# Collagen IV of basement membranes: I. Origin and diversification of COL4 genes enabling animal evolution.

Patrick S. Page-McCaw[#,1,2], Elena N. Pokidysheva[2], Carl E. Darris[3], Sergei Chetyrkin[13], Aaron L. Fidler[5], Prayag Murawala[10,11], Julianna Gallup[+], Aspirnauts[+], Julie K. Hudson[12], Billy G. Hudson[#2-9]

+Aspirnauts: Jillian Balser, Cullen Curbow, Ricard Harris, Cody Stothers, Keith Wade.

Aspirnaut is a K-20 Science, Technology, Engineering and Math (STEM) pipeline for increasing the diversity and wellness of the STEM workforce. The holistic training approach features guided discovery science that is augmented with guided professional skills development, guided self-discovery, and wellness training. High school students from rural America and diverse backgrounds engage in hands-on discovery science for six weeks while in residence at Vanderbilt University Medical Center, and diverse undergraduate students engage for ten weeks. The experience provides students with the tools and empowerment to effect positive change in themselves, their families, and communities for generations to come.

# To whom correspondence should be addressed: Patrick S. Page-McCaw, Billy G. Hudson: Division of Nephrology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232

**Affiliations:**

[1] Department of Cell and Developmental Biology, Vanderbilt University, Nashville TN, 37232

[2] Division of Nephrology and Hypertension, Department of Medicine,Vanderbilt University Medical Center, Nashville, TN 37232

[3] Center for Matrix Biology, Vanderbilt University Medical Center, Nashville, TN 37232

[4] Department of Biochemistry, Vanderbilt University, Nashville, TN 37232

[5] Vanderbilt-Ingram Cancer Center, Vanderbilt University, Nashville, TN 37232

[6] Vanderbilt Institute of Chemical Biology, Vanderbilt University, Nashville, TN 37232

[7] Department of Biological Sciences, and Vanderbilt University, Nashville, TN 37232

[8] Department of Cell and Developmental Biology, Vanderbilt University, Nashville, TN 37232;

[9] Department of Pathology, Microbiology, and Immunology, Vanderbilt University Medical Center, Nashville, TN 37232 Vanderbilt University Medical Center, Vanderbilt University, Nashville TN, 37232

[10] Mount Desert Island Biological Laboratory, Bar Harbor, ME, 04609

[11] Clinic for Kidney and Hypertension Diseases, Hannover Medical School, 30625 Hannover, Germany

[12] Department of Medical Education and Administration, Vanderbilt University Medical Center, Nashville, TN 37232

[13] Mass Spectrometry Research Center, Vanderbilt University, Nashville, TN, USA.

**Short Title:** Evolutionary origin and diversification of the COL4 genes

**Keywords:** Collagen IV, Basement membrane, Metazoan evolution, Glomerular filtration barrier, Alport Syndrome

## Abstract

Collagen IV is a primordial component of basement membranes, a specialized form of extracellular matrix that enabled multi-cellular epithelial tissues. In mammals, collagen IV assembles from a family of six α-chains (α1 to α6), encoded by six genes (COL4A1 to COL4A6), into three distinct scaffolds: the α121, the α345 and a mixed scaffold containing both α121 and α565. The six mammalian COL4A genes occur in pairs that occur in a head-to-head arrangement on three distinct chromosomes. In Alport syndrome, variants in the COL4A3, 4 or 5 genes cause either loss or defective assembly of the collagen IV$^{\alpha 345}$ scaffold which results in a dysfunctional glomerular basement membrane, proteinuria and progression to renal failure in millions of people worldwide. Here, we determine the evolutionary emergence and diversification of the COL4A genes using comparative genomics and biochemical analyses. Using syntenic relationships to genes closely linked to the COL4A genes, we determine that the COL4A3 and COL4A4 gene pair appeared in cyclostomes (hagfish and lampreys) while the COL4A5 and COL4A6 gene pair emerged in gnathostomes, jawed vertebrates. The more basal chordate species, lancelets and tunicates, do not have discrete kidneys and have a single COL4A gene pair, though often with single isolated COL4 genes similar to those found in *C elegans*. Remarkably, while the six COL4A genes are conserved in vertebrates, amphibians have lost the COL4A3 and COL4A4 genes. Our findings of the evolutionary emergence of these genes, together with the amphibian double-knockout, opens an experimental window to gain insights into functionality of the Col IV$^{\alpha 345}$ scaffold.

## Introduction

Collagen IV is the primordial component of basement membranes, a specialized form of extracellular matrix that enabled the genesis and evolution of multicellular epithelial tissues (Fidler *et al.*, 2017). Collagen IV was discovered 50 years ago in pioneering studies of the glomerular basement membrane (GBM), a prominent feature of the kidney filtration barrier (Kefalides, 1973). Subsequent studies of the pathobiology of three GBM diseases (Goodpasture's disease, Alport syndrome and diabetic nephropathy) revealed that mammalian collagen IV is a family of six α-chains (α1 to α6). The chains are encoded by six genes (COL4A1, COL4A2, COL4A3, COL4A4, COL4A5, COL4A6) and are located pairwise in a head-to-head arrangement on three chromosomes (Figure 1). The α-chains co-assemble into protomers of three distinct molecular compositions: α121, α345 and α565, which in turn assemble into three distinct supramolecular scaffolds, referred to as Collagen IV$^{\alpha121}$, Collagen IV$^{\alpha345}$, and Collagen IV$^{\alpha556—\alpha121}$, that have distinct distributions within the nephron (Figure 1).

The tissue distribution, structure, and assembly of Collagen IV$^{\alpha121}$ is well studied. It is a ubiquitous component of metazoan basement membranes, beginning with basal animals (Fidler *et al.*, 2014, 2017, 2018; Boudko *et al.*, 2023). The scaffold confers tensile strength and tether diverse macromolecules, including laminin nidogen, proteoglycan, and growth factors, forming the supramolecular complexes that interact with cell surface receptors Wang 2008, Parkin 2011 (Emsley *et al.*, 2004; McCall *et al.*, 2014; Brown *et al.*, 2017; Fidler *et al.*, 2018). Disrupting the Collagen IV$^{\alpha121}$ scaffold causes basement membrane destabilization and tissue dysfunction in mice, zebrafish, flies, and nematodes (Borchiellini *et al.*, 1996; Gupta *et al.*, 1997; Pöschl *et al.*, 2004; Fidler *et al.*, 2014). Moreover, genetic defects cause Gould syndrome and HNAC in humans (Labelle-Dumais and Gould, 2012). Collectively, the findings reveal that Collagen IV$^{\alpha121}$ is essential for tissue architecture and function across animalia, yet the molecular mechanisms of function and dysfunction remain obscure.

In contrast to Collagen IV$^{\alpha121}$, there is a paucity of knowledge about the Collagen IV$^{\alpha345}$ scaffold despite its known dysfunction in prominent kidney diseases. Of note, this scaffold is the major constituent of the glomerular basement membrane (GBM) (Figure 1), a critical morphological feature that functions as an ultrafilter of proteins in the kidney. It has long been hypothesized that Collagen IV$^{\alpha345}$ plays a critical role in ultrafiltration of serum limiting the entry of serum proteins into urine. The Collagen IV$^{\alpha345}$ scaffold is affected in three GBM diseases: Alport syndrome, Goodpasture's disease and diabetic nephropathy (Figure 1). In Alport syndrome, affecting millions of people worldwide, nearly two thousand genetic variants occur in the COL4A3, COL4A4, and COL4A5 genes. Variants cause either loss of the scaffold from the GBM or its defective assembly resulting in proteins leaking through the GBM, proteinuria, and progression to kidney failure. In Goodpasture's disease, affecting thousands of people worldwide, autoantibodies target neoepitopes in the Collagen IV$^{\alpha345}$ scaffold causing low-grade proteinuria rapidly progressing to renal failure. Finally, in diabetic nephropathy, this scaffold is also involved in the thickening of the GBM, proteinuria, and progressive kidney failure. How these structural alterations cause kidney failure raises a fundamental unanswered question: "How does the Collagen IV$^{\alpha345}$ scaffold enable the GBM to function as a protein ultrafilter?"

Knowledge of COL4 gene evolution may shed light on the functionality of Collagen IV scaffolds in the evolution of multicellular tissues and on disease mechanisms. Here, we used comparative genomics, to determine the evolutionary origin and lineage of the six vertebrate collagen IV genes. Our findings, together with biochemical analyses, revealed the lineage of the genes, and unveiled a naturally- occurring double-knockout of the COL4A3, COL4A4 genes in amphibians, opening a window to gain insights into the functionality of the Collagen IV$^{\alpha345}$ scaffold.

## Results

### Synteny at the COL4 loci in vertebrate genomes

The genes which encode Collagen IV appear most commonly as duplicated genes in a head-to-head arrangement (Pöschl *et al.*, 1988). The divergent orientation of this gene pair was first noted as the COL4A genes were cloned (Burbelo *et al.*, 1988; Pöschl *et al.*, 1988) and the possible function of this arrangement in regulating gene expression was noted. Remarkably, this gene arrangement was found to be conserved in diverse species, though not all. It has previously been reported that three of the four Insulin Receptor Substrate (IRS)[1] gene family members are found in close proximity to each COL4A gene pair (Al-Salam and Irwin, 2017). This observation suggests that the three gene clusters including both COL4A and the IRS gene share common ancestry. In order to understand how the three COL4A gene pairs evolved in vertebrates, we investigated the genomic locus and microsyntenies of the three COL4A gene pairs COL4A〈1|2〉, COL4A〈3|4〉 and COL4A〈5|6〉 found in vertebrates.

To investigate the conservation of this gene cluster, we generated maps of the syntenic regions adjacent to each of the COL4A gene pairs from diverse species within each of the vertebrate classes which have high quality genome assemblies. We first generated syntenic maps of the COL4A〈1|2〉[2] gene pair (Figure 2). Remarkably, the gene order surrounding this collagen IV encoding gene pair is highly conserved. Immediately downstream of COL4A1 the IRS2 gene and the MYO16 genes are observed in all Euteleostomi ("bony" vertebrates), however homologues of both IRS2 and MYO16 are absent in the Chondrichthyes (cartilaginous vertebrates). IRS2B, MYO16, ABHD13, LIG4 and TNFSF13B are linked and found on a distinct contig in Chondrichthian chimera shark strongly suggesting that there was a Chondrichtyes specific chromosomal rearrangement secondary to the formation of this gene cluster. It is not possible to rule out an error in contig assembly to explain this finding; however, because this break in synteny is observed in multiple Chondrichthyes species it is likely due to a chromosomal rearrangement in the Chondrichthyes ancestor rather than a mis-assembly. The gene cluster extends to include TNFSF13B, ABHD13, LIG4 and NALF1 in most vertebrates. However, the more distal genes are not as well conserved as the more proximal genes. For example, NALF1 is not found in the microsynteny block of three of the four amphibian genomes we assessed. LIG4, while present in ray-finned fish genomes, is absent from the microsyntenic region of the COL4A gene pair in each ray-finned finned fish we have examined.

Downstream of the COL4A2 gene similar conservation of gene position is observed, however in this case the COL4A2 microsynteny is conserved between Chondrichthyes and Euteleostomi including the RAB20, NAXD, CARS2, ING1, and ANKRD10 genes. The gene cluster conservation extends beyond our investigated region. One gene, UBE2L, is found in both some Chondrichthyes and Actinopterygii (ray-finned fishes) and the Sarcopterygii (Coelacanth), but not in Tetrapods (terrestrial vertebrates). Similar to the chromosomal changes observed downstream of COL4A1 where the genes downstream of MYO16 are found elsewhere in the genomes of birds, the CARS2 gene has been lost from the COL4A〈1|2〉 microsynteny domain in Actinopterygii. In the case of *Danio rerio* (zebrafish), the three genes ANKRD10, UBE2L and ING1 have been inverted along with an additional gene (white arrow in Figure 2). The TNFSF13B and ABHD13 genes remain on the same chromosome as the COL4A〈1|2〉 gene pair (Chromosome 9), however they are located over 5 Mbp away, likely this represents a chromosomal inversion in the zebrafish lineage. The Axolotl genome represents a special case in our data set, we wished to include a member of the sequenced genomes from salamanders and newts (Caudata), however the available genomes from these species are extremely large which makes assembly of

---

[1] The genes in the syntenic groups described here are identified using a consensus gene name. As the nature and function of these genes is not important for this study, we do not further define them here.
[2] The nomenclature COL4A〈1|2〉 is used here to refer to the paired COL4A genes which are transcribed from opposite strands. Where only one gene is referred to, we use the standard gene nomenclature, eg COL4A1.

large contigs difficult. The edges of the Axolotl contig on which each COL4A⟨1|2⟩ gene pair is found are marked by curly brackets. The extent of the synteny in these species is difficult to assess as the contigs, though long, have sparse genes and so are incomplete. These results demonstrate a remarkable conservation of the COL4A⟨1|2⟩ syntenic region throughout gnathostome evolution.

### Synteny at the COL4A⟨3|4⟩ locus

We next investigated the microsyntenies present at the COL4A⟨3|4⟩ locus. In mammals, an IRS gene homolog, IRS1, is found in close proximity to the COL4A⟨3|4⟩ gene pair as IRS2 is for COL4A⟨1|2⟩ (Figure 3); it is located downstream of COL4A4 and the RHBDD1 gene intervenes. Although the COL4A⟨3|4⟩ is likely to have arisen as a gene duplication from COL4A⟨1|2⟩ suggesting that additional genes homologous to those found in the COL4A⟨1|2⟩ microsynteny region would be found in the COL4A⟨3|4⟩microsynteny region, there are no other gene duplications from COL4A⟨1|2⟩ in the region other than the IRS1 gene. In general, there is reduced conservation at the COL4A⟨3|4⟩ microsynteny region than that found in COL4A⟨1|2⟩.

Two significant chromosomal rearrangements are apparent within Gnathostomes at this locus: first, the COL4A⟨3|4⟩ locus is absent in Amphibia (discussed below) and ray-finned fishes have rearranged the chromosome immediately downstream of COL4A3 to place a pair of SST1 homologous genes immediately downstream of COL4A3. The MFF, MRPL4, AGFG1 and SLC19A3 genes are found inverted 5 genes downstream of the COL4A4 locus. In the stickleback, likely orthologs of MRPL4, AGFG1 and a duplication of SLC19A3 occur about 5.5 Mbp centromeric on the same chromosome. Similar rearrangements are found in other ray-finned fishes (not shown). These observations suggest that the COL4A⟨3|4⟩ locus is under active selection and the chromosomal locus is rearranging in ray-finned fishes. This is perhaps not surprising given the extraordinary morphological and functional diversity in the teleosts (Volff, 2005). Given the functional role of COL4A⟨3|4⟩ in the renal filtration barrier (Figure 1) and the differences in renal physiology in fish which live in highly stable, high osmolarity marine environments, highly variable estuarine osmotic environments, and low osmolarity freshwater environments, it is perhaps not surprising that the COL4A⟨3|4⟩ synteny group is diverse due to diverse selections pressures at the locus.

