Collagen type IV in the Ctenophore *Pleurobrachia bachei*

Nathan Churches[1,2], Billie J. Swalla[1,2], Leonid Moroz[1,3], Andrea Kohn[1,3]

Marine Genomics Research Apprenticeship
Spring 2012

[1]University of Washington, Friday Harbor Laboratories, Friday Harbor, WA 98250
[2]Department of Biology, University of Washington, Seattle, WA 98195
[3]Department of Biology, The Whitney Laboratory for Marine Bioscience, St. Augustine, FL 32080

Contact Information:
Nathan Churches
nthnchrchs@yahoo.com
(360) 870 8589

Collagen type IV α-chains in *Pleurobrachia bachei*

Nathan Churches

ABSTRACT

The evolution of multicellular animals required the development of epithelial tissues that function in controlling the transport of molecules from one environment to another. Collagen proteins are crucial to the formation of epithelial tissues, and are therefore critical in understanding the origins of multicellularity and Metazoa. We looked for collagen type IV proteins in the recently sequenced Ctenophore *Pleurobrachia bachei* genome, to assess whether this most basal of Metazoa phylum could contain conserved collagen proteins. We used *Homo sapiens* collagen type IV sequences and Blasted against the Moroz lab *P.bachei* genome server to search for collagen type IV proteins. We found that *P. bachei* possesses 6 distinct type IV alpha chains along three scaffolds, two of which were aligned in a head-to-head fashion, indicating both traditional and inverted gene duplication events. Given recent evidence suggesting that Ctenophores are possibly the most basal of the Metazoans, our findings suggest that the common ancestor to all Metazoa contained a much more developed collagen profile than previously appreciated.

INTRODUCTION

The advent of multicellularity involved the evolution of epithelial tissues that were capable of controlling the transfer of molecules from one environment to another. This task is carried out by epithelia in modern multicelled animals. Presumably, a proto-

epithelia evolved in colonial single celled animals, allowing two or more organisms to forge and control their surroundings, thus forming distinguished multicellular organisms (Leys and Reisgo, 2011). Collagens are the backbone of epithelial tissues, playing a crucial role in animal morphology and physiology. Therefore, understanding collagen origins and evolution is an essential piece to the puzzle of metazoan phylogeny. In vertebrate chordates, including mammals, collagens are critical in formation of skin, bone, teeth, and other tissues. In humans collagens are the most abundant protein produced, making up 25% of body mass (Alberts et al., 1994). Many diseases are associated with mutations in collagens or breakdown of the collagen molecular pathway (Leitinger and Hohenester, 2006), and collagen pro-peptides may have anti-tumor and anti-angiogenic regulation properties (Richard-Blum and Ballut, 2011). These facts reinforce their status as an important molecule for research concerning the evolution and development of metazoans.

All collagens are secretory proteins, and are characterized by having three protein polymers wound into a diagnostic collagen triple helix. Though the triple helix is the most dominant feature of the collagen protein, non collagen motifs surrounding the triple helix are thought to be the more functionally important part, playing a key role in receptor-protein recognition (Leitinger and Hohenester, 2006). Each individual polymer in the triple helix is called an alpha chain, and a given collagen can be either a homotrimer (having three genetically similar alpha chains) or heterotrimer (a combination of two or three genetically distinct alpha chains). These polymers are composed with Glycine at every third position, followed by two amino acids (X-Y). These triplets, called Gly-X-Y repeats, allow for the formation of a tightly wound triple

helix, as Glycine is the smallest amino acid and thus allows for tight packing formation. Collagens can contain over 1500 amino acids in their Gly-X-Y domains (Boot-Handford and Tuckwell, 2004). On the C terminal ends of collagen triple helixes are Non-collagen (NC1) domains called C-propeptides, while the N terminus contains a short helical triple helical domain (minor helix) and a second Non-collagen domain (NC2) (Boot-Hanford and Tuckwell, 2003). These terminal ends, C-propeptides and NC2, are insoluble and facilitate both the retention of the soluble inner triple helix collagen domain in solution and in the localization of collagen in tissues. These propeptides are cleaved as they exit the cell by specific enzymes (C- and N-proteinases), leaving the triple helix domain intact. Once extracellular and relieved of their propeptides, individual collagen helices will covalently bind to each other in a lateral formation, allowing for a variety of secondary structures including networks, microfilaments, and collagen fibrils. Collagen fibrils are the backbone of all epithelial tissues, crisscrossing the space between nearly each of our cells, giving them structure and tensile strength (Alberts, et al. 1994).

