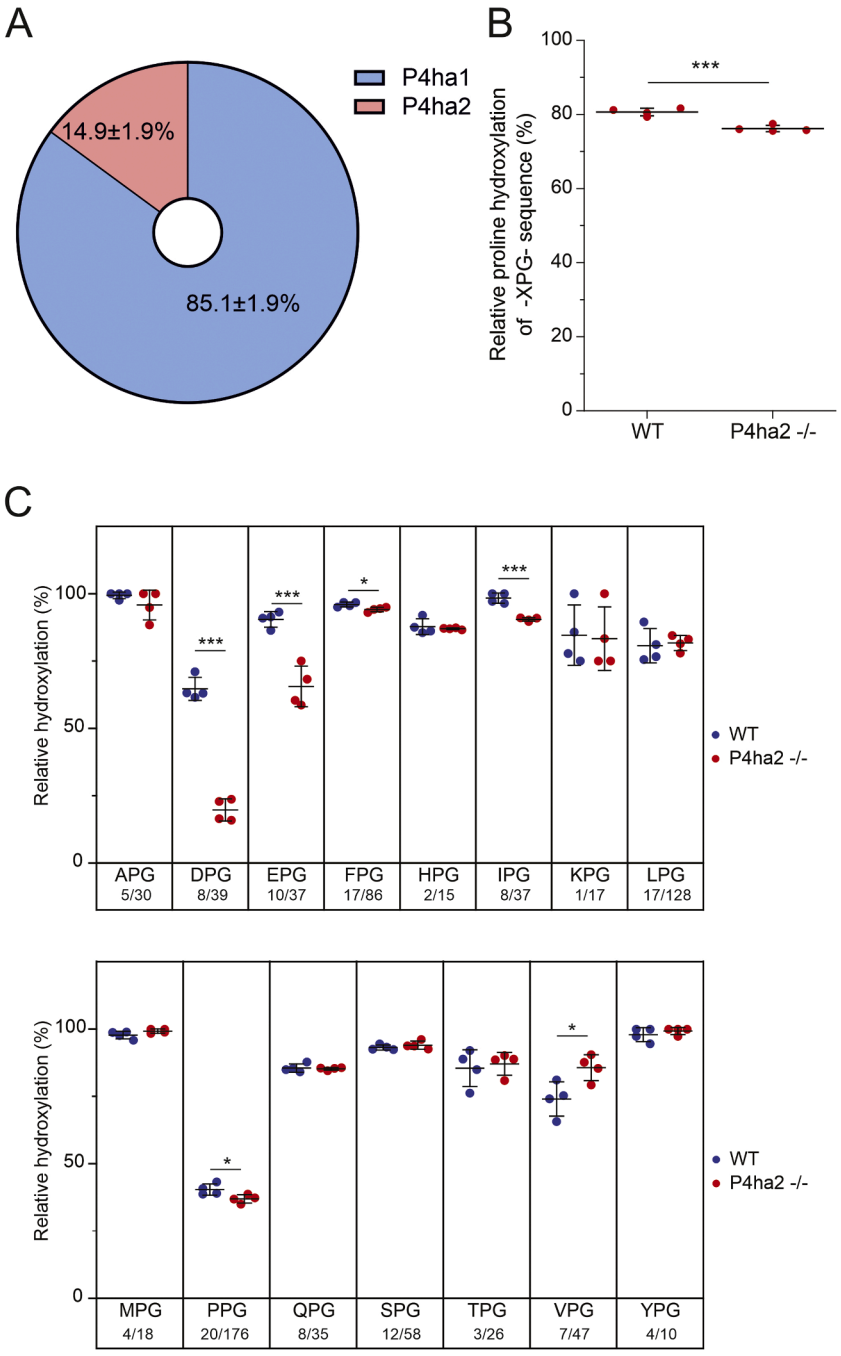

Fig. 3. Hydroxylation of prolines in XPG triplets of type IV collagen mainly consisting of the α1(IV) and α2(IV) chains from *P4ha2*$^{-/-}$ mouse kidney. (A) ddPCR analysis of the relative abundance of *P4ha1, P4ha2* and *P4ha3* transcripts in WT mouse kidney. Data is from 5 WT mouse kidneys. (B) MS analysis of relative hydroxylation of all XPG sequences. (C) MS analysis of relative hydroxylation of different XPG sequences analyzed separately. Only the spectra matching to type IV collagen were counted. The number of XPG sites identified in each genotype per the number of XPG sites in mouse type IV collagen α1, α2, α3, and α4 chains is shown below each XPG triplet. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$. Data is from 4 WT and 4 *P4ha2*$^{-/-}$ mice.

50 % hydroxylation (Fig. S5) suggesting that the amino acid preceding the proline is a major but not the only determinant for hydroxylation.

Reduced collagen melting temperature, fibril diameter and basement membrane abnormalities in P4ha2$^{-/-}$ mice

To study the potential importance of hydroxylation of EPG and DPG sites in collagens, we first analyzed their prevalence. We calculated the number of all different XPG triplets in all human and mouse collagen α-chains as well as the collagen-like proteins C1QA (complement C1q A chain) and COLQ (collagen-like tail subunit of asymmetric acetylcholinesterase) (Tables S3, S4). The results showed that both EPG and DPG triplets exist in essentially all collagen α-chains (Fig. 4A). The number of EPG and DPG triplets is typically lower than that of the abundant PPG triplet, with the exception of type VI collagen that has an unusually low number of PPG triplets. In general, EPG is among the most frequent triplets in collagens, whereas the number of DPG triplets varies more between collagen types, being notably low in the most abundant type I, II and III fibrillar collagens and relatively high in type IV collagen (Fig. 4A, Tables S3, S4). For example, type I collagen α-chains have about 1000 amino acids in the helical region that contains 116 and 95 XPG repeats in COL1A1 and COL1A2, respectively. Of these, both chains

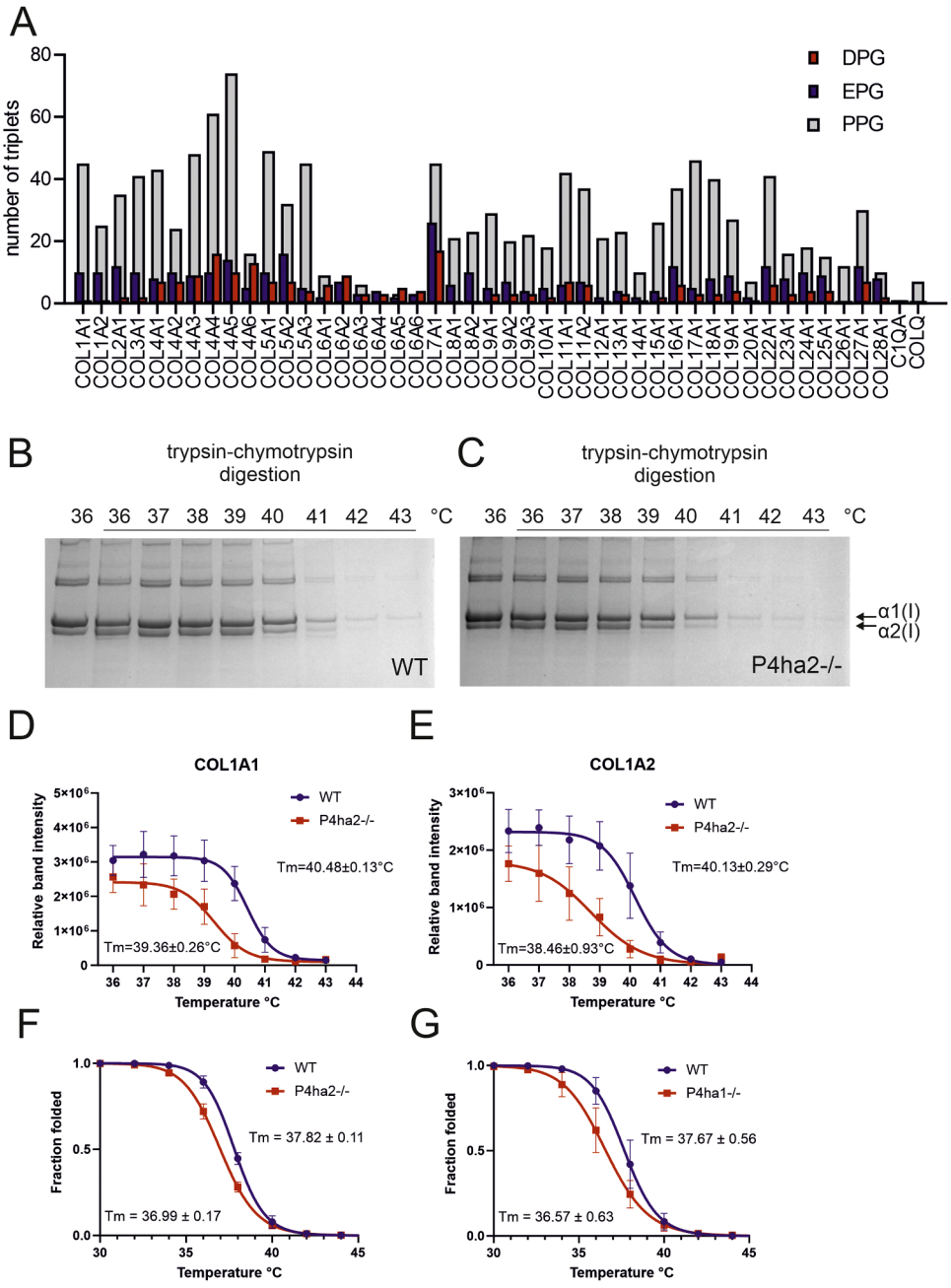

Fig. 4. Lack of C-P4H-II-catalyzed hydroxylation of EPG and DPG sites reduces the thermal stability of type I collagen. (A) The number of EPG and DPG sites in human collagens and the collagen-like proteins C1QA and COLQ. The number of PPG sites is given as a reference. Uniprot accession numbers and full data for the human and mouse proteins is given in Tables S3 and S4. (B-C) Trypsin-chymotrypsin digestion in various temperatures followed by SDS-PAGE to analyze the melting temperature of type I collagen extracted from WT (B) and *P4ha2*$^{-/-}$ (C) mouse skin. Undigested sample at 36 °C is in the first lane. Data is from 4 WT and 4 *P4ha2*$^{-/-}$ mice. A representative SDS-PAGE is shown. Quantification is shown as a mean ±SD in (D) and (E), respectively. (F) CD analysis of WT and *P4ha2*$^{-/-}$ mouse skin collagen. Data is shown as a mean ±SD from 4 WT and 4 *P4ha2*$^{-/-}$ mice. (G) CD analysis of collagen from WT and *P4ha1*$^{-/-}$ MEFs. Data is shown as a mean ±SD from 4 WT and 4 *P4ha1*$^{-/-}$ cell lines.

contain 10 EPGs each, but there is only a single DPG in COL1A1 and none in COL1A2.