### Loss of the COL4A⟨3|4⟩ gene pair from Amphibia

Retrieval of the COL4A⟨3|4⟩ gene pair locus from genomes from Vertebrata revealed the presence of the COL4A⟨3|4⟩ gene pair in cartilaginous fishes, ray-finned fishes, reptiles, birds and mammals. In contrast to the situation in ray-finned fish and amphibians, the COL4A⟨3|4⟩ microsynteny region is highly conserved between cartilaginous fishes, the coelacanth and non-amphibian tetrapods (Figure 3A). The nine gene synteny group NYAP2A, IRS1, RHBDD1, COL4A⟨3|4⟩, MFF, TM4SF20, AGFG1A and SLC19A3 genes are all found in cartilaginous fish, sarcopterygian, reptiles, birds and mammals. However, we were unable to identify the COL4A⟨3|4⟩ gene pair in amphibian genomes. To determine whether these genes were lost in amphibians, we identified the syntenic region by identification of the two neighboring genes RHBDD1 and MFF found in both fishes and reptiles in each amphibian genome examined. Remarkably, there is a precise deletion of the COL4A⟨3|4⟩ gene pairs in both the Caecilian and Axolotl genomes (Figure 3B). In the Caecilian genome, a small, single exon testes-specific kinase domain gene replaces the COL4A⟨3|4⟩ gene pair. In the Axolotl genome, no genes are found between the COL4A⟨3|4⟩ gene pair, though 2.6 Mbp of genomic sequence is present. In the two frog genomes investigated, there is a chromosomal inversion of this region which moves the MFF, MRPL44, TMFSF20 and AGFG1 genes distal on the chromosome with an intervening sequence of 10 and 32 Mbp and additional genes. A parsimonious explanation for these observations is that the ancestral amphibian had a precise excision of the COL4A⟨3|4⟩ as seen in the Caecilian genome. Subsequently, the locus was inverted between the MFF and distal ANKUB1 genes in frogs.

### Synteny at the COL4A⟨5|6⟩ gene locus

Similar to the COL4A⟨1|2⟩ locus, the COL4A⟨5|6⟩ locus is part of a conserved synteny block in gnathostome evolution. Immediately downstream of COL4A5, but on the opposite strand, the IRS paralog IRS2B is found

(Figure 4). This arrangement is conserved in nearly all gnathostomes investigated save for teleostans which, like the COL4A⟨3|4⟩ synteny group in teleosts, is less conserved. Unlike the case of COL4A⟨3|4⟩, an additional gene is found to have a conserved homolog in addition to the COL4A gene pair at the COL4A⟨5|6⟩ locus. TNFSF13B is found downstream of COL4A1 in the Coelacanth and tetrapod genomes, though not in the teleost genomes. This larger microsyntenic block provides evidence that the COL4A⟨5|6⟩ locus is derived from the COL4A⟨1|2⟩ locus and not the COL4A⟨3|4⟩ locus consistent with its appearance later in vertebrate evolution.

**Origin of the COL4A⟨1|2⟩, COL4A⟨3|4⟩ and COL4A⟨5|6⟩ gene pairs in basal Chordates**

We have identified conserved synteny at the three vertebrate COL4A gene pair loci. To attempt to determine the evolutionary origins of these gene pairs, we next looked within the Deuterostome lineage (Figure 5). Deuterostomes are found in three distinct phyla: Echinodermata, Hemichordata and Chordata. Within the available genomes of both Echinodermata and Hemichordata, we observe a single COL4A gene pair, as seen in Protostomes and other Eumetazoan animals, suggesting that the two additional vertebrate COL4A gene pairs are specific for the chordate lineage. In order to identify the evolutionary origins of the two COL4A gene pair duplications found in Chordates, we first examined the COL4A gene pairs in the basal chordates: the lancelet, *Branchiostoma floridae*, and in tunicates. In these basal chordates, we do not observe conserved synteny in the immediate vicinity of the COL4A gene pair. Notably, an IRS ortholog is not linked to the COL4A locus in these species in stark contrast to the tight linkage observed between an IRS homolog and each of the COL4A gene pairs found in vertebrates. For example, there are 3 likely IRS homologs in *Branchiostoma floridae*, however none of them are on the chromosome containing the single COL4A gene pair. In addition, there are several instances of unpaired COL4A gene homologs, we refer to COL4A genes in this configuration as COL4S (S for singleton) genes. For example, *Branchiostoma floridae* chromosome 16 has two COL4S genes with short collagen repeat domains and NC1-like C-terminal domains, but are separated by many unrelated protein coding genes and approximately 15 Mbp (LOC118432728 and LOC118403826). A similar COL4S gene arrangement is observed on chromosome 18. Chromosome 17 has a single COL4S gene (LOC118404283) with a short collagen domain and a complete NC1 domain. Two models for the origins of the vertebrate COL4A genes are suggested by these findings: First, the COL4S genes are likely specific to the pre-chordate lineages. Second, the COL4S genes may represent the ancestral gene duplications leading to the vertebrate COL4A. Notably, a conserved feature of the COL4S genes in the lancelet are unusual in having a short collagen repeat domain. This suggests that the originating duplication that created COL4S resulted in a truncated collagen domain. This gene was then re-duplicated. Alternatively, distinct duplication events may have occurred but each new COL4S gene was under selection for short collagen domains.

We next turned to the cyclostomes, the sister group to the jawed vertebrates, to determine if these species had multiple COL4A gene pairs. At least one whole genome duplication event is thought to have occurred prior to the emergence of the cyclostomes, predicting that the cyclostomes may have two COL4A gene pairs. However, recent evidence demonstrates that there were two cyclostome-specific genome duplications in the lineage which suggests that any conserved duplications may also result in ambiguous syntenic relationships. Consistent with the whole genome duplication hypothesis, two copies of a COL4A gene pair are found in each cyclostome genome investigated. Moreover, as for the jawed vertebrate COL4A gene pairs described earlier, these gene pairs occur adjacent to an IRS gene homolog and either MYO16 or NYAP2. This result strongly suggests that the two gene pairs found in cyclostomes share a common evolutionary origin to the COL4A⟨1|2⟩ and COL4A⟨3|4⟩ gene pairs.

While there is little synteny between the COL4A gene pairs in cyclostomes and gnathostomes beyond the presence of the IRS and NYAP2/MYO16 loci, we did find evidence that the region downstream of lamprey COL4A2 is retained as a syntenic block in the gnathostome lineage. The syntenic block DOCK9, SLC15A1, STK24, FARP1, and IPO5 are found together approximately 3 Mbp away from the COL4A⟨1|2⟩ gene pair on chromosome 6 of the thorny skate, a cartilaginous fish, suggesting that there was a gene inversion event in

vertebrates. The DOCK9, SLC15A1, STK24, FARP1, and IPO5 synteny block also appears on the same chromosome as the COL4A⟨1|2⟩ gene pair in the early ray-finned fish *Lepisosteus oculatus* and in the human genome, albeit approximately 20 and 12 Mbp away, respectively. These results suggest a common evolutionary ancestry between the COL4A⟨3|4⟩ gene pairs in cyclostomes and the COL4A⟨1|2⟩ gene pair found in gnathostomes.

**Origin of NYAP2 and MYO16**

The hagfish COL4A⟨1|2⟩ synteny group contains both an IRS homolog and NYAP2. We note that in hagfish the MYO16 gene is also present. The assignment of the novel cyclostome COL4A gene pair to COL4A⟨3|4⟩, COL4A⟨5|6⟩ or a distinct duplication event was not initially clear from an examination of the genes present on this chromosome. However, we noted that NYAP2 is found in hagfish immediately downstream of and on the opposite strand from the IRS homolog. This same gene arrangement, while not found in lampreys, is found downstream of COL4A4 in vertebrates suggesting a gene inversion event occurred in the vertebrate lineage. Further support for this assignment can be found in an examination of the NYAP2 and MYO16 gene sequence (Figure 6). It has been previously noted that NYAP2 and MYO16 are homologous at their C-terminal halves. This suggests that during the duplication of COL4A⟨1|2⟩ to generate the ancestral cyclostome COL4A gene pairs, three genes were duplicated COL4A1, COL4A2 and IRS. Additionally, the 3' half of MYO16 was included in this gene duplication and this 'half gene' was retained as NYAP2. Moreover, a possible MYO16 gene is found in both lancelets and tunicates demonstrating that the ancestral gene is likely the MYO16 gene which was partially duplicated and retained generating the NYAP2 locus. We then hypothesize that when the vertebrate and cyclostome lineages diverged, the IRS and NYAP2 genes subsequently appear to have rearranged, moving them from downstream of COL4A3 to downstream of COL4A4. These results support the hypothesis that the COL4A⟨1|2⟩ gene was duplicated in the common ancestor of vertebrates generating the COL4A⟨3|4⟩ gene pair.

We next examined genes syntenic to COL4A⟨5|6⟩ to determine whether the origins of this gene pair could be determined. We hypothesized that there may be gene duplication artifacts, like that seen at the NYAP2/MYO16 locus, present near the COL4A⟨5|6⟩ locus that could facilitate tracing COL4A⟨5|6⟩ back to either the COL4A⟨1|2⟩ or COL4A⟨3|4⟩ gene pairs. We note that downstream of the COL4A⟨1|2⟩-IRS2-MYO16 gene group is TNFSF13B. TNFSF13B is remarkably similar to the gene EDA which is found immediately downstream of COL4A6. The N-terminal half of TNFSF13B and EDA from the thorny skate, *Amblyraja radiata*, is 30% identical and 57% similar while the C-terminal TNF domain half is 25% identical and 49% similar. Intervening between these two domains EDA has a collagen-triple helix domain. We suggest the possibility that during the gene duplication event resulting in the creation of COL4A⟨5|6⟩, IRS2, MYO16 and TNFSF13B were duplicated as well as COL4A⟨1|2⟩ and inserted between the ATG4A and GUCY2F loci. The COL4A genes, IRS2 and TNFSF13B were retained, but MYO16 was lost during subsequent chromosomal rearrangements that inverted the IRS homolog and inserted exons encoding collagen domain sequences from the COL4A loci in the midst of the TNFSF13B gene generating EDA. While the scars left from this genome rearrangement are more difficult to interpret than that creating the COL4A⟨3|4⟩ gene pair, it seems possible that the COL4A⟨5|6⟩ derives from the COL4A⟨1|2⟩ locus and not the COL4A⟨3|4⟩ locus based on this speculative reasoning.

**COL4A⟨1|2⟩ gene pair structure in Animalia**

To determine whether this gene arrangement is a common feature of COL4A genes, we have sampled genomes spanning Metazoan evolution. In many cases where the genome is well annotated and supplemented with transcriptome sequence, the 5' end of these genes are closely juxtaposed and the transcription start site of one gene occurs within the transcription unit of its partner. These gene pairs are highly homologous and referred to as COL4A1 and COL4A2; as in the vertebrate lineage, we refer to the combined locus as COL4A⟨1|2⟩ (Figure 1). In contrast to the COL4A⟨1|2⟩ gene locus observed in bilaterian animals, in *C. elegans*, the genes which express Collagen IV α1 and α2 (*emb-9* and *let-2*, respectively) are

found on distinct chromosomes, Chr III and X, and so resemble the arrangements seen in early chordates with COL4S gene. Within nematodes, the unlinked locus structure is found also in *Ascaris suum*, suggesting that this gene structure is conserved within nematodes. Moreover, the COL4A⟨1|2⟩ gene pair is not present, as it is in the early chordates. Chromosomal rearrangement has been estimated to be very rapid in *C. elegans* (Coghlan and Wolfe, 2002) suggesting that this locus arrangement, in which the COL4S1 and COL4S2 loci are unlinked, may be a common feature in nematodes.

**COL4A locus structure in Ecdysozoans**

As a first step in understanding the evolution of this conserved Metazoan gene, we have turned our attention to the locus structure to determine whether the paired genes in the COL4A⟨1|2⟩ locus observed in vertebrates or the unlinked loci observed in *C. elegans* and the early chordates is ancestral and conserved. The two best studied invertebrate model organisms typify the two COL4 gene structures. In *Drosophila*, the COL4A gene locus is in the COL4A⟨1|2⟩ linked configuration where the COL4A1 and COL4A2 genes are arranged head-to-head with transcription start sites less than 1500 bp apart. In contrast, in C. elegans the two Collagen IV expressing genes are not linked. To determine whether the Drosophila COL4A⟨1|2⟩ paired locus structure is conserved within Dipterans, we inspected the genomes of other *Drosophilids* and *Apis mellifera* and observed the conserved COL4A⟨1|2⟩ configuration (not shown). To determine whether this gene structure is conserved within insects we next inspected the genome of the short-germ band insect *Tribolium castaneum*. The *T. castaneum* contig AAJJ02001976.1 contains complete copies of both COL4A⟨1|2⟩ in a head-to-head orientation with closely apposed apparent transcription start site. Similarly, the genome of the Hymenoptera, *Aphidius gifuensis*, demonstrates the presence of the linked COL4A⟨1|2⟩ locus with no evidence for unlinked COL4S loci.

To determine whether the COL4A⟨1|2⟩-paired gene locus structure was conserved within Arthropoda (Figure 10), we next investigated the gene structure in the Chelicerate *Dermatophagoides pteronyssinus*. We identified a contig MQNO02000001.1 which contains the 5’ ends of both COL4A1 and COL4A2 homologs in head-to-head orientation. We were unable to identify unlinked COL4A loci in this genome. A second Chelicerate, the common house spider, *Parasteatoda tepidariorum*, similarly had a single COL4A⟨1|2⟩ locus found on contig ML632983.1. However, this genome also contains two copies of COL4S-like genes. A genome of a basal Chelicerate, *Limulus polyphemus,* was queried for the presence of COL4A genes. Partial COL4A1 and COL4A2 homologs were identified in this species; however, contigs containing these COL4A genes were clearly fragmented. For example, the likely 3’ end of the COL4A2 homolog was found less than 1 kb from the end of the K1894497.1 contig. Based on these results, it was not possible to confidently assign a COL4A locus structure in this species.