There are a total of 29 different types of collagens described in humans to date, consisting of different combinations of the 48 described alpha chains. Collagens are named and numbered in the order they are discovered in vertebrate organisms, while there is some inconsistency in naming collagens in invertebrates. Types I, II, III, V, and XI are called the fibrillar collagens, which are responsible for skeletogenesis in vertebrate organisms. The two most abundant of the fibrillar collagens, types I and II, are responsible for proper formation of bone & teeth, and cartilage, respectively. Types III, V, and XI are less abundant than types I and II, but are essential for bone and cartilage tissues as well (Boot-Hanford and Tuckwell, 2003). Interestingly, fibrillar collagens are

found across Metazoa, from Porifera & Cnidaria through Chordates. Few exceptions to this case are found, the occurrences being in the arthropods and annelids. This Metazoa-wide collagen signature indicates that the metazoan common ancestor had the molecular toolkit for developing some sort of fibrillar collagen structures, even if they were rudimentary (Boot-Hanford and Tuckwell, 2003).

Surprising new evidence suggests that collagen IV may also be conserved across the metazoan lineage. Collagen IV is a key component of the Basement Membrane (BM), a layer of tissue which connects outer epithelial layers to tissues below. BM formation is considered by some to be indicative and a necessary hallmark of 'true epithelia' (Exposio et al., 2009). There is a debate among scientists as to where 'true epithelia' evolved, or indeed if the current definition of true epithelia is adequate. In one review on origins of epithelia, it was suggested that epithelia evolved in its first true form in Cnidaria, stating that Porifera do not possess the necessary hallmarks of epithelial tissue (Tyler, 2003). It has been shown that Cnidaria use collagen IV in development and healing (Fowler, et al. 2000), and possess a basal lamina-like membrane containing collagen IV (Shimizu, et al., 2007). Literature in science will often use molecular markers to determine if a trait is present in a specific animal, and some argue that this widens the criteria for a given morphology (Leys and Reisgo, 2011). An example of this is the fact that cnidarians seem to lack a mesoderm, yet have all of the genes for muscle formation and, by association, mesoderm. Another example includes a recent paper which suggests that Porifera, one of the most basal metazoans, have many of the molecular components related to collagen IV, to make a primordial BM like structure (Leys and Reisgo, 2011). Therefore the sponges, which have historically been considered to lack true epithelial tissues, can serve

as an argument for the loosening or broadening of the term 'true epithelia'. This considering that the molecular signatures for collagen IV-associated proteins are present and perform a common function.

Collagen IV has recently been proposed to have evolved from sponging short-chain collagen (SSCC). SSCC's are defined as having 2 collagenous domains with 79 Gly-X-Y repeats and 3 NC domains. A hypothesis put forth by Aouacheria et al. 2006, is that early multicellular animals first evolved SSCC in order to attach to some substrate in the marine environment, with collagen IV and BM formation evolving from co-adaptation of SSCC genetic pathways. Porifera still contain these SSCC's, and SSCC-related molecules are still present up through the non-vertebrate chordates. It is thought that the split between vertebrates and the rest of Metazoa was facilitated in some way by a duplication event within the SSCC family, allowing eventually the development of true collagen IV proteins. Phylogenetic comparison of SSCC to type IV NC domains supports this theory, showing that NC1 domain in *E. mulleri* (a sponge in class Demospongiae) is homologous to the corresponding domain in type IV collagens, and that human ColVα1 and ColVα2 contain more than 75% homology with *P. jarrei* sponge SSCC domains (Aouacheria, et al, 2006). It has been shown that collagen molecules have evolved using standard gene duplication events, and this fact has been exploited to study evolution of associated HOX genes (Bailey et al., 1997). In vertebrate chordates, collagens are aligned in the genome in a pair-wise 'head-to-head' fashion along the chromosomes. That is to say that the NC2 domains are abutting, with one collagen coded in a forward direction and the other on the complimenting opposite strand in the reverse direction, indicating an inverted duplication in the ancestor. This inverted duplication has

happened 6 times in vertebrate chordates, allowing for 6 alpha chains in type IV collagens (Hudson, et al. 1993). These 6 alpha chains are categorized into two types; α1-like and α2-like. In humans, the α1-like types are α1, α3, and α5, while the α2-like types are α2, α4, and α6. This 'type' patterning can be seen across the rest of Metazoa, where most of the more basal organisms have only α1-like and one α2-like chain (Figure 5/table 3).