Next, we investigated consequences of the reduced EPG and DPG hydroxylation upon the loss of C-P4H-II. We subjected the skin type I collagen isolated from WT and *P4ha2*$^{-/-}$ mice to trypsin-chymotrypsin digestion at various temperatures to assess its melting temperature (Fig. 4B-C). The reduction in 4Hyp in these sites in *P4ha2*$^{-/-}$ mice resulted in about 1 °C reduction in the melting temperature of the type I collagen (Fig. 4D-E). We also performed circular dichroism (CD) spectroscopy to show that the reduced melting temperature results from an overall reduction in the helix stability rather than from local breathing of the triple helix. The results showed a similar, about 1 °C reduction in the melting temperature (Fig. 4F). In addition, CD analysis of the type I collagen from *P4ha1*$^{-/-}$ MEFs (Fig. 4G) showed that its melting temperature was quite similar to that of the *P4ha2*$^{-/-}$ collagen. We then determined potential consequences for extracellular collagen-rich matrices. As the *P4ha2*$^{-/-}$ mice have no overt phenotype and only bone and cartilage tissue have been previously studied in detail in these mice [30,31], we used transmission electron microscopy to study the potential effects of the observed reduced hydroxylation of the EPG and DPG sites on the collagen fibrils (Fig. 5A-D) and basement membranes (Fig. 5E-J) in the dermis of *P4ha2*$^{-/-}$ mice. Results showed that *P4ha2*$^{-/-}$ mice exhibited an average fibril diameter reduction (Fig. 5C) from 65.5 nm to 56.9 nm and there was a shift in distribution towards thinner fibrils (Fig. 5D). In addition, basement membrane defects were observed in the *P4ha2*$^{-/-}$ dermis. The average capillary basement membrane thickness was increased from 34.9 nm to 50.6 nm (Fig. 5E-G), while the dermal-epidermal basement membrane thickness was unaltered (Fig. 5H-J). These data show that even subtle underhydroxylation of collagen due to reduced hydroxylation of EPG and DPG sites upon the absence of C-P4H-II has consequences for different collagen assemblies.

C-P4H isoenzymes have no major collagen type specificity

Next, we wanted to study how C-P4H isoforms hydroxylate different procollagen chains. We produced full-length procollagen chains (with the exception of COL5A1 being a fragment) by reticulocyte *in vitro* transcription/translation kit, followed by analysis of the formation of 4Hyp catalyzed by recombinant C-P4H-I and C-P4H-II (Fig. 6A). Expectedly, C-P4H-I was able to efficiently hydroxylate *in vitro* almost all the procollagen chains studied. Hydroxylation by C-P4H-II was quite similar to C-P4H-I in the case of many procollagen chains, including COL1A1, COL3A1, COL5A1 fragment, COL17A1, COL19A1 and COL22A1. The procollagen chains COL4A3, COL8A1, COL13A1, COL15A1 and COLQ were hydroxylated more with C-P4H-I than with C-P4H-II. Surprisingly COL6A1, COL6A2 and C1QA were poorly hydroxylated by both enzymes. We then analyzed hydroxylation of COL3A1 and COL6A1 further using increasing amounts of the enzymes. The results showed that maximum hydroxylation of COL3A1 was achieved already with 1 μg of either C-P4H-I or C-P4H-II (Fig. 6B). However, even with the highest enzyme amount tested, the hydroxylation level of COL6A1 remained lower than that observed for most procollagen chains with both C-P4H-I and C-P4H-II (Fig. 6C).

Catalytic and structural analyses show sequence specific hydroxylation and substrate selectivity in the active sites of the C-P4H isoenzymes

To study if the substrate X-position amino acid plays a direct role in the catalysis we performed *in vitro* activity assays using recombinant human C-P4H-I and C-P4H-II and short collagenous $(XPG)_5$ peptides as substrates. The 15-mer peptide substrate length was chosen because such peptides do not reach the PSB domain from the active site of the CAT domain in the α subunit as proposed by the current structural model of the C-P4H-II tetramer [15], and they are thus suitable to limit the analysis only to the intrinsic effects on catalysis occurring in the active site. In agreement with the tissue MS data, $(EPG)_5$ and $(DPG)_5$ were very poor substrates for C-P4H-I with Vmax values of only about 1 % of that obtained with $(PPG)_5$ (Table 1). Conversely, C-P4H-II hydroxylated these peptides markedly more efficiently than C-P4H-I. The Vmax of C-P4H-II with $(EPG)_5$ was twice that obtained with $(PPG)_5$, while the Vmax with $(DPG)_5$ was about 10 % of that obtained with $(PPG)_5$, but still 10-fold higher than the corresponding Vmax value of C-P4H-I, further supporting the specific role of C-P4H-II in the hydroxylation of these sites. Interestingly, the Km values of the $(EPG)_5$ and $(DPG)_5$ peptides were markedly higher than that of $(PPG)_5$ for both isoenzymes. Furthermore, again in agreement with the MS data, detectable hydroxylation of $(KPG)_5$ was obtained by C-P4H-I only, while $(VPG)_5$ and $(NPG)_5$ were hydroxylated by both isoenzymes in the *in vitro* activity assays.

The solved structure of the CAT domain of the α(II) subunit in complex with the complete PDI/β subunit does not contain any ligands in the active site of the CAT domain [15]. The two flexible loop regions, the hairpin and the βII-βIII loops, next to the active site were disordered and therefore we used the RoseTTAFold structure in the docking calculations [34]. First, we modeled the peptide substrate GPPGPP to the CAT domain of C-P4H-I by using the GalaxyPepDock program [35]. The calculated RoseTTAFold structure of the CAT domain with the model peptide GPPGPP was very similar to the CrP4H crystal structure with the model peptide PSPSPS (Fig. S6A-B). The GPPGPP took the poly-L-proline type II helix conformation [7] similarly to the bound PSPSPS peptide in the CrP4H crystal structure. GalaxyPepDock modeled the hairpin and the βII-βIII loops in such a way that they are folded on top of the GPPGPP peptide similarly as seen in the CrP4H structure. Next, we docked GKPGKPG and GEPGEPG peptides to C-P4H-I and C-P4H-II, respectively. These peptides also took the poly-L-proline type II helix conformation and were similarly buried in the tunnel with the hairpin and βII-βIII loops surrounding the peptide in both C-P4H-I and II docking studies (Fig. 7). Both GKPGKPG (Fig. 7A-B) and GEPGEPG (Fig. 7C-D) adopted a structure where both X-position amino acids K and E pointed towards the loop regions. The enzyme active site was occupied with the first Y-position proline. The X-position amino acid of the first triplet (referred as X_1) was located at the entrance of the peptide binding tunnel in a large pocket, while the X-position amino acid of the second triplet (referred to as X_2), was bound to the exit site of the tunnel (Fig. 7). Based on this model, the X_1 amino acid is more buried within the hairpin and/or the βII-βIII loops, whereas X_2 is more outside the tunnel especially in C-P4H-I. We calculated also the electrostatic surface potentials of the peptide binding tunnels of both C-P4H-I and C-P4H-II docking models. We observed that the surface potential was slightly negative in C-P4H-I (Fig. S6C-D) and strongly positive in C-P4H-II (Fig. S6E-F).

C-P4H and procollagen gene expression is not coordinated

We then wanted to study if C-P4H and its substrate gene expression is co-regulated. We employed Tabula Sapiens [36] and Tabula Muris [37] datasets to analyze C-P4H and procollagen gene expression at the single cell level. First, we studied the expression of the C-P4H α subunit mRNAs (Figs S7, S8). We categorized cells to groups according to the presence of P4HA transcripts. Three groups were single positive cells (P4HA1+, P4HA2+ or P4HA3+), where the two other isoforms are not detected. In three double positive groups two isoforms are present, but one is missing (P4HA1+P4HA2+, P4HA1+P4HA3+ or P4HA2+P4HA3+). The two remaining groups were triple negative (P4HA1-P4HA2-P4HA3-) and triple positive cells (P4HA1+P4HA2+P4HA3+). To simplify evaluation of the expression patterns, we manually grouped the 33 murine and 172 human cell types from different organs (32,743 and 218,317 single cells in total, respectively) into “epithelial”, “endothelial” and “stromal” categories (Fig. 8). All three human cell categories and the mouse epithelial and stromal cell categories contained all P4HA expression groups (Fig. 8). Notably, cells that contained the P4HA3 transcript alone or in combination with the other two isoforms were present in very low quantities and were completely absent from the mouse endothelial cell