A genome for the basal Ecdysozoan group Scalidophora is available for *Priapulus caudatus*. A contig from this assembly (KQ714952.1) contains both COL4A1 and at least two exons from the 5’ end of COL4A2 in the head-to-head orientation. No other COL4 genes are apparent in this species. This result strongly suggests that Scalidophora, like Drosophila, have the paired COL4A⟨1|2⟩ locus without apparent COL4S genes. The phylogenetic evidence within Ecdysozoans suggests that the unlinked locus structure is common within nematodes, and rare in other arthropods, but is found in spiders where the linked COL4A⟨1|2⟩ locus is also found. This suggests that within the Ecdysozoan lineage the presence of COL4S loci have arisen at least twice and the COL4A⟨1|2⟩ locus can be lost as well. Notably, all genomes within Ecdysozoa have either or both the linked and unlinked loci; no organisms have been identified that lack COL4A genes.

**COL4A gene structure in Spiralia**

Both the linked COL4A⟨1|2⟩ and unlinked locus structures are present in Ecdysozoa. To determine whether both the linked and unlinked locus structures are common more broadly in Protostoma, we turned to the Ecdysozoan paraphyletic group Spiralia. In the Lophotrochozoan, *Lingula anatina,* COL4A1 and COL4A2 transcription units are found in the typical linked configuration. However at least two other COL4A genes are

found in the genome.  These genes are found on distinct scaffolds and are apparently not linked to each other or to the COL4A⟨1|2⟩ locus. Similarly, within the sister clade Nemertea, the species *Phoronis australis* has both the COL4A⟨1|2⟩-linked structure and the unlinked COL4S1 and COL4S2-like structures.

More broadly spiralian species including Molluscs, Platyhelminthes and Rotifer species show diverse COL4A gene configuration(s) suggesting that the ancestral Spiralia species may have contained both the paired and unpaired gene configuration. Notably, schistosomes have the paired COL4A⟨1|2⟩ locus which has an additional COL4S gene immediately downstream of the COL4A1 putative homolog. This species also has a single COL4S gene that is unlinked.

**Six COL4A genes are expressed in the shark kidney**

Based on our metagenomics analysis of the COL4A gene composition of vertebrates, we sought to confirm the presence of six COL4A genes in sharks and determine which of these genes were expressed in the kidney of these animals. At the time this study began genomic data for sharks were unavailable. Therefore, we isolated total RNA from the dogfish kidneys and performed next generation RNA sequencing (RNA-Seq). De novo assembly of the dogfish transcriptome was performed and used to design specific primers to PCR COL4A transcripts that were subsequently sequenced to confirm the transcriptomics results. Results of these experiments demonstrated that all six chains (COL4A1 through COL4A6) were expressed in the dogfish kidney.

To analyze Collagen IV chain composition on the protein level, we solubilized non-collagenous NC1 domains from dogfish tissues by collagenase digestion. The NC1 domains were then purified using a previously described protocol involving size exclusion chromatography (Boudko *et al.*, 2018). Characterization of the NC1 domains from the shark and hagfish species by SDS/PAGE revealed the presence of dimer subunits, indicative of sulfilimine cross-links (Fidler *et al.*, 2014; McCall *et al.*, 2014). NC1 domains isolated from hagfish and shark run as dimers and monomers similar to the mammalian control (Figure 9). However, unlike in mammals where the NC1 domain dimers run as two distinct bands, there are three prominent dimer bands in shark.

To investigate tissue distribution of the different Collagen IV isoforms in shark, we performed two-dimensional non-equilibrium pH gel electrophoresis (2D-NEPHGE). NC1 domains isolated from the shark ocular lens were labeled with the Cy2 fluorescent dye while NC1 domains isolated from shark kidneys were labeled with Cy3. Samples conjugated with fluorescent dyes were mixed and resolved on 2D-NEPHGE (Figure 5B). The two fluorescent dyes produced distinct maps indicating tissue specific differences in the NC1 domain composition. Individual spots from separate 2D-NEPHGE gel of shark kidney samples were excised and analyzed with high resolution liquid chromatography and tandem mass spectrometry (LC-MS/MS) at Vanderbilt's Proteomics Core. A custom dogfish protein database, based on the transcriptomics data, was used to identify which Collagen IV-NC1 domains were present in each 2D spot.  Based on spot identification, we labeled the chains in (Supplemental Figure 1, Figure 9). Intriguingly, the α5 and α6 chains of Collagen IV were found predominantly in shark eye while all six chains (α1 through α6) were identified in the kidney.

## Discussion

The basement membrane is a conserved proteinaceous structure of multicellular animals. A key component of the basement membrane is Collagen IV, a heterotrimeric protein, the protomer, consisting of a long central collagen domain flanked by N-terminal and C-terminal domains which mediate Collagen IV-Collagen IV interactions. The network of Collagen IV proteins, referred to as a scaffold, is present in all known multicellular animals and the protein encoding genes are highly conserved in their peptide encoding, their domain structure and even in the genomic structure. Within diverse animal species from Ctenophores, Placozoans and Cnidarians to Bilaterians including insects and vertebrates, the Collagen IV encoding genes (the COL4 genes) commonly occur as a head-to-head paired gene. In some species however, including the nematode *C. elegans*, the Collagen IV genes are not found in head-to-head gene pairs and are instead individual, unlinked genes (which we refer to as COL4S in contrast to the paired genes which are referred to as COL4A genes). In this study, we have investigated the evolution of the COL4 genes with particular emphasis placed on the evolution of the three pairs of these genes that occur in the vertebrate lineage.

While head-to-head gene orientations are not uncommon in metazoan animals, it is highly unusual for this gene organization to be highly conserved. For COL4 this locus structure is conserved in species spanning all multicellular eumetazoan animals. This result suggests that these genes share cis-acting regulatory sequences which assist in the coordination of gene expression, likely facilitating the expression of these genes and maintaining expression at the appropriate stoichiometry. However, exceptions to this gene arrangement are observed in species found in many clades. For example, in nematodes the COL4 genes appear as two single, unlinked genes and represent the prototype of the COL4S genes. Nematode genomes are unusual in bilaterian animals in having a fraction of genes occurring in cistron-like, head-to-tail, gene arrangements. However, the COL4 genes, whose expression is coordinately regulated in other species, are not cistronic in *C. elegans*. Other examples of single or multiple COL4S genes are found in pre-vertebrate chordates, spiralia, and some Ecdysozoans. In Ctenophores and in Schistosoma the COL4A gene pair has additional copies of the COL4 gene lying downstream of the COL4A gene or genes; we refer to these linked gene clusters as COL4T to distinguish them from both the COL4A and COL4S gene structures. In many cases, though not in all, with locus structures that differ from the canonical COL4A locus structure the structural features of the protein coding regions of the COL4 genes are distinct from those observed in COL4A genes. For example the COL4T genes in the Ctenophore *Mnemiopsis leidyi* have an additional non-collagenous domain adjacent to the C-terminal NC1 domain. Similarly, the length of the collagenous domain in the COL4S genes observed in the lancelet is generally shorter than that observed in vertebrate collagenous domains of COL4 genes.

The centrality of the head-to-head gene arrangements for nearly all COL4A genes identified throughout metazoan species raises the question of how such a gene arrangement might occur. Head-to-head gene duplications have been observed in other eukaryotic genomes (Ouellette *et al.*, 1991; Butler *et al.*, 1995; Qin and Cohen, 2000; Tanaka and Yao, 2009; Hermetz *et al.*, 2014). The rearranged chromosomes resulting in the head-to-head arrangement are often observed under conditions of strong selection for the gene duplication (Ouellette *et al.*, 1991; Tanaka and Yao, 2009; Marotta *et al.*, 2017). The analysis of these systems provides a plausible mechanism by which the COL4A gene pair could arise in evolution. Generation of the head-to-head COL4A gene pair would likely involve a mechanism similar to the breakage-fusion-bridge model (Tanaka and Yao, 2009; Hermetz *et al.*, 2014; Marotta *et al.*, 2017). In this model, a chromosome break is repaired during mitosis to generate an inverted duplication close to the chromosomal breakpoint (Supplemental Figure 2).

The most diverged group of multicellular animals in terms of the structure of the COL4 loci are found in Spiralia; in Spiralia many species have multiple lone COL4S genes as well as COL4A gene pairs. Perhaps the most divergent pattern is found in Schistosoma which has a COL4S gene immediately downstream of the COL4A gene pair as well as a second COL4S gene that is not tightly linked to the COL4A locus. While Spiralian genomes show conserved syntenies across large spans of their genome (macrosynteny), some

species within Spiralia show considerable loss of local synteny. For example, the Hox gene cluster is fragmented in *Helobdella robusta* (Simakov *et al.*, 2013) similar to the COL4A locus in this species. The regulation of COL4 gene expression is poorly understood and has been little studied; the diversity in locus structure in the Spiralia suggests opportunities to identify key COL4 regulatory sequences using informatics-based approaches as well as to investigate the functions of both the COL4A and COL4S genes.

**Vertebrate specific duplications of the COL4A genes**

While other metazoans have multiple copies of the COL4 genes, it is rare to observe duplication of the COL4A gene pair. This is striking as gene duplication and divergence is commonly observed in many gene families. In the case of the hox gene cluster, these genes often remain within a gene cluster facilitating their regulation. However, the vertebrate lineage contains three unlinked copies of the COL4A gene locus. The evolutionary origin of these gene pairs is of great interest as the origin of these genes may provide essential clues to the functional differences between these COL4A gene pairs. Remarkably, the unusual locus structure is retained in each gene pair and, more strikingly, a larger synteny that is shared between all three COL4A gene pairs. This is the presence of an IRS gene homolog adjacent to each COL4A gene pair (Al-Salam and Irwin, 2017). This observation suggested that synteny at the COL4A loci in vertebrates may lead to a model for the evolution of these unique loci.

**Synteny of the COL4A and IRS paralogs in the vertebrate lineage**

The syntenic relationship of the COL4A gene pairs and IRS paralogs has been previously noted(Al-Salam and Irwin, 2017). In all vertebrate species investigated, we observe that downstream of the COL4A1 gene a paralog of the IRS family, IRS2, is found. Similarly, downstream of each COL4A4 gene is the gene RHBDD1 and IRS1 in all tetrapods and sarcopterygians investigated. Interestingly, even in Amphibians which have lost the COL4A〈3|4〉 gene pair, the amphibian IRS paralog has maintained synteny with DOCK10, NYAP2, and RHBDD1. An IRS homolog, IRS4, is also found immediately downstream of COL4A5. The synteny between each COL4A gene and its IRS homolog does not extend further between the COL4A loci. There are no other clear paralogous gene pairs between the three COL4A gene loci; however, each COL4A locus exists within a larger syntenic region. We have investigated the syntenic relationships of each COL4A gene pair and note the generally high conserved synteny of the COL4A〈1|2〉 locus while both COL4A〈3|4〉 and COL4A〈5|6〉 show less syntenic gene conservation. These syntenies are conserved from the cartilaginous fishes to mammals. The synteny is observed in fishes from the early vertebrate cartilaginous fishes to sarcopterygian fishes (primitive tetrapods), but is less conserved in the teleostean lineages. Notably, teleostean fishes have undergone a genome duplication which likely led to large-scale genome rearrangements. For example, downstream of COL4A3, there has been a significant genome rearrangement in teleostean fishes resulting in the presence of multiple copies of a somatostatin homolog. While the genome duplication present in teleosts may explain some of the syntenic diversity observed in COL4A〈3|4〉 and COL4A〈5|6〉 the ray-finned, but non-teleostean gar fish which diverged prior to the whole genome duplication, also has extensive breaks in these syntenic regions. The synteny downstream of COL4A4 is well preserved in zebrafish, NYAP2, IRS1, and RHBDD1 are all present in this species. The relatively rapid changes to the COL4A〈3|4〉 syntenic region may be consistent with the incredible diversity present in teleostean fishes and may also be affected by the highly diverse osmotic environments in which these animals live.

Syntenies stretching over small clusters of genes have been observed for many genes, perhaps most famously for the Hox complex. These conserved cluster are thought to occur due to the pairing of a gene regulatory sequence and several protein coding genes which are largely co-regulated.

**Evolutionary origins of the COL4A〈3|4〉 and COL4A〈5|6〉 loci**

Examination of the jawed vertebrate COL4A syntenic loci described above did not reveal the origins of the COL4A loci duplications. To determine the evolutionary origins of these duplications we extended our investigation into the cyclostome lineage and from there into the early chordate lineages. Lampreys which

have both COL4A⟨1|2⟩ and COL4A⟨3|4⟩, but not COL4A⟨5|6⟩, also have IRS paralogs adjacent to both COL4A loci. Facilitating identification of the COL4A⟨1|2⟩ gene pair is the existence of the MYO16 gene downstream of COL4A1. This result strongly suggests that the original gene duplication resulted in COL4A⟨1|2⟩ and COL4A⟨3|4⟩. MYO16 is also found downstream of COL4A1 and the IRS paralog commonly in vertebrates, for example in Gar, *Coelacanth* and terrestrial vertebrates. Similarly NYAP2 is found downstream of COL4A3 in lamprey and is found in the COL4A⟨3|4⟩ syntenic group of jawed vertebrates, albeit on the other side of the COL4A⟨3|4⟩ gene pair. Consistent with the hypothesis that the lamprey COL4A⟨1|2⟩ and COL4(3,4) share a common origin, the FARP1 and FARP2 genes are found upstream of COL4A2. In hagfish, the IRS paralog adjacent to the COL4A⟨1|2⟩ gene pair has apparently been lost. Intriguingly, in the early chordates *Branchiostoma floridae* and the tunicates, a single COL4A⟨1|2⟩ locus is observed, though in these species COL4S genes are also present. Together these results suggest that in an ancestral chordate the COL4A⟨1|2⟩ gene pair occurred adjacent to the ancestral IRS paralog. This region was duplicated generating COL4A⟨3|4⟩. Subsequently, in cartilaginous fishes the COL4A⟨5|6⟩ gene pair appeared. Scars of the original locus duplication giving rise to COL4A⟨3|4⟩ appear in the synteny observed with the IRS paralogs and the COL4A pair specific MYO16 and NYAP2 syntenies.