Ctenophores, a group of jelly bodied animals that use ciliated comb rows for locomotion, are one of the most basal metazoans on the animal tree. Scientists argue as to their proper place on the universal animal tree, especially with respect to Porifera or Cnidaria. Some groups place Porifera as the most basal metazoan, mostly based on morphological evidence. Others argue that Ctenophores are the true metazoan ancestor, stating that their possession of the smallest mitochondrial genome yet found is a highly derived character state, indicating a pre-Porifera split (Kohn, et al. 2011). Recently, it has been shown that two separate sponges, namely *Sycon coactum* and *Corticum candelabrum*, have two distinct type IV collagens in each organism. These findings indicate that it is possible that α1 and α2 chains had diverged before Porifera (Leys and Riesgo, 2011). Therefore, the presence or absence of collagen IV or SSCC in ctenophores could shed light on their place in the metazoan tree, as well as give further information as to the evolution of epithelia and BM tissues. Furthermore, *P. bachei* may be a model organism to use for the study of collagen in human diseases, given their basal state. This paper focuses on collagen family types found in the *Pleurobrachia bachei* genome, recently sequenced by the Moroz lab at the University of Florida. We show here that *P. bachei* in fact have at least six distinct full length α-chains (average = 2200 aa) of type IV

collagen, each with multiple stretches of Gly-X-Y repeats (average = 13). Interestingly, two of the three collagen IV α1-like proteins found in *P. bachei* were aligned in a 'head-to-head' alignment typical of vertebrate chordates (Figure 3).

METHODS

We used collagen IV *Homo sapiens* sequences, obtained from the Nucleotide database at National Center for Biotechnology Information (NCBI) (http://ncbi.nlm.nih.gov/) to pull up conserved proteins in *P. bachei*. We then used TBMastN on the Moroz lab's *P. bachei* gene database server to obtain conserved protein sequences. With these sequences we ran BMastP against the NCBI database to gain conserved sequences from across Metazoa. Some of the more basal organism sequences were obtained in the same way from the BROAD institute database (http://www.broadinstitute.org/) and from the Joint Genome Institute (JGI) database (http://www.jgi.doe.gov/). We used GeneDoc, MEGA, and ClustalX software to obtain our aligned sequences. To obtain gene trees, we used MEGA tree software ML. Bootstrap support of less than 70 was omitted. We trimmed our sequences to the same region as described in Aouacheria et al. 2006, in order to get the most accurate phylogenetic trees, because Gly-X-Y repeats are highly divergent across all collagen families. To obtain intron/exon information, as well as general protein family domains, we used the Simple Modular Architecture Research Tool (SMART) server (http://smart.emBM-heidelberg.de/smart/set_mode.cgi?NORMAL=1). Please contact me for PCR sequences.

Animals for in situ hybridizations were collected off the docks at Friday Harbor Labs in San Juan Island, Washington between the months of April, May, and June 2012. Animals were kept in circulating salt water tanks for up to two weeks prior to fixation. In situ protocol follows.

IN SITU PROTOCOL
Pleurobrachia *In situ* hybridization – Modified 2011
Adapted from Derelle and Manuel 2007 and Moroz

Day 1

Fix whole specimen in 4% paraformaldehyde in Filtered Sea Water (FSW) overnight (O/N) at 4^0C

Day2

Rinse 3 x for 10 min in PTW (PBST) at Room Temperature
Wash in 1:1 Methanol (MeOH)/PTW (to equilibrate to MeOH) 10 min at Room Temperature
Store in 100% MeOH at -20C for 2 hours up to a week

Day 3

Rehydrate specimen for 10 minutes in MeOH/PTW 3:1, 1:1, 1:3, 0:1 at Room Temperature
Wash in 1:1 solution of hybridization buffer (HB) and PTW for 15 minutes at Room Temperature
Incubate (prehybridize) in HB buffer for 1 hours at 60^0C
Incubate (hybridize) in HB with DIG-RNA probe O/N at 60^0C

Day 4

Wash in HB for 30 min at 60^0C
Wash in 1:1 HB/PTW for 30 min at 60^0C
Wash in PTW for 30 min at Room Temperature
Block in 10% Goat Serum (GS) for 60 min at Room Temperature
Incubate in anti-DIG 1/2000 at 4^0C O/N

Day 5

Wash 4 x 30 min in PBS Room Temperature
Make detection buffer and aliquot 1mL into clean well for each sample. When ready to develop, add 20uL of NBT/BICP mix until dissolved. Should be yellow in color! **NOW** add samples. **Put on ICE and cover with tin foil.**
Watch for appropriate color development Stop in 4% paraformaldehyde in MeOH
Stop in PBS leave days with several changes at 4^0C
Wash in 4% paraformaldehyde in MeOH 30 min Room Temperature
Wash 3x 10 min in Ethanol (EtOH) Room Temperature
Store in 100% EtOH
Mount
Add animals to Methylsalicylate, until they sink
Put animal onto microscope glass, clean, absorb methylsalicylate leftovers, add a drop of Permount, put on the cover slip