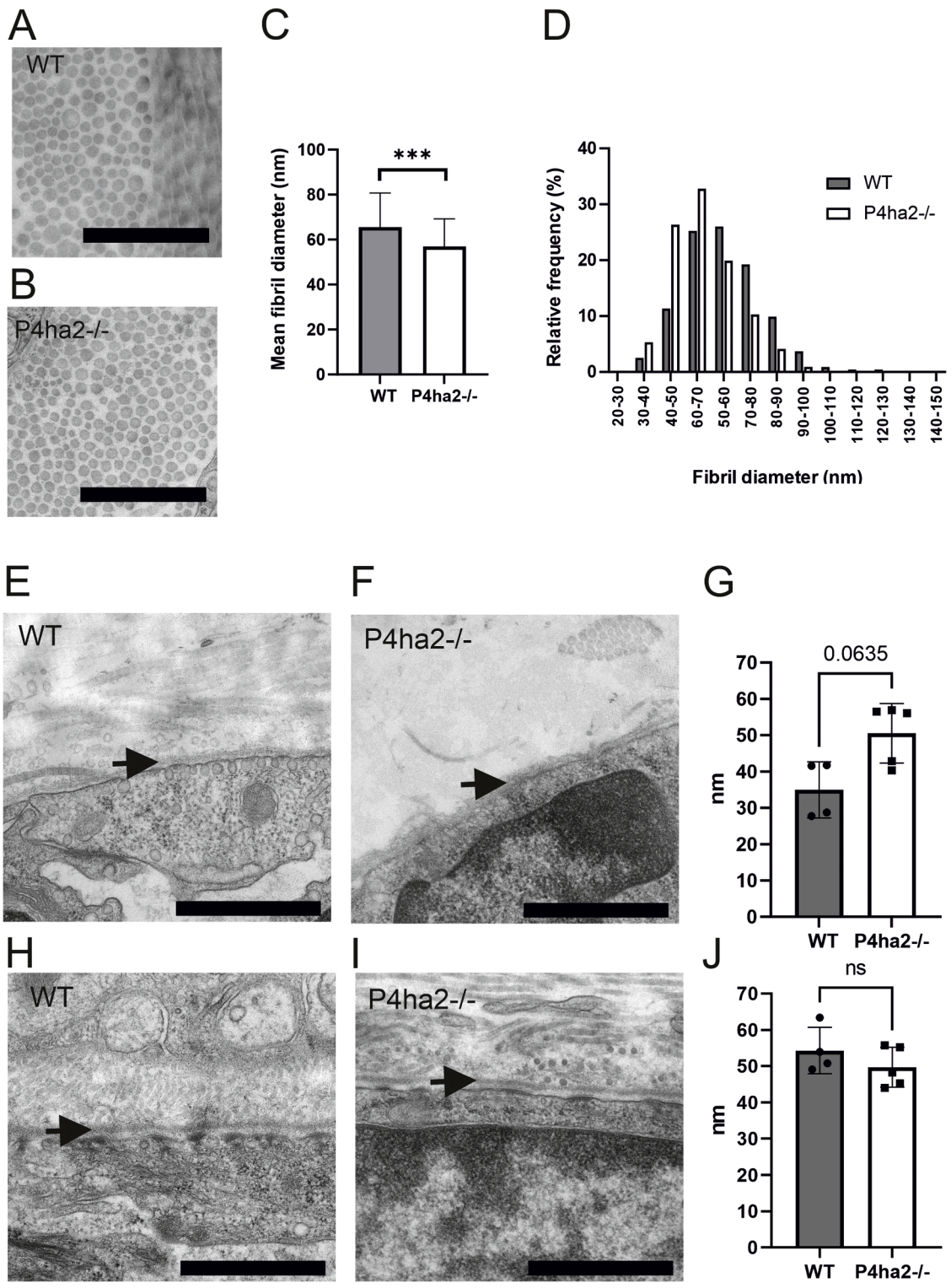

Fig. 5. Transmission electron microscopy analysis of collagen fibrils and basement membranes in skin dermis. Representative images of skin collagen fibrils in WT (A) and *P4ha2*$^{-/-}$ (B) mice. Mean fibril diameter (C) and frequency distribution of the diameter (D) of 2115 and 2618 collagen fibrils counted from 4 WT and 5 *P4ha2*$^{-/-}$ mice, respectively. ***$P < 0.001$, Student's t-test. Representative images of skin capillary basement membranes of WT (E) and *P4ha2*$^{-/-}$ (F) mice, and average capillary basement membrane thickness (G). Representative images of dermal-epidermal basement membranes of WT (H) and *P4ha2*$^{-/-}$ (I) skin, and average dermal-epidermal BM thickness (J). Data in (G) and (J) are from 4 WT and 5 *P4ha2*$^{-/-}$ mice. Statistical analysis was done using Mann-Whitney U test. Arrows indicate the basement membrane. Scale bar is 1 μm in all images.

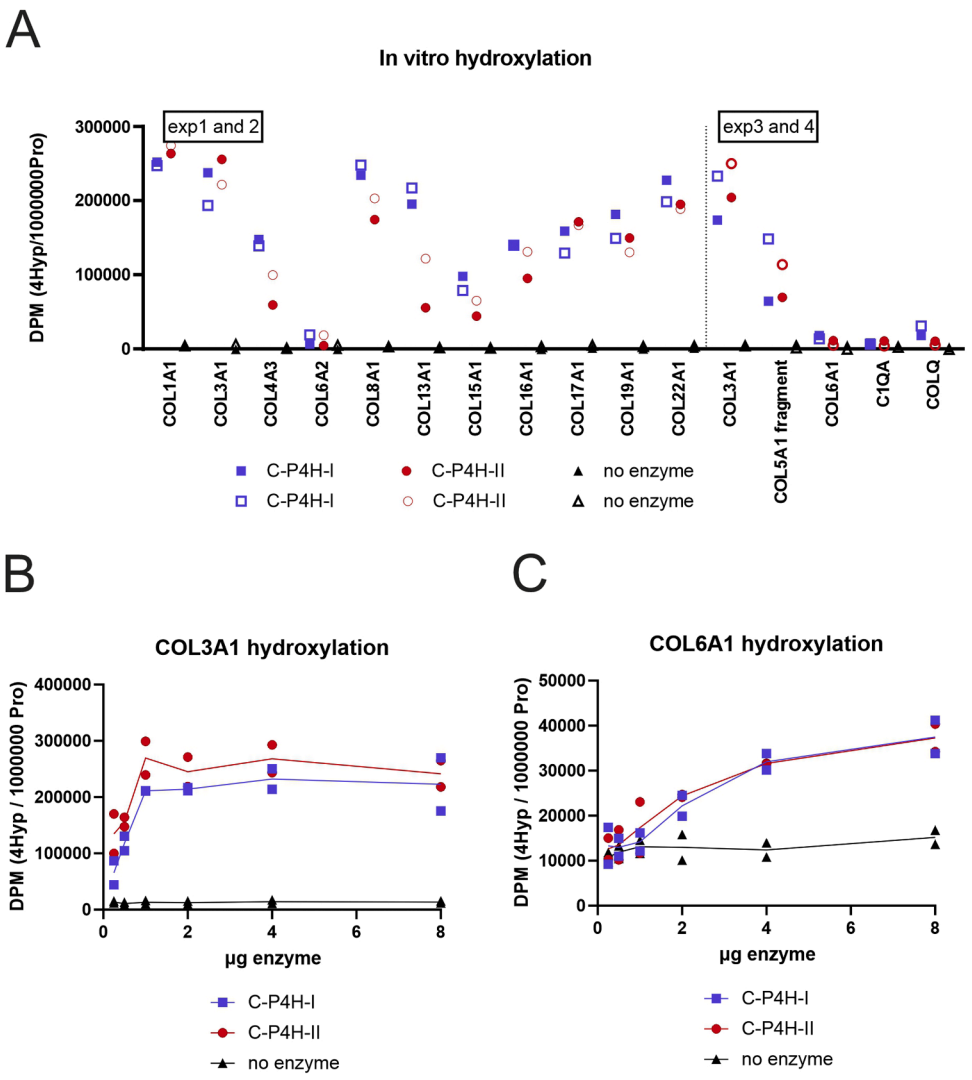

Fig. 6. *In vitro* hydroxylation of various human procollagen chains with recombinant human C-P4Hs. (A) Procollagen chains were produced by *in vitro* transcription/translation and used as substrates for C-P4H-I and C-P4H-II in an activity assay. A reaction with no enzyme was used as a negative control. Two independent assays were conducted for each substrate and enzyme combination and are shown as squares for C-P4H-I and circles for C-P4H-II, the negative control reaction is indicated with triangles, and the first and second assays are shown with closed and open symbols, respectively. *In vitro* hydroxylation of proα1(III) (B) and proα1(VI) (C) chains with increasing C-P4H amount. Two independent assays were performed.

category. The single positive P4HA1+ and P4HA2+ and the double positive P4HA1+P4HA2+ groups were clearly the most abundant ones among the P4HA expressing cells in all cell categories, and typically in this order, except for human stromal cells where the abundance of the P4HA1+ and the double positive P4HA1+P4HA2+ groups was almost identical. Interestingly, cells that do not express any of the P4HA isoforms was the most abundant group in all three cell categories in both species (Fig. 8).

Co-positivity for the P4HA isoforms was significantly different in human and mouse cells (Figs. 8, S7, S8). Regarding P4HA1 and P4HA2, 29,571 cells were P4HA1+P4HA2+ out of 73,635 being either P4HA1+ or P4HA2+ (~29 %) in humans (Fig. S7), the corresponding values being 1434 out of 6018 (~20 %) in mice (Fig. S8). Co-positivity of P4HA1 or P4HA2 with P4HA3 was markedly rare, only 1472 cells were either P4HA1+P4HA3+ or P4HA2+P4HA3+ out of 74,771 being either P4HA1+, P4HA2+ or P4HA3+ (~2 %) in humans and 29 out of 9158 (~0.3 %) in mice. In both cases, these differences gave a p-value (from Chi-square test with Yates correction) lower than 2×10^{-16}, suggesting that co-occurrence of P4HA1 and P4HA2 across cell types might depend on a co-regulatory mechanism that does not apply to P4HA3.

We then analyzed procollagen expression in the same cells and noticed major differences between the cell types as expected (Fig. S9). We then inspected procollagen expression in the different P4HA groups. Surprisingly, procollagen expression was remarkably similar between the P4HA groups, although clear cell-of-origin effects were observed across the groups, for example, higher expression of collagen I and VI in stromal/mesenchymal cells, collagen IV in epithelial and endothelial cells, and collagen XVII in skin epithelial cells (Fig. S9). Thus, no clear correlation between the procollagen expression pattern and the P4HA positivity groups was observed, suggesting that procollagen expression is independent of cellular P4HA subunit transcript status but dependent on cell origin at least in the transcript level.