Careful examination of MYO16 and NYAP2 suggests a speculative hypothesis further confirming that COL4A⟨3|4⟩ and COL4A⟨1|2⟩ together with an IRS gene share a common origin. MYO16 and NYAP2 appear to be derived from each other: specifically, the NYAP2 is highly homologous to the C-terminal domain of MYO16 (Yokoyama *et al.*, 2011; Al-Salam and Irwin, 2017). This suggests that NYAP2 and MYO16 share a common evolutionary origin: either NYAP2 formed from the C-terminal half of the MYO16 gene during the duplication of the COL4A locus or, alternatively, the NYAP2 gene inserted into an ancestral MYO16-like gene generating the new MYO16 gene. Confounding any analysis of this sort is the low level of synteny in the cyclostome, tunicate and lancelet genomes near the COL4A genes. The lack of synteny makes evolutionary trajectories uncertain, moreover, cyclostomes have undergone at least one additional whole genome duplication after they diverged from the chordate lineage leading to vertebrates . These whole genome duplications likely led to recombination events which scrambled the synteny.

**COL4A⟨3|4⟩ deletion in Amphibia**

Unexpectedly, we found that the COL4A⟨3|4⟩ gene pair was not present in the genome of either frog species we investigated. To determine whether this was a common feature to amphibians, we investigated the genomes from each of the 3 extant classes of Amphibia: Salienta (frogs and toads), Caudata (salamanders and newts), and Gymnophiona (caecilians). In each genome examined we found that COL4A⟨3|4⟩ was missing. Moreover in each genome the synteny group downstream of COL4A4 was intact. The synteny group downstream of COL4A3 was found intact and adjacent to the COL4A4 synteny group in the caecilians and axolotl genomes. In other amphibians a variably sized intervening sequence containing a variable number of genes was identified between these synteny groups. Because caecilians are thought to have diverged relatively early in amphibian evolution (San Mauro, 2010; Pyron and Wiens, 2011; San Mauro *et al.*, 2014), it is not surprising that this locus shows an apparent ancestral-like sequence. It is notable that the deletion of the COL4A⟨3|4⟩ loci is conservative, neighboring genes are unaffected by the deletion. It is interesting to note that the loss of COL4A⟨3|4⟩ must have occurred after the divergence of the amphibian lineage from that leading to reptiles and mammals, the amniotes.

## Experimental Procedures

***Data Availability.*** Data deposition: Dogfish Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GKOS00000000. The version described in this paper is the first version, GKOS01000000. Proteomics data has been submitted to the ProteomeXchange, projects PXD0441912 and PXD042111.

***Next Generation RNA Sequencing.*** Total RNA isolated from animal tissues was sequenced at the Vanderbilt Technologies for Advanced Genomics Core Facility. *De novo* assembly of transcriptomes was performed using Velvet/Oases, and Trinity software packages with default settings (Zerbino and Birney, 2008; Grabherr *et al.*, 2011; Schulz *et al.*, 2012). The accuracy of de novo assembly was checked in a parallel next generation RNA-Seq experiment using RNA from mouse PFHR9 cells. De novo assembled transcripts were used to generate BLAST databases to search for Collagen IV hits using tblastn (Altschul *et al.*, 1990) with e-value cutoff set to $10^{-15}$. Multiple sequence alignments and conserved domain searches were performed with the Geneious v5-6 software (Biomatters).

***Cloning Dogfish collagen IV*** To confirm accuracy of dogfish COL4 cDNA sequences obtained from RNAseq-based sequences, we performed a series of RT-PCR cloning experiments using primers designed to NGS-detected COL4 candidates. RNA was prepared using the QIAGEN One-Step RT-PCR Kit and PCR products were cloned using the QIAGEN PCR Cloning Kit.

***Isolation, purification, and analysis of collagen IV NC1 hexamers*** Tissues were frozen in liquid nitrogen, pulverized in a mortar and pestle and then homogenized in 2.0 ml $g^{-1}$ digestion buffer and 0.1 mg $ml^{-1}$ Worthington Biochemical bacterial collagenase and allowed to digest at 37°C, with spinning for 24 hr. Liquid chromatography purification of solubilized NC1 varied by species based on protein yield. All ctenophore NC1s were purified by gel-exclusion chromatography (GE Superdex 200 10/300 GL). For reduction and alkylation of collagen IV NC1 hexamers, fractions containing high-molecular-weight complex from size-exclusion chromatography were concentrated by ultrafiltration and reduced in TBS buffer with various concentrations of DTT. After incubation for 30 min at 37°C, samples were alkylated with twofold molar excess of iodoacetamide for 30 min at room temperature in the dark. After mixing with SDS loading buffer, samples were heated for 5 min in a boiling water bath and analyzed by non-reducing SDS-PAGE. Collagenase-solubilized NC1 hexamers were analyzed by one-dimensional SDS-PAGE in 12% *bis*-acrylamide mini-cells with Tris-Glycine-SDS running buffer. Sample buffer was 62.5 mM Tris-HCl, pH 6.8, 2% SDS (w/v), 25% glycerol (w/v), 0.01% bromophenol blue (w/v). Western blotting of SDS-dissociated NC1 hexamer was developed with JK-2, rat monoclonal antibody (kindly provided by Dr. Yoshikazu Sado, Shigei Medical Research Institute, Okayama, Japan). All Western blotting was done with Thermo-Scientific SuperSignal West Femto chemiluminescent substrate and digitally imaged with a Bio-Rad GelDoc. Two-dimensional NEPHGE electrophoresis was performed according to the original protocol developed (O'Farrell *et al.*, 1977) with slight modifications developed in Hudson lab and used successfully to separate NC1 domains of Collagen IV (Timoneda *et al.*, 1990).

***Proteomics.*** All proteomics experiments were done at Vanderbilt's MSRC Proteomics Core facility. Major protein spots from Coomassie stained 2D-NEPHGE separated dogfish Collagen IV preparations isolated from various tissues were cut out of the gel and proteins were identified and LFQ-quantified using MaxQuant (Cox and Mann, 2008; Cox *et al.*, 2014). The heatmap was created with heatmapper.ca based on LFQ-data (Babicki *et al.*, 2016).

***Synteny analysis.*** Well-characterized genomes were selected from those appearing in the NCBI and Ensembl databases (Rangwala *et al.*, 2021; Martin *et al.*, 2023) and COL4 gene sequences were identified based on text and blastp (Altschul *et al.*, 1990) searches. Syntenic genes were then manually identified and their orientation and order on the chromosome annotated and plotted with matplotlib (Hunter, 2007). Protein domains were mapped to their ORFs using the Conserved Domain Database (Lu *et al.*, 2020). The COL4 gene

phylogeny was derived from the NCBI Taxonomy (Federhen, 2012; Schoch *et al.*, 2020) and plotted using iTOL (Letunic and Bork, 2021).

## Author Contributions

Patrick S. Page-McCaw acquired and analyzed the data, generated the synteny and metagenomics figures and wrote the draft manuscript;

Carl E. Darris conceived and initiated the synteny analysis project;

Sergei Chetyrkin performed the hagfish and shark transcriptomics, cDNA cloning, Collagen IV protein preparation and NC1 domain generation, mass spectrometry, and analysis;

Elena N. Pokidysheva acquired data, generated figures, wrote and edited with particular contributions to the hagfish and shark biochemistry project;

Julianna Gallup assisted in the genomics analysis, developed automated analysis tools to facilitate synteny analysis;

Julie K. Hudson supervised and managed the Aspirnaut program and students;

Billy G. Hudson conceived and designed the work, interpreted data and substantively wrote and edited the work.

## Funding

This work was supported by the National Institutes of Health grants R01DK018381 to B.G.H.. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The Aspirnaut high-school students were supported by Springer-Lu Family Foundation and the undergraduate students by NIH grant R25DK096999 to B.G.H.

## Abbreviations

AS, Alport syndrome; BM, basement membrane; COL4A, the Collagen IV gene where found in the head-to-head orientation; COL4S, the Collagen IV gene where found as a single gene; GBM, glomerular basement membrane.

## Conflicts of Interest

The authors declare that there are no conflicts of interest.

## Bibliography


Al-Salam, A, and Irwin, DM (2017). Evolution of the vertebrate insulin receptor substrate (Irs) gene family. BMC Evol Biol 17, 148.

Altschul, SF, Gish, W, Miller, W, Myers, EW, and Lipman, DJ (1990). Basic local alignment search tool. J Mol Biol 215, 403–410.

Babicki, S, Arndt, D, Marcu, A, Liang, Y, Grant, JR, Maciejewski, A, and Wishart, DS (2016). Heatmapper: web-enabled heat mapping for all. Nucleic Acids Res 44, W147–W153.

Borchiellini, C, Coulon, J, and Le Parco, Y (1996). The function of type IV collagen during Drosophila embryogenesis. Rouxs Arch Dev Biol 205, 468–475.

Boudko, SP et al. (2023). Collagen IV of basement membranes: III. Chloride pressure is a primordial innovation that drives and maintains the assembly of scaffolds. J Biol Chem, 105318.

Boudko, SP, Danylevych, N, Hudson, BG, and Pedchenko, VK (2018). Basement membrane collagen IV: Isolation of functional domains. Methods Cell Biol 143, 171–185.

Brown, KL, Cummings, CF, Vanacore, RM, and Hudson, BG (2017). Building collagen IV smart scaffolds on the outside of cells. Protein Sci 26, 2151–2161.

Burbelo, PD, Martin, GR, and Yamada, Y (1988). Alpha 1(IV) and alpha 2(IV) collagen genes are regulated by a bidirectional promoter and a shared enhancer. Proceedings of the National Academy of Sciences 85, 9679–9682.

Butler, DK, Yasuda, LE, and Yao, MC (1995). An intramolecular recombination mechanism for the formation of the rRNA gene palindrome of Tetrahymena thermophila. Mol Cell Biol 15, 7117–7126.

Coghlan, A, and Wolfe, KH (2002). Fourfold Faster Rate of Genome Rearrangement in Nematodes Than in *Drosophila*. Genome Research 12, 857–867.

Cox, J, Hein, MY, Luber, CA, Paron, I, Nagaraj, N, and Mann, M (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol Cell Proteomics 13, 2513–2526.

Cox, J, and Mann, M (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26, 1367–1372.

Emsley, J, Knight, CG, Farndale, RW, and Barnes, MJ (2004). Structure of the integrin alpha2beta1-binding collagen peptide. J Mol Biol 335, 1019–1028.

Federhen, S (2012). The NCBI Taxonomy database. Nucleic Acids Res 40, D136–D143.

Fidler, AL et al. (2014). A unique covalent bond in basement membrane is a primordial innovation for tissue evolution. Proceedings of the National Academy of Sciences 111, 331–336.

Fidler, AL, Boudko, SP, Rokas, A, and Hudson, BG (2018). The triple helix of collagens - an ancient protein structure that enabled animal multicellularity and tissue evolution. J Cell Sci 131.

Fidler, AL, Darris, CE, Chetyrkin, SV, Pedchenko, VK, Boudko, SP, Brown, KL, Gray Jerome, W, Hudson, JK, Rokas, A, and Hudson, BG (2017). Collagen IV and basement membrane at the evolutionary dawn of metazoan tissues. Elife 6.

Grabherr, MG et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29, 644–652.

Gupta, MC, Graham, PL, and Kramer, JM (1997). Characterization of alpha1(IV) collagen mutations in

Caenorhabditis elegans and the effects of alpha1 and alpha2(IV) mutations on type IV collagen distribution. J Cell Biol 137, 1185–1196.

Hermetz, KE, Newman, S, Conneely, KN, Martin, CL, Ballif, BC, Shaffer, LG, Cody, JD, and Rudd, MK (2014). Large inverted duplications in the human genome form via a fold-back mechanism. PLoS Genet 10, e1004139.

Hunter (2007). Matplotlib: A 2D Graphics Environment. Comput Sci Eng 9, 90–95.

Kefalides, NA (1973). Biochemical properties of human glomerular basement membrane in normal and diabetic kidneys. Adv Metab Disord 2, Suppl 2:167–177.

Labelle-Dumais, C, and Gould, DB (2012). COL4A1 and COL4A2 mutations and disease: insights into pathogenic mechanisms and potential therapeutic targets. Hum Mol Genet.

Letunic, I, and Bork, P (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49, W293–W296.

Lu, S et al. (2020). CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res 48, D265–D268.

Marotta, M, Onodera, T, Johnson, J, Budd, GT, Watanabe, T, Cui, X, Giuliano, AE, Niida, A, and Tanaka, H (2017). Palindromic amplification of the ERBB2 oncogene in primary HER2-positive breast tumors. Sci Rep 7, 41921.

Martin, FJ et al. (2023). Ensembl 2023. Nucleic Acids Res 51, D933–D941.

McCall, AS, Cummings, CF, Bhave, G, Vanacore, R, Page-McCaw, A, and Hudson, BG (2014). Bromine is an essential trace element for assembly of collagen IV scaffolds in tissue development and architecture. Cell 157, 1380–1392.

O'Farrell, PZ, Goodman, HM, and O'Farrell, PH (1977). High resolution two-dimensional electrophoresis of basic as well as acidic proteins. Cell 12, 1133–1141.

Ouellette, M, Hettema, E, Wüst, D, Fase-Fowler, F, and Borst, P (1991). Direct and inverted DNA repeats associated with P-glycoprotein gene amplification in drug resistant Leishmania. EMBO J 10, 1009–1016.

Pöschl, E, Pollner, R, and Kühn, K (1988). The genes for the alpha 1(IV) and alpha 2(IV) chains of human basement membrane collagen type IV are arranged head-to-head and separated by a bidirectional promoter of unique structure. The EMBO Journal 7, 2687–2695.

Pöschl, E, Schlötzer-Schrehardt, U, Brachvogel, B, Saito, K, Ninomiya, Y, and Mayer, U (2004). Collagen IV is essential for basement membrane stability but dispensable for initiation of its assembly during early development. Development 131, 1619–1628.

Pyron, RA, and Wiens, JJ (2011). A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. Mol Phylogenet Evol 61, 543–583.

Qin, Z, and Cohen, SN (2000). Long palindromes formed in Streptomyces by nonrecombinational intra-strand annealing. Genes Dev 14, 1789–1796.

Rangwala, SH et al. (2021). Accessing NCBI data using the NCBI Sequence Viewer and Genome Data Viewer (GDV). Genome Res 31, 159–169.

San Mauro, D (2010). A multilocus timescale for the origin of extant amphibians. Mol Phylogenet Evol 56, 554–561.

San Mauro, D, Gower, DJ, Müller, H, Loader, SP, Zardoya, R, Nussbaum, RA, and Wilkinson, M (2014). Life-history evolution and mitogenomic phylogeny of caecilian amphibians. Mol Phylogenet Evol 73, 177–189.

Schoch, CL et al. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database 2020.

Schulz, MH, Zerbino, DR, Vingron, M, and Birney, E (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28, 1086–1092.

Simakov, O et al. (2013). Insights into bilaterian evolution from three spiralian genomes. Nature 493, 526–531.