RESULTS

We used *Homo sapiens* collagen type IV-α1-6 (*COLIVα1, COLIVα2, COLIVα3,* etc.) sequences obtained from NCBI to pull up conserved proteins in *Pleurobrachia bachei* using the Moroz lab genome server. We found 6 distinct homologues to type IV α-chains, and phylogenetic analysis showed that three were specific to α1-like families and three were specific to α2-like families. Furthermore, the collagen chains found in *P. bachei* are more diverse when compared intra-genically than *H sapiens* collagen chains (Figure 5). It is interesting to note that only the α2-like chains had repeat domains (RPT). Our gene alignment showed that the 12 cysteine residues found in NC1 domains were conserved from the *H. sapiens* sequence in *P. bachei* (Figure 5). We also note considerable homology in the β-hairpin regions of the NC1 domains. We did not find any SSCC homologues in the *P. bachei* genome, only full length type IV collagen homologues, which we dubbed homologues *PbalphaA-F* (*PbalphaA:* 2272aa, *PbalphaB:* 2448aa, *PbalphaC:* 2021aa, *PbalphaD:* 2395aa, *PbalphaE:* 1753aa, *PbalphaF:* 2402aa). SMARTing showed that each of these full length sequences contained a minimum of 9 Gly-X-Y regions, and a maximum of 21. Each homologue also contained two NC1 domain sequences conserved to specific collagen IV proteins (Figures 1 and 2).

We found that two of our α1-like proteins, *PbalphaB* and *PbalphaC*, were aligned on our scaffold in a pair-wise head-to-head fashion, similar to the arrangement found in

vertebrate chordates (Figure 3). Furthermore, we found that an apparent in-line gene duplication (traditional gene duplication) event happened with *P. bachei's* α1-like *PbalphaA* and *PbalphaB*, as well as with α2-like *PbalphaD* and *PbalphaE*. We deduced this as they are directly in line along the scaffold. Lastly, we found a single collagen IV α2-like homologue on a separate scaffold, *PbalphaF*. This α2-like sequence ran nearly the entire length of our scaffold, which raises the possibility that there are more α2-like in *P. bachei* if the in-line duplication pattern was present in this particular gene also. We have insufficient data to prove this, however. In situ hybridization of *PbalphaC* shows diverse expression across tissues, most notably at the base of the comb rows, along the underlying meridional canals, in the tissues holding the statolith, and in the dermal tissues. Single cell expression was also notable.

DISCUSSION

It has recently been suggested the sponge *Pseudocorticium jarrei* has type IV collagens, with recent evidence unveiling two other distinct collagen IV sequences in *Sycon coactum* and *Corticum candelabrum*. These findings indicate that the evolution of α1-like and α2-like type IV collagens could have predated the sponge ancestor (Leys and Riesgo, 2011). We found a total of 6 distinct type IV collagens in *Pleurobrachia bachei*, half which aligned to α1-like chains (*PbalphaA-C*) and half to α2-like chains (*PbalphaD-F).* In situ hybridizations of *PbalphaC* revealed expression patterns in a variety of tissues. This is what we would predict considering the nature of collagen proteins in the rest of metazoa (Figure 5). Most interestingly is the even expression patterns in single cells in the dermal layers of the Ctenophore. This indicates that it is possible that sponges in fact lost several collagen IV alpha chain types, considering recent evidence that Ctenophores

are likely more basal than Porifera. The fact that a pair of collagens were found in a pair-wise head-to-head fashion (*PbalphaB, PbalphaC*), and all except one were found to be associated with some sort of gene duplication event along the *P. bachei* genome, further support the hypothesis that our metazoan ancestor had a collagen profile more like that of what we would consider derived. This ancestor therefore could have had the ability to synthesize more developed tissues and could have performed a larger variety of physiological functions than the current 'Porifera first' theory allows.