Discussion

Collagens rely on the C-P4H-catalyzed co- and post-translational formation of numerous 4Hyp residues in their α-chains to provide thermal stability for the triple-helical collagen structure. The aim of this work was to provide a molecular level explanation for the need of multiple C-P4H isoenzymes and their specific functions. The existence of three C-P4H isoforms has been now known for two decades, but their individual roles have remained largely unclear. The isoenzymes have

Table 1
C-P4H-I and C-P4H-II K_m and V_{max} values for $(XPG)_5$ peptide substrates.

	K_m (μM)					
	C-P4H-I			C-P4H-II		
(EPG)5	2900	±	2000	3500	±	1000
(DPG)5	1100	±	400	4500	±	1100
(VPG)5	4100	±	2900	4800	±	2100
(KPG)5	1200	±	700	not detectable		
(NPG)5	4700	±	1200	2600	±	800
(PPG)5	340	±	290	100	±	47
(PPG)10	25	±	9	20	±	3

results are average values (±SD) from 3 experiments (except 4 for $(VPG)_5$ and 6 for $(PPG)_{10}$)

	V_{max} (DPM)					
	C-P4H-I			C-P4H-II		
(EPG)5	200	±	200	50,000	±	0
(DPG)5	200	±	100	2900	±	1800
(VPG)5	12,900	±	6700	45,800	±	37,000
(KPG)5	3400	±	1500	not detectable		
(NPG)5	800	±	200	2000	±	500
(PPG)5	23,300	±	8800	23,300	±	8800
(PPG)10	20,700	±	7500	25,600	±	12,500

results are average values (±SD) from 3 experiments (except 4 for $(VPG)_5$ and 6 for $(PPG)_{10}$)

been shown to have some differences in expression levels and tissue distribution [10,22,31], although also strikingly similar patterns have been described [11]. Our expression data showed that *P4ha1* is the most abundant isoform in mouse skin, kidney and in MEFs (Figs. 1A, 2A, 3A). However, MS analysis of collagen showed that deletion of either *P4ha1* or *P4ha2* reduced the overall collagen 4Hyp level in the same magnitude relative to WT (Figs. 1B, 2B). However, these genetic deletions have very different consequences for life. The *P4ha1* null mice exhibit a massive type IV collagen defect and early embryonic lethality [29] in contrast to *P4ha2* knockout mice, which develop to term and the adult mutant mice are only mildly affected [30,31].

Our current data shows that deletion of either *P4ha1* or *P4ha2* led to a reduction in 4Hyp that was not uniform across the collagen molecules, but instead showed site specific differences in the hydroxylation of XPG triplets. Employing MS analyses, we found out that the X-position amino acid played a significant role in the hydroxylation of the following proline by C-P4H-I or C-P4H-II (Figs. 1C, 2C, 3C). Genetic *P4ha1* deletion in MEFs led to a hydroxylation defect particularly in XPG sites where the X-position amino acid is a positively charged side chain or a polar uncharged side chain (Fig. 1C). Unlike the other triplets, these affected sites were not fully hydroxylated by the remaining C-P4H-II or III that were expressed in the MEFs in a reasonable level (Fig. 1A). This is in agreement with the finding that the remaining two isoforms cannot compensate for the loss of *P4ha1* in mouse [29]. Furthermore, a human

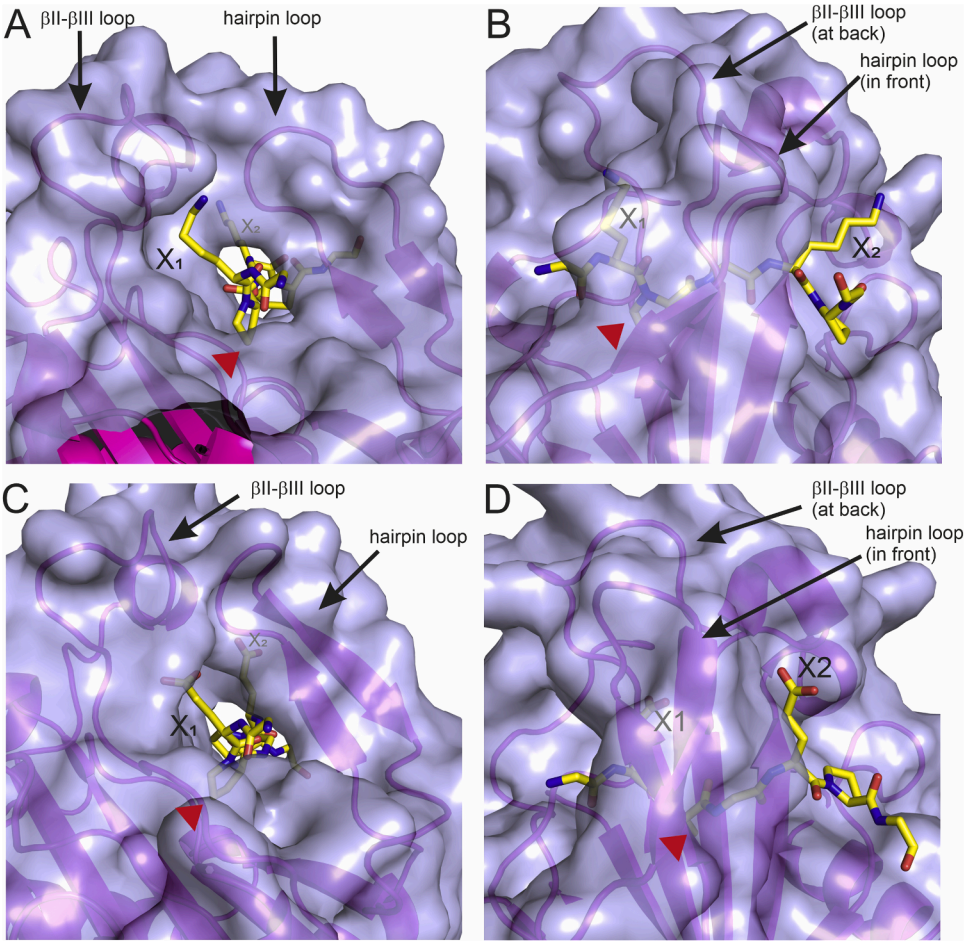

Fig. 7. Peptide substrate docking to the active site of C-P4H I and II. (A-B) Docking model of the GKPGKPG peptide substrate binding to the CAT domain of C-P4H-I from two different views with the peptide N-terminus (A) and C-terminus (B) shown. (C-D) Docking model of the GEPGEPG peptide substrate binding to the CAT domain of C-P4H-II from two different views with the peptide N-terminus (C) and C-terminus (D) shown. In all panels, the protein part is shown with surface (blue, transparent) as well as in ribbon (magenta) representation and the peptide is shown with yellow sticks. The two flexible loops are named in all panels. X_1 is the N-terminal X-position amino acid preceding the peptidyl proline that is to be hydroxylated (marked with a red arrowhead) and sits in the active site, whereas X_2 is the X-position amino acid of the next triplet.

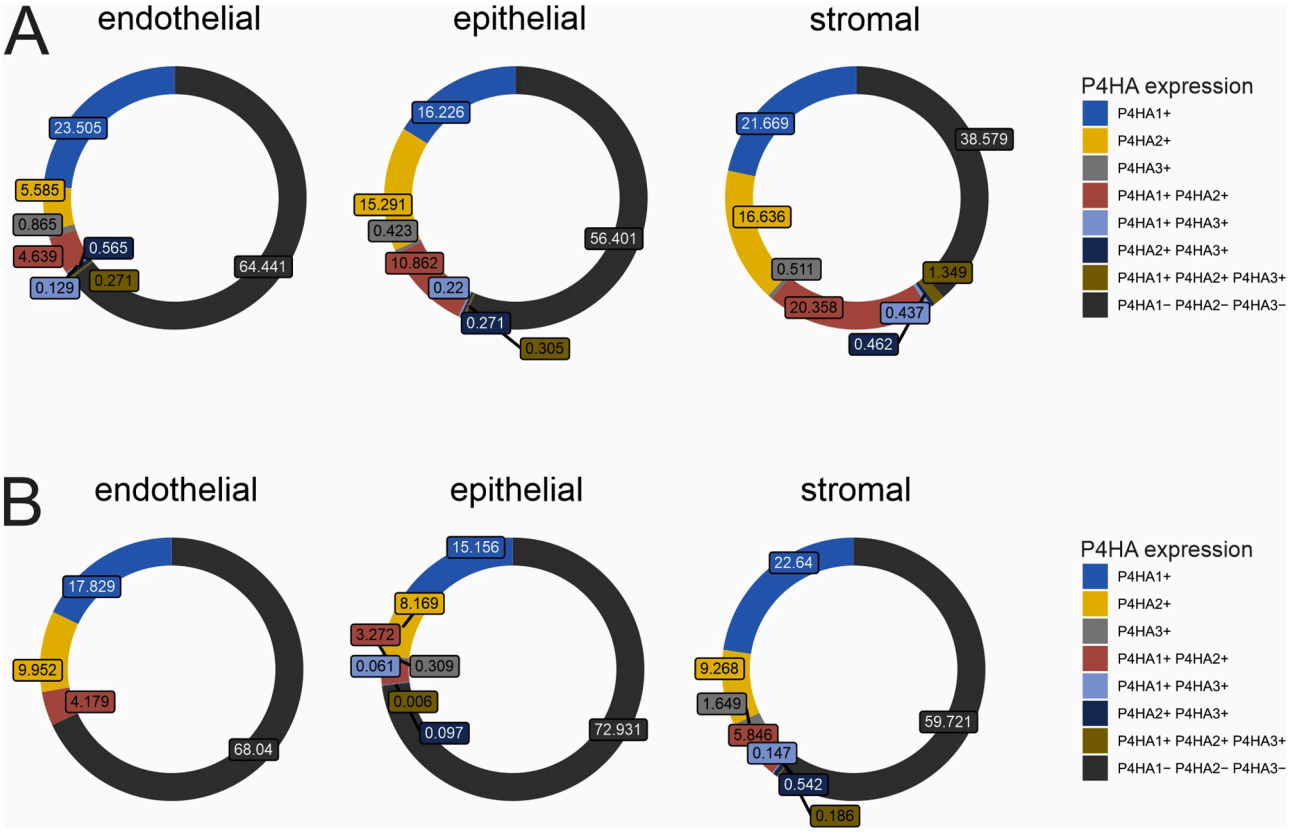

Fig. 8. Analysis of C-P4H mRNA expression in single cell human and mouse data. Percentage of cells expressing either *P4HA1, P4HA2, P4HA3*, or any of their two- and three-way combinations and of cells with no *P4HA* expression at all in (A) human cells from the Tabula Sapiens and (B) mouse cells from the Tabula Muris datasets. The total cell number was 218,317 and 32,743, respectively. The different cell types were manually assigned into the "epithelial", "endothelial" and "stromal" group to aid visualization of the results.

connective tissue disease patient with reduced P4HA1 protein amount and total C-P4H activity likewise shows the inability of the remaining C-P4H isoforms to fully compensate for the reduced function of C-P4H-I [23].