Tanaka, H, and Yao, M-C (2009). Palindromic gene amplification--an evolutionarily conserved role for DNA inverted repeats in the genome. Nat Rev Cancer 9, 216–224.

Timoneda, J, Gunwar, S, Monfort, G, Saus, J, Noelken, ME, and Hudson, BG (1990). Unusual dissociative behavior of the noncollagenous domain (hexamer) of basement membrane collagen during electrophoresis and chromatofocusing. Connect Tissue Res 24, 169–186.

Volff, J-N (2005). Genome evolution and biodiversity in teleost fish. Heredity 94, 280–294.

Yokoyama, K et al. (2011). NYAP: a phosphoprotein family that links PI3K to WAVE1 signalling in neurons. EMBO J 30, 4739–4754.

Zerbino, DR, and Birney, E (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18, 821–829.

# Figures and Figure Legends

Figure 1 Introductory slide

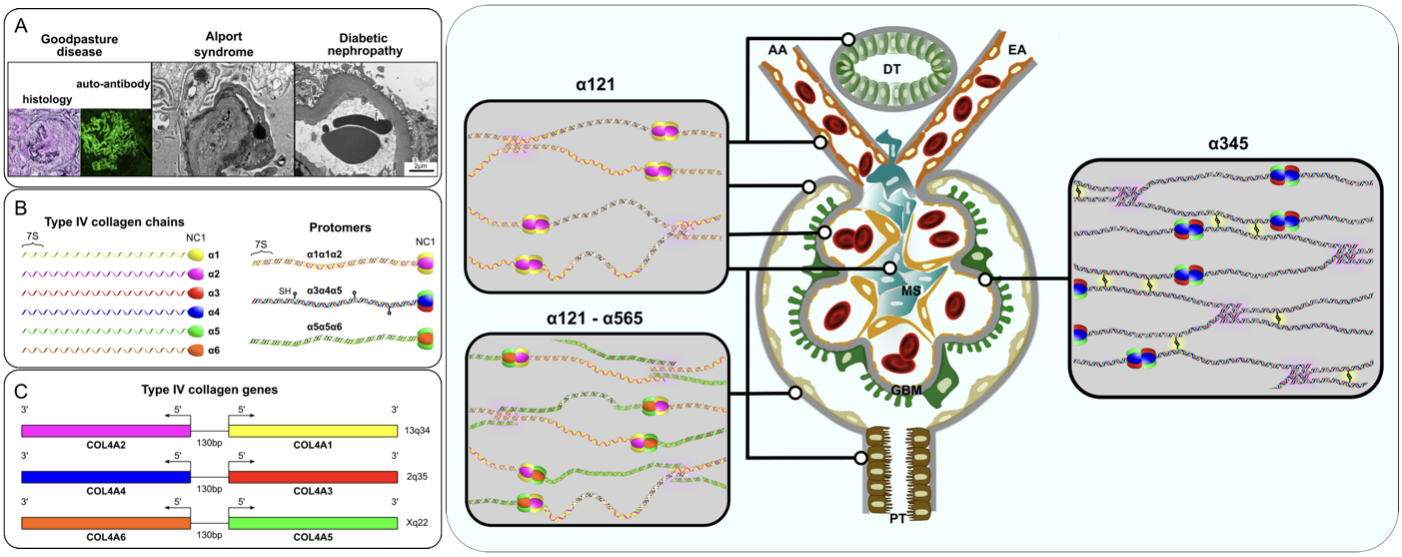



**Figure 1. Collagen IV composition of the glomerular basement membrane.**

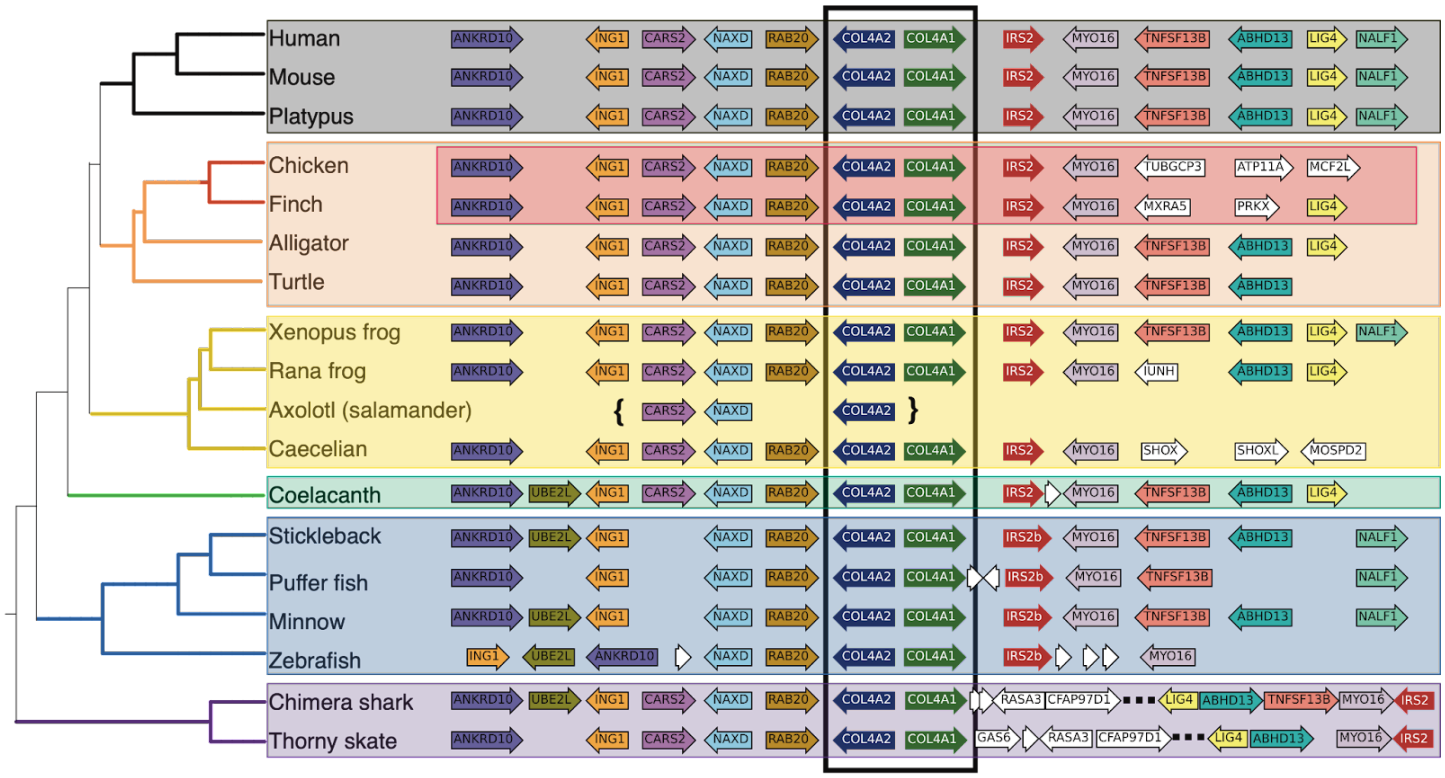
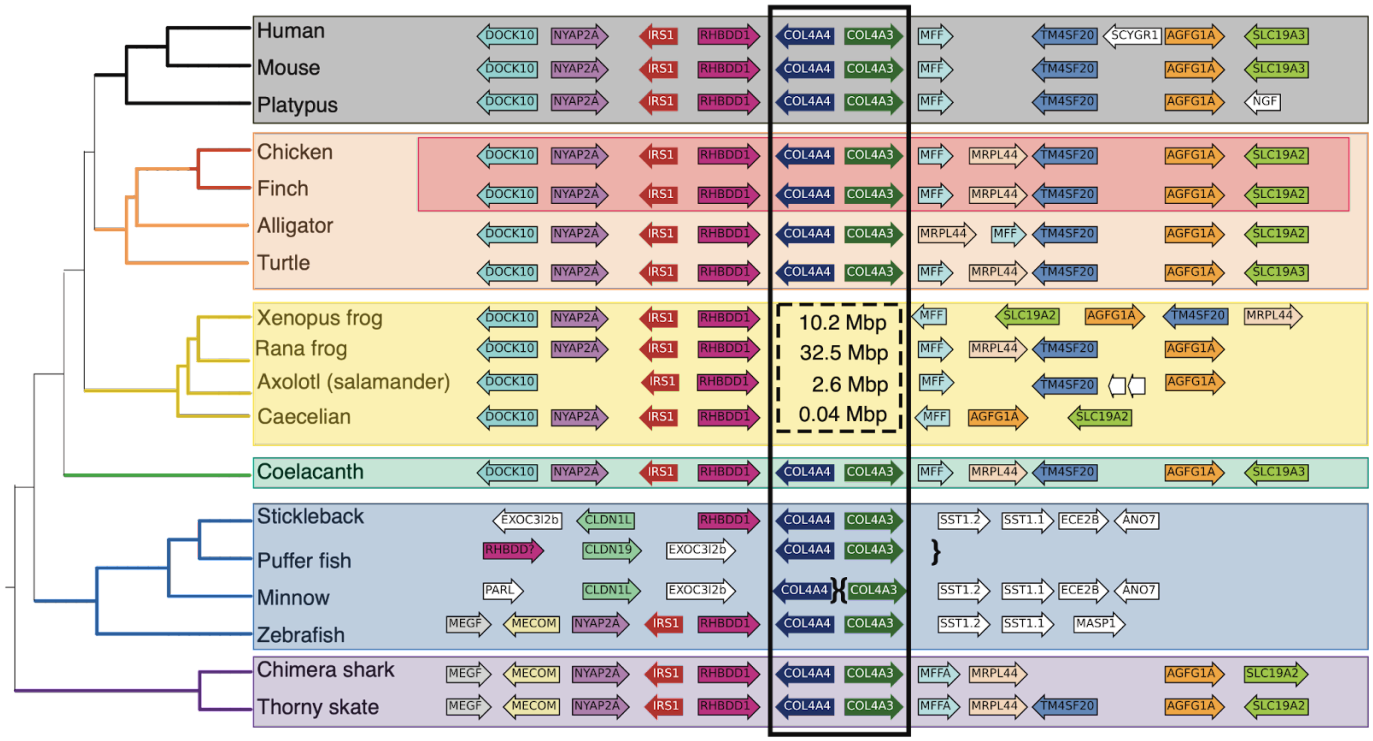



**Figure 2. Vertebrate synteny of the COL4A⟨1|2⟩ locus.** Genome sequences from selected and diverse vertebrates were examined surrounding the COL4A⟨1|2⟩ gene pair and protein coding genes and their orientations were identified. These are indicated as rows next to their gene names for each species. The standard ncbi phylogenetic tree is indicated to the right of each species. The axolotl genome is very large and the contig containing COL4A2 is shown; curly braces ({,}) mark the boundaries of the contig. Genes indicated in white are either low confidence gene annotations or genes that are not syntenic to other vertebrates. Note that there is an apparent chromosome inversion in the chondrichthyes fishes and intervening genes that are not shown are indicated by the ellipses (…).

Figure 3 COL4A⟨3|4⟩ Synteny



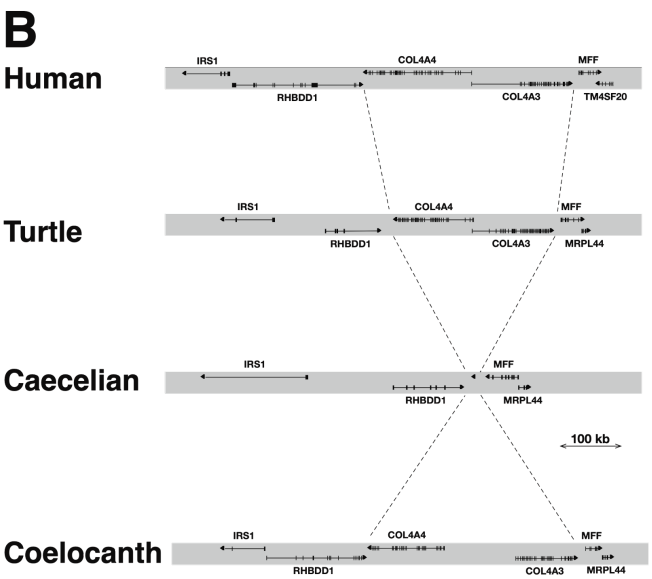




**Figure 3. Vertebrate synteny of the COL4A⟨3|4⟩ locus.** (A) Genome sequences from selected and diverse vertebrates were examined surrounding the COL4A⟨3|4⟩ gene pair and protein coding genes and their orientations were identified as for Figure 2. In zebrafish, the genes found downstream of COL4A3 in other vertebrates, including SLC19A3, AGFG1, MRPL4, and MFF are found upstream of col4a1, consistent with a chromosome inversion. The absence of the COL4A⟨3|4⟩ locus in amphibians is indicated with the dashed box. The size of the chromosome fragment in each species replacing the COL4A⟨3|4⟩ locus is indicated in the box. For both the Caecilian genome, a single gene encoded by a single exon encoding a kinase similar to TSSK6 in humans is found in this intergenic region, while in Axolotl no genes are observed in the intergenic gap, and in the two frog species multiple genes are found in the intergenic region. (B) The exon-intron structure of selected vertebrate genomes reveals the precise deletion of COL4A⟨3|4⟩ in the Caecelian genome: neither neighboring genes are affected save that MFF is inverted in some amphibians.