CONCLUSIONS & FURTHER RESEARCH

Our findings suggest that the ancestor to the metazoan tree had a much more derived and diverse collagen profile than previously thought. We found 6 distinct collagen type IV α-chains, 5 of which were associated with gene duplication events. Our conclusion is further supported by the pair-wise head-to-head fashion in which the α1-like *PbalphaB* and *PbalphaC* are aligned on the genome, which has previously been seen only in the highly derived vertebrate chordates. Further research needs to be done on collagen profiles in *Pleurobrachia bachei*, so that we may glean a more complete understanding of the origins of Metazoa. Our lab is currently working on isolating several collagen types and matrix related proteins via cDNA library cloning, including collagen types I & II, Netrin, and Integrin. We plan to continue work with in situ hybridizations, with a goal of expressing all 6 alpha chains.

TABLES & FIGURES

Collagen IV α1-like chains in P. bachei

Pbalpha-A	Begin	End
Gly-X-Y	1	57
Gly-X-Y	47	112
Gly-X-Y	97	165
Gly-X-Y	203	273
Gly-X-Y	392	455
Gly-X-Y	442	514
Gly-X-Y	500	558
Gly-X-Y	557	612
Gly-X-Y	611	668
Gly-X-Y	653	721
Gly-X-Y	699	775
Gly-X-Y	759	828
Gly-X-Y	812	876
Gly-X-Y	918	975
Gly-X-Y	974	1035
Gly-X-Y	1182	1248
Gly-X-Y	1241	1301
Gly-X-Y	1300	1365
Gly-X-Y	1448	1515
Gly-X-Y	1630	1696
Gly-X-Y	1684	1745
Gly-X-Y	1833	1890
Gly-X-Y	1887	1937
Gly-X-Y	1936	1992
Gly-X-Y	2118	2178
Gly-X-Y	2178	2244
NC1	2249	2367
NC1	2368	2444

Pbalpha-B	Begin	End
Gly-X-Y	332	395
Gly-X-Y	381	452
Gly-X-Y	439	498
Gly-X-Y	496	552
Gly-X-Y	551	610
Gly-X-Y	817	876
Gly-X-Y	921	986
Gly-X-Y	959	1026
Gly-X-Y	1080	1140
Gly-X-Y	1136	1188
Gly-X-Y	1186	1241
Gly-X-Y	1576	1636
Gly-X-Y	1708	1749
Gly-X-Y	1756	1819
NC1	1822	1944
NC1	1982	2071

Pbalpha-C	Begin	End
C-SP	1	19
Gly-X-Y	64	117
Gly-X-Y	90	164
Gly-X-Y	141	202
Gly-X-Y	264	323
Gly-X-Y	614	673
Gly-X-Y	674	731
Gly-X-Y	728	784
Gly-X-Y	750	808
Gly-X-Y	875	934
Gly-X-Y	1070	1142
Gly-X-Y	1114	1175
Gly-X-Y	1174	1231
Gly-X-Y	1223	1289
Gly-X-Y	1264	1343
Gly-X-Y	1319	1380
Gly-X-Y	1573	1637
Gly-X-Y	1675	1743
Gly-X-Y	1718	1787
Gly-X-Y	1774	1837
Gly-X-Y	1836	1897
Gly-X-Y	1871	1949
NC1	2049	2172
NC1	2173	2270

Table 1: **Above are the conserved protein domains in α1-like chains (*Pbalpha-A, Pbalpha-B, Pbalpha*-C) found in *Pleurobrachia bachei* gene models, recently sequenced at the Moroz lab at the University of Florida. 'Begin' and 'End' columns show the amino acid position along the length of the protein. Most notable are the Gly-X-Y protein families, which are diagnostic of true collagen alpha chains and have only recently been found in other basal Metazoans. Each α1-like chain also contains two conserved NC1 domains, highlighted in red, which we later used to deduce gene trees and infer relationships across metazoa. One C-terminal signal peptide domain was found in *Pbalpha-C*, marked as C-SP. These tables were constructed using the Simple Modular Architecture Research Tool (SMART) server (http://smart.emBM-heidelberg.de/smart/set_mode.cgi?NORMAL=1). Total amino acid counts: *Pbalpha-A: 2448, Pbalpa-B: 2021, Pbalpha-C: 2272.***

Collagen IV α2-like chains in P. bachei

Pbalpha-D	Begin	End
Gly-X-Y	62	126
Gly-X-Y	85	160
Gly-X-Y	234	290
Gly-X-Y	289	352
Gly-X-Y	341	409
Gly-X-Y	942	1003
Gly-X-Y	1049	1113
Gly-X-Y	1109	1170
Gly-X-Y	1156	1214
Gly-X-Y	1535	1592
Gly-X-Y	1581	1639
Gly-X-Y	1639	1697
Gly-X-Y	1720	1778
Gly-X-Y	1826	1883
Gly-X-Y	1855	1908
Gly-X-Y	1907	1965
Gly-X-Y	1953	2015
Gly-X-Y	2041	2100
Gly-X-Y	2092	2152
NC1	2163	2273
NC1	2275	2391