MS analysis of *P4ha2* null mouse skin fibrillar collagen showed that lack of C-P4H-II leads to marked underhydroxylation of XPG triplets where the proline follows a negatively charged amino acid glutamate or aspartate (Fig. 2C). Underhydroxylation of EPG and DPG sites was also present in type IV collagen isolated from the kidney (Fig. 3C) and thus expands the C-P4H-II specific function on these sites beyond fibrillar collagens. Based on these observations, the remaining isoforms, mainly C-P4H-I, as C-P4H-III is expressed only in low quantity in the skin and is absent from the kidney (Figs. 2A, 3A), cannot efficiently hydroxylate EPG and DPG sites. The C-P4H-II specificity towards these sites was recently independently found also in ATDC5 mouse teratocarcinoma cells [33].

Although *P4ha2* null mice have no overt phenotype, the lack of full compensation by the remaining C-P4H isoforms was plausible, as histological abnormalities have been detected in their bone development and structure, which were enhanced to a visible growth retardation and chondrodysplasia phenotype upon concomitant loss of one *P4ha1* allele [30,31]. Here we show that the effects of the hydroxylation defect in EPG and DPG sites upon *P4ha2* knockout extend to also other tissues and different collagen assemblies, as abnormal BM structures and altered collagen fibril diameter were observed in the kidney and skin, respectively (Fig. 5). At the molecular level, the number of DPG sites is very low in mouse type I collagen, but it has several EPG triplets and their underhydroxylation upon loss of *P4ha2* led to about 1 °C decrease in the melting temperature of type I collagen (Fig. 4). The melting temperature of type I collagen from the *P4ha2* knockout mouse skin was quite similar to that of the type I collagen secreted by *P4ha1* knockout MEFs. The most probable reason for the lack of any major difference in the melting temperature between the *P4ha1* and *P4ha2* knockout genotypes is most likely caused by the tight quality control of collagen secretion, *i.e.* only collagen molecules sufficiently hydroxylated to retain the triple-helical conformation at a physiological temperature are secreted. Taken together, the obtained data suggested that the C-P4H isoenzymes do not exhibit collagen type or α-chain specificity, but instead their hydroxylation capability is largely determined by the X-position amino acid of the XPG triples. This was evident from the kinetic analysis performed with synthetic peptides as a substrate (Table 1) and was further supported by the finding that most of the procollagen chains analyzed were hydroxylated to approximately similar extent by both C-P4H-I and C-P4H-II in an *in vitro* hydroxylation assay (Fig. 6A). A few procollagen chains (COL4A3 and COL13A1) were hydroxylated somewhat more efficiently by C-P4H-I, but no collagen had a clear preference towards C-P4H-II. This is in accordance with the abundance of the DPG and EPG triplets in collagens. Interestingly, a few procollagen chains, namely COL6A1, COL6A2 and the collagen-like proteins C1Q and COLQ showed poor hydroxylation by both C-P4H-I and C-P4H-II (Fig. 6A). In case of COL6A1, 4Hyp formation remained poor, about 15 % of that obtained in COL3A1, even with the highest amount of C-P4H used (8-fold to that required for maximum hydroxylation of COL3A1) (Fig. 6B,C). It will be interesting to study in future whether additional proteins are needed for efficient hydroxylation of these proteins or if C-P4H-III has a role in their hydroxylation. The former possibility is supported by cell-based studies showing that C-P4H-I knockdown results in reduced C1q secretion [38]. Notably, there is a lack of consecutive PPG triples in these proteins (Fig. 4A), which could be relevant for efficient binding to the C-P4H PSB-domain [39].

As the loss of *P4ha1* was found to affect XPG triples with positively charged amino acids at the X-position, while loss of *P4ha2* affected triplets with negatively charged X-position amino acids, we wanted to study binding of such peptides to the enzyme active site in detail. The results were clearly in line with the kinetic analyses (Table 1) and indicated that optimal peptide substrates are different for C-P4H-I and C-P4H-II. Our peptide docking analyses of C-P4H-I and C-P4H-II with KPG and EPG triplet peptides, respectively showed that the substrate specificity dictated by the X-position amino acid preceding the hydroxylated Y-position proline is determined by two flexible loop regions, the hairpin loop and the βII-βIII loop, that fully shield the peptide substrate bound to the CAT domain (Figs. 7, S6). These loop regions have areas with low sequence similarity between the isoenzymes suggesting differences in their binding properties (15,17), which is here experimentally shown to lead to distinct differences in the catalytically competent binding mode of the peptide substrate. The PSB domain locates near the CAT domain in the current C-P4H model structure [15]. However, the distance of the peptide binding sites of the PSB and CAT domains is about 30 Å, which means that the PSB domain binds to a substrate polypeptide at least 3 to 4 triplets away from the hydroxylation site of the CAT domain [14,15]. Based on our structural modeling calculations, the peptide binding tunnel of the C-P4H-I CAT domain has an overall negative charge, whereas that of C-P4H-II has an overall positive charge (Fig. S6C-F). This fully agrees with the results that C-P4H-I prefers positively charged amino acid residues in the X-position, whereas C-P4H-II prefers negatively charged residues, suggesting that the substrate peptide selection could be based on these charges. However, there are also differences in the size (bigger in C-P4H-I) and polarity (more polar in C-P4H-II) of the peptide binding tunnel, both of which may have additional effects on the specificity. The larger pocket available for the X-position residue in C-P4H-I is consistent with the finding that hydroxylation of triplets with a large and non-charged X-position amino acid such as FPG, HPG, NPG and QPG, are reduced in the absence of C-P4H-I, whereas generally the smaller amino acids in the X-position are also accepted by C-P4H-II. However, some Y-position prolines with bulky amino acids in the X-position, like methionine and tyrosine, are still hydroxylated normally in the absence of C-P4H-I. It is also possible that the two flexible loop regions can adopt different conformations depending on the X-position amino acids. To understand further the substrate specificity differences of the CAT domains, experimental structural data in complex with various XPG triplet peptides will be needed in the future. In conclusion, our *in vitro* activity assays and peptide-enzyme docking studies strongly suggest that the C-P4H active sites have isoenzyme-specific selectivity on substrates.

Our data show that although both C-P4H-I and C-P4H-II can efficiently hydroxylate many different collagens, they both have XPG site specificity that cannot be fully compensated by the other one. This suggests that C-P4H-I and C-P4H-II might often co-exist in a cell. Our single cell gene expression analyses showed that of P4HA expressing cells, the majority typically express only P4HA1 or P4HA2 (Fig. 8). On the other hand, the significantly higher frequency of co-occurrence of P4HA1 and PH4A2 in cells *vs*. P4HA1 and P4HA3 or P4HA2 and P4HA3 across species seems to suggest the existence of a co-regulatory mechanism between P4HA1 and P4HA2 expression, but its identity is yet to be found. Strikingly, the gene expression analysis suggested that procollagen and C-P4H expression is not co-regulated at least in the mRNA level, being dependent largely on cell-of-origin patterns, as already noticed in healthy and diseased tissues [40,41]. The finding that procollagen mRNA is produced even without C-P4H mRNA expression, even accounting for dropout effect, is surprising and more studies are needed to reveal if there are for example temporal and half-life differences or involvement of additional protein level regulation in the hydroxylation of different procollagen chains. Heterogeneity in the prolyl 4-hydroxylation of the Y-position prolines has been observed [6] and we also observed variability in the hydroxylation status of collagen from different sources. This may be relevant for example in the interactions of collagen molecules, the lower hydroxylation potentially affecting for instance the integrin binding [42].

Our data shows that C-P4H-I and C-P4H-II can together hydroxylate the majority, if not all, of the sites that are to be hydroxylated at least in type I collagen. Future studies are therefore needed to elucidate the biological role of C-P4H-III. The expression level of *P4HA3* in several adult and fetal human tissues is much lower than that of *P4HA1* [12] and a similar finding was made here in mouse skin and kidney, the *P4ha3* transcript completely lacking from the latter. However, the *P4ha3* mRNA was relatively highly expressed in MEFs. Relatively high expression levels of *P4ha3* mRNA have also been observed in mouse growth plate and tibia at early post-natal time points [31]. Furthermore, *P4ha3* has been shown to be expressed in a different bone marrow cell population (myeloid supportive cluster) than the largely co-expressed *P4ha1* and *P4ha2* transcripts that are abundant in mature osteoblasts [31]. It is thus possible that C-P4H-III is needed for hydroxylation of some specific sites of specific collagens other than type I and IV, it is needed only in specific times and/or locations in tissues during development or that it is not involved in collagen hydroxylation at all and may instead have non-collagenous substrates.

In conclusion, our results clearly show that the biological function of C-P4H-II is to hydroxylate collagen EPG and DPG sites that C-P4H-I cannot hydroxylate. Conversely, C-P4H-I preferentially hydroxylates XPG triplets with positive and polar uncharged amino acids in the X-position. Our data also shows that this selectivity arises from intrinsic differences in the active sites of the two C-P4H isoenzymes. Surprisingly, we also show that procollagen transcription is not co-regulated with C-P4H transcription raising interesting questions for future studies in the regulation of collagen synthesis.