Figure 4 COL4A⟨5|6⟩ synteny

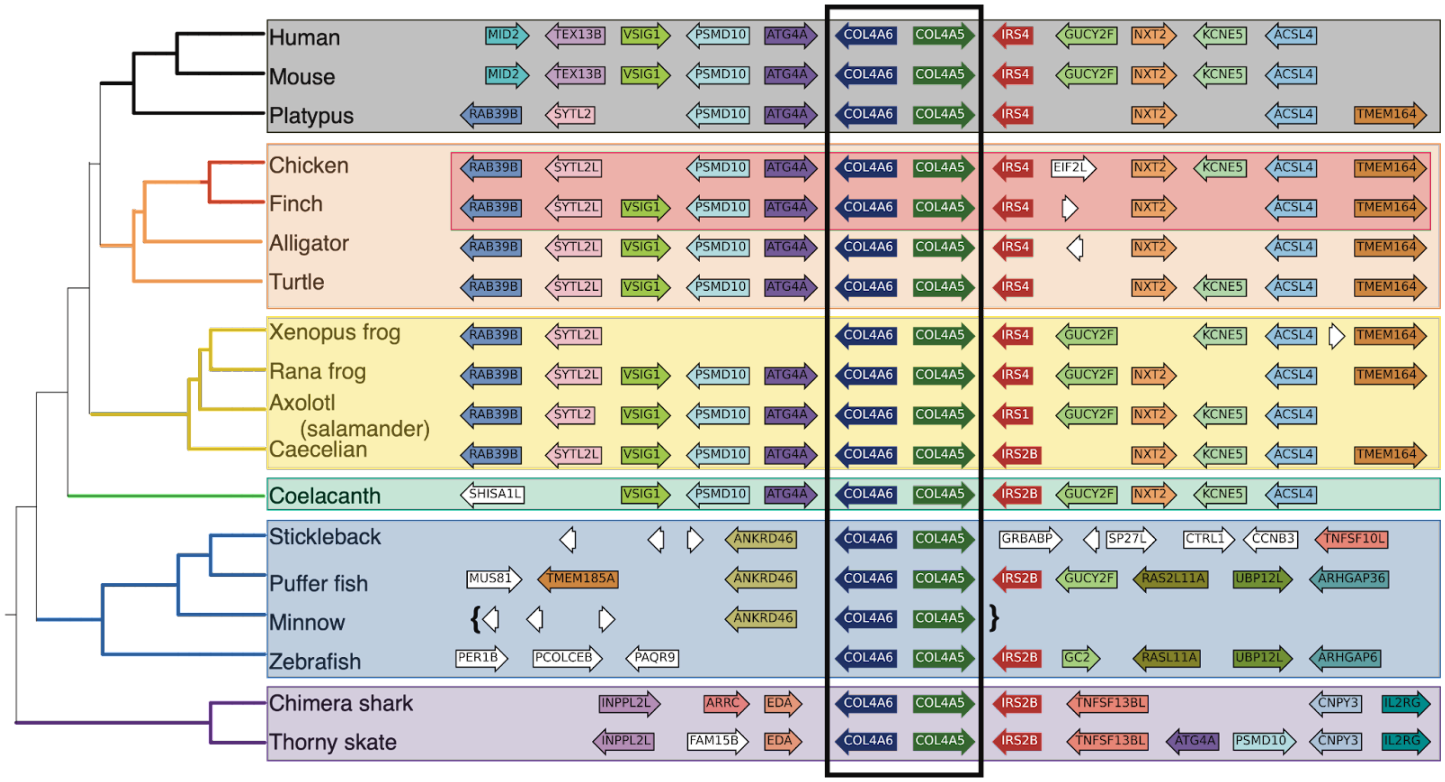
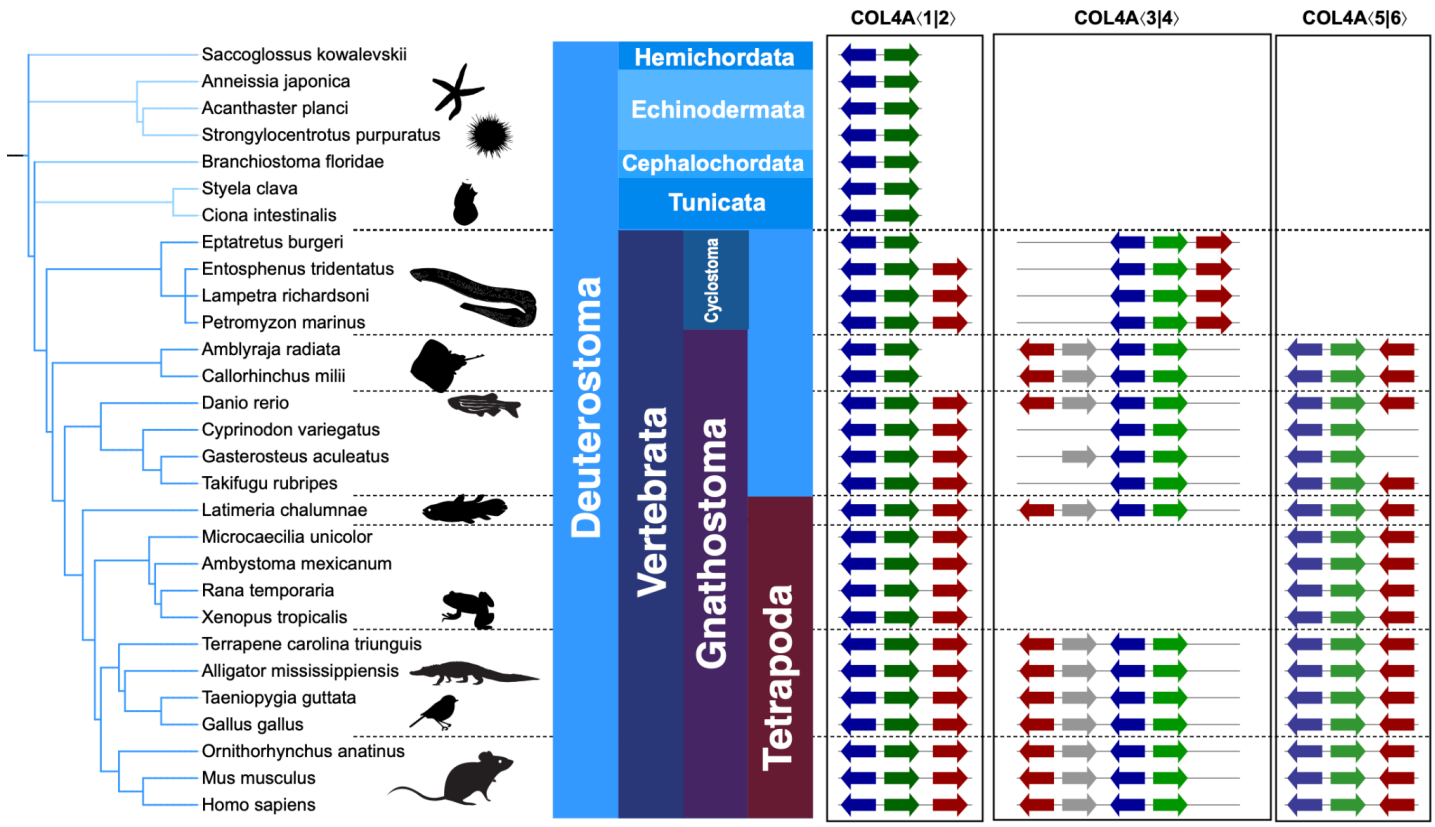



**Figure 4. Vertebrate synteny of the** COL4A⟨5|6⟩ **locus.** Genome sequences from selected and diverse vertebrates were examined surrounding the COL4A⟨5|6⟩ gene pair and protein coding genes and their orientations were identified as for Figure 2. Note that the synteny is not well maintained in the teleost (blue) lineage perhaps due to the teleostean specific whole genome duplication.

Figure 5A.COL4A genes in Deuterostome evolution

A



B

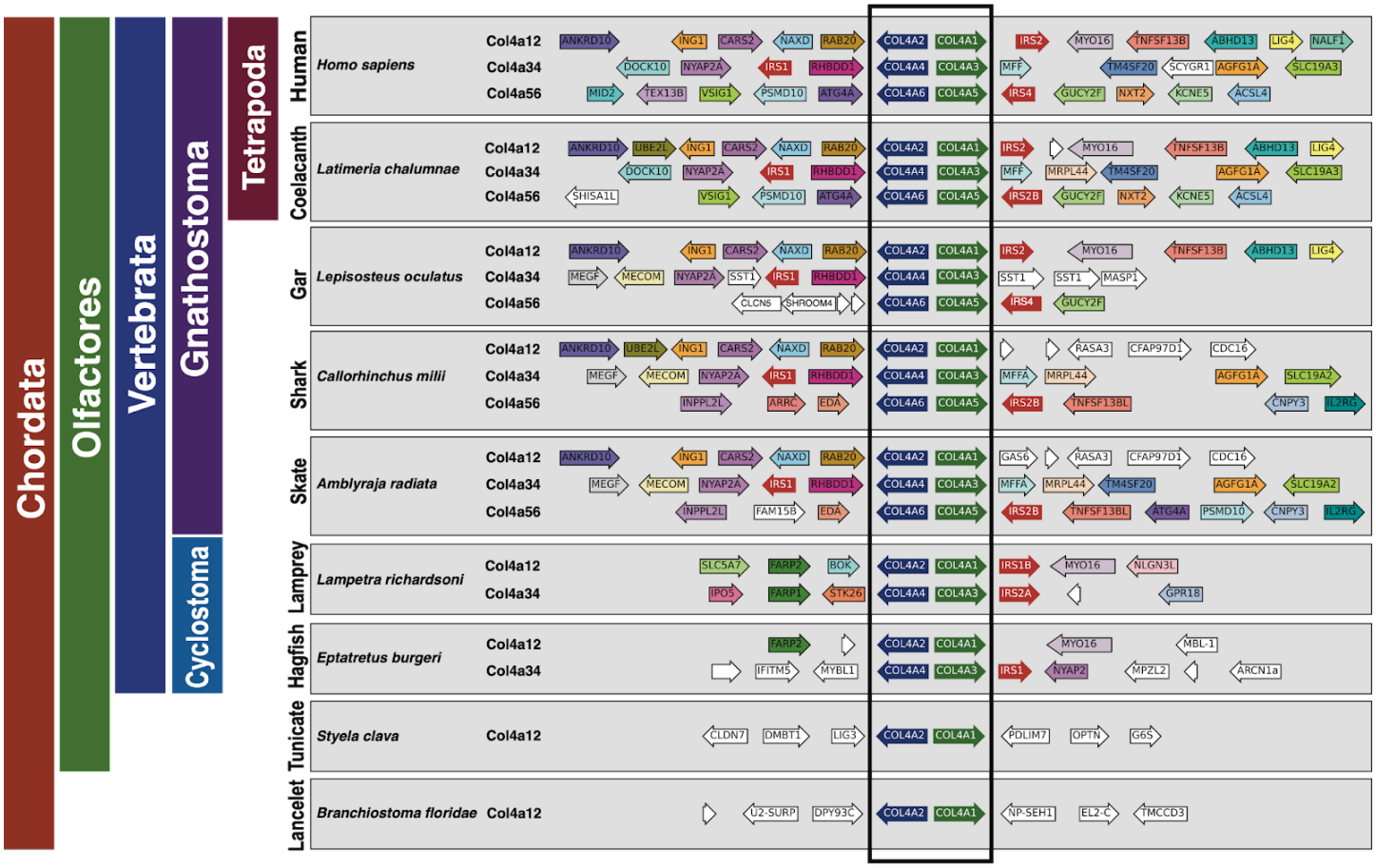



**Figure 5. COL4A gene locus evolution in the Deuterostomes.** We sampled the genomes from diverse phyla within Deuterostomes for instances of COL4A gene pairs and single COL4S genes in order to identify a possible ontogeny of the COL4A⟨1|2⟩ gene pair and its homologs. (A) Examination of the COL4A-IRS locus in Deuterostomes is shown, note that single copy COL4S genes are found in some ambulacraria species, but appear to be lineage specific innovations. (B) The microsyntenic regions at each COL4A⟨1|2⟩ gene pair (boxed) from selected species are shown. The first species to demonstrate synteny beyond the COL4A⟨1|2⟩ gene pair is the hagfish which is also the first species to demonstrate duplication of the COL4A⟨1|2⟩ gene pair into COL4A⟨3|4⟩. Lampreys demonstrate increased regional synteny as well as two COL4A gene pairs. Gnathostomes, except amphibia, show three COL4A gene pairs.

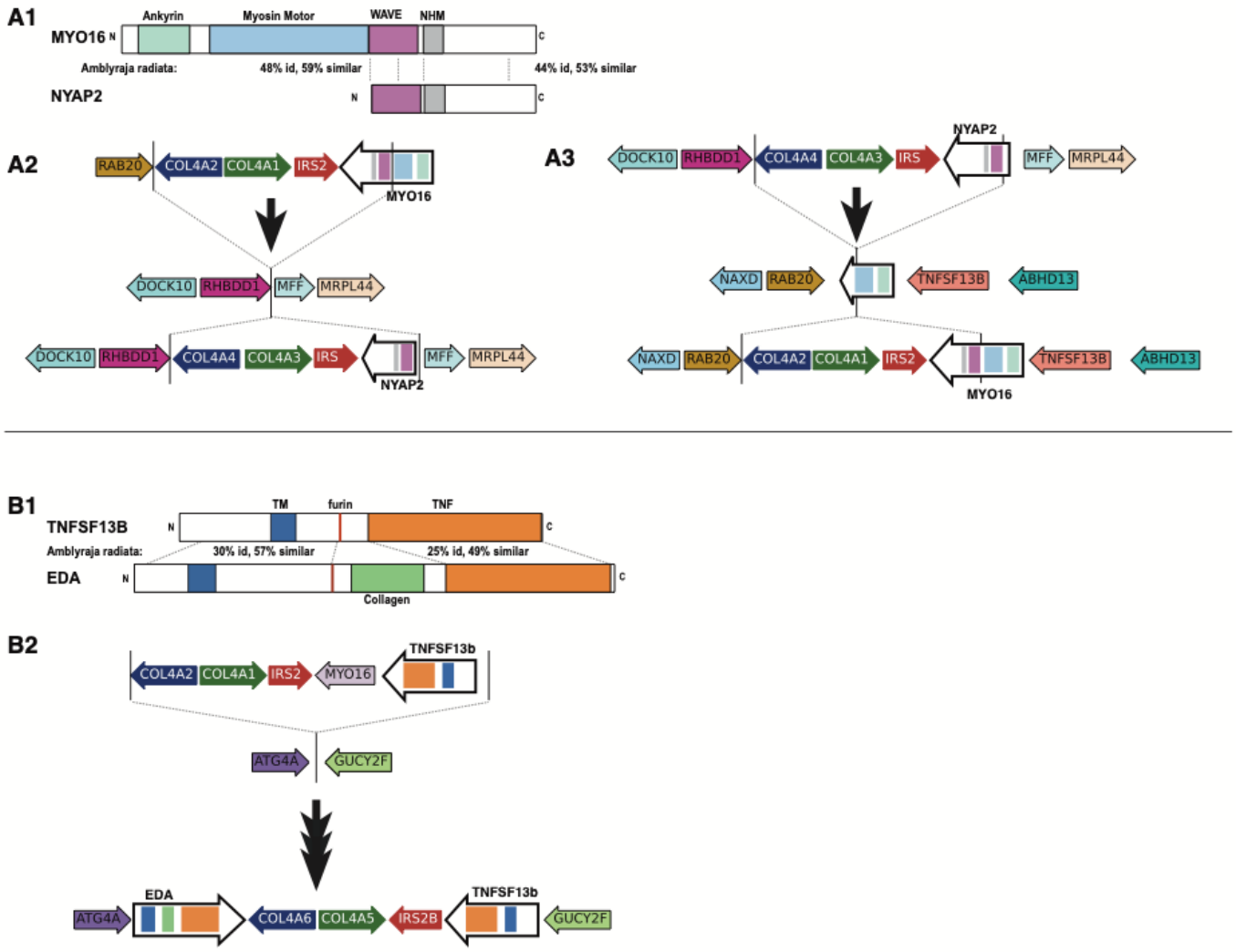



**Figure 6. Hypothesized chromosome rearrangements generating the COL4A⟨3|4⟩ and COL4A⟨5|6⟩ gene pairs.** The gene pair duplications and their subsequent genome reorganzations that generated COL4A⟨3|4⟩ and COL4A⟨5|6⟩ have left possible molecular scars in the form of new protein coding genes that are revealed by the synteny at these loci. While these hypothesized genome rearrangements are necessarily speculative, they suggest that both COL4A⟨3|4⟩ and COL4A⟨5|6⟩ are derived from COL4A⟨1|2⟩. The presence of protein coding genes that contain fragments of COL4A⟨1|2⟩ proximal genes in rearranged form (A1, B1) suggest that both COL4A⟨3|4⟩ and COL4A⟨5|6⟩ are derived from COL4A⟨1|2⟩. The C-terminus of NYAP2 is highly homologous to MYO16 suggesting either that the C-terminal fragment of MYO16 was carried with the copied COL4A⟨1|2⟩ IRS2 microsynteny into a new chromosomal region generating NYAP2 (A2). Alternatively, the microsyntenic region containing ancestral COL4A⟨1|2⟩ IRS2 and NYAP2 inserted into a myosin locus creating MYO16 from NYAP2 and the myosin gene (A3). Similar, rearrangements may have led to the generation of the highly similar gene pair found in chondrichthyes TNFSF13B and EDA (A1) from an ancestral COL4A⟨1|2⟩ IRS2 locus. Subsequent rearrangements in the vertebrate lineages left TNFSF13B on the COL4A⟨1|2⟩ microsynteny while EDA was retained with COL4A⟨5|6⟩ (B2).