Pbalpha-E	Begin	End
Gly-X-Y	1	59
Gly-X-Y	140	209
Gly-X-Y	200	264
Gly-X-Y	241	314
Gly-X-Y	529	603
Gly-X-Y	837	914
Gly-X-Y	891	963
Gly-X-Y	945	1008
Gly-X-Y	1007	1066
Gly-X-Y	1040	1123
Gly-X-Y	1175	1239
Gly-X-Y	1226	1290
Gly-X-Y	1282	1343
Gly-X-Y	1372	1433
Gly-X-Y	1425	1486
NC1	1495	1605
NC1	1606	1722

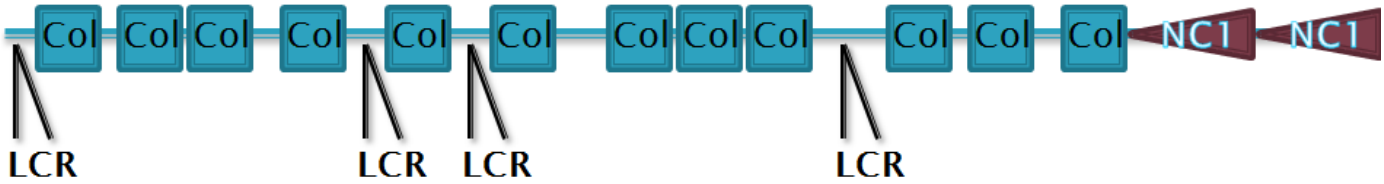
Pbalpha-F	Begin	End
C-SP	1	26
Gly-X-Y	178	234
Gly-X-Y	220	279
Gly-X-Y	278	337
Gly-X-Y	324	389
Gly-X-Y	422	485
Gly-X-Y	475	540
Gly-X-Y	535	595
Gly-X-Y	1078	1151
Gly-X-Y	1422	1486
Gly-X-Y	1479	1542
Gly-X-Y	1529	1587
Gly-X-Y	1743	1818
Gly-X-Y	1950	2008
Gly-X-Y	1994	2057
NC1	2137	2248
NC1	2249	2364

Table 2: **Above are the conserved protein domains in α2-like chains (*Pbalpha-D, Pbalpha-E, Pbalpha*-F) found in *Pleurobrachia bachei* gene models, recently sequenced at the Moroz lab at the University of Florida. 'Begin' and 'End' columns show the amino acid position along the length of the protein. Most notable are the Gly-X-Y protein families, which are diagnostic of true collagen alpha chains and have only recently been found in other basal Metazoans. Each α1-like chain also contains two conserved NC1 domains, highlighted in red, which we later used to deduce gene trees and infer relationships across metazoa. One C-terminal signal peptide domain was found in *Pbalpha-F*, marked as C-SP. These tables were constructed using the Simple Modular Architecture Research Tool (SMART) server (http://smart.emBM-heidelberg.de/smart/set_mode.cgi?NORMAL=1). Total amino acid counts: *Pbalpha-D: 2395, Pbalpa-E: 1753, Pbalpha-F: 2402.***

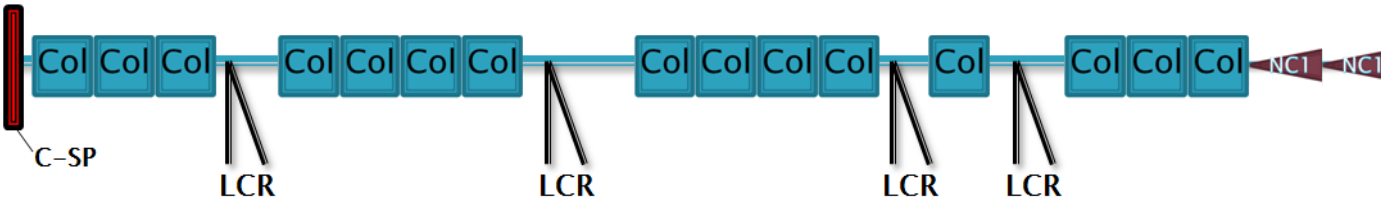
α1-Like chain motifs in P. bachei

Pbalpha-A (2448 amino acids)

Pbalpha-B (2021 amino acids)



Pbalpha-C (2272 amino acids)

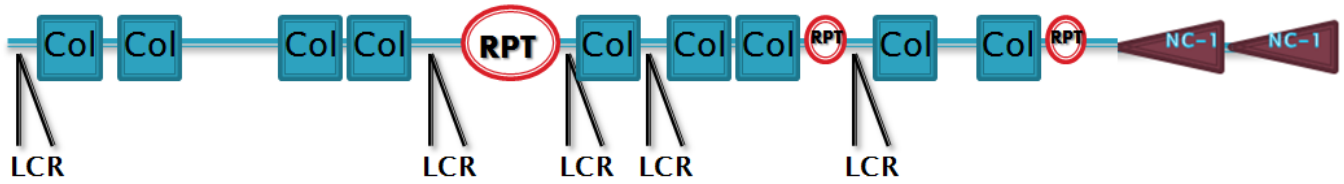


Symbol	Meaning
COL	Gly-X-Y domain
NC1	NC1 domain
LCR	Low Complexity Region
C-SP	C-terminal Signal Peptide

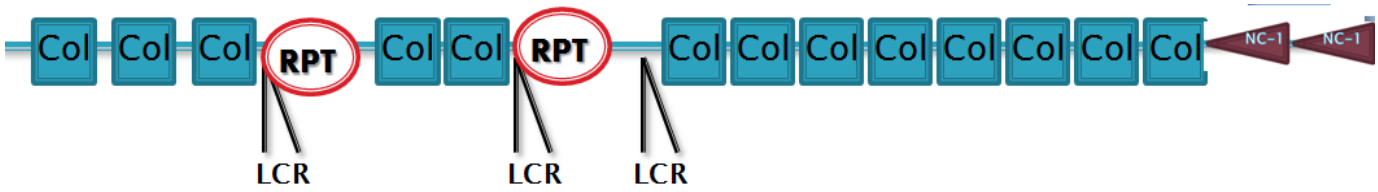
Figure 1: **Collagen IV α1-Like chain motifs found in *P. bachei*. Note the large amount of collagen Gly-X-Y motifs in each gene (*Pbalpha-A: 21 Gly-X-Y domains, Pbalpha-B: 12 Gly-X-Y domains, Pbalpha-C: 15 Gly-X-Y domains).* Each gene also contained two NC-1 domains, and several Low Complexity Regions (LCR). These LCR are places of high amino acid repeats or scaffold gaps in sequence. All figures were generated with information from the Simple Modular Architecture Research Tool (SMART) server (http://smart.emBM-heidelberg.de/smart/set_mode.cgi?NORMAL=1).**

α2-Like chain motifs in P. bachei

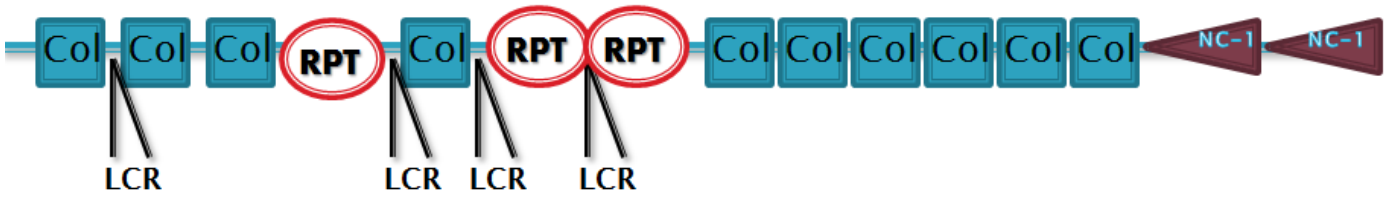
Pbalpha-F (2402 amino acids)



Pbalpha-D (2395 amino acids)



Pbalpha-E (1753 amino acids)



Symbol	Meaning
COL	Gly-X-Y domain
NC1	NC1 domain
RPT	Internal repeat
LCR	Low Complexity Region
C-SP	C-terminal Signal Peptide

Figure 2: **Collagen IV α1-Like chain motifs found in *P. bachei*. Note the large amount of collagen Gly-X-Y motifs in each gene (*Pbalpha-A: 21 Gly-X-Y domains, Pbalpha-B: 12 Gly-X-Y domains, Pbalpha-C: 15 Gly-X-Y domains).* Each gene also contained two NC-1 domains, and several Low Complexity Regions (LCR). These LCR are places of high amino acid repeats or scaffold gaps in sequence. All figures were generated with information from the Simple Modular Architecture Research Tool (SMART) server (http://smart.emBM-heidelberg.de/smart/set_mode.cgi?NORMAL=1).**

Scaffold organization in type IV alpha chains

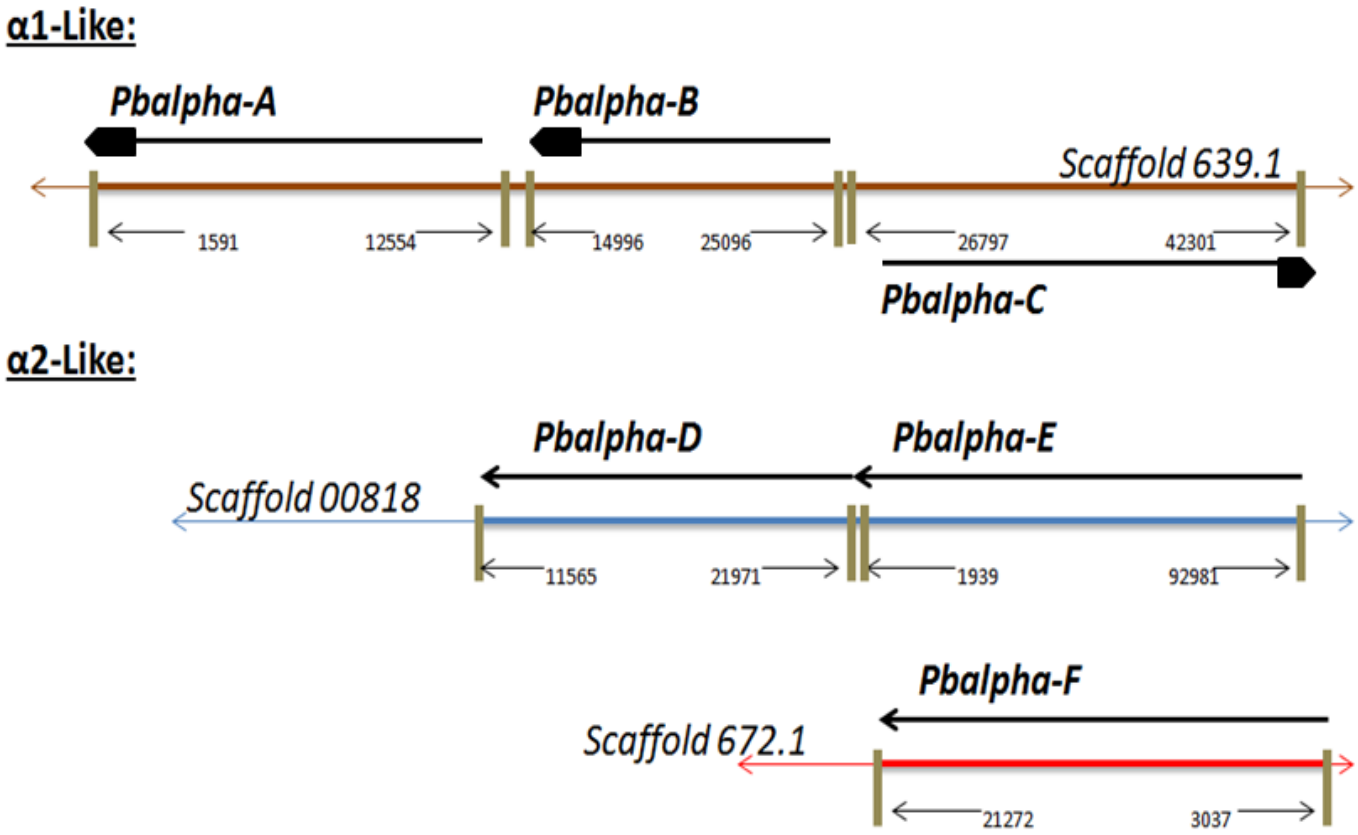


Figure 3: **Scaffold alignment in *Pleurobrachia bachei* alpha chains. Thick colored shading indicates protein coding region along scaffold, while thinner like-colored arrows indicate entire scaffolds. Thick black arrows beneath gene names (*Pbalpa-A*, *Pbalpha-B* etc.) indicate directionality along scaffold. Small numbers with arrows pointing to vertical tan bars below the scaffolds represent amino acid start and stop positions. Notice the 'head-to-head' alignment between *Pbalpha-B* and *Pbalpha-C*, indicating an inverted gene duplication. Note also the traditional gene duplications indicated by position along scaffold between *Pbalpha-A&B* and *Pbalpha-D&E*. Figure generated using the Moroz lab *Pleurobrachia bachei* gene model server.**

Type IV Collagen and Spongin Short Chain Collagen Gene Alignments

