Evidence for the major role of PH4αEFB in the prolyl 4-hydroxylation of *Drosophila* collagen IV

Yoshihiro Ishikawa[1]*, Melissa A. Toups[2,3], Marwan Elkrewi[4], Allison L. Zajac[5], Sally Horne-Badovinac[5] and Yutaka Matsubayashi[2]*

1: Department of Ophthalmology, University of California, San Francisco, School of Medicine, CA USA
2: Department of Life and Environmental Sciences, Faculty of Science and Technology, Bournemouth University, UK
3: Department of Biology, University of Louisiana at Lafayette, Lafayette, LA 70503, USA.
4: Institute of Science and Technology Austria (ISTA), Klosterneuburg, Austria
5: Department of Molecular Genetics and Cell Biology, the University of Chicago, Chicago, IL, USA

* Corresponding authors:
Yoshihiro Ishikawa, Ph.D.,
Department of Ophthalmology, University of California, San Francisco, School of Medicine, CA USA
E-mail: yoshihiro.ishikawa@ucsf.edu
ORCID-ID: https://orcid.org/0000-0003-2013-0518

Yutaka Matsubayashi
Department of Life and Environmental Sciences, Faculty of Science and Technology, Bournemouth University, UK
E-mail: ymatsubayashi@bournemouth.ac.uk
ORCID-ID: https://orcid.org/0000-0003-2196-8099

Keywords: *Drosophila*, collagen IV, prolyl 4-hydroxylase, Basement membrane, Endoplasmic reticulum.

Abstract

Collagens are fundamental components of extracellular matrices, requiring precise intracellular post-translational modifications for proper function. Among the modifications, prolyl 4-hydroxylation is critical to stabilise the collagen triple helix. In humans, this reaction is mediated by collagen prolyl 4-hydroxylases (P4Hs). While humans possess three genes encoding these enzymes (P4Hαs), *Drosophila melanogaster* harbour at least 26 candidates for collagen P4Hαs despite its simple genome, and it is poorly understood which of them are actually working on collagen in the fly. In this study, we addressed this question by carrying out thorough bioinformatic and biochemical analyses. We demonstrate that among the 26 potential collagen P4Hαs, PH4αEFB shares the highest homology with vertebrate collagen P4Hαs. Furthermore, while collagen P4Hs and their substrates must exist in the same cells, our transcriptomic analyses at the tissue and single cell levels showed a global co-expression of *PH4αEFB* but not the other P4Hα-related genes with the collagen IV genes. Moreover, expression of *PH4αEFB* during embryogenesis was found to precede that of collagen IV, presumably enabling efficient collagen modification by PH4αEFB. Finally, biochemical assays confirm that PH4αEFB binds collagen, supporting its direct role in collagen IV modification. Collectively, we identify PH4αEFB as the primary and potentially constitutive prolyl 4-hydroxylase responsible for collagen IV biosynthesis in *Drosophila*. Our findings highlight the remarkably simple nature of *Drosophila* collagen IV biosynthesis, which may serve as a blueprint for defining the minimal requirements for collagen engineering.

Introduction

Collagens are one of the most abundant protein superfamilies and composed of 28 different types of collagen proteins in humans [1, 2]. They play essential functions such as imparting biophysical properties to tissues and acting as signalling scaffolds [3-5]. For example, collagens I, II, III, V, and XI are classified

as fibrillar collagens, forming structural frameworks in fibril-rich connective tissues, including those in bone, skin, and the vasculature. In contrast, collagens IV, XVII and XVIII and others form flexible networks present in the basement membranes that support all epithelia and many other tissues such as muscles. For a single type of collagen, multiple isoforms can exist. For instance, there are six collagen IV genes in mammals (*COL4A1 – COL4A6*), encoding six collagen IV proteins known as α chains (collagen IV α1 – α6). These six α chains assemble into three heterotrimeric collagen IV isoforms, collagens α1α1α2(IV), α3α4α5(IV) and α5α5α6(IV) [3, 6].

Collagen biosynthesis is a highly orchestrated process and requires more than 20 enzymes and chaperones (hereafter referred to as the 'collagen molecular ensemble') residing in the endoplasmic reticulum (ER) [7-11]. Interestingly, similar to the diversity of collagen types and isoforms, many components of this ensemble, particularly collagen modifying enzymes, also exhibit remarkable diversity: for example, in humans there are three prolyl 4-hydroxylases (P4Hs), three prolyl 3-hydroxylases (P3Hs), three lysyl hydroxylases (LHs), and two glycosyl transferase 25 domain enzymes (GLT25Ds) [12-14]. Therefore, collagen post-translational modifications (PTMs) can serve as signatures that distinguish collagen types [8, 9, 15, 16]. This diversity of collagen types, collagen modifying enzymes, and PTM patterns suggest that collagen biosynthesis must facilitate a complex set of parameters essential for the correct combinations of α-chain isoforms and PTMs across the 28 different collagen types. Therefore, understanding collagen functions and designing therapies against collagen-related diseases requires research not only on collagens themselves but also on the components of the collagen molecular ensemble [17, 18].

The fruit fly *Drosophila melanogaster* is a well-established model organism for several reasons: 1) it has a relatively compact genome yet shares a significant portion of its genes with humans, including those associated with diseases, 2) its genome is well-mapped and extensively annotated, providing a wealth of genetic tools and resources, including numerous mutant strains and transgenic lines, and 3) it boasts a strong and supportive research community with extensive collaboration and resources sharing, facilitated by accessible database and stock centres [19]. This toolset would also be useful to boost research on collagen biosynthesis, considering the reasons described below. First, *Drosophila* possesses only a minimum set of collagens, lacking any fibrillar collagens [20]. Second, its predominant collagen is the basement membrane collagen α1α1α2(IV), with Multiplexin (a homologue of human collagen XV/XVIII) and Pericardin (a collagen IV-like protein) being other collagen(-related) molecules [21, 22]. While mammals have six collagen IV α chain genes, *Drosophila* has only two collagen IV α1 and α2, encoded by *Col4a1* and *viking* (*vkg)* genes, respectively [23, 24]. Lastly, the collagen molecular ensemble for *Drosophila* collagen biosynthesis is also simple. For example, *Drosophila* has only one isoform each of LH and GLT25D (FlyBase IDs FBgn0036147 and FBgn0051915), compared to three and two in humans, respectively. Additionally, a critical collagen molecular chaperone HSP47 has not been identified in *Drosophila*, and it lacks a gene encoding P3H. This suggests that *Drosophila* represents a minimal system for collagen biosynthesis, making it an ideal model to study the core mechanisms of this essential process.

Prolyl 4-hydroxylases (P4Hs) are enzymes that hydroxylate proline residues to 4-hydroxylproine (4Hyp). In humans, there are seven P4Hs, three of which work on collagens while the others target different substrates including the transcription factors known as hypoxia-inducible factors (HIFs) [25-28]. Collagen P4Hs exist in three isoforms, each of which is a tetrameric complex composed of two α and two β subunits. The catalytic α subunits (P4Hαs) are specific to each isoform, and encoded by the genes P4HA1, P4HA2, and P4HA3, respectively. The β subunit, which is identical to the enzyme protein disulfide isomerase (PDI), is shared by all three isoforms [29-31]. The formation of 4Hyp by this tetramer confers thermal stability to collagens: without 4Hyp, collagens are unable to form a stable triple helical structure at body temperature [32-35]. Surprisingly, in contrast to the aforementioned cases with collagens, LHs, and GLT25Ds, *Drosophila* possesses a far larger number of genes potentially encoding collagen P4Hαs than mammals do: 26 genes are currently annotated with a Gene Ontology term suggesting collagen P4H activity,

mainly based on their sequences (see Results). This raises the question: how many P4Hs are involved in *Drosophila* collagen α1α1α2(IV) biosynthesis? One possibility is that the large number of P4H enzymes have evolved to modify collagens in a manner that is specific to *Drosophila,* making them poor models for the action of the human collagen P4Hs. Alternatively, only a small number of the P4Hs modify collagens and the others have different substrates, which might make the *Drosophila* P4Hs good models for the human enzymes.

This question has been partially addressed in previous studies. In 2002, Abrams and Andrew reported that their BLAST search identified nineteen P4Hα-related genes in the *Drosophila* genome. They analysed the expression patterns of ten of them in the embryo and found that only PH4αEFB was expressed in hemocytes and the fat body [36], the major collagen IV-producing tissues [37]. Since prolyl 4-hydroxylation of collagen occurs intracellularly [32, 38], expression in these tissues suggests PH4αEFB is a likely collagen-modifying P4Hα. Later, genetic evidence supported this: RNAi knockdown of PH4αEFB disrupted collagen IV-secretion [39, 40]. In contrast, biochemical studies showed that a recombinant form of another *Drosophila* PH4α that was later named PH4αMP [36] can hydroxylate collagen peptides *in vitro* [41]. Although both PH4αEFB and MP are strong candidates for collagen modifying P4Hαs, it remains unclear whether one, both, or additional P4Hαs contribute to collagen biosynthesis directly in *Drosophila*, as their roles are still not fully defined.

In this study, we carried out bioinformatic and biochemical analyses to identify the *Drosophila* P4Hαs responsible for collagen modification. Our results suggest that PH4αEFB is a central collagen modifying P4Hα among the 26 potential candidates in *Drosophila,* highlighting the simplicity of the collagen molecular ensemble in this model organism.

Results

The Drosophila genome encodes 26 P4Hα-related genes

To comprehensively identify fly collagen P4Hα-related genes, we searched FlyBase for the genes annotated with the Gene Ontology (Molecular Function) term 'procollagen-proline 4-dioxygenase activity', which is associated with human P4HA1 (Prolyl 4-hydroxylase subunit α-1, NCBI Gene ID 5033). This analysis revealed 26 genes (Table 1), including all the 10 that have been analysed previously [36]. At the time of the search, no experimental data had been recorded as the evidence of the GO term annotation, which was instead based solely on the sequence of the genes (the reports about PH4αMP and EFB [34-36, 41] had not been curated yet). Moreover, the majority of these genes had not been named and are still referred to by 'CG' symbols. These facts indicate that the genes have been only minimally characterised.

Domain structure is conserved between human collagen P4Hα and 25 Drosophila homologues

Each human collagen P4Hα protein has four parts: the N-terminal (N) domain, the peptide-substrate-binding (PSB) domain, a linker (L) region, and the C-terminal catalytic (CAT) domain (Figure 1A). This structure enables a P4Hα protein to bind to another P4Hα via the N-domain and to the β subunit PDI via the CAT domain, forming the α2β2 tetramer[30, 31, 42, 43]. The P4HA-PDI interaction is essential for maintaining the solubility and ER retention of P4HA required for its enzymatic activity [44, 45]. The formation of this tetramer has been proposed to be evolutionarily conserved in a previously identified *Drosophila* collagen P4Hα-related protein PH4αMP [29, 36, 41]. Thus, we examined whether the 'N-PSB-L-CAT' structure that enables collagen P4H formation is present in the proteins encoded by the 26 fly collagen P4Hα-related genes identified in this study. We compared the sequences of the fly proteins with that of human collagen P4HA2; as a control, we also examined human PHD3, a prolyl hydroxylase whose substrate is the transcription factor hypoxia-inducible factor (HIF) rather than collagens. It is known that the structure outside the catalytic domain is unrelated between collagen and HIF P4Hs [26, 46] (Figure 1B). While the sequence alignment confirmed that the structures of P4HA2 and PHD3 are conserved only within

the catalytic domains (Table S1), we found that almost all the *Drosophila* collagen P4Hα-related proteins have the same domain organisation as human collagen P4HAs except for the following two cases. First, about 70% of the N-terminal side of the N domain of CG15539-PA is truncated, while the other protein isoforms encoded by the same gene (CG15539-PB and PC) contain a full-length N-domain (https://flybase.org/reports/FBgn0039782) (Figure 1C, Table S1). Second, all the three products of *CG34041* (PD, PE, and PF) lack the catalytic domain; FlyBase shows that they contain one or two N-domains (https://flybase.org/reports/FBgn0054041) (Figure 1 D and E, Table S1). Thus, *CG34041* does not encode a prolyl hydroxylase. However, for the other 25 fly collagen P4Hα-related genes, the 'N-PSB-L-CAT' domain organisation is conserved between their products and human collagen P4HAs, suggesting the ability of the fly proteins to form tetramers with PDI. This necessitates further investigation to identify which of the fly enzymes work on collagen IV.

Comprehensive phylogenetic analysis confirms that PH4αEFB is closest to human P4HAs

Subsequently, we examined the sequence similarity between human and *Drosophila* collagen P4Hα-related proteins. Since phylogenetic analysis revealed that PH4αEFB was closest to human collagen P4H1A1/2 among the 8 P4Hα-related proteins previously tested [36], we conducted a more comprehensive analysis to determine if unexamined fly proteins show a similar or higher homology to human collagen P4HAs. We constructed a phylogenetic tree using the sequences of human collagen P4HA1, 2, and 3, and the proteins encoded by 25 out of the 26 *Drosophila* collagen P4Hα-related genes: *CG34041* was excluded from the assay as it was found not to encode a catalytic domain. As a control, we also examined three human and one *Drosophila* HIF prolyl hydroxylases, PHD1-3 [26] and Hph [47], respectively. This analysis first revealed that the HIF prolyl hydroxylases—human PHD1-3 and *Drosophila* Hph—formed a distinct clade (Figure 1F top, green and marked with asterisks), suggesting a correspondence between sequence similarity and substrate specificity. Building on this, the three human collagen P4HA proteins and *Drosophila* PH4αEFB formed another clade comprising only these proteins, with the fly enzyme occupying a central position (Figure 1F, magenta and highlighted in bold; arrow points PH4αEFB-PA, the only annotated product of the *PH4αEFB* gene). These results strengthen the previous notion that PH4αEFB is closest to human collagen P4HAs among all the fly collagen P4Hα-related proteins [36].

PH4αEFB shows a spatiotemporal expression pattern compatible with it being the major P4Hα for Drosophila collagen IV

As the proline 4-hydroxylation of collagens occurs intracellularly [32, 38], an enzyme mediating this reaction must be co-expressed in the same cell as its substrate collagens. As described in the phylogenetic analysis above, Abrams and Andrew also analysed the expression patterns of the 10 P4Hα-related genes to identify which enzymes are responsible for the proline 4-hydroxylation in the fly embryo using mRNA in situ hybridisation [36]. Their analysis revealed that *PH4αEFB* but not the other nine genes is expressed in hemocytes (macrophages) and the fat body, which are the major sources of collagen IV in the embryo [24, 37]. Here, we validate whether PH4αEFB remains the sole *Drosophila* PH4α targeting collagen IV as previously suggested [36], utilising newly available materials.