Experimental procedures

Transgenic mouse and cell lines

Heterozygous *P4ha1* mice [29] in C57BL/6JOlaHsd background were subjected to timed matings to obtain 10.5 dpc embryos. The embryos were placed into DMEM (Gibco), containing 20 % fetal bovine serum (BioWest), penicillin-streptomycin (Sigma-Aldrich) and Glutamax (Gibco), and disintegrated by pipetting up and down. The cells were cultured in air with 5 % CO_2 at +37 °C, were let to attach and dead cells were removed by washing with PBS. The attached wild-type (WT) and *P4ha1*$^{-/-}$ MEFs were immortalized by retroviral introduction of large T antigen. *P4ha2*$^{+/-}$, *P4ha1*$^{+/-}$;*P4ha2*$^{+/-}$, *P4ha2*$^{-/-}$ and *P4ha1*$^{+/-}$; *P4ha2*$^{-/-}$ mice were obtained by breeding as described earlier [30]. Cells and mice were genotyped as described before [29,30].

Collagen extraction

To partially purify type I collagen from the WT and *P4ha1*$^{-/-}$ MEFs for the MS analysis, the cells were cultured in air with 5 % CO_2 at +37 °C with DMEM (Gibco) containing 10 % fetal bovine serum (BioWest), penicillin-streptomycin (Sigma-Aldrich) and Glutamax (Gibco). The cells were then washed 4 times with PBS followed by serum-free culture with 150 μg/ml ascorbic acid phosphate (Wako). The culture medium was collected at 24 h, replaced and collected again after 24 h. The two medium samples were combined and filtered using a 0.22-μm filter (Sartorius) to remove cell debris. Ammonium sulphate was added to 176 mg/ml and the samples were incubated with gentle mixing for 24 h at +4 °C. Collagen pellet was obtained by centrifugation at 10,000 g for 1 h at +4 °C and dissolved into 0.1 M acetic acid.

Triple-helical part of type IV collagen was partially purified from female mouse kidneys. The kidneys were homogenized using 0.5 M acetic acid and two 5-mm steel beads in TissueLyser LT (Qiagen), followed by addition of pepsin to 0.1 mg/ml and incubation for 3 days. The supernatant was collected after centrifugation at 20,000 g for 40 min, NaCl was added to 1 M final concentration and the sample was

incubated overnight. Supernatant was then collected by centrifugation at 30,000 g for 40 min, NaCl was added to a final concentration of 1.8 M and collagen pellet was obtained by centrifugation at 30,000 g for 40 min. The collagen fractions were separated in SDS-PAGE followed by Coomassie Blue staining. The gel bands of type I and IV collagen were cut and dried with acetonitrile.

Skin collagen (consisting mainly of type I collagen) from WT and *P4ha2-/-* mice was prepared as previously for CD analysis and trypsin-chymotrypsin digestion [42]. For CD analysis, collagen secreted into the culture medium was obtained from WT and *P4ha1*$^{-/-}$ MEFs cultured in DMEM (Gibco) supplemented with 10 % Panexin CD (Pan Biotech), 150 μg/ml ascorbic acid phosphate (Wako), penicillin-streptomycin (Sigma-Aldrich) and Glutamax (Gibco). The culture medium was collected and centrifuged at 4000 rpm for 30 min to remove dead cells. Collagen was precipitated by adding 176 mg/ml of ammonium sulphate and incubated overnight with mixing. All steps were done in +4 °C. Pellet was collected by centrifugation and dissolved in 0.25 mg/ml pepsin (Roche Diagnostics GmbH) in 0.5 M acetic acid and incubated overnight with mixing to digest any non-collagenous material. Samples were centrifuged at 20,000 x g for 30 min to collect the supernatant that was then precipitated by adding NaCl to a final 0.7 M concentration. After o/n incubation with mixing, samples were centrifuged at 20,000 x g for 30 min to get the collagen pellet that was dissolved in 0.1 M acetic acid.

Mass spectrometry and data analysis

The dried gel pieces were washed with 200 μl of 40 mM NH_4HCO_3 in 50 % acetonitrile at +37 °C for 15 min and shrunk with 200 μl of 100 % acetonitrile. Reduction, alkylation and in-gel digestion were performed as described [43] except that Trypsin/LysC mix (Promega) was used instead of trypsin. The peptides were vacuum-dried, dissolved in 1 % formic acid and loaded on a nanoflow HPLC system (Easy-nLC1000, Thermo Fisher Scientific) coupled to a Q Exactive (Q Exactive HF for collagen IV samples) Hybrid Quadrupole-Orbitrap Mass Spectrometer (Thermo Fisher Scientific) equipped with a nano-electrospray ionization source. The peptides were first loaded on a trapping column and subsequently separated inline on a 15-cm C18 column (75 μm x 15 cm, ReproSil-Pur 5 μm 200 Å C18-AQ, Dr. Maisch HPLC GmbH). The mobile phase consisted of 0.1 % formic acid (solvent A) and acetonitrile/water (95:5 (v/v)) with 0.1 % formic acid (solvent B). The peptides were separated with a 10-min gradient from 5 to 43 % of solvent B. Before the end of the run, the percentage of solvent B was raised to 100 % in 2 min and kept there for 8 min. Full MS scan over the mass-to-charge (*m/z*) range of 300–2000 was performed. The top 5 (top 10 for collagen IV samples) ions were selected with an isolation window of 2.0 *m/z* and a dynamic exclusion time of 10 s and fragmented by higher energy collisional dissociation with a normalized collision energy of 27 and scanned over the *m/z* range of 200–2000. After the MS2 scan for each of the top 5 ions had been obtained, a new full mass spectrum scan was acquired and the process repeated until the end of the 20-min run.

Tandem mass spectra were searched against mouse Swissprot sequences (release 2018_01) with Proteome Discoverer software, version 1.4 (2.3 for collagen IV samples) (Thermo Fischer Scientific) using Mascot 2.6.1 search engine (Matrix Science) allowing for 7 (5 for collagen IV samples) ppm precursor mass tolerance and 0.02 (0.2 for collagen IV samples) Da fragment mass tolerance. Carbamidomethyl (C) as a fixed modification, and oxidation (Met, Lys, Pro) (additionally galactosyl-Lys-and glucosylgalactosyl-Lys-for collagen IV samples) as dynamic modifications were included. Maximum of two (four for collagen IV samples) missed cleavages were allowed. Decoy database search using reversed mouse SwissProt sequences was used to assess false discovery rate. Only the peptides with false discovery rate < 0.05 and determined as "rank 1″ by Proteome Discoverer software were accepted for further analysis. Only the triplets with more than 5 PSMs in each sample were analyzed.

Relative hydroxylation for each XPG sequence was obtained by using the following formula: HypPSM / allPSM, where HypPSM is the number of MS/MS spectra matching to any peptide containing at least one hydroxylated XPG sequence, and allPSM is the number of MS/MS spectra matching to any peptide containing at least one hydroxylated or non-hydroxylated XPG sequence. For example, for relative hydroxylation of APG in one sample, we calculated all spectra matching to any peptide containing at least one hydroxylated APG sequence to get HypPSM for APG. Then we calculated all MS/MS spectra matching to any peptide containing at least one APG regardless of its hydroxylation state to get allPSM for APG. Relative hydroxylation of all XPG sequences was calculated by dividing the sum of HypPSM values of all XPG sequences with the sum of allPSM values of all XPG sequences. Only the spectra matching peptides that contain an XPG site detected in all genotypes within a sample type were used for calculations.

All statistical tests were performed using IBM SPSS Statistics software, version 28. Shapiro-Wilk test was used to test the normality assumption, homogeneity of variance was tested by Levene's test. Differences between two genotypes were tested by either Student's t-test or non-parametric Mann-Whitney U test if normality assumption was rejected. Differences between multiple genotypes were tested by ANOVA test followed by Tukey test, if normality and homogeneity of variances assumptions were not rejected. Only two-tailed *P* values were considered.

Melting temperature analyses

After neutralization, soluble collagen was digested with a mixture of trypsin and chymotrypsin [44] for 2 min at temperatures from 36 °C to 43 °C with 1 °C increments The samples were analyzed by 8 % SDS-PAGE with Coomassie Blue staining and the collagen α1(I) and α2(I) chains were quantified using Image Lab 6.1.0 (Bio-Rad). Tm value was calculated with GraphPad Prism 9.3.1 (GraphPad Software, LLC. CD spectroscopy was performed using a Chirascan CD spectrometer (Applied Photophysics, Leatherhead, UK). The collagen samples were diluted with 0.1 M HAc to a concentration of 0.1 mg/ml. The concentration of each sample was verified with absorbance at 205 nm. CD data was collected between 280 and 190 nm at 20 °C using a 0.1 cm path-length quartz cuvette. CD measurements were acquired every 1 nm with 0.5 s as an integration time and repeated three times with baseline correction. Data were processed using Chirascan Pro-Data Viewer (Applied Photophysics). The direct CD measurements (θ; mdeg) were converted into mean residue molar ellipticity ([θ]MR) by Pro-Data Viewer. Thermal denaturation of the protein samples was monitored by measuring the CD spectra in the same setup with a temperature range from 20 °C to 96 ºC at a rate of 1 °C/minute using a Peltier Temperature Control TC125 (Quantum Northwest, Liberty Lake, WA). The CD data was recorded at every 2 °C. The Tm was calculated with GraphPad Prism 9.3.1 (GraphPad Software, LLC).

Transmission electron microscopy

Transmission electron microscopy (TEM) was performed at the Biocenter Oulu Electron Microscopy Core Facility. For TEM analysis, 12-week-old WT and *P4ha2*$^{-/-}$ female mouse skin was fixed in 1 % glutaraldehyde - 4 % formaldehyde mixture in 0.1 M phosphate buffer, post-fixed in 1 % osmium tetroxide, dehydrated in acetone and embedded in Epon LX 112 (Ladd Research Industries). Thin sections (70 nm) were cut with Leica Ultracut UCT ultramicrotome (Leica Microsystems), stained in uranyl acetate and lead citrate and examined in Tecnai G2 Spirit 120 kV transmission electron microscope (FEI Europe). Images were captured by Quemesa CCD camera and analyzed using iTEM software (Olympus Soft Imaging Solutions GMBH). Basement membrane thickness was calculated with iTEM from 4 WT and 5 *P4ha2*$^{-/-}$ mice. Collagen fibril diameter was measured using ImageJ 1.53 s. 80–100 fibrils were measured from six different locations of the

dermis from each mouse resulting in a total of 2115 and 2618 fibrils measured from 4 WT and 5 *P4ha2*$^{-/-}$ mice, respectively. All locations analyzed were within 10 μm from the epidermis.