Figure 7. all Col4 gene structurecures

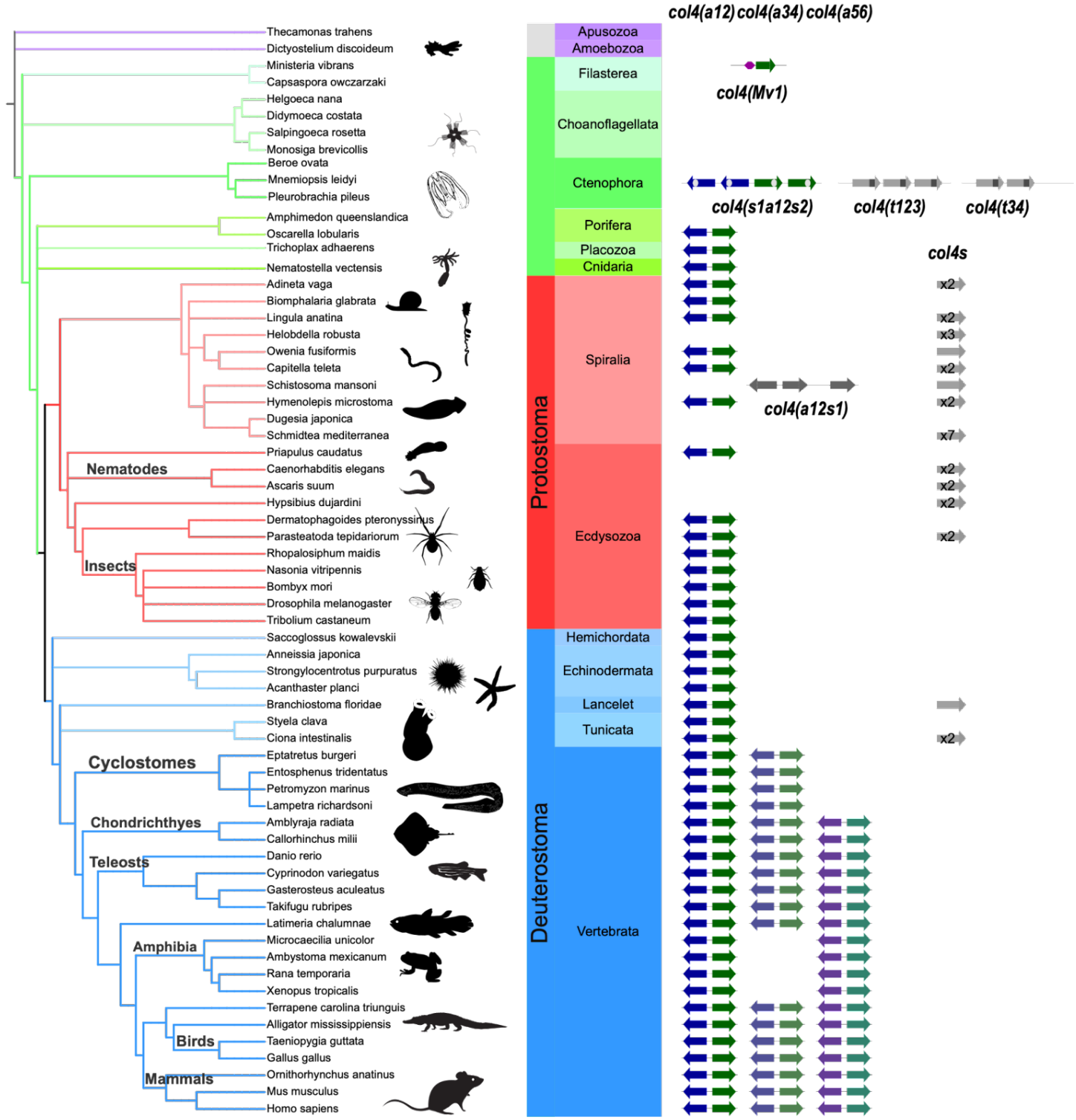



**Figure 7. COL4 gene pairs and single COL4 genes are found throughout evolutionary history in multicellular animals**, however, the COL4A⟨1|2⟩ gene pair is retained through most animals. As has been previously discussed (Fidler *et al.*, 2017) Ctenophores encode several COL4A like gene pairs that contain both an additional domain proximal to the NC1 domain, but also have tandem gene duplications of COL4 genes. The most primitive animal cell to encode a COL4 like gene, *Ministeria vibrans*, has a unique COL4

related structure and occurs as a single gene. It is unclear whether this gene is related to the ancestral COL4 locus directly or may represent the generation of a COL4 like gene through horizontal gene transfer from some other animal cell.

Figure 8 COL4 gene pair evolution

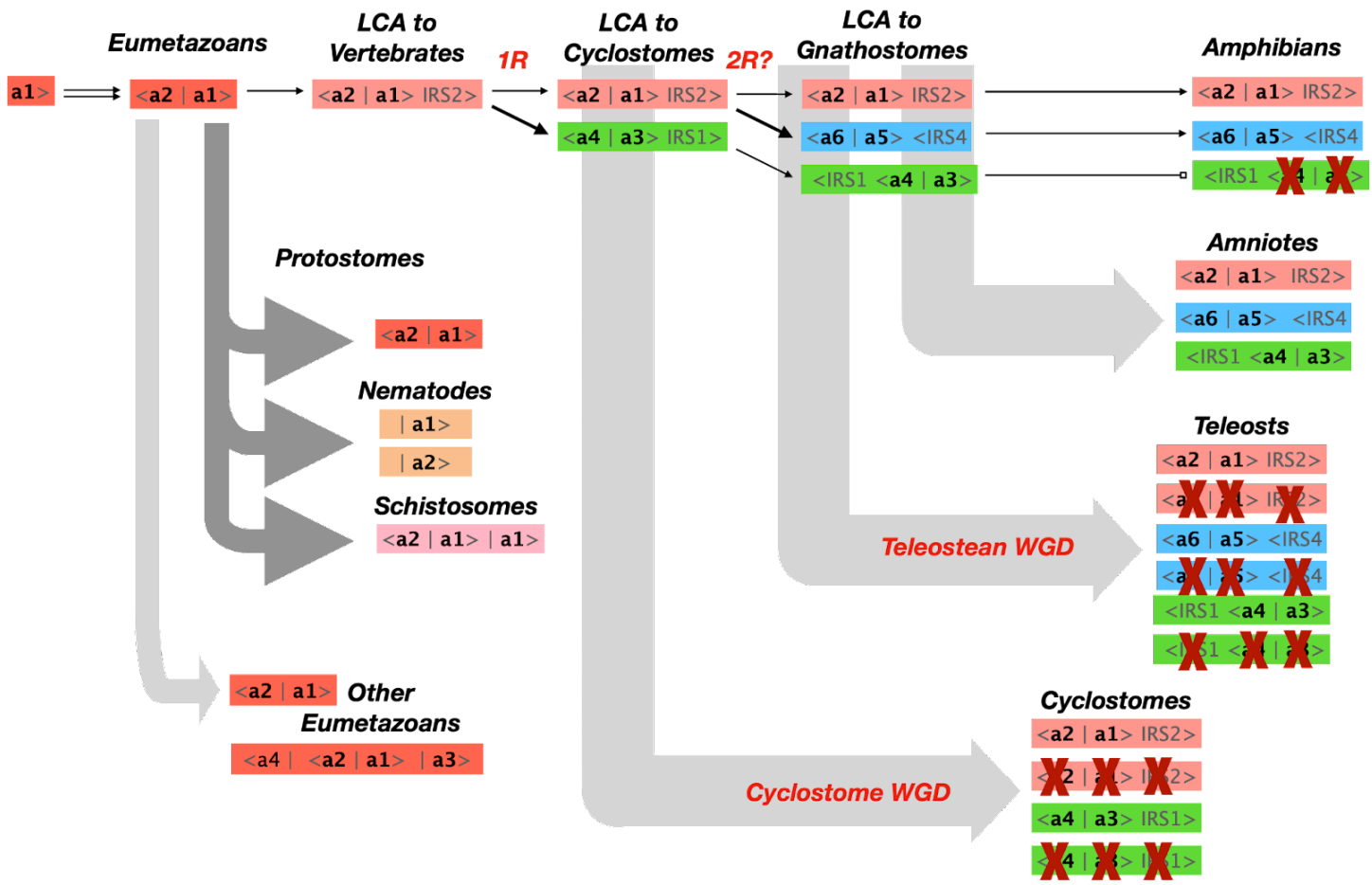



**Figure 8. Evolution of the COL4 loci with respect to the hypothesized whole genome duplications.** The analysis of many animal genomes suggests an evolutionary trajectory for the COL4 gene family. We propose that first a single COL4 gene evolved. This gene became the COL4A⟨1|2⟩ gene pair that is ancestral to all COL4 genes in an early multicellular animal. The COL4A⟨1|2⟩ gene pair was retained in many organisms including many protostomes, but single COL4S genes appeared in different lineages. Additionally, duplications within the COL4A⟨1|2⟩ locus are also observed. The last common ancestor to vertebrates appears to have had a linked COL4A⟨1|2⟩ and an IRS homolog. Likely during the first whole genome duplication (1R) that occurred in early vertebrate evolution the COL4A⟨1|2⟩ IRS locus was duplicated generating the COL4A⟨3|4⟩ IRS locus. We can observe this locus organization in cyclostomes; however cyclostomes have a lineage-specific duplication in which the novel COL4A genes were lost. The COL4A⟨1|2⟩ IRS locus was duplicated again in the evolution of gnathostomes resulting in the appearance of the COL4A⟨5|6⟩ IRS locus. It is unclear if this event correlates with the second vertebrate whole genome duplication (2R). One gnathostome branch, the teleost fishes, also had a whole genome duplication event, but the presumably three additional COL4A loci were subsequently lost. In amphibians, the COL4A⟨3|4⟩ gene pair was lost, but amniotes retain the three COL4A gene pairs.