First, we examined the co-expression of all 26 collagen P4Hα-related genes (Table 1) and the two collagen IV genes, *Col4a1* and *vkg*, in many different postembryonic tissues, cultured cells, and individuals under various environmental perturbations utilising five publicly available transcriptome datasets: FlyAtlas Anatomy Microarray, FlyAtlas2 Anatomy RNA-Seq, and modENCODE Anatomy RNA-Seq for the data from the tissues, modENCODE Cell Lines RNA-Seq for the cultured cells, and modENCODE Treatments RNA-Seq for the environmental perturbations [48-50] (Figure 2, Table 2, Table S2). To quantify the similarity of the expression patterns of *Col4a1* and each of the other genes, we measured the correlation coefficient (r), which takes a value between -1 and 1. The value of r = 1, 0, and -1 indicate identical, uncorrelated, and complementary expression patterns to *Col4a1*, respectively. In the first four datasets, *vkg* showed a nearly identical expression pattern to *Col4a1* ($r > 0.95$); in the modENCODE Treatments dataset,

202 the correlation was slightly lower but still high (r = 0.79). These high correlations presumably reflect the
203 evolutionarily conserved co-expression of the two collagen IV subunit genes under the control of a common
204 promoter [51], supporting the reliability of our assay. Regarding the collagen P4Hα-related genes,
205 *PH4αEFB* showed the largest and outstanding r values in three datasets (FlyAtlas, FlyAtlas2, and
206 modENCODE Cell Lines) (Figures 2A and B). Indeed, the expression levels of *Col4a1* and *PH4αEFB*
207 often change in parallel between samples in these datasets (e.g., thick solid bracket in Figure 2C, FlyAtlas2).
208 Even in the remaining two datasets (modENCODE Anatomy and Treatments), *Col4a1* and *PH4αEFB*
209 levels change almost in parallel, although the graphs do not necessarily exhibit identical numbers of peaks
210 and troughs (e.g., thin solid brackets in Figure 2C). In addition, apart from only two exceptions (Figure 2C,
211 closed arrows, both from the larval salivary gland) out of a total of 145 samples in all the five datasets,
212 *PH4αEFB* expression was visible in the samples expressing *Col4a1*.
213
214 Moreover, we also examined the co-expression of *Col4a1*, *vkg* and *P4Hα*-related genes using single-
215 nucleus transcriptome data available from the Fly Cell Atlas [52] in two target tissues ovary and fat body,
216 which are known to express collagen IV [24, 53], as well as the whole-body dataset. To reduce the sparsity
217 of the data, we aggregated neighbouring cells into metacells, where the expression levels of each gene were
218 averaged. Below, all the assays were done using these metacells unless otherwise stated. As in the
219 transcriptome datasets above, we examined the co-expression of *Col4a1*, *vkg* and all the 26 collagen P4Hα-
220 related genes. In all the three single cell datasets, we confirmed the strong co-expression of *Col4a1* and *vkg*
221 (r > 0.90) (Figure 3A, open squares; raw values are in Table S3) and the prominent co-expression of *Col4a1*
222 with *PH4αEFB* among the 26 P4Hα-related genes (Figure 3A, magenta and black circles) consistent with
223 Figure 2. Collectively, these results indicate a high level of co-expression between *PH4αEFB* with *Col4a1*
224 at the tissue and single cell levels.
225
226 Next, to further confirm the co-expression of *PH4αEFB* with *Col4a1* in these single metacell datasets, we
227 implemented a high-dimensional weighted correlation network analysis in the R package hdWGCNA [54],
228 which detects co-expression modules for each cell type within a given tissue. Across all analyses, the only
229 *P4Hα*-related gene that shared a co-expression module with the collagen subunits was *PH4αEFB*. To
230 determine whether the co-occurrence was statistically significant, we performed a Monte Carlo permutation
231 test: module assignments for our candidate and collagen IV subunits for co-expression were randomised
232 10,000 times, and the number of co-assigned modules was recorded or a randomly selected candidate gene
233 in each iteration. The resulting p-values are the proportion of randomisations where the number of matched
234 module assignments was equal to or exceeded the empirical value. In all three datasets, we detected a
235 significant enrichment of *PH4αEFB* (Table 3), suggesting potential co-regulation with collagen IV subunits.
236 Therefore, we plotted the expression levels of *Col4a1* vs. *vkg* or *PH4αEFB* for each single metacell to
237 visualise their co-expression (Figures 3B and S1). In all the fat body, ovary, and whole-body datasets, we
238 found that the plotted points for *Col4a1* vs. *vkg* largely aligned along a single regression line with a positive
239 slope, indicating a prominent positive correlation in their expression (Figures 3B top and Figure S1).
240 Similarly, most of the points for *Col4a1* vs. *PH4αEFB* also aligned along one regression line with a positive
241 slope in all the three datasets (Figure 3B bottom and Figure S1). These positive correlations between *Col4a1*
242 and *vkg/PH4αEFB* were also observed in many individual cell types separately analysed (Figs. S2-4). We
243 also visualised the co-expression of the collagen IV genes and PH4αEFB on the uniform manifold
244 approximation and projections (UMAPs). Co-expression of *Col4a1* and *vkg* or *PH4αEFB* was confirmed
245 in cells such as hemocytes, follicle cells, and glia in the fat body, ovary, and whole-body datasets,
246 respectively (closed arrows in Figure S5, Figure 4, and Figure S6, respectively). In UMAPs, co-expression
247 of *Col4a1* and *vkg* or *PH4αEFB* in cells with lower gene expression (e.g., 'adult fat body cells' in the fat
248 body dataset, indicated by open arrows in Figures. 4 and S5) was less clear than in metacell data (Figure
249 3B), likely due to lower expression levels and higher noise in raw single-cell data. Taken together, we
250 conclude that these metacell plots and UMAPs indicate a high level of co-expression between *PH4αEFB*
251 with both *Col4a1* and *vkg* at the single cell level in various tissues.

It is worth noting that not all cells showed the *Col4a1-PH4αEFB* correlation as mentioned above, although the number of such outliers was not high enough to noticeably affect the overall regression lines or correlation coefficients. For example, there existed 'low (or no) *PH4αEFB* but high *Col4a1*' expression cells in the fat body and whole-body datasets (Figure 3B bottom, dashed-edge rectangles). Conversely, in the metacell plot of the ovary dataset, there was a group of metacells with 'high *PH4αEFB* but low *Col4a1*' expression, many of which are unannotated cells (green interior rectangle in Figure 3B bottom centre; Figure S3 bottom right, 'unannotated'). These metacells should correspond to the cluster of 'unannotated' cells with the area of 'high *PH4αEFB* but low collagen genes expression' in the UMAP (Figure 4, open arrows). Potential mechanisms of collagen IV modification and functions of PH4αEFB in these 'low *PH4αEFB*-high *Col4a1*' and 'high *PH4αEFB*-low *Col4a1*' expression cells will be explored further in the Discussion.

As the final transcriptome analysis, we compared the expression time courses of the collagen P4Hα-related and collagen IV genes using the modENCODE Development RNA-Seq data [55]. The expression of *PH4αEFB* was found to precede that of the collagen IV genes and continue throughout life, suggesting that PH4αEFB is always 'ready' to hydroxylate newly synthesised collagen IV cains (Figure 5 and Table S4). In conclusion, these spatial and temporal expression patterns are highly suggestive that PH4αEFB plays the major and potentially a constitutive role in the proline 4-hydroxylation of collagen IV in *Drosophila*.

PH4αEFB is a collagen binding Drosophila P4Hα

Finally, we examined the collagen-binding ability of PH4αEFB using a gelatine pulldown assay. *Drosophila* D17 cells express 6 potential collagen P4Hα-related genes including *PH4αEFB* (≥ 1 RPKM in sample 15, modENCODE_Cell_Lines, Sup_Table_1). We incubated D17 cell lysate with gelatine-coupled beads and identified gelatine-bound proteins by protein ID LC-MS analysis. As a positive control, we used mouse PFHR9 cells, which are known to express collagen α1α1α2(IV) and its binding proteins [9, 56] (Figures 6A-C). From the PFHR9 cell extract, many collagen-interacting proteins such as HSP47 and collagen P4HAs were pulled down, demonstrating the reliability of this assay. Importantly, PH4αEFB was among the 19 proteins pulled down from the D17 extract (Figures 6C and D), indicating its collagen-binding ability. Importantly, while the structure of PH4αEFB predicted its binding to PDI [29] (Table S1, Figure 1A), *Drosophila* PDI indeed co-precipitated PH4αEFB (Figures 6C and E). Thus, our result strongly supports that PH4αEFB directly binds to collagen and forms a complex with PDI under physiological conditions.

Discussion

In *Drosophila*, despite its relatively simple genome, there are at least 26 potential collagen P4Hαs, and their roles have not been fully elucidated. PH4αEFB and PH4αMP have been suggested to mediate collagen IV hydroxylation [36, 39-41], which is the only major collagen in *Drosophila*. However, almost nothing is known about the functions of the other potential collagen P4Hαs, and some of them may also modify collagen IV. In this study, our thorough bioinformatic and biochemical analyses corroborate that PH4αEFB plays the major and potentially a constitutive role in the proline 4-hydroxylation of collagen IV. This conclusion is also supported by several other studies. First, there are multiple reports of the enriched expression of *PH4αEFB* in embryonic macrophages [36, 57], which are the major source of collagen IV in the *Drosophila* embryo [37, 56]. Consistently, mutants lacking macrophages show reduced expression of *PH4αEFB* [58]. Moreover, knockdown of PH4αEFB alone is sufficient to disrupt collagen IV secretion from the fat body and ovarian follicle cells [39, 40].

However, our findings do not entirely rule out the role of P4Hαs other than PH4αEFB in the proline 4-hydroxylation of *Drosophila* collagen IV in specific tissues or contexts. Notably, our sequence analysis

302 revealed that apart from *CG34041*, all the other 25 *Drosophila* collagen P4Hα-related genes encode
303 proteins that harbour the 'N-PSB-L-CAT' domain organisation, a hallmark structure of collagen P4Hs [30,
304 31, 42, 43] (Figure 1A). Therefore, the products of any of these 25 genes may work on collagen IV (or/and
305 other collagen(-related) molecules such as Multiplexin or Pericardin as discussed later). Indeed, PH4αMP
306 hydroxylates collagen peptides *in vitro* [36, 41] and clusters near PH4αEFB and human collagen P4HAs in
307 the phylogenetic tree (shaded dark grey in Figure 1F), suggesting potential similarity in substrate
308 specificities. If PH4αMP targets collagen IV, other related P4Hαs including PH4αSG1, SG2, and NE1
309 (shaded light grey in Figure. 1F) might also do so. Consistently, several observations support their possible
310 roles in collagen IV hydroxylation *in vivo*. First, *PH4αSG1* and *SG2* are highly expressed in the larval
311 salivary gland, where *PH4αEFB* is nearly absent despite *Col4a1* expression (closed arrows in Figure 2C).
312 While PH4αSG1 and SG2 are suggested to have other substrate(s) in the embryonic salivary gland [59],
313 they may target collagen IV in later stages. Moreover, in the *Drosophila* embryo, the expression of
314 *PH4αSG1* (but not *PH4αEFB*) is reported to be enriched in caudal visceral mesoderm (CVM) cells, another
315 collagen IV source [57]. These findings suggest a possible role for PH4αSG1 and 2 in the salivary gland
316 and for PH4αSG1 in CVM cells. Furthermore, in the modENCODE Treatments dataset, *PH4αSG1, SG2,*
317 and *NE1* expressions show high correlations with *Col4a1*expression (grey arrows in Figures. 2A and B,),
318 similarly to *PH4αEFB* in other datasets. In Figure 2C, when *Col4a1* expression rises (sample 16 – 20),
319 *PH4αEFB* also rises (dotted bracket). However, from sample 20 to 21 (open arrow), *Col4a1* rises sharply
320 while *PH4αEFB* slightly decreases. Here, *PH4αSG1*, *SG2*, and *NE1* expression increase instead. In the
321 modENCODE Treatments dataset, gene expression changes in response to various environmental
322 perturbations were examined. For example, cadmium (sample 4 in Figure 2C), ethanol (sample 13), and
323 rotenone (sample 20) induced *PH4αSG1* expression, while sindbis virus (sample 22) induced *PH4αSG2*
324 and *NE1* expression. Thus, PH4αSG1, SG2, and NE1 induced by various stresses may compensate for or
325 support PH4αEFB in collagen IV hydroxylation. Nevertheless, we did not detect PH4αSG1 binding to
326 gelatine beads in our biochemical analysis using *Drosophila* D17 cells, despite reported *PH4αSG1*
327 expression in this cell line (Figure 2A and Table S2, sample 15). Possible explanations include loss of
328 *PH4αSG1* expression in our cell sub-strain, lack of translation, or inability to bind gelatine. These
329 possibilities warrant future investigation.
330
331 In the single nucleus data, we detected a small population of metacells with low (or no) *PH4αEFB* but high
332 *Col4a1* (Figure. 3B). We could not easily pinpoint these cells in UMAPs due to dispersion and noise
333 (Figures. S5 and S6). We hypothesised that other enzymes might be modifying collagen IV in these cells,
334 but could not identify clear candidates, as most of these cells did not express any of the 25 other PH4α-
335 related genes (Table S5). We speculate that in these cells, collagen P4Hα mRNA, regardless of whether it
336 is for PH4αEFB, might have been degraded while the translated enzyme remains active, or collagen IV
337 mRNAs are not translated: the levels of an mRNA and the protein it encodes do not always correlate [60].
338
339 While PH4αEFB might not be the sole P4Hα for *Drosophila* collagen IV, collagen IV might not be the
340 only substrate for PH4αEFB either. The correlation between *Col4a1* and *PH4αEFB* expression is weaker
341 than that between *Col4a1* and *vkg*, which encodes the other subunit of *Drosophila* collagen IV (Figures. 2-
342 4 and S2-S6). Figure 2C shows several *PH4αEFB* expression peaks accompanied with low *Col4a1*
343 expression (asterisks). Similarly, single nucleus analysis revealed a group of cells with high *PH4αEFB* but
344 low *Col4a1* (Figures. 3, 4, and S3). Consistent with this, the follicle cells of the egg chamber stop making
345 collagen IV at the end of stage 8 [53, 61, 62], and yet a subset of the follicle cells (border and centripetal
346 cells) continue to express *PH4αEFB* into stages 9 and 10, where it is required for their migration [63].
347 Likewise, *PH4αEFB* is expressed in the embryonic epidermis [36], where collagen IV expression is not
348 detected [37, 56]. These findings suggest that PH4αEFB may target other proteins in these cells.
349 Alternatively, some cells might have ceased collagen IV expression and degraded its mRNA, while
350 PH4αEFB mRNA persists after fulfilling its function. To explore these possibilities, future studies should

351 examine the temporal expression patterns of *PH4αEFB* and *Col4a1* at higher resolution and determine
352 PH4αEFB's full substrate spectrum.
353
354 In addition to collagen IV, *Drosophila* harbours several collagen(-related) molecules such as Pericardin and
355 Multiplexin [21, 22]. It is intriguing to examine whether PH4αEFB also modifies them as a 'pan *Drosophila*
356 collagen prolyl hydroxylase', and/or whether any other enzymes modify these proteins. Future experiments
357 exploring the correlation between the phylogeny and substrate specificities of the fly collagen P4Hα-related
358 proteins in Figure 1F will provide evolutionary insights into the diversification of the 'collagen molecular
359 ensemble'.
360
361 From an evolutionary perspective, it is also noteworthy that some molecules exhibit non-canonical domain
362 organisations. While typical collagen P4Hαs comprise the N, PSB, L, and CAT (catalytic) domains [30, 31,
363 42, 43] (Figure 1A), CG15539-PA possesses a truncated N-domain. Moreover, the three isoforms of
364 CG34041 all lack the catalytic domain: CG34041-PD contains one N-domain, whereas CG34041-PE and -
365 PF harbour two tandem N-domains (Figure 1A–E). Currently, the functions of these proteins are totally
366 unknown. Because the N-domain is the interface for the dimerisation of P4Hα proteins and the CAT domain
367 is used not only for the enzymatic activity but also for the interaction with PH4β/PDI [31], we speculate
368 that CG15539-PA may have lost the ability to dimerise while still being able to bind to PDI, potentially
369 altering its catalytic activity and/or substrate specificity. Regarding CG34041 isoforms, they may interact
370 with other P4Hα molecules via their N-domain(s) either homo- or heterotypically; the consequences of
371 such interactions are difficult to predict. Exploring the functions of these non-canonical collagen P4Hα-
372 related proteins would shed light on the diversification of regulatory mechanisms governing prolyl
373 hydroxylation reactions.
374
375 Our results may also provide some hints toward the search for non-collagen substrates of the collagen
376 P4Hα-related proteins. For example, 18 out of the 26 *Drosophila* collagen P4Hα-related genes are highly
377 expressed in the male accessory gland (red arrows in Figure 2A), which produces seminal fluid components
378 and is analogous to the human prostate [64]. Interestingly, a seminal fluid component 'Sex Peptide (SP)' is
379 known to be 4-prolyl hydroxylated. SP is transferred to the female during copulation and elicits a wide
380 range of post-mating changes in female physiology and behaviour, such as rejection of further mating,
381 increased food intake, enhanced oviposition, and the augmentation of immune response [65-71]. Among
382 these changes, at least that of the immune response is mediated by the central region of SP containing 4Hyps
383 [67]. Therefore, the accessory gland P4Hαs may include the hitherto unidentified enzyme(s) responsible
384 for the 4-prolyl hydroxylation of SP. Indeed, in the phylogenetic tree, all 18 accessory gland P4Hαs (Figure
385 1F, blue and marked with daggers) fall outside the clade containing PH4αEFB (magenta) and the proteins
386 implicated in collagen modification earlier in this study and by others [41] (grey). Moreover, 16 of the
387 accessory gland P4Hα-like proteins cluster in the region most distant from the magenta and grey clades of
388 (potential) collagen P4Hαs (Figure 1F, bracket). This pattern may reflect differences in substrate specificity
389 between accessory gland P4Hα-related proteins and collagen P4Hs. The roles of the accessory gland P4Hα-
390 related proteins in reproduction represent intriguing targets of future research.
391
392 In summary, our research identifies PH4αEFB as the primary P4Hα for *Drosophila* collagen IV,
393 highlighting a remarkably simple enzymatic toolkit for collagen IV biosynthesis in *Drosophila,* involving
394 likely single isoforms of P4H, LH, and GLT25D. Notably, while mammals require additional modifications
395 such as proline 3-hydroxylation, which plays critical roles in the interactions between other ECM proteins
396 with collagen IV and in basement membrane integrity [16, 72-75], *Drosophila* collagen IV biosynthesis
397 proceeds without them, underscoring its minimalistic nature. This minimal *Drosophila* collagen IV
398 biosynthetic machinery offers an exciting avenue for future research. A key direction is to explore whether
399 this simplified system can successfully produce mammalian collagens, especially collagen IV. Conversely,
400 investigating if mammalian machinery can accommodate *Drosophila* collagen IV will reveal crucial

insights into the compatibility of these systems. This comparative approach could help clarify which factors, collagen sequences, ER molecular compositions, or both, are critical for optimising *in vitro* collagen biosynthesis. Ultimately, such studies could define the minimal requirements for collagen engineering, a field with significant biomedical and biomaterial implications.

Experimental procedures

Identification of Drosophila P4Hα-related genes

In the 'QuickSearch' menu at the top page of FlyBase (http://flybase.org/), gene ontology ('GO') tab was clicked, 'molecular function' was selected from the 'Data field' pull-down menu, and the keyword 'procollagen-proline 4-dioxygenase activity' was typed into the 'Enter term' window. Subsequently, the 'Search' button was pressed, to obtain a single match CV (controlled vocabulary) term 'procollagen-proline 4-dioxygenase activity'. The link on this term was clicked to access its CV Term Report page, which shows that there are 26 genes annotated with this term (Table 1).

Phylogenetic analysis

Protein sequence comparison and phylogenetic analysis were conducted using Clustal Omega [76]. SwissProt IDs for the annotated protein products of the 26 *Drosophila* P4Hα-related genes were obtained from FlyBase (Table 1; URLs follow the format: 'http://flybase.org/reports/' + 'FlyBase ID'). For genes annotated with multiple polypeptides, the isoform listed first on the FlyBase page was used in the initial analysis. The same procedure was applied to the *Drosophila* HIF hydroxylase Hph (FlyBase ID: FBgn0264785; SwissProt ID: Q9VN98). SwissProt IDs for the human proteins included in the analysis were as follows: PH4A1 (P13674), PH4A2 (O15460), PH4A3 (Q7Z4N8), PHD1 (Q9GZT9), PHD2 (Q9H6Z9), and PHD3 (Q96KS0).

Preliminary analysis of the 33 proteins above revealed non-canonical domain structures in the protein products of CG15539 and CG34041. To examine whether other annotated products of these genes possess the canonical 'N-PSB-L-CAT' structure, a second Clustal Omega analysis was performed including all annotated isoforms for CG15539 and CG34041 (Table 1). For clarity, P4HA1, P4HA3, PHD1, PHD2, and Hph were excluded from the analysis. The resulting multiple sequence alignment was visualised in Table S1 according to this website (https://ouchidekaiseki.com/align.php - note that this website is written in Japanese). Briefly, the file '*Alignment in CLUSTAL format with base/residue numbering'* was downloaded from the 'Results Files' tab of Clustal Omega and opened in TextEdit (Apple). The contents were copied into Microsoft Excel using the 'Text Import Wizard'. In the first window of the Wizard, 'Fixed width' was selected; in the next, break lines were inserted between residues manually. Upon completion, each amino acid was placed in a separate cell. The cell dimensions were adjusted to form squares, and amino acids were colour-coded using Excel's 'Conditional Formatting'. Sequences separated to multiple rows were consolidated into single lines. Human P4HA2 was repositioned to the top of the table, and its domain structure was annotated. Additional formatting (e.g., row numbering, highlights, and colour codes) was applied to enhance clarity.

To construct the phylogenetic tree shown in Figure 1F, a third Clustal Omega run was performed, this time including P4HA1, P4HA3, PHD1, PHD2, and Hph. For clarity, only one protein product with a canonical domain structure was analysed for each *Drosophila* P4Hα-related gene. Specifically, all CG34041 isoforms were excluded, and only the PB isoform of CG15539 was included. The resulting dendrogram was downloaded in SVG format and edited using Adobe Illustrator and Microsoft PowerPoint, preserving branch topology and relative branch lengths.

Obtaining microarray/RNA-seq data

For each P4Hα-related gene (Table 1), *Col4a1* (FlyBase ID FBgn0000299) and *vkg* (FBgn0016075), gene report was obtained on FlyBase (http://flybase.org/reports/FBID, where 'FBID' is the FlyBase ID of the

gene [Table 1]). In the Expression Data > High-Throughput Expression Data section, the following five datasets were opened:

1. FlyAtlas Anatomy Microarray [49]
2. FlyAtlas2 Anatomy RNA-Seq [50]
3. modENCODE Anatomy RNA-Seq [48]
4. modENCODE Development RNA-Seq [55]
5. modENCODE Cell Lines RNA-Seq [48]
6. modENCODE Treatments RNA-Seq [48]

Subsequently, gene expression data were downloaded from the 'download data (TSV)' links. From the obtained files, the Mean Affy2 Probeset Expression Values (for FlyAtlas Anatomy Microarray) and RPKM values (for the others) were summarised and analysed in Supplementary Tables 1 (spatial data, 1-3, 5 and 6 above) and 2 (temporal data, 4).

Brightness coding of gene expression patterns

Normalised gene expression levels (values marked cyan in Supplementary Tables YMS1 and 2) were saved as text files and opened as images with ImageJ/Fiji, using File > Import > 'Text Image…' command.

Single-cell data acquisition and analysis

Drosophila melanogaster single-nucleus RNA-sequencing (snRNA-seq) datasets for fat body, ovaries and whole body were obtained from the Fly Cell Atlas [52]. Loom files and H5AD files were integrated and converted to RDS format using SCANPY [77]. Quality control and subsequent analysis were implemented in the R package Seurat [78]. Specifically, we retained cells that expressed a minimum of 200 and a maximum of 2000-3500 genes, depending on tissue type. We then normalised the gene expression measurement for each cell by the total expression, multiplied this result by 10,000 and log-transformed the result. The 2000 most variable genes were selected to aid in detecting biological signal in downstream analysis. These were used for linear transformation of the data (scaling), followed by principal components analysis, retaining the first 50 components. Unsupervised clustering of cells was conducted using the standard Seurat pipeline. This constructs a K-nearest neighbour graph (KNN) based on the Euclidean distance in the PCA space. Clusters were defined using the Louvain algorithm with a resolution of 0.5 and visualised using the non-linear dimensional reduction technique uniform manifold approximation and projection (UMAP). To visualise co-expression, we overlaid expression of *Col4a1*, *vkg*, and *PH4αEFB* on tissue-specific UMAPs.

For the other analyses, we processed the data as described below to remove noise and gain clearer information. First, we limited the investigation to genes expressed in at least 5% of cells. To reduce the sparsity of the data, we aggregated neighbouring cells with similar expression levels using the MetacellsByGroups function in the R package hdWGCNA (parameters k=25 and max_shared=10, reduction = pca) [54]. We normalised expression values in the metacell expression matrix for downstream analysis. For each cell type in each tissue, we then set the expression matrix. For each cell type, the soft power threshold was automatically selected using the lowest power that meets 0.8 scale-free topology fit [54]. Co-expression networks were then constructed using default parameters. We then computed module eigengenes as well as eigengene connectivity to define co-expression structure. Obtained data were plotted and used to compute Pearson correlations either for the entire dataset or within each cell type[79].
To identify which candidate gene(s) were consistently co-expressed with the two known collagen subunits (*Col4a1* and *vkg*), we performed a high-dimensional weighted correlation network analysis in the R package hdWGCNA [54]. We first identified co-expression modules that included both *Col4a1* and *vkg*. We reasoned that if our gene was involved in the proline 4-hydroxylation of collagen IV, it would co-occur within those modules. For each tissue, we counted the number of times each candidate gene was co-expressed in the same module as *Col4a1*, and *vkg*. We then assessed statistical significance using a Monte Carlo permutation test. Module assignments were randomised across 10,000 replicates. We then selected a

503 single gene at random and computed the number of matches for those modules. The resulting p-values are
504 the proportion of randomisations where the number of matched module assignments was equal to or
505 exceeded the empirical value.
506
507 ***Drosophila D17 Cell Culture*** – A detailed protocol for the culture of *Drosophila* D17 cells is presented in
508 the following reference [80].
509
510 ***Mouse PFHR9 Cell Culture*** – PFHR9 cells were cultured following the referenced protocol [9]. After
511 reaching 80-90 % confluency, to stimulate procollagen biosynthesis, ascorbic acid phosphate (100 µg/ml;
512 Wako Chemicals, 013-12061) was supplemented in Dulbecco's modified Eagle's medium (DMEM)/high
513 glucose/pyruvate (Gibco, 11995065) containing 10% (v/v) foetal bovine serum (R&D systems, S11150),
514 penicillin streptomycin glutamine 100× (Gibco, 10378016), and 5 mM Hepes for 1 day. The medium was
515 replaced to the fresh DMEM with ascorbic acid, and the cells were cultured for 2 days. After washing with
516 PBS twice, the cells were scraped and transferred into 15 mL falcon tube. The cell pellets were collected
517 by centrifugation with 4000 rpm using TX-400 rotor (Thermo scientific) for 5 min and stored in -20 C.
518
519 ***Gelatine coupled beads pull-down assay*** **–** D17 and PFHR9 cell pellets (0.5 g each) were lysed with 7.0
520 mL of pre-cooled M-PER (Thermo Scientific, 78501) containing Halt Protease Inhibitor Cocktail, EDTA-
521 free (Thermo Scientific, 87785) at 4 °C. Following the manufacturer's instructions, soluble proteins in the
522 supernatant were incubated with 1.5 mL of gelatine coupled beads, gelatine Sepharose 4B (Cytiva,
523 17095603), for 2 hours at 4 °C. After washing twice with 10 mL TBS buffer followed by one wash with 25
524 mM Tris buffer, pH 7.4, containing 1 M NaCl and two additional washes with TBS buffer, the proteins
525 tightly bound to gelatine sepharose beads were eluted with 2X SDS-PAGE sample buffer containing DTT
526 by heating to 95°C for 5 minutes. The eluted proteins were separated on a Bolt 4-12 % Bis-Tris Plus gel
527 (Thermo Scientific) using MES running buffer (Thermo Scientific, B0002). Gels were stained with
528 GelCode Blue Stain Reagent (Thermo Scientific, 24592) and SilverXpress Silver Staining Kit (Thermo
529 Scientific, LC6100). The stained gel images were taken by ChemiDoc MP imaging system (Bio-Rad) using
530 the software Image Lab version 4.0.1 (Bio-Rad). Individual gel bands were carefully excised and analysed
531 by protein identification LC-MS performed by the peptide core facility in the research Department of
532 Shriners Hospitals in Portland OR.
533
534 **Data processing, statistics, and presentation**
535 ImageJ/Fiji, R (public domain software), TextEdit (Apple), Excel, PowerPoint (Microsoft), Prism
536 (GraphPad), Illustrator (Adobe), and Inkscape (a free and open-source vector graphics editor) were used.
537

Acknowledgements
Computational analyses of single-nucleus transcriptome data were performed on the high performance computer (HPC) at Bournemouth University, the HPC at Institute of Science and Technology Austria, and the high-performance computational resources provided by the Louisiana Optical Network Infrastructure (http://www.loni.org). The authors are grateful to the researchers who published the transcriptome datasets [48, 49, 52, 55] that became the essential bases for this study, to FlyBase for curating the datasets in an easily accessible format, and the Drosophila Genomics Resource Center (DGRC), supported by NIH grant 2P40OD010949, for providing the D17 cell line used in this research. The authors thank Kristian Koski (University of Oulu, Finland) for crucial advice on the domain structure of collagen P4Hαs, and Ryusuke Niwa and Ryo Hoshino (University of Tsukuba, Japan) for helpful discussions on SP.

Author contributions
YI and YM were responsible for the overall design of the study. YI, MAT, ME, ALZ, and YM conducted and analysed all experiments, and also prepared the figures. YI, MAT, ME, and YM wrote the original draft. All authors prepared essential materials, reviewed and discussed the results, and were involved in editing the manuscript.

Additional information
Competing Interests: The authors declare that they have no competing interests related to this work.

Funding information
This project was supported by the All May See Foundation 7031182 to YI, American Heart Association 16POST2726018 and American Cancer Society 132123-PF-18-025-01-CSM postdoctoral fellowships to ALZ, National Institutes of Health R01 GM136961 and R35 GM148485 to SH-B, and the Academy of Medical Sciences/the Wellcome Trust/ the Government Department of Business, Energy and Industrial Strategy/the British Heart Foundation/Diabetes UK Springboard Award SBF008\1115 to YM.

565 **References**
566
567 [1] K. Tarnutzer, D. Siva Sankar, J. Dengjel, C.Y. Ewald, Collagen constitutes about 12% in females and
568 17% in males of the total protein in mice, Sci Rep 13(1) (2023) 4490.
569 [2] A. Naba, Mechanisms of assembly and remodelling of the extracellular matrix, Nat Rev Mol Cell Biol
570 (2024).
571 [3] M.D. Shoulders, R.T. Raines, Collagen structure and stability, Annu Rev Biochem 78 (2009) 929-58.
572 [4] M.S. Osidak, E.O. Osidak, M.A. Akhmanova, S.P. Domogatsky, A.S. Domogatskaya, Fibrillar, fibril-
573 associated and basement membrane collagens of the arterial wall: architecture, elasticity and remodeling
574 under stress, Curr Pharm Des 21(9) (2015) 1124-33.
575 [5] J. Bella, D.J. Hulmes, Fibrillar Collagens, Subcell Biochem 82 (2017) 457-490.
576 [6] H.P. Bächinger, K. Mizuno, J.A. Vranka, S.P. Boudko, 5.16 - Collagen Formation and Structure, in: H.-
577 W. Liu, L. Mander (Eds.), Comprehensive Natural Products II, Elsevier, Oxford, 2010, pp. 469-530.
578 [7] Y. Ishikawa, H.P. Bachinger, A substrate preference for the rough endoplasmic reticulum resident
579 protein FKBP22 during collagen biosynthesis, J Biol Chem 289(26) (2014) 18189-201.
580 [8] Y. Ishikawa, Y. Taga, K. Zientek, N. Mizuno, A.M. Salo, O. Semenova, S.F. Tufa, D.R. Keene, P.
581 Holden, K. Mizuno, D.B. Gould, J. Myllyharju, H.P. Bachinger, Type I and type V procollagen triple helix
582 uses different subsets of the molecular ensemble for lysine posttranslational modifications in the rER, J
583 Biol Chem 296 (2021) 100453.
584 [9] Y. Ishikawa, Y. Taga, T. Coste, S.F. Tufa, D.R. Keene, K. Mizuno, E. Tournier-Lasserve, D.B. Gould,
585 Lysyl hydroxylase 3-mediated post-translational modifications are required for proper biosynthesis of
586 collagen alpha1alpha1alpha2(IV), J Biol Chem 298(12) (2022) 102713.
587 [10] S. Ito, K. Nagata, Quality Control of Procollagen in Cells, Annu Rev Biochem 90 (2021) 631-658.
588 [11] C. Onursal, E. Dick, I. Angelidis, H.B. Schiller, C.A. Staab-Weijnitz, Collagen Biosynthesis,
589 Processing, and Maturation in Lung Ageing, Front Med (Lausanne) 8 (2021) 593874.
590 [12] R.A. Gjaltema, R.A. Bank, Molecular insights into prolyl and lysyl hydroxylation of fibrillar collagens
591 in health and disease, Crit Rev Biochem Mol Biol 52(1) (2017) 74-95.
592 [13] T. Hennet, Collagen glycosylation, Curr Opin Struct Biol 56 (2019) 131-138.
593 [14] A.M. Salo, J. Myllyharju, Prolyl and lysyl hydroxylases in collagen synthesis, Exp Dermatol 30(1)
594 (2021) 38-49.
595 [15] D.M. Hudson, M. Weis, J. Rai, K.S. Joeng, M. Dimori, B.H. Lee, R. Morello, D.R. Eyre, P3h3-null
596 and Sc65-null Mice Phenocopy the Collagen Lysine Under-hydroxylation and Cross-linking Abnormality
597 of Ehlers-Danlos Syndrome Type VIA, J Biol Chem 292(9) (2017) 3877-3887.
598 [16] D.M. Hudson, K.S. Joeng, R. Werther, A. Rajagopal, M. Weis, B.H. Lee, D.R. Eyre, Post-
599 translationally abnormal collagens of prolyl 3-hydroxylase-2 null mice offer a pathobiological mechanism
600 for the high myopia linked to human LEPREL1 mutations, J Biol Chem 290(13) (2015) 8613-22.
601 [17] Y. Ishikawa, Collagen Biosynthesis and Its Molecular Ensemble: What Remains Unexplored,
602 Biochemistry (2025).
603 [18] Y. Ishikawa, R. Lennon, F. Forneris, J. Myllyharju, A.M. Salo, Collagen IV Biosynthesis: Intracellular
604 Choreography of Post-Translational Modifications, Matrix Biol (2025).
605 [19] M.D. Adams, S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer,
606 P.W. Li, R.A. Hoskins, R.F. Galle, R.A. George, S.E. Lewis, S. Richards, M. Ashburner, S.N. Henderson,
607 G.G. Sutton, J.R. Wortman, M.D. Yandell, Q. Zhang, L.X. Chen, R.C. Brandon, Y.H. Rogers, R.G. Blazej,
608 M. Champe, B.D. Pfeiffer, K.H. Wan, C. Doyle, E.G. Baxter, G. Helt, C.R. Nelson, G.L. Gabor, J.F. Abril,
609 A. Agbayani, H.J. An, C. Andrews-Pfannkoch, D. Baldwin, R.M. Ballew, A. Basu, J. Baxendale, L.
610 Bayraktaroglu, E.M. Beasley, K.Y. Beeson, P.V. Benos, B.P. Berman, D. Bhandari, S. Bolshakov, D.
611 Borkova, M.R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K.C. Burtis, D.A. Busam, H. Butler, E. Cadieu,
612 A. Center, I. Chandra, J.M. Cherry, S. Cawley, C. Dahlke, L.B. Davenport, P. Davies, B. de Pablos, A.
613 Delcher, Z. Deng, A.D. Mays, I. Dew, S.M. Dietz, K. Dodson, L.E. Doup, M. Downes, S. Dugan-Rocha,
614 B.C. Dunkov, P. Dunn, K.J. Durbin, C.C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler,
615 A.E. Gabrielian, N.S. Garg, W.M. Gelbart, K. Glasser, A. Glodek, F. Gong, J.H. Gorrell, Z. Gu, P. Guan,

M. Harris, N.L. Harris, D. Harvey, T.J. Heiman, J.R. Hernandez, J. Houck, D. Hostin, K.A. Houston, T.J. Howland, M.H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G.H. Karpen, Z. Ke, J.A. Kennison, K.A. Ketchum, B.E. Kimmel, C.D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A.A. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T.C. McIntosh, M.P. McLeod, D. McPherson, G. Merkulov, N.V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S.M. Mount, M. Moy, B. Murphy, L. Murphy, D.M. Muzny, D.L. Nelson, D.R. Nelson, K.A. Nelson, K. Nixon, D.R. Nusskern, J.M. Pacleb, M. Palazzolo, G.S. Pittman, S. Pan, J. Pollard, V. Puri, M.G. Reese, K. Reinert, K. Remington, R.D. Saunders, F. Scheeler, H. Shen, B.C. Shue, I. Siden-Kiamos, M. Simpson, M.P. Skupski, T. Smith, E. Spier, A.C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A.H. Wang, X. Wang, Z.Y. Wang, D.A. Wassarman, G.M. Weinstock, J. Weissenbach, S.M. Williams, WoodageT, K.C. Worley, D. Wu, S. Yang, Q.A. Yao, J. Ye, R.F. Yeh, J.S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X.H. Zheng, F.N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H.O. Smith, R.A. Gibbs, E.W. Myers, G.M. Rubin, J.C. Venter, The genome sequence of Drosophila melanogaster, Science 287(5461) (2000) 2185-95.

[20] R.O. Hynes, The evolution of metazoan extracellular matrix, J Cell Biol 196(6) (2012) 671-9.

[21] A. Chartier, S. Zaffran, M. Astier, M. Semeriva, D. Gratecos, Pericardin, a Drosophila type IV collagen-like protein is involved in the morphogenesis and maintenance of the heart epithelium during dorsal ectoderm closure, Development 129(13) (2002) 3241-53.

[22] F. Meyer, B. Moussian, Drosophila multiplexin (Dmp) modulates motor axon pathfinding accuracy, Dev Growth Differ 51(5) (2009) 483-98.

[23] J.E. Natzle, J.M. Monson, B.J. McCarthy, Cytogenetic location and expression of collagen-like genes in Drosophila, Nature 296(5855) (1982) 368-71.

[24] S. Yasothornsrikul, W.J. Davis, G. Cramer, D.A. Kimbrell, C.R. Dearolf, viking: identification and characterization of a second type IV collagen in Drosophila, Gene 198(1-2) (1997) 17-25.

[25] P. Rappu, A.M. Salo, J. Myllyharju, J. Heino, Role of prolyl hydroxylation in the molecular interactions of collagens, Essays Biochem 63(3) (2019) 325-335.

[26] J. Myllyharju, P. Koivunen, Hypoxia-inducible factor prolyl 4-hydroxylases: common and specific roles, Biol Chem 394(4) (2013) 435-48.

[27] P. Koivunen, P. Tiainen, J. Hyvarinen, K.E. Williams, R. Sormunen, S.J. Klaus, K.I. Kivirikko, J. Myllyharju, An endoplasmic reticulum transmembrane prolyl 4-hydroxylase is induced by hypoxia and acts on hypoxia-inducible factor alpha, J Biol Chem 282(42) (2007) 30544-52.

[28] T. Ala-Nisula, R. Halmetoja, H. Leinonen, M. Kurkela, H.R. Lipponen, S. Sakko, M. Karpale, A.M. Salo, N. Sissala, T. Roning, G.S. Raza, K.A. Makela, J. Thevenot, K.H. Herzig, R. Serpi, J. Myllyharju, H. Tanila, P. Koivunen, E.Y. Dimova, Metabolic characteristics of transmembrane prolyl 4-hydroxylase (P4H-TM) deficient mice, Pflugers Arch 476(9) (2024) 1339-1351.

[29] J. Myllyharju, K.I. Kivirikko, Collagens, modifying enzymes and their mutations in humans, flies and worms, Trends Genet 20(1) (2004) 33-43.

[30] M.K. Koski, J. Anantharajan, P. Kursula, P. Dhavala, A.V. Murthy, U. Bergmann, J. Myllyharju, R.K. Wierenga, Assembly of the elongated collagen prolyl 4-hydroxylase alpha(2)beta(2) heterotetramer around a central alpha(2) dimer, Biochem J 474(5) (2017) 751-769.

[31] A.V. Murthy, R. Sulu, A. Lebedev, A.M. Salo, K. Korhonen, R. Venkatesan, H. Tu, U. Bergmann, J. Janis, M. Laitaoja, L.W. Ruddock, J. Myllyharju, M.K. Koski, R.K. Wierenga, Crystal structure of the collagen prolyl 4-hydroxylase (C-P4H) catalytic domain complexed with PDI: Toward a model of the C-P4H alpha(2)beta(2) tetramer, J Biol Chem 298(12) (2022) 102614.

[32] 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) 115-20.

[33] S. Jimenez, M. Harsch, J. Rosenbloom, Hydroxyproline stabilizes the triple helix of chick tendon collagen, Biochem Biophys Res Commun 52(1) (1973) 106-14.

[34] J. Rosenbloom, M. Harsch, S. Jimenez, Hydroxyproline content determines the denaturation temperature of chick tendon collagen, Arch Biochem Biophys 158(2) (1973) 478-84.

666 [35] T. Holster, O. Pakkanen, R. Soininen, R. Sormunen, M. Nokelainen, K.I. Kivirikko, J. Myllyharju,
667 Loss of assembly of the main basement membrane collagen, type IV, but not fibril-forming collagens and
668 embryonic death in collagen prolyl 4-hydroxylase I null mice, J Biol Chem 282(4) (2007) 2512-9.
669 [36] E.W. Abrams, D.J. Andrew, Prolyl 4-hydroxylase alpha-related proteins in Drosophila melanogaster:
670 tissue-specific embryonic expression of the 99F8-9 cluster, Mech Dev 112(1-2) (2002) 165-71.
671 [37] C. Mirre, J.P. Cecchini, Y. Le Parco, B. Knibiehler, De novo expression of a type IV collagen gene in
672 Drosophila embryos is restricted to mesodermal derivatives and occurs at germ band shortening,
673 Development 102(2) (1988) 369-76.
674 [38] A.R. Walmsley, M.R. Batten, U. Lad, N.J. Bulleid, Intracellular retention of procollagen within the
675 endoplasmic reticulum is mediated by prolyl 4-hydroxylase, J Biol Chem 274(21) (1999) 14884-92.
676 [39] J.C. Pastor-Pareja, T. Xu, Shaping cells and organs in Drosophila by opposing roles of fat body-
677 secreted Collagen IV and perlecan, Dev Cell 21(2) (2011) 245-56.
678 [40] D.W. Lerner, D. McCoy, A.J. Isabella, A.P. Mahowald, G.F. Gerlach, T.A. Chaudhry, S. Horne-
679 Badovinac, A Rab10-dependent mechanism for polarized basement membrane secretion during organ
680 morphogenesis, Dev Cell 24(2) (2013) 159-68.
681 [41] P. Annunen, P. Koivunen, K.I. Kivirikko, Cloning of the alpha subunit of prolyl 4-hydroxylase from
682 Drosophila and expression and characterization of the corresponding enzyme tetramer with some unique
683 properties, J Biol Chem 274(10) (1999) 6790-6.
684 [42] J. Myllyharju, K.I. Kivirikko, Identification of a novel proline-rich peptide-binding domain in prolyl
685 4-hydroxylase, EMBO J 18(2) (1999) 306-12.
686 [43] J. Anantharajan, M.K. Koski, P. Kursula, R. Hieta, U. Bergmann, J. Myllyharju, R.K. Wierenga, The
687 structural motifs for substrate binding and dimerization of the alpha subunit of collagen prolyl 4-
688 hydroxylase, Structure 21(12) (2013) 2107-18.
689 [44] D.C. John, M.E. Grant, N.J. Bulleid, Cell-free synthesis and assembly of prolyl 4-hydroxylase: the role
690 of the beta-subunit (PDI) in preventing misfolding and aggregation of the alpha-subunit, EMBO J 12(4)
691 (1993) 1587-95.
692 [45] P. Annunen, T. Helaakoski, J. Myllyharju, J. Veijola, T. Pihlajaniemi, K.I. Kivirikko, Cloning of the
693 human prolyl 4-hydroxylase alpha subunit isoform alpha(II) and characterization of the type II enzyme
694 tetramer. The alpha(I) and alpha(II) subunits do not form a mixed alpha(I)alpha(II)beta2 tetramer, J Biol
695 Chem 272(28) (1997) 17342-8.
696 [46] J. Myllyharju, Prolyl 4-hydroxylases, key enzymes in the synthesis of collagens and regulation of the
697 response to hypoxia, and their roles as treatment targets, Ann Med 40(6) (2008) 402-17.
698 [47] J.M. Acevedo, L. Centanin, A. Dekanty, P. Wappner, Oxygen sensing in Drosophila: multiple isoforms
699 of the prolyl hydroxylase fatiga have different capacity to regulate HIFalpha/Sima, PLoS One 5(8) (2010)
700 e12390.
701 [48] J.B. Brown, N. Boley, R. Eisman, G.E. May, M.H. Stoiber, M.O. Duff, B.W. Booth, J. Wen, S. Park,
702 A.M. Suzuki, K.H. Wan, C. Yu, D. Zhang, J.W. Carlson, L. Cherbas, B.D. Eads, D. Miller, K. Mockaitis,
703 J. Roberts, C.A. Davis, E. Frise, A.S. Hammonds, S. Olson, S. Shenker, D. Sturgill, A.A. Samsonova, R.
704 Weiszmann, G. Robinson, J. Hernandez, J. Andrews, P.J. Bickel, P. Carninci, P. Cherbas, T.R. Gingeras,
705 R.A. Hoskins, T.C. Kaufman, E.C. Lai, B. Oliver, N. Perrimon, B.R. Graveley, S.E. Celniker, Diversity
706 and dynamics of the Drosophila transcriptome, Nature 512(7515) (2014) 393-9.
707 [49] V.R. Chintapalli, J. Wang, J.A. Dow, Using FlyAtlas to identify better Drosophila melanogaster
708 models of human disease, Nat Genet 39(6) (2007) 715-20.
709 [50] D.P. Leader, S.A. Krause, A. Pandit, S.A. Davies, J.A.T. Dow, FlyAtlas 2: a new version of the
710 Drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data, Nucleic Acids
711 Res 46(D1) (2018) D809-D815.
712 [51] A.L. Fidler, C.E. Darris, S.V. Chetyrkin, V.K. Pedchenko, S.P. Boudko, K.L. Brown, W. Gray Jerome,
713 J.K. Hudson, A. Rokas, B.G. Hudson, Collagen IV and basement membrane at the evolutionary dawn of
714 metazoan tissues, Elife 6 (2017).
715 [52] H. Li, J. Janssens, M. De Waegeneer, S.S. Kolluru, K. Davie, V. Gardeux, W. Saelens, F.P.A. David,
716 M. Brbic, K. Spanier, J. Leskovec, C.N. McLaughlin, Q. Xie, R.C. Jones, K. Brueckner, J. Shim, S.G.

Tattikota, F. Schnorrer, K. Rust, T.G. Nystul, Z. Carvalho-Santos, C. Ribeiro, S. Pal, S. Mahadevaraju, T.M. Przytycka, A.M. Allen, S.F. Goodwin, C.W. Berry, M.T. Fuller, H. White-Cooper, E.L. Matunis, S. DiNardo, A. Galenza, L.E. O'Brien, J.A.T. Dow, F.C.A.C.s. sign, H. Jasper, B. Oliver, N. Perrimon, B. Deplancke, S.R. Quake, L. Luo, S. Aerts, D. Agarwal, Y. Ahmed-Braimah, M. Arbeitman, M.M. Ariss, J. Augsburger, K. Ayush, C.C. Baker, T. Banisch, K. Birker, R. Bodmer, B. Bolival, S.E. Brantley, J.A. Brill, N.C. Brown, N.A. Buehner, X.T. Cai, R. Cardoso-Figueiredo, F. Casares, A. Chang, T.R. Clandinin, S. Crasta, C. Desplan, A.M. Detweiler, D.B. Dhakan, E. Dona, S. Engert, S. Floc'hlay, N. George, A.J. Gonzalez-Segarra, A.K. Groves, S. Gumbin, Y. Guo, D.E. Harris, Y. Heifetz, S.L. Holtz, F. Horns, B. Hudry, R.J. Hung, Y.N. Jan, J.S. Jaszczak, G. Jefferis, J. Karkanias, T.L. Karr, N.S. Katheder, J. Kezos, A.A. Kim, S.K. Kim, L. Kockel, N. Konstantinides, T.B. Kornberg, H.M. Krause, A.T. Labott, M. Laturney, R. Lehmann, S. Leinwand, J. Li, J.S.S. Li, K. Li, K. Li, L. Li, T. Li, M. Litovchenko, H.H. Liu, Y. Liu, T.C. Lu, J. Manning, A. Mase, M. Matera-Vatnick, N.R. Matias, C.E. McDonough-Goldstein, A. McGeever, A.D. McLachlan, P. Moreno-Roman, N. Neff, M. Neville, S. Ngo, T. Nielsen, C.E. O'Brien, D. Osumi-Sutherland, M.N. Ozel, I. Papatheodorou, M. Petkovic, C. Pilgrim, A.O. Pisco, C. Reisenman, E.N. Sanders, G. Dos Santos, K. Scott, A. Sherlekar, P. Shiu, D. Sims, R.V. Sit, M. Slaidina, H.E. Smith, G. Sterne, Y.H. Su, D. Sutton, M. Tamayo, M. Tan, I. Tastekin, C. Treiber, D. Vacek, G. Vogler, S. Waddell, W. Wang, R.I. Wilson, M.F. Wolfner, Y.E. Wong, A. Xie, J. Xu, S. Yamamoto, J. Yan, Z. Yao, K. Yoda, R. Zhu, R.P. Zinzen, Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly, Science 375(6584) (2022) eabk2432.

[53] S.L. Haigo, D. Bilder, Global tissue revolutions in a morphogenetic movement controlling elongation, Science 331(6020) (2011) 1071-4.

[54] S. Morabito, F. Reese, N. Rahimzadeh, E. Miyoshi, V. Swarup, hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data, Cell Rep Methods 3(6) (2023) 100498.

[55] B.R. Graveley, A.N. Brooks, J.W. Carlson, M.O. Duff, J.M. Landolin, L. Yang, C.G. Artieri, M.J. van Baren, N. Boley, B.W. Booth, J.B. Brown, L. Cherbas, C.A. Davis, A. Dobin, R. Li, W. Lin, J.H. Malone, N.R. Mattiuzzo, D. Miller, D. Sturgill, B.B. Tuch, C. Zaleski, D. Zhang, M. Blanchette, S. Dudoit, B. Eads, R.E. Green, A. Hammonds, L. Jiang, P. Kapranov, L. Langton, N. Perrimon, J.E. Sandler, K.H. Wan, A. Willingham, Y. Zhang, Y. Zou, J. Andrews, P.J. Bickel, S.E. Brenner, M.R. Brent, P. Cherbas, T.R. Gingeras, R.A. Hoskins, T.C. Kaufman, B. Oliver, S.E. Celniker, The developmental transcriptome of Drosophila melanogaster, Nature 471(7339) (2011) 473-9.

[56] J.M. Davis, B.A. Boswell, H.P. Bachinger, Thermal stability and folding of type IV procollagen and effect of peptidyl-prolyl cis-trans-isomerase on the folding of the triple helix, J Biol Chem 264(15) (1989) 8956-62.

[57] Y.K. Bae, F. Macabenta, H.L. Curtis, A. Stathopoulos, Comparative analysis of gene expression profiles for several migrating cell types identifies cell migration regulators, Mech Dev 148 (2017) 40-55.

[58] B. Stramer, M. Winfield, T. Shaw, T.H. Millard, S. Woolner, P. Martin, Gene induction following wounding of wild-type versus macrophage-deficient Drosophila embryos, EMBO Rep 9(5) (2008) 465-71.

[59] E.W. Abrams, W.K. Mihoulides, D.J. Andrew, Fork head and Sage maintain a uniform and patent salivary gland lumen through regulation of two downstream target genes, PH4alphaSG1 and PH4alphaSG2, Development 133(18) (2006) 3517-27.

[60] S.H. Payne, The utility of protein and mRNA correlation, Trends Biochem Sci 40(1) (2015) 1-3.

[61] B. Knibiehler, C. Mirre, Y. Le Parco, Collagen type IV of Drosophila is stockpiled in the growing oocyte and differentially located during early stages of embryogenesis, Cell Differ Dev 30(2) (1990) 147-57.

[62] A.J. Isabella, S. Horne-Badovinac, Dynamic regulation of basement membrane protein levels promotes egg chamber elongation in Drosophila, Dev Biol 406(2) (2015) 212-21.

[63] L. Manning, J. Sheth, S. Bridges, A. Saadin, K. Odinammadu, D. Andrew, S. Spencer, D. Montell, M. Starz-Gaiano, A hormonal cue promotes timely follicle cell migration by modulating transcription profiles, Mech Dev 148 (2017) 56-68.

[64] A. Rambur, M. Vialat, C. Beaudoin, C. Lours-Calet, J.M. Lobaccaro, S. Baron, L. Morel, C. de Joussineau, Drosophila Accessory Gland: A Complementary In Vivo Model to Bring New Insight to Prostate Cancer, Cells 10(9) (2021).
[65] E. Kubli, Sex-peptides: seminal peptides of the Drosophila male, Cell Mol Life Sci 60(8) (2003) 1689-704.
[66] J. Peng, S. Chen, S. Busser, H. Liu, T. Honegger, E. Kubli, Gradual release of sperm bound sex-peptide controls female postmating behavior in Drosophila, Curr Biol 15(3) (2005) 207-13.
[67] E.V. Domanitskaya, H. Liu, S. Chen, E. Kubli, The hydroxyproline motif of male sex peptide elicits the innate immune response in Drosophila females, FEBS J 274(21) (2007) 5659-68.
[68] P. Cognigni, A.P. Bailey, I. Miguel-Aliaga, Enteric neurons and systemic signals couple nutritional and reproductive status with intestinal homeostasis, Cell Metab 13(1) (2011) 92-104.
[69] T. Ameku, Y. Yoshinari, M.J. Texada, S. Kondo, K. Amezawa, G. Yoshizaki, Y. Shimada-Niwa, R. Niwa, Midgut-derived neuropeptide F controls germline stem cell proliferation in a mating-dependent manner, PLoS Biol 16(9) (2018) e2005004.
[70] S. Sturm, A. Dowle, N. Audsley, R.E. Isaac, The structure of the Drosophila melanogaster sex peptide: Identification of hydroxylated isoleucine and a strain variation in the pattern of amino acid hydroxylation, Insect Biochem Mol Biol 124 (2020) 103414.
[71] S. Sturm, A. Dowle, N. Audsley, R.E. Isaac, Mass spectrometric characterisation of the major peptides of the male ejaculatory duct, including a glycopeptide with an unusual zwitterionic glycosylation, J Proteomics 246 (2021) 104307.
[72] R.M. Gryder, M. Lamon, E. Adams, Sequence position of 3-hydroxyproline in basement membrane collagen. Isolation of glycyl-3-hydroxyprolyl-4-hydroxyproline from swine kidney, J Biol Chem 250(7) (1975) 2470-4.
[73] N.T. Montgomery, K.D. Zientek, E.N. Pokidysheva, H.P. Bachinger, Post-translational modification of type IV collagen with 3-hydroxyproline affects its interactions with glycoprotein VI and nidogens 1 and 2, J Biol Chem 293(16) (2018) 5987-5999.
[74] E. Pokidysheva, S. Boudko, J. Vranka, K. Zientek, K. Maddox, M. Moser, R. Fassler, J. Ware, H.P. Bachinger, Biological role of prolyl 3-hydroxylation in type IV collagen, Proc Natl Acad Sci U S A 111(1) (2014) 161-6.
[75] H. Aypek, C. Krisp, S. Lu, S. Liu, D. Kylies, O. Kretz, G. Wu, M. Moritz, K. Amann, K. Benz, P. Tong, Z.M. Hu, S.M. Alsulaiman, A.O. Khan, M. Grohmann, T. Wagner, J. Muller-Deile, H. Schluter, V.G. Puelles, C. Bergmann, T.B. Huber, F. Grahammer, Loss of the collagen IV modifier prolyl 3-hydroxylase 2 causes thin basement membrane nephropathy, J Clin Invest 132(9) (2022).
[76] F. Madeira, N. Madhusoodanan, J. Lee, A. Eusebi, A. Niewielska, A.R.N. Tivey, R. Lopez, S. Butcher, The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024, Nucleic Acids Res 52(W1) (2024) W521-W525.
[77] F.A. Wolf, P. Angerer, F.J. Theis, SCANPY: large-scale single-cell gene expression data analysis, Genome Biol 19(1) (2018) 15.
[78] Y. Hao, T. Stuart, M.H. Kowalski, S. Choudhary, P. Hoffman, A. Hartman, A. Srivastava, G. Molla, S. Madad, C. Fernandez-Granda, R. Satija, Dictionary learning for integrative, multimodal and scalable single-cell analysis, Nat Biotechnol 42(2) (2024) 293-304.
[79] M.T. Tang, How to do gene correlation for single-cell RNAseq data (part 2) using meta-cell, 2023. https://divingintogeneticsandgenomics.com/post/how-to-do-gene-correlation-for-single-cell-rnaseq-data-part-2-using-meta-cell/.
[80] J.D. Currie, S.L. Rogers, Using the Drosophila melanogaster D17-c3 cell culture system to study cell motility, Nat Protoc 6(10) (2011) 1632-41.

Gene Symbol	FlyBase ID	Analysed isoform(s)	SwissProt ID
CG11828	FBgn0039616	PD	Q9VAR7
CG15539	FBgn0039782	PA	Q4V443
		PB	Q0KHY7
		PC	A0A0B4LHW6
CG15864	FBgn0040528	PB	Q9VHU7
CG18231	FBgn0036796	PB	Q9VVQ7
CG18233	FBgn0036795	PB	Q9VVQ6
CG18234	FBgn0265268	PC	Q9VVQ5
CG18749	FBgn0042182	PB	Q8MSK0
CG31013	FBgn0051013	PA	Q9VA50
CG31016	FBgn0051016	PB	Q9VA52
CG31021	FBgn0051021	PB	Q8IMH8
CG31371	FBgn0051371	PB	Q8IMI4
CG31524	FBgn0051524	PA	Q8IMI2
CG32199	FBgn0052199	PB	Q9VVQ9
CG32201	FBgn0052201	PB	Q8IQS7
CG34041	FBgn0054041	PD	D3DMS2
		PE	A0A0B4KHI9
		PF	Q2PDP5
CG34345	FBgn0085374	PA	A8DYR4
CG4174	FBgn0036793	PD	Q9VVQ4
CG9698	FBgn0039784	PA	Q9VA60
PH4αEFB	FBgn0039776	PA	Q9VA69
PH4αMP	FBgn0026190	PA	Q9VA65
PH4αNE1	FBgn0039780	PA	Q9VA64
PH4αNE2	FBgn0039783	PA	Q9VA61
PH4αNE3	FBgn0051017	PA	Q961I8
PH4αPV	FBgn0051015	PA	Q8T5S8
PH4αSG1	FBgn0051014	PA	Q9VA63
PH4αSG2	FBgn0039779	PA	Q9I7H5

Table 1. *Drosophila* collagen P4Hα-related genes. Genes annotated with the GO (Molecular Function) term 'procollagen-proline 4-dioxygenase activity' were searched for on FlyBase and listed. For their protein products, the isoforms used for the phylogenetic analyses in this study are shown with their SwissProt IDs. Bold letters indicate the genes that have been analysed previously [36]. For all the genes, the GO term was recorded to have been inferred from 'electronic annotation with InterPro:IPR013547' and 'biological aspect of ancestor with PANTHER:PTN004202971'.

821
822
823 **Table 2. Legend for the sample numbers in the horizontal axes of each panel in Fig. 2A and C.**

Tissue	Number of modules (cell types) with co-expression between candidate and collagen subunits	Total number of modules (cell types) with co-expression of collagen subunits	p-value
Ovary	2	5	$p = 0.0351$
Fat Body	3	8	$p = 0.0039$
Whole Body	6	15	$p < 0.0001$

824
825 **Table 3. Statistical significance of co-occurrence of *PH4αEFB*, *Col4a1* and *vkg* in co-expression modules.**
826

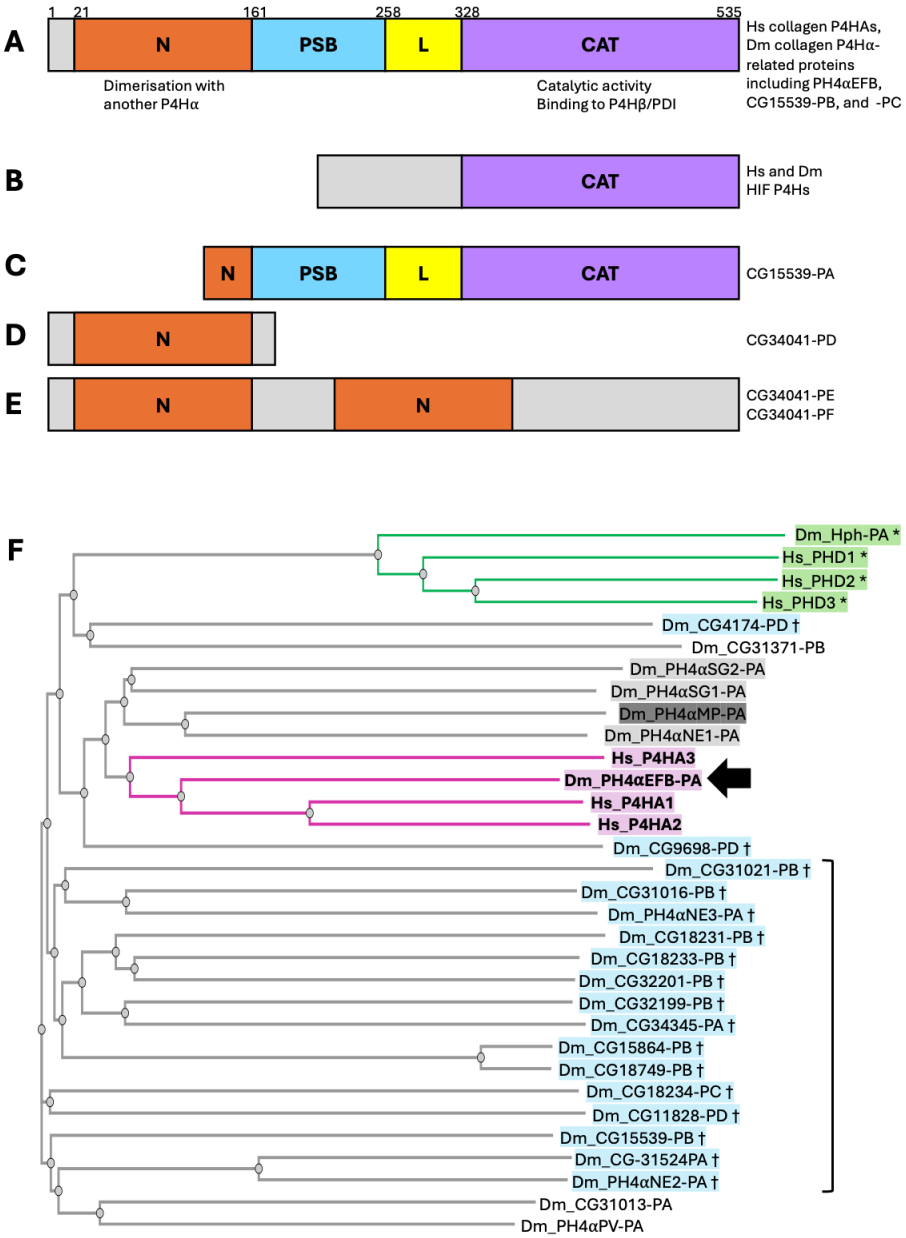


Figure. 1. Domain organisation and phylogeny of human (Hs) and *Drosophila* (Dm) prolyl 4-hydroxylases. **(A–E)** Schematic representations of the protein domain structures analysed. Diagrams are not drawn to scale. For full details, see Table S1.

(A) Canonical domain structure of collagen P4Hα proteins, comprising N-terminal (N), peptide-substrate-binding (PSB), linker (L), and catalytic (CAT) domains. This organisation is conserved in all three human collagen P4HAs and in proteins encoded by 25 of the 26 *Drosophila* P4Hα-related genes including those mentioned at the right. Numbers above the schematic indicate domain boundary positions for human P4HA2.

(B) Domain structure of HIF prolyl 4-hydroxylases (PHD1–3 and Hph). These are homologous to collagen P4HAs only within the catalytic domain.

(C) The CG15539-PA isoform has a truncated N-domain, while the other two isoforms (PB and PC) exhibit the full domain organization as in **(A)**.

839 **(D, E)** All three annotated isoforms of CG34041 (PD, PE, and PF) lack the catalytic domain. CG34041-PD contains
840 an N-domain, whereas CG34041-PE and -PF contain two tandem N-domains.
841 **(F)** Phylogenetic tree of P4Hαs. At the top of the tree, *Drosophila* Hph and human PHD1–3 (green, marked with
842 asterisks [*]) form a distinct clade. In the middle, *Drosophila* PH4αEFB (arrow) clusters with the three human
843 collagen prolyl 4-hydroxylases (P4HA1–3), forming a separate clade (magenta, highlighted in bold). PH4αMP
844 (dark grey shading) and neighbouring *Drosophila* enzymes (light grey shading) are implicated in tissue- or context-
845 specific collagen modification (see Discussion). Proteins shaded in blue and marked with daggers (†) are highly
846 expressed in the male accessory gland and may hydroxylate the seminal fluid protein Sex Peptide (SP). Sixteen of
847 these 18 accessory gland P4Hα-related proteins cluster in the most phylogenetically distant region from the magenta
848 and grey clades (bracket; see also Discussion).
849

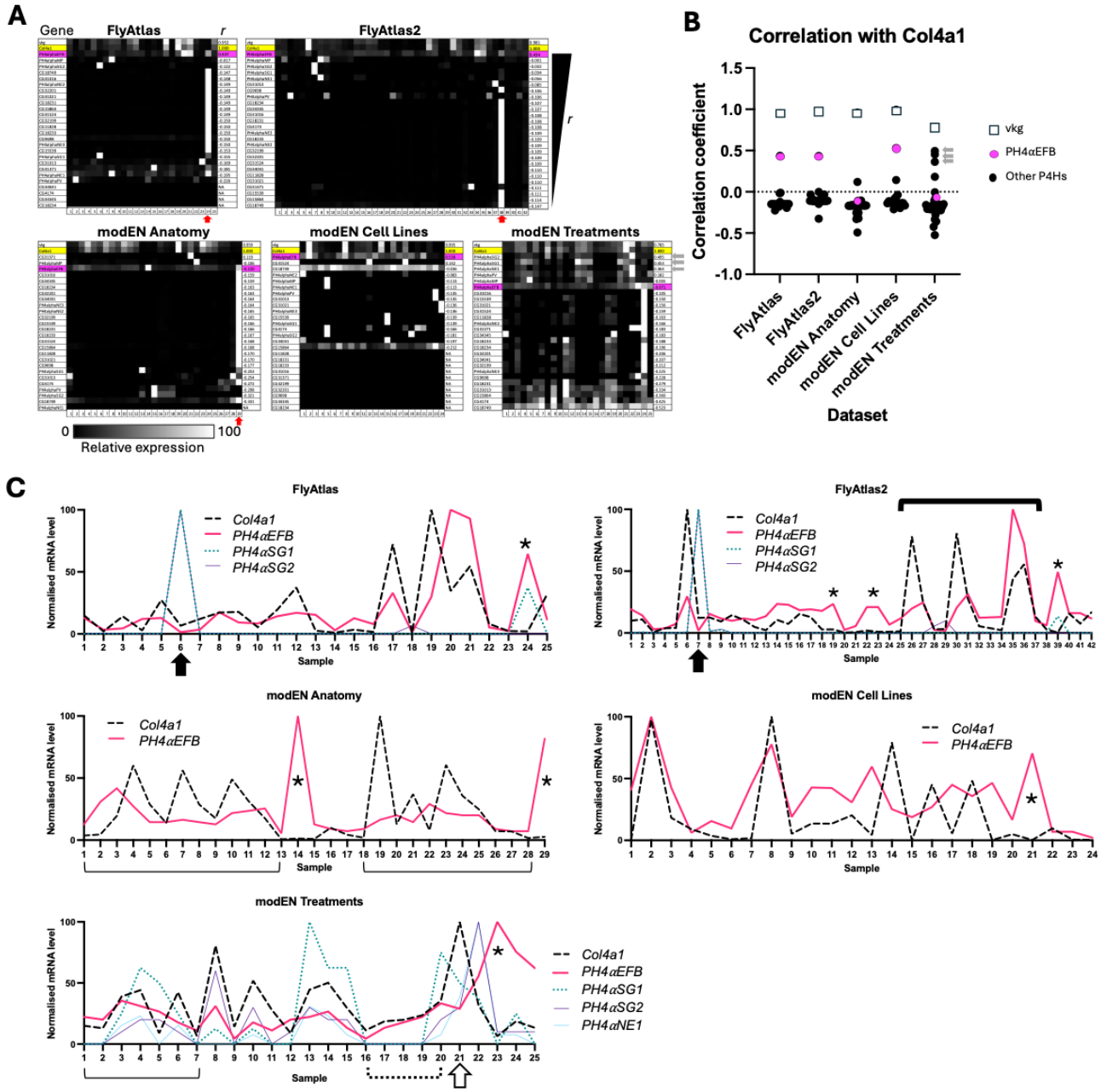


Figure. 2. Spatial expression patterns of *Drosophila* collagen IV and PH4α-related genes.
(A) The expression levels of each gene from the transcriptome dataset indicated are normalised for the maximum level and brightness-coded as in the bar at the bottom. modEN, modENCODE. In all panels, *vkg* is shown at the top. Yellow, *Col4a1*; magenta, *PH4αEFB*. The PH4α-related genes are sorted in the descending order of the values of the correlation coefficient *r* between the expression patterns of *Col4a1* and each of the other genes. For the legends for the sample numbers in **(A)** and **(C)**, see Table 2. For enlarged gene names and the values of gene expression levels, see Table S2, where the genes are sorted in the same order. Grey arrows beside the bottom right panel point the data with *PH4αSG2*, *SG1*, and *NE1*. Red arrows point accessory gland samples in which multiple PH4α-related genes are highly expressed.
(B) Plot of the *r* values for the genes in each dataset in **(A)**. Grey arrows point the data with *PH4αSG2*, *SG1*, and *NE1* from top to bottom, respectively.

862 **(C)** For *Col4a1*, *PH4αEFB, SG1, SG2,* and *NE1*, the values in **(A)** are displayed as line scattered plots. Thick solid
863 bracket, examples of the parallel peaks and troughs of the expression levels of *Col4a1* and *PH4αEFB*. Thin solid
864 brackets, examples of the cases where *Col4a1* and *PH4αEFB* levels change largely in parallel, although the graphs
865 do not exhibit identical numbers of peaks and troughs. Closed arrows, samples in which the expression levels of
866 *PH4αEFB* is low while *Col4a1* is detected. Asterisks, samples in which *PH4αEFB* level is high while *Col4a1* level
867 is low. Dotted bracket, parallel increase of *Col4a1* and *PH4αEFB* expression levels. Open arrow, between samples
868 20 and 21, while the expression level of *Col4a1* increases steeply, that of *PH4αEFB* slightly decreases. For detail,
869 see Discussion.
870

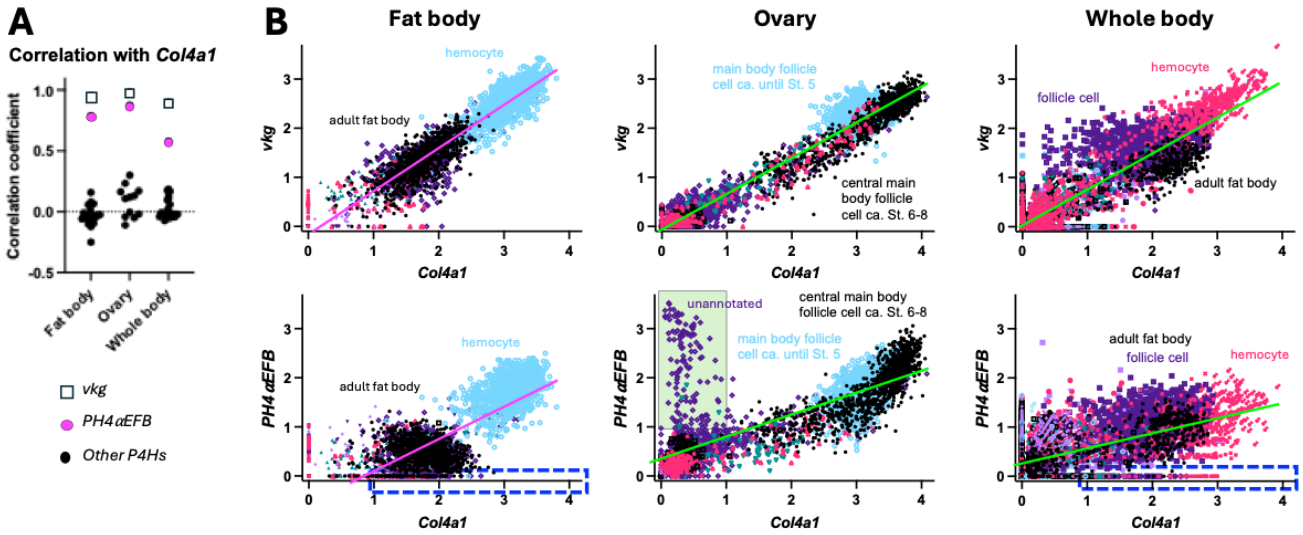


Figure. 3. Co-expression of *Col4a1* and *PH4αEFB* in single cells

(A) Plot of correlation coefficient *r* between *Col4a1* and *vkg* or each of the 26 PH4α-related genes in the indicated datasets.

(B) The expression levels of *Col4a1* vs. *vkg* (top) or *PH4αEFB* (bottom) in each single metacell were plotted for the indicated datasets. Annotations of representative cell types are shown. For full annotations, see Fig. S1. Regression lines for all the data points in each panel are shown; their colours are altered only for the sake of visibility. n = 3331 (fat body), 3408 (ovary), and 10186 (whole body). Blue dashed-edge rectangles, the area of high *Col4a1* (> 1) and low *PH4αEFB* (< 0.1) expression; green interior rectangle, the area of low *Col4a1* (< 1) and high *PH4αEFB* (> 1) expression. The latter rectangle contains 133 metacells, 129 of which are unannotated and the remaining 4 are stretch follicle cells (Cf. Table S3).

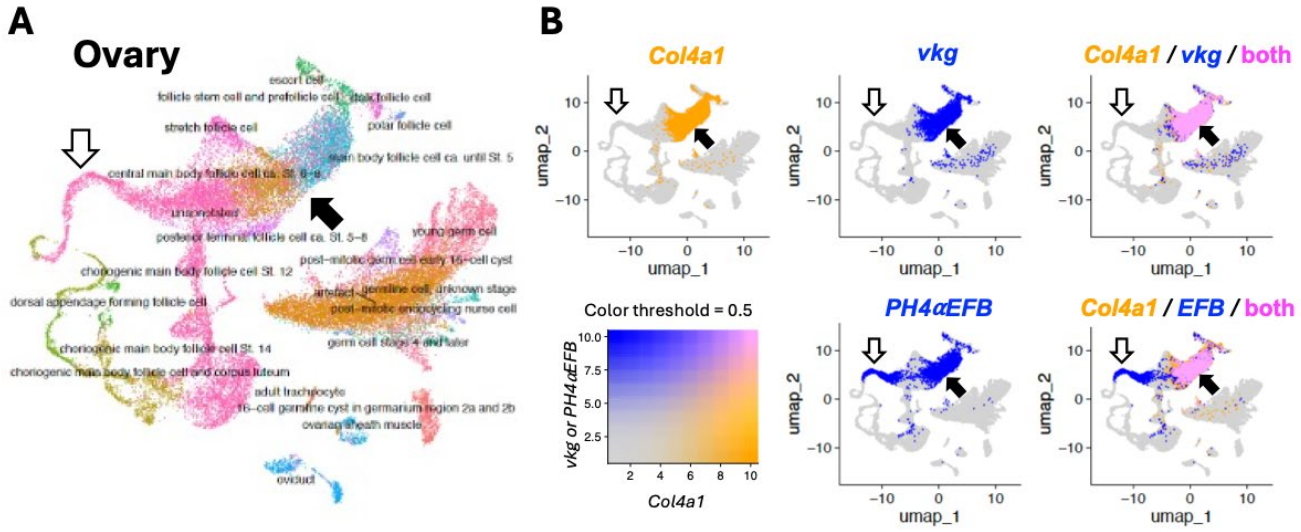


Figure. 4. Uniform manifold approximation and projection (UMAP) for the ovary single cell data. **(A)** UMAP showing the entire cells with annotations, with different cell types coded by different colour. **(B)** Expression of *Col4a1*, *vkg,* and *PH4αEFB (EFB)* colour coded as in the bottom left panel. Top left and middle panels show single gene expression; right panels show overlap. Closed arrows, follicle cells in which the three genes are co-expressed; open arrows, cells that express *PH4αEFB* but not the collagen IV genes; these cells should correspond to the metacells within the green interior rectangle in Figure. 3B.

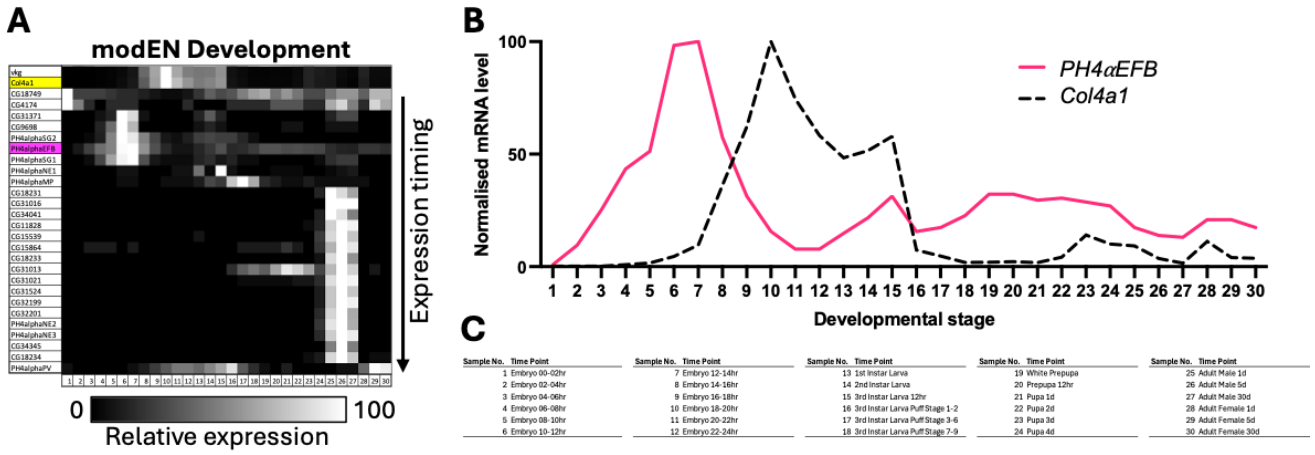
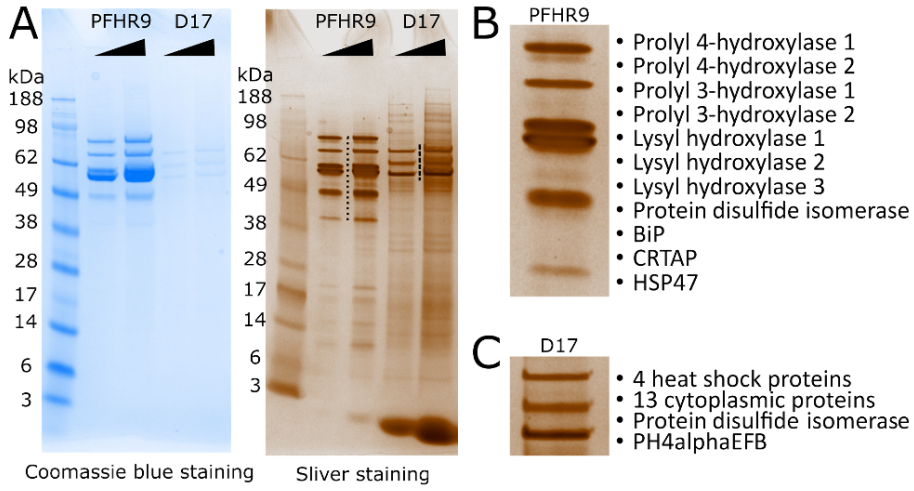


Figure. 5. Expression time courses of *Drosophila* collagen IV and PH4α-related genes.
(A) The expression levels of each gene from the modENCODE (modEN) Development transcriptome dataset are normalised for the maximum level and brightness-coded as in the bar at the bottom. *vkg* is at the top. Yellow, *Col4a1*; magenta, *PH4αEFB*. The PH4α-related genes are sorted according to the timing of expression peak: genes with earlier peak are located higher. For enlarged gene names and the values of gene expression levels, see Table S4, where the genes are sorted in the same order.
(B) Expression time courses of *PH4αEFB* and *Col4a1* extracted from (A) and displayed in the 2D scattered plot format.
(C) Legend for the sample numbers in the horizontal axes of **(A)** and **(B)**.



D **PH4alphaEFB (UniPort Entry:Q9VA69)**

*MAADWRLMLLLGILLLVGGPANG*EVYTALAEMEELLETESVLITNLEGYIRV
QEDKLNFLKNKMDEYQREHSDASHDITAYVSNPINAYLLTKRLTTDWRQVEN
LMEHDVGTDFLQNITQYRSLLKFPSDEDLNGAAVALLR**LQDTYQLDTSSVAR**
GKLNGIQYSTEMSSDDCFELGRQSYVNHDYYHTVLWMNEAMARMLEEPTNHT
QSFTKADILEYLAFSTYKEGNIESALTMTNELLQLLPHHERANGNKRFYEKE
IAQQLQLRKMK**GDDGTDEMPK**SDLPVAKSDPAIFDMTERRAYEMLCRGELKP
SPSDLRSLRCRYVTNRVPFLRLGPLKLEEVHADPYIVIYHDAMYDSEIDLIK
RMARPRFRRATVQNSVTGALETANYRISKSAWLKTQEDRVIETVVQRTADMT
GLDMDSAEELQVVNYGIGGHYEPHFDFARKEEQR**AFEGLNLGNR**IATVLFYM
SDVEQGGATVFTSLHTALFPKKGTAAFWMNLHRDGQGDVRTRHAACPVLTGT
KWVSNKWIHERGQEFR**RPCDLEEDHGEFAI**

E **Protein disulfide isomerase (UniPort Entry:P54399)**

*MKFLICALFLAASYVAAS*AEAEVK**VEEGVLVATVDNFK**QLIADNEFVLVEFY
APWCGHCKALAPEYAKAAQQLAEKESPIKLAKVDATVEGELAEQYAVRGYPT
LKFFRSGSPVEYSGGR**QAADIIAWVTK**KTGPPAK**DLTSVADAEQFLK**DNEIA
IIGFFKDLESEEAKTFTKVANALDSFVFGVSSNADVIAKYEAKDNGVVLFKP
FDDKK**SVFEGELNEENLKKFAQVQSLPLIVDFNHESAS**KIFGGSIKSHLLFF
VSREGGHIEKYVDPLKEIAKKYRDDILFVTISSDEEDHTRIFEFFGMNKEEV
PTIRLIKLEEDMAKYKPESDDLSAETIEAFLKKFLDGKLKQHLLSQELPEDW
DKNPVK**VLVSSNFESVALDK**SKSVLVEFYAPWCGHCK**QLAPIYDQLAEK**YKD
NEDIVIAK**MDSTANELESIK**ISSFPTIKYFRKEDNKVIDFNLDRTLDDFVKF
LDANGEVADSEPVEETEEEEEAPKKDEL

Figure. 6. PH4αEFB is captured by gelatine Sepharose from Drosophila D17 cells lysate.
(A) Mouse and *Drosophila* collagen binding proteins extracted from gelatine Sepharose with SDS sample buffer after extensive NaCl washes. The eluted samples were loaded onto the gels in two different volumes and stained with coomassie blue and silver staining.
(B and C) The magnified images of the area annotated with dots and dash lines for Mouse **(B)** and *Drosophila* **(C)** collagen binding proteins in the sliver stained gel. Protein names were identified by protein ID LC-MS analyses. The details of protein ID LC-MS results are in Fig. S7. In **(B)**, 'Prolyl 4 hydroxylase 1' and '2' correspond to P4HA1 and 2, respectively.
(D and E) The identified peptides from *Drosophila* PH4αEFB **(D)** and protein disulfide isomerase **(E)**, as determined by protein ID LC-MS. The underlined italic and red colour fonts indicate the signal peptide and identified peptides, respectively. The lists of identified proteins by MS (Mascot data) are in Supplementary Source Data.

Supplementary Table Legends (separated Excel and PDF files)

Table S1. Alignment of human P4HA2, PHD3, and *Drosophila* P4Hα-related proteins.
For human P4HA2, the domain structure and residue numbers at domain boundaries are shown at the top, using the same colour scheme as in Figure 1A. The amino acid length of each protein is indicated at the right end of the sequence. *Hs, Homo sapiens*; Dm, *Drosophila melanogaster*. Colour codes for amino acid residues are shown at the bottom left. Human P4HA2 and PHD3 (1 and 2, highlighted in blue and yellow, respectively) are homologous only within the catalytic domain. *Drosophila* CG15539-PA (22, magenta) lacks approximately the first 70% of the N-domain, whereas the other CG15539 isoforms (23 and 24) contain a complete N-domain. The three CG34041 isoforms (bottom of table, green) share homology with other P4Hα-related proteins in the N-domain and part of the PSB domain, but lacks the catalytic domain.

Table S2. Spatial expression patterns of *Drosophila* collagen IV and P4Hα-related genes.
Each sheet shows the values from the transcriptomics data indicated. Raw values of gene expression levels are summarised at the top left, with the *r* values with *Col4a1* shown in blue letters at the right. Sample legend is at the top right, in the cells shaded grey. *Col4a1* and *PH4αEFB* are marked yellow and magenta, respectively. In the cells marked cyan at the bottom, the values are normalised for the maximum expression level of each gene. The genes are sorted in the same order as in Fig. 2A.

Table S3. Expression of the Collagen IV and PH4α-related Genes in Single Metacells.
Each sheet shows the results from the indicated dataset. Expression levels of the two collagen IV genes (*Col4a1*, *vkg*) and the 26 collagen PH4α-related genes in each metacell are shown, together with other information indicated in the top row.

Table S4. Expression time courses of *Drosophila* collagen IV and P4Hα-related genes.
Values from the modENCODE Development data. Raw values of gene expression levels are summarised at the top left. Sample legend is at the top right, in the cells shaded grey. *Col4a1* and *PH4αEFB* are marked yellow and magenta, respectively. In the cells marked cyan at the bottom, the values are normalised for the maximum expression level of each gene. The genes are sorted in the same order as in Fig. 5A.

Table S5. Expression of the Collagen IV and PH4α-related Genes in 'High *Col4a1*, Low *PH4αEFB*' Single Metacells.
From Table S3, the data of the 374 metacells with high *Col4a1* (> 1) and low *PH4αEFB* (< 0.1) expression were extracted. Column AE shows the maximum expression level of the 25 PH4α-related genes excluding *PH4αEFB*. The value was zero for 240 metacells (Cell AL2), i.e., no expression of non-*EFB* PH4α-related genes was detected in these metacells.

Supplementary Source Data. Identification of gelatine binding proteins by mass spectrometry.
Below follows the full MASCOT search result file that includes all proteins found in the gel band from sliver staining SDS-PAGE gel showing top-left.

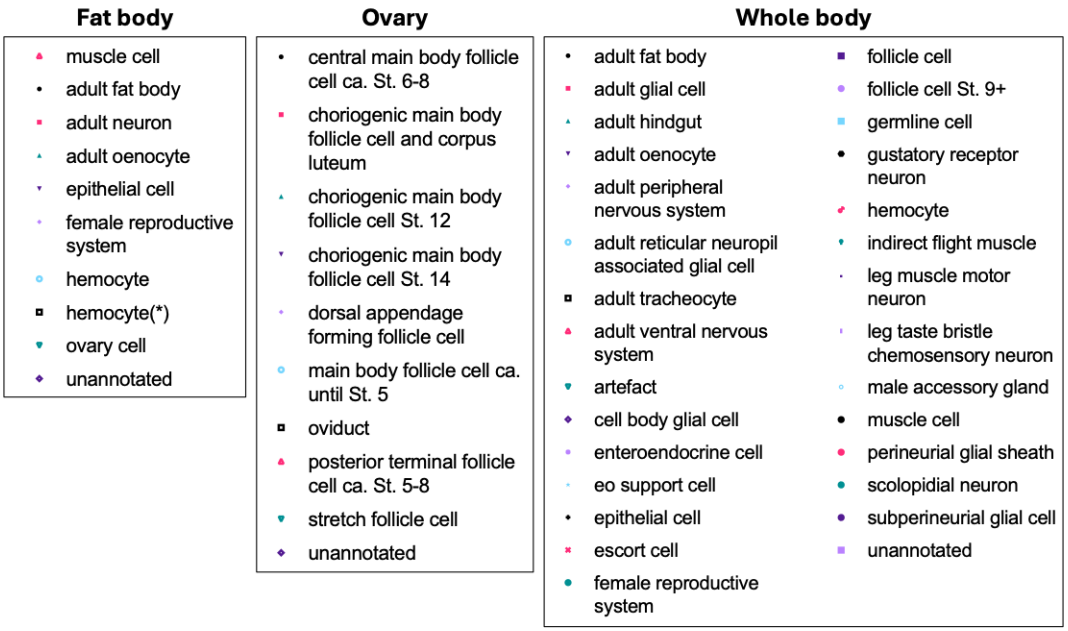


Figure S1. Full legend for the symbols in Figure 3B.

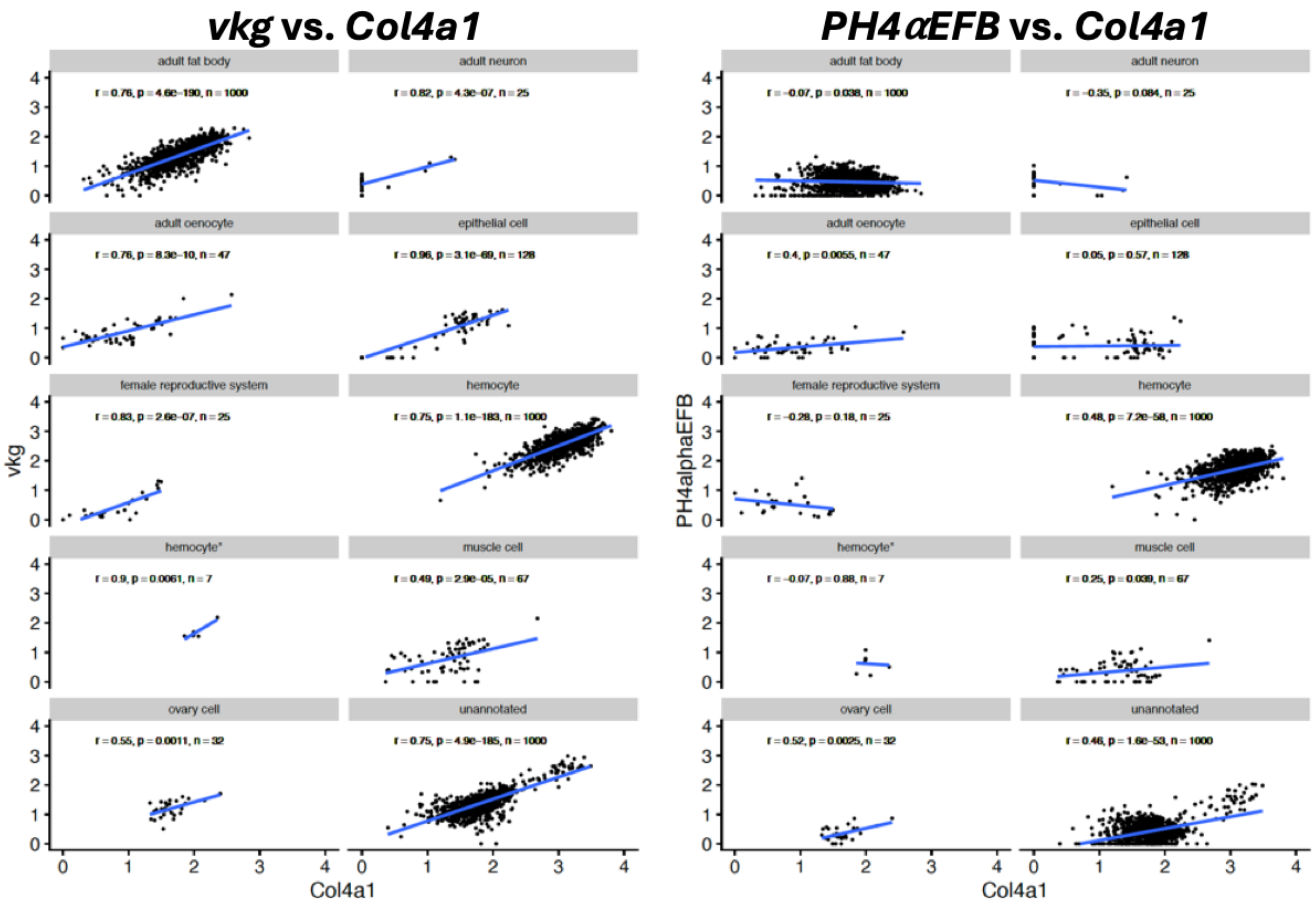


Figure S2. Expression of *Col4a1*, *vkg*, or *PH4αEFB* separately analysed for each different cell type in the fat body dataset.

Regression lines of the plots, correlation coefficients between the two genes plotted (*r*), the *p* values to obtain the results from the null hypothesis that the slope of the regression lines are zero, and the numbers of metacells analysed (n) are shown.

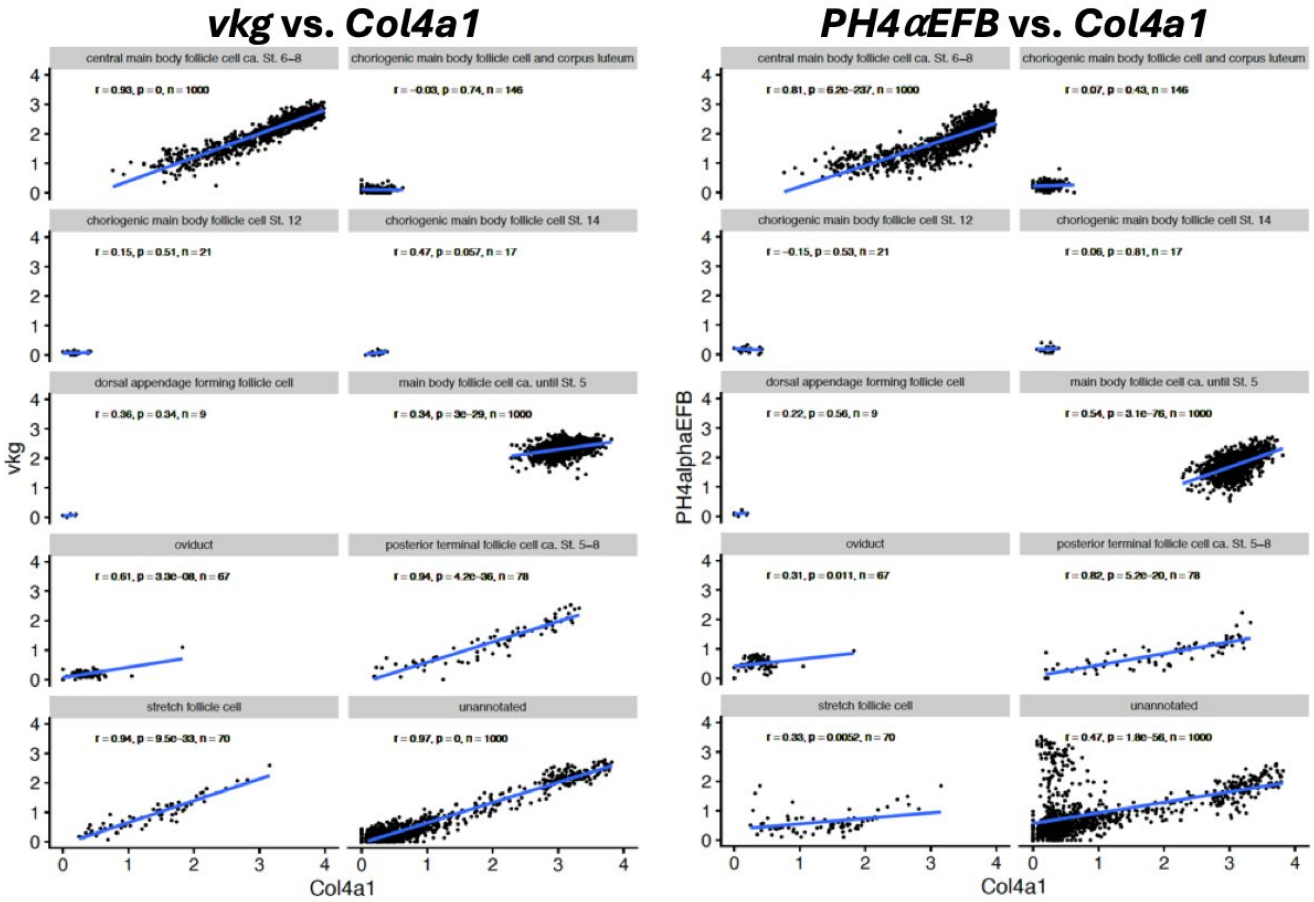


Figure S3. Expression of *Col4a1*, *vkg,* or *PH4αEFB* separately analysed for each different cell type in the ovary dataset.

Regression lines of the plots, correlation coefficients between the two genes plotted (*r*), the *p* values to obtain the results from the null hypothesis that the slope of the regression lines are zero, and the numbers of metacells analysed (n) are shown.

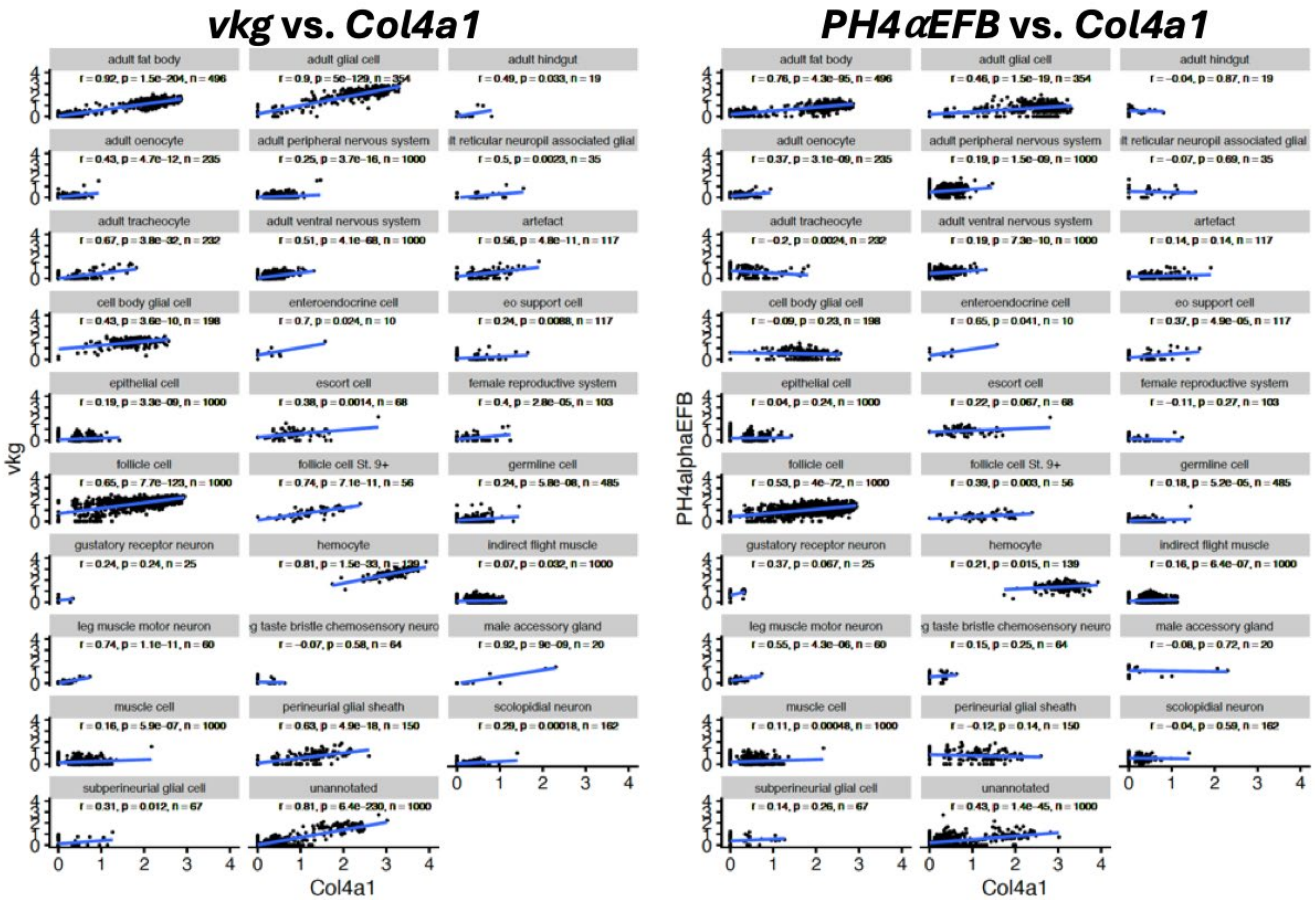


Figure S4. Expression of *Col4a1*, *vkg,* or *PH4αEFB* separately analysed for each different cell type in the whole-body dataset.

Regression lines of the plots, correlation coefficients between the two genes plotted (*r*), the *p* values to obtain the results from the null hypothesis that the slope of the regression lines are zero, and the numbers of metacells analysed (n) are shown.

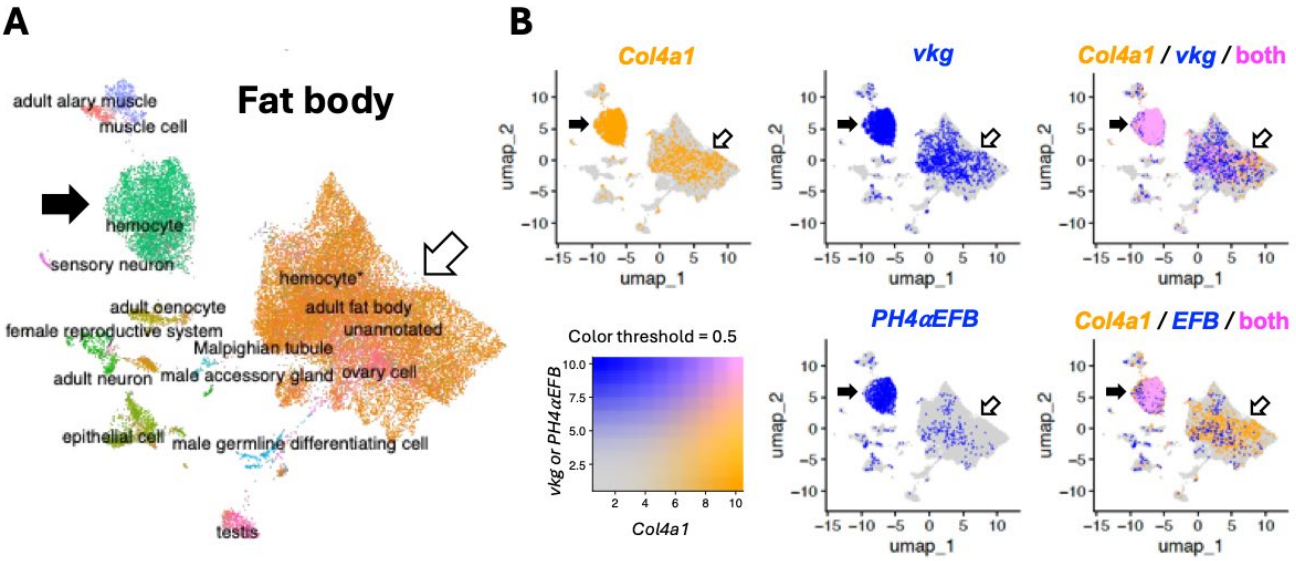


Figure S5. UMAP for the fat body single cell data.
(A) UMAP showing the entire cells with annotations, with different cell types coded by different colour.
(B) Expression of *Col4a1*, *vkg,* and *PH4αEFB (EFB)* colour coded as in the bottom left panel. Top left and middle panels show single gene expression; right panels show overlap. Closed arrows, hemocytes in which the three genes are co-expressed; open arrows, adult fat body cells in which the co-expression of the collagen IV genes and *PH4αEFB* is not as clear as that in the metacell data in Fig. 3B.

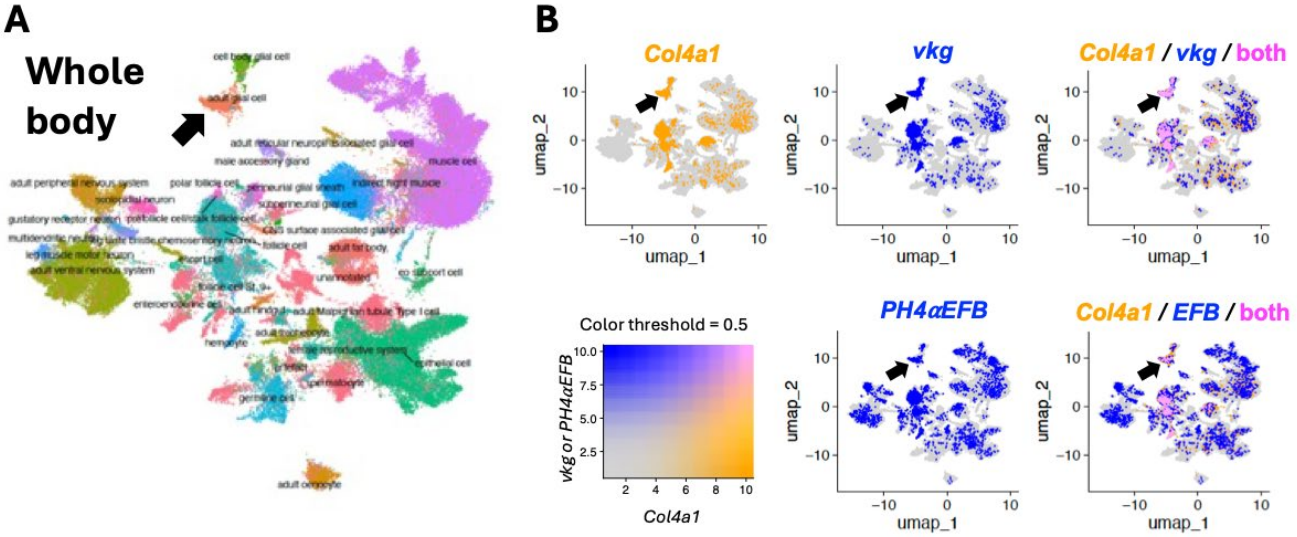


Figure S6. UMAP for the whole-body single cell data.
(A) UMAP showing the entire cells with annotations, with different cell types coded by different colour.
(B) Expression of *Col4a1*, *vkg,* and *PH4αEFB (EFB)* colour coded as in the bottom left panel. Top left and middle panels show single gene expression; right panels show overlap. Closed arrows, glial cells in which the three genes are co-expressed.