In vitro peptide and procollagen prolyl 4-hydroxylation assays

Recombinant human C-P4H-I and C-P4H-II were produced in Sf9 insect cells and purified as described previously [10,14]. Expression plasmids for various collagens and the collagen-like proteins C1qA and ColQ, and their sources are listed in Table S5. The ethanol:water component of 70 μCi L-[2,3,4,5–^{3}H]-Proline (Perkin Elmer) was removed by evaporation using speed-vac and the residue was used to produce L-[2,3,4,5–^{3}H]-proline labeled procollagen chains and the collagen-like proteins in the TnT® Coupled Reticulocyte Lysate Systems (Promega) using either SP6 or T7 promoters. Unincorporated radiolabeled proline was removed by 3–4 rounds of dialysis (Visking Dialysis Tubing 12 - 14,000 Da, Medicell Membranes Ltd) against H_2O at +4 °C. Each proline-labeled protein sample was divided to three and used as a substrate for recombinant human C-P4H-I and II, and the 4-[^{3}H]-hydroxyproline formed in the substrate was analyzed as described [45]. Reaction with no enzyme was used as a control.

$(X\text{-Pro-Gly})_5$ peptides were synthesized by Innovagen and used as C-P4H substrates in an assay where C-P4H activity was determined based on hydroxylation-coupled decarboxylation of 2-oxo[1–^{14}C]glutarate followed by measurement of the formed $^{14}CO_2$ [46]. Lineweaver-Burk blot was used to calculate kinetic values.

Droplet digital PCR

ddPCR was performed for absolute quantification of *P4ha* transcript levels. RNA was extracted from WT MEFs, mouse skin and kidneys using E.Z.N.A total RNA kit I (Omega Bio-Tek) for cells and TRIzol (Invitrogen) for tissues. Residual DNA was removed by RNAase-free DNAase I from Omega Bio-Tek for cell samples and from Thermo Scientific for tissue samples. cDNA was prepared by reverse transcription with iScript cDNA synthesis kit (Bio-Rad). TaqMan® Gene expression assays (ThermoFisher Scientific) Mm00803137_m1, Mm00477940_m1 and Mm00622868_m1 were used for *P4ha1, P4ha2* and *P4ha3*, respectively. Droplet digital PCR QX200 (Bio-Rad) was used with manual droplet generator (Bio-Rad) and ddPCR Supermix for probes (no dUTP) (Bio-Rad) according to the manufacturer's instructions. Relative abundance of transcript copy numbers was calculated.

Peptide docking

Peptide docking experiments with various X-Pro-Gly-triplet peptides were done with GalaxyPepDock server [35]. This program uses template-based docking method and builds the models by energy-based scoring function. In all docking runs, the program used the CrP4H crystal structure (PDB code 3GZE) and its bound Ser-Pro-repeat peptide as a template for the protein and peptide parts, respectively. Only the CAT domains of the C-P4H-I and C-P4H-II α subunits were used as input coordinate models for the docking experiments. The CAT domain models were based on the AlphaFold2 [47] and the RoseTTAFold of the Robetta protein structure prediction server [34] models of the complete α subunits of C-P4H-I and C-P4H-II. These models were almost identical with the CAT domain of the experimentally determined crystal structure of a truncated C-P4H-II [15], with the exception of two flexible loop regions, a hairpin loop and βII-βIII loop, which are not seen or built to the fragmented electron density in the crystal structure. Structural figures were made with Pymol (https://pymol.org/2/) and CCP4mg molecular graphics software [48].

Single cell RNA-seq and bioinformatics analyses

Single cell RNAseq data from human and mouse were downloaded from the Tabula Sapiens [36] and the Tabula Muris [37] consortia repositories, respectively. For both, immune cell-specific data were removed prior to the analysis to focus on the epithelial and stromal compartments of any organ. Data were imported into R (version 4.0.2), reconstructed using Seurat [49] and reanalyzed. Positive expression of *P4HA* genes (*P4HA1, P4HA2* and *P4HA3*) was defined by gene count values greater or equal to 1. Collagen triplets were defined by the X-Pro-Gly-sequence, and all possible combinations with any of the 20 natural amino acids in position X were searched. Collagen gene symbols (human and mouse) were mapped onto Ensemble IDs which were used to retrieve Uniprot canonical sequence identifiers via the UniprotR [50] package. IDs were then manually reviewed and are provided in Tables S3 and S4.

Data availability

The MS proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [51] partner repository with the dataset identifiers PXD008802 (generated previously by Sipilä and co-workers [42]) and PXD035945.

Declaration of Competing Interest

J. M. owns equity in FibroGen Inc, which develops P4H inhibitors as potential therapeutics. This company has supported research in the J.M. group.

Data availability

Data will be made available on request.

Acknowledgements

Minna Siurua is acknowledged for expert technical assistance. MS analyses were performed at the Turku Proteomics Facility and peptide docking experiments at the Biocenter Oulu Structural Biology core facility (part of Instruct-ERIC Centre Finland and FINStruct). EM sample preparation and imaging was done at the Biocenter Oulu EM core facility (part of Finnish Advanced Microscopy Node of Euro-BioImaging Finland). We thank Dr. Hongmin Tu for CD analyses that were done at the Biocenter Oulu proteomics and protein analysis core facility. All the facilities are supported by Biocenter Finland. Part of the work was carried out with the support of The Oulu Laboratory Animal Centre Research Infrastructure, University of Oulu, Finland. This research is connected to the DigiHealth-project (VI), a strategic profiling project at the University of Oulu, and the Infotech Institute (VI) of the University of Oulu. This work was funded by the Academy of Finland [project grants 296498 (JM), 259769 (JH) 329742 (VI)], the Academy of Finland Center of Excellence 2012–2017 [grant 251314 (JM)], the Sigrid Jusélius Foundation (JM and JH), the Jane and Aatos Erkko Foundation (JM) and Cancer Foundation Finland (JH and VI).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.matbio.2023.12.001.

References

[1] J. Myllyharju, K.I. Kivirikko, Collagens, modifying enzymes and their mutations in humans, flies and worms, Trends Genet. 20 (1) (2004) 33–43.
[2] J. Myllyharju, Collagen hydroxylases. RSC Metallobiology. The Royal Society of Chemistry, 2015, pp. 149–168, https://doi.org/10.1039/9781782621959-00149 (2-Oxoglutarate-Dependent Oxygenases; vols. 2015-Janua). Available from:.
[3] S Ricard-Blum, The collagen family, Cold Spring Harb. Perspect. Biol. 3 (1) (2011), a004978. Jan.

[4] A.M. Salo, J. Myllyharju, Prolyl and lysyl hydroxylases in collagen synthesis, Exp. Dermatol. 30 (1) (2021) 38–49. Jan.

[5] J.A. Ramshaw, N.K. Shah, B. Brodsky, Gly-X-Y tripeptide frequencies in collagen: a context for host-guest triple-helical peptides, J. Struct. Biol. 122 (1–2) (1998) 86–91.

[6] M. Kirchner, H. Deng, Y. Xu, Heterogeneity in proline hydroxylation of fibrillar collagens observed by mass spectrometry, PLoS One 16 (8) (2021), e0250544.

[7] M.D. Shoulders, R.T. Raines, Collagen structure and stability, Annu. Rev. Biochem. 78 (2009) 929–958.

[8] R.A. Berg, D.J. Prockop, 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 (1) (1973) 115–120. May.

[9] T. Helaakoski, K. Vuori, R. Myllylä, K.I. Kivirikko, T. Pihlajaniemi, Molecular cloning of the alpha-subunit of human prolyl 4-hydroxylase: the complete cDNA-derived amino acid sequence and evidence for alternative splicing of RNA transcripts, Proc. Natl. Acad. Sci. U. S. A. 86 (12) (1989) 4392–4396. Jun.

[10] P. Annunen, T. Helaakoski, J. Myllyharju, J. Veijola, T. Pihlajaniemi, K.I. Kivirikko, Cloning of the human prolyl 4-hydroxylase alpha subunit isoform alpha(II) and characterization of the type II enzyme tetramer. The alpha(I) and alpha(II) subunits do not form a mixed alpha(I)alpha(II)beta2 tetramer, J. Biol. Chem. 272 (28) (1997) 17342–17348. Jul.

[11] T. Helaakoski, P. Annunen, K. Vuori, I.A. MacNeil, T. Pihlajaniemi, K.I. Kivirikko, Cloning, baculovirus expression, and characterization of a second mouse prolyl 4-hydroxylase alpha-subunit isoform: formation of an alpha 2 beta 2 tetramer with the protein disulfide-isomerase/beta subunit, Proc. Natl. Acad. Sci. U. S. A. 92 (10) (1995) 4427–4431. May.

[12] L. Kukkola, R. Hieta, K.I. Kivirikko, J. Myllyharju, Identification and characterization of a third human, rat, and mouse collagen prolyl 4-hydroxylase isoenzyme, J. Biol. Chem. 278 (48) (2003) 47685–47693. Nov.

[13] Diepstraten C Van Den, K. Papay, Z. Bolender, A. Brown, J.G Pickering, Cloning of a novel prolyl 4-hydroxylase subunit expressed in the fibrous cap of human atherosclerotic plaque, Circulation 108 (5) (2003) 508–511. Aug.

[14] M.K. Koski, J. Anantharajan, P. Kursula, P. Dhavala, A.V. Murthy, U. Bergmann, et al., Assembly of the elongated collagen prolyl 4-hydroxylase alpha2beta2 heterotetramer around a central alpha2 dimer, Biochem. J. 474 (5) (2017) 751–769. Feb.

[15] A.V. Murthy, R. Sulu, A. Lebedev, A.M. Salo, K. Korhonen, R. Venkatesan, et al., Crystal structure of the collagen prolyl 4-hydroxylase (C-P4H) catalytic domain complexed with PDI: toward a model of the C-P4H α(2)β(2) tetramer, J. Biol. Chem. 298 (12) (2022), 102614. Dec.

[16] R. Chowdhury, M.A. McDonough, J. Mecinovic, C. Loenarz, E. Flashman, K. S. Hewitson, et al., Structural basis for binding of hypoxia-inducible factor to the oxygen-sensing prolyl hydroxylases, Structure 17 (7) (2009) 981–989. Jul.

[17] M.K. Koski, R. Hieta, M. Hirsilä, A. Rönkä, J. Myllyharju, R.K. Wierenga, The crystal structure of an algal prolyl 4-hydroxylase complexed with a proline-rich peptide reveals a novel buried tripeptide binding motif, J. Biol. Chem. 284 (37) (2009) 25290–25301. Sep.

[18] J. Myllyharju, K.I. Kivirikko, Characterization of the iron- and 2-oxoglutarate-binding sites of human prolyl 4-hydroxylase, EMBO J. 16 (6) (1997) 1173–1180. Mar.

[19] J. Myllyharju, K.I. Kivirikko, Identification of a novel proline-rich peptide-binding domain in prolyl 4-hydroxylase, EMBO J. 18 (2) (1999) 306–312. Jan.

[20] J. Anantharajan, M.K. Koski, P. Kursula, R. Hieta, U. Bergmann, J. Myllyharju, et al., The structural motifs for substrate binding and dimerization of the alpha subunit of collagen prolyl 4-hydroxylase, Structure 21 (12) (2013) 2107–2118. Dec.

[21] A.V. Murthy, R. Sulu, M.K. Koski, H. Tu, J. Anantharajan, S.K. Sah-Teli, et al., Structural enzymology binding studies of the peptide-substrate-binding domain of human collagen prolyl 4-hydroxylase (type-II): high affinity peptides have a PxGP sequence motif, Protein Sci. 27 (9) (2018) 1692–1703. Sep.

[22] P. Annunen, H. Autio-Harmainen, K.I. Kivirikko, The novel type II prolyl 4-hydroxylase is the main enzyme form in chondrocytes and capillary endothelial cells, whereas the type I enzyme predominates in most cells, J. Biol. Chem. 273 (11) (1998) 5989–5992. Mar.

[23] Y. Zou, S. Donkervoort, A.M. Salo, A.R. Foley, A.M. Barnes, Y. Hu, et al., P4HA1 mutations cause a unique congenital disorder of connective tissue involving tendon, bone, muscle and the eye, Hum. Mol. Genet. 26 (12) (2017) 2207–2217. Jun.

[24] H. Guo, P. Tong, Y. Liu, L. Xia, T. Wang, Q. Tian, et al., Mutations of P4HA2 encoding prolyl 4-hydroxylase 2 are associated with nonsyndromic high myopia, Genet. Med. 17 (4) (2015) 300–306. Apr.

[25] F.D. Carmona, A. Vaglio, S.L. Mackie, J. Hernández-Rodríguez, P.A. Monach, S. Castañeda, et al., A genome-wide association study identifies risk alleles in plasminogen and P4HA2 associated with giant cell arteritis, Am. J. Hum. Genet. 100 (1) (2017) 64–74. Jan.

[26] D.M. Gilkes, G.L. Semenza, D. Wirtz, Hypoxia and the extracellular matrix: drivers of tumour metastasis, Nat. Rev. 14 (6) (2014) 430–439. Jun.

[27] H.M. Hanauske-Abel, Prolyl 4-hydroxylase, a target enzyme for drug development. Design of suppressive agents and the *in vitro* effects of inhibitors and proinhibitors, J. Hepatol. (1991), 13 Suppl 3:S8-15; discussion S16.

[28] J. Myllyharju, Prolyl 4-hydroxylases, key enzymes in the synthesis of collagens and regulation of the response to hypoxia, and their roles as treatment targets, Ann. Med. 40 (6) (2008) 402–417.

[29] T. Holster, O. Pakkanen, R. Soininen, R. Sormunen, M. Nokelainen, K.I. Kivirikko, et al., Loss of assembly of the main basement membrane collagen, type IV, but not fibril-forming collagens and embryonic death in collagen prolyl 4-hydroxylase I null mice, J. Biol. Chem. 282 (4) (2007) 2512–2519. Jan.

[30] E. Aro, A.M. Salo, R. Khatri, M. Finnila, I. Miinalainen, R. Sormunen, et al., Severe extracellular matrix abnormalities and chondrodysplasia in mice lacking collagen prolyl 4-hydroxylase isoenzyme II in combination with a reduced amount of isoenzyme I, J. Biol. Chem. 290 (27) (2015) 16964–16978. Jul.

[31] J.-P. Tolonen, A.M. Salo, M. Finnilä, E. Aro, E. Karjalainen, V.-P. Ronkainen, et al., Reduced bone mass in collagen prolyl 4-hydroxylase P4ha1 (+/-); P4ha2 (-/-) compound mutant mice, JBMR plus 6 (6) (2022) e10630. Jun.

[32] K.I. Kivirikko, R. Myllylä, T. Pihlajaniemi, Hydroxylation of Proline and Lysine Residues in Collagens and Other Animal and Plant Proteins, CRC Press, 1991, pp. 1–51. In: Harding JJ, Crabbe MJC, editorsPost-translational modifications of proteins.

[33] D. Wilhelm, A. Wurtz, H. Abouelfarah, G. Sanchez, C. Bui, J.-B. Vincourt, Tissue-specific collagen hydroxylation at GEP/GDP triplets mediated by P4HA2, Matrix Biol. (2023). Mar.

[34] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G.R. Lee, et al., Accurate prediction of protein structures and interactions using a three-track neural network, Science 373 (6557) (2021) 871–876. Aug.

[35] H. Lee, L. Heo, M.S. Lee, C. Seok, GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization, Nucleic. Acids. Res. 43 (W1) (2015) W431–W435. Jul.

[36] Tabula Sapiens Consortium*, Jones RC, J. Karkanias, M.A. Krasnow, A.O. Pisco, S. R. Quake, et al., The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans, Science (80-) 376 (6594) (2022) eabl4896. Available from: https://www.science.org/doi/abs/10.1126/science.abl4896.

[37] N. Schaum, J. Karkanias, N.F. Neff, A.P. May, S.R. Quake, T. Wyss-Coray, et al., Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris, Nature 562 (7727) (2018) 367–372, https://doi.org/10.1038/s41586-018-0590-4. Available from:.

[38] S. Kiriakidis, S.S. Hoer, N. Burrows, G. Biddlecome, M.N. Khan, C.C. Thinnes, et al., Complement C1q is hydroxylated by collagen prolyl 4 hydroxylase and is sensitive to off-target inhibition by prolyl hydroxylase domain inhibitors that stabilize hypoxia-inducible factor, Kidney Int. 92 (4) (2017) 900–908. Oct.

[39] R. Hieta, L. Kukkola, P. Permi, P. Pirilä, K.I. Kivirikko, I. Kilpeläinen, et al., The peptide-substrate-binding domain of human collagen prolyl 4-hydroxylases. Backbone assignments, secondary structure, and binding of proline-rich peptides, J. Biol. Chem. 278 (37) (2003) 34966–34974. Sep.

[40] N. Loyfer, J. Magenheim, A. Peretz, G. Cann, J. Bredno, A. Klochendler, et al., A DNA methylation atlas of normal human cell types, Nature 613 (7943) (2023) 355–364, https://doi.org/10.1038/s41586-022-05580-6. Available from:.

[41] K.A. Hoadley, C. Yau, T. Hinoue, D.M. Wolf, A.J. Lazar, E. Drill, et al., Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer, Cell 173 (2) (2018) 291–304. Apre6.

[42] K.H. Sipilä, K. Drushinin, P. Rappu, J. Jokinen, T.A. Salminen, A.M. Salo, et al., Proline hydroxylation in collagen supports integrin binding by two distinct mechanisms, J. Biol. Chem. 293 (20) (2018) 7645–7658. May.

[43] A. Shevchenko, H. Tomas, J. Havlis, J.V. Olsen, M. Mann, In-gel digestion for mass spectrometric characterization of proteins and proteomes, Nat. Protoc. 1 (6) (2006) 2856–2860.

[44] P. Bruckner, D.J. Prockop, Proteolytic enzymes as probes for the triple-helical conformation of procollagen, Anal. Biochem. 110 (2) (1981) 360–368. Jan.

[45] K. Juva, D.J. Prockop, Modified procedure for the assay of H-3-or C-14-labeled hydroxyproline, Anal. Biochem. 15 (1) (1966) 77–83. Apr.

[46] K.I. Kivirikko, R. Myllylä, Posttranslational enzymes in the biosynthesis of collagen: intracellular enzymes, Methods Enzymol. 82 (1982) 245–304. Pt A.

[47] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, et al., Highly accurate protein structure prediction with AlphaFold, Nature 596 (7873) (2021) 583–589. Aug.

[48] S. McNicholas, E. Potterton, K.S. Wilson, M.E.M. Noble, Presenting your structures: the CCP4mg molecular-graphics software, Acta. Crystallogr. D Biol. Crystallogr. 67 (Pt 4) (2011) 386–394. Apr.

[49] Y. Hao, S. Hao, E. Andersen-Nissen, W.M. Mauck, S. Zheng, A. Butler, et al., Integrated analysis of multimodal single-cell data, Cell 184 (13) (2021) 3573–3587, e29Available from: https://www.sciencedirect.com/science/article/pii/S0092867421005833.

[50] M. Soudy, A.M. Anwar, E.A. Ahmed, A. Osama, S. Ezzeldin, S. Mahgoub, et al., UniprotR: retrieving and visualizing protein sequence and functional information from universal protein resource (UniProt knowledgebase), J. Proteomics 213 (2020), 103613. Available from: https://www.sciencedirect.com/science/article/pii/S1874391919303859.

[51] Y. Perez-Riverol, J. Bai, C. Bandla, D. García-Seisdedos, S. Hewapathirana, S. Kamatchinathan, et al., The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences, Nucleic. Acids. Res. 50 (D1) (2022) D543–D552. Jan.