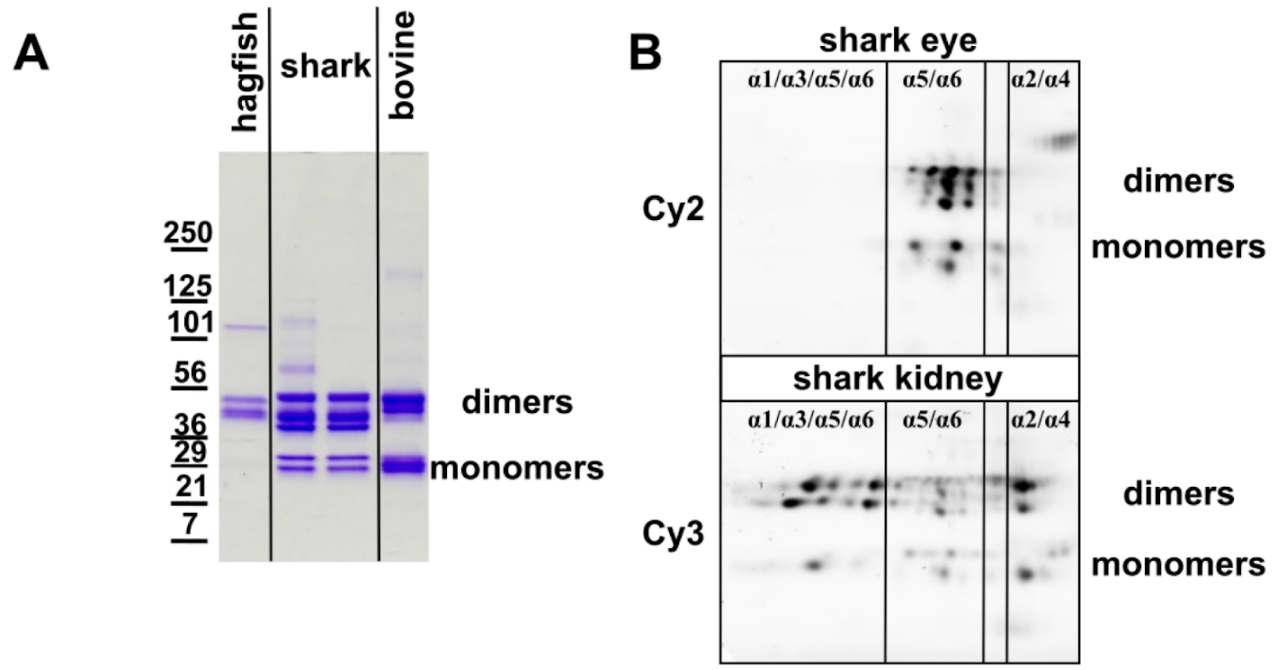
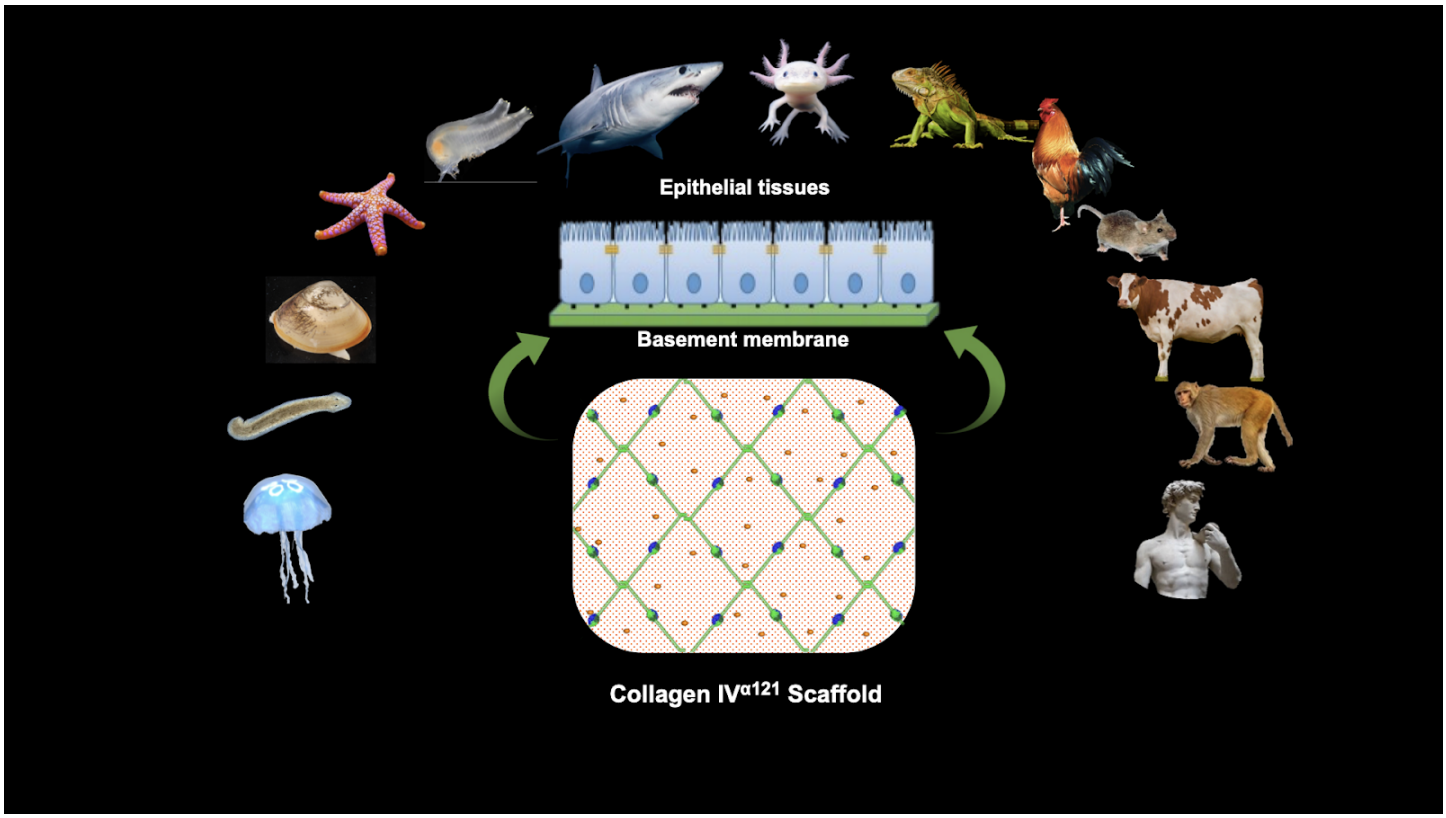



**Figure 10. Tissue specific Collagen IV chain composition. A.** SDS Polyacrylamide Gel Electrophoresis of purified NC1 domains of Collagen IV derived from hagfish kidney, shark ocular lens and kidney, and bovine kidney. The shark and bovine NC1 domains run as both uncrosslinked monomers and crosslinked dimers while the hagfish protein does not have detectable monomers. The shark dimer band runs as a set of three bands likely reflecting the different scaffolds. **B**. Fluorescent images of the same 2D-NEPHGE gel loaded with purified NC1 domains from the shark ocular lens and kidney showing that the shark ocular lens consists primarily of α565 while the shark kidney has less α565 than other protomers.

Figure 10 Summary figures



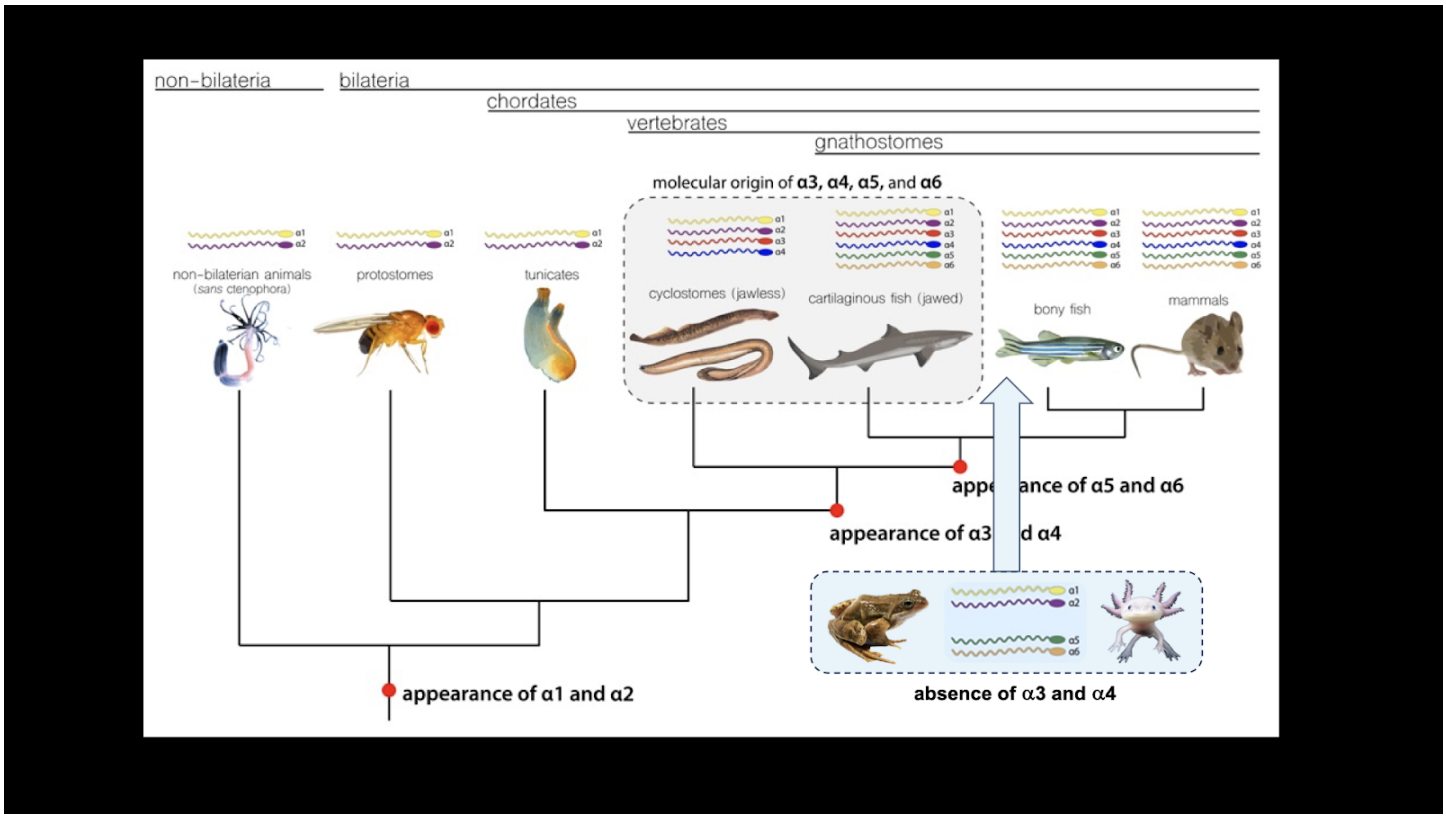
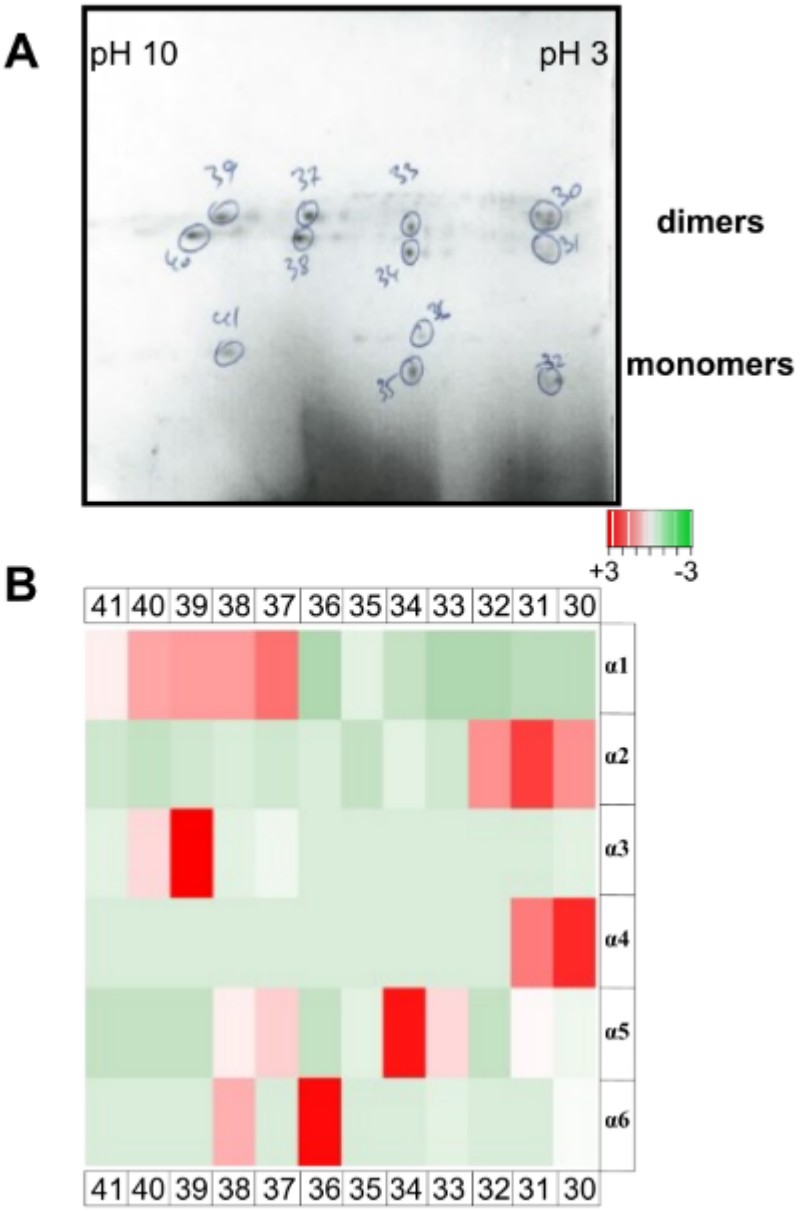




Figure 10.

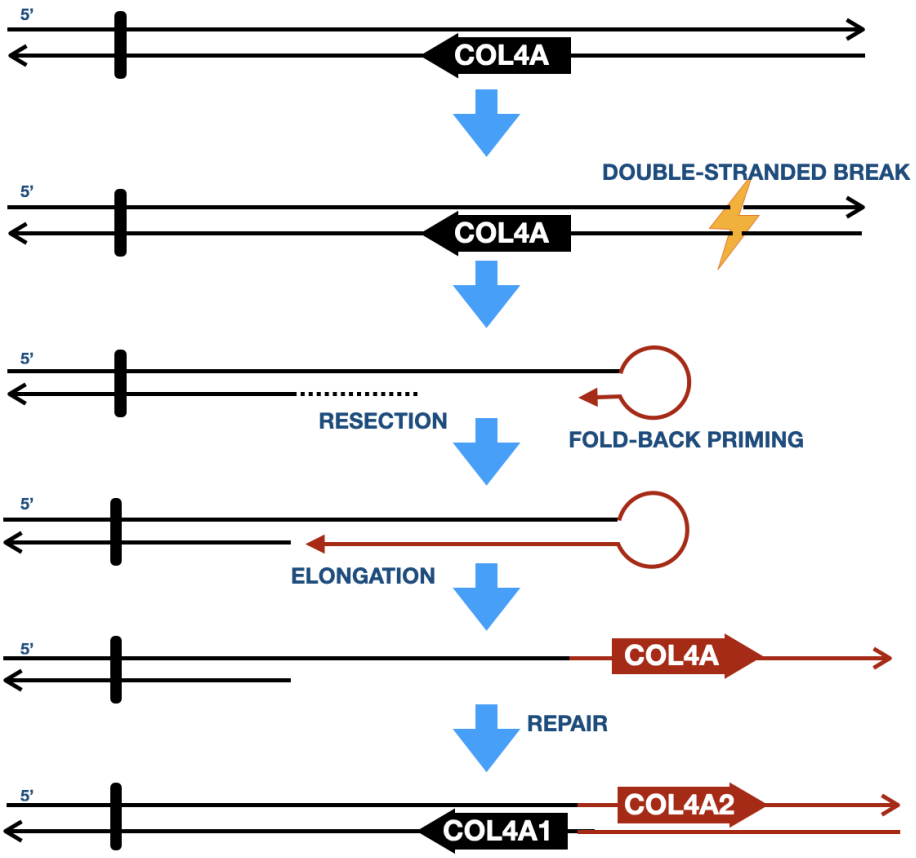
## Supplemental Figures

Supplemental Figure 1.




**Supplemental Figure 1.**
**A.** Colloidal Coomassie stained 2D-NEPHGE gel with numbered spots that were excised for mass spectroscopy.
**B** A heat map created with heatmapper.ca build from label free quantification (LFQ) of proteins data. A custom dogfish protein database was used to define chain composition in each 2D spot.

Supplemental Figure 2.




Supplemental Figure 2. Hypothesized mechanism for the creation of the ancestral COL4A⟨1|2⟩ gene pair following a double-strand break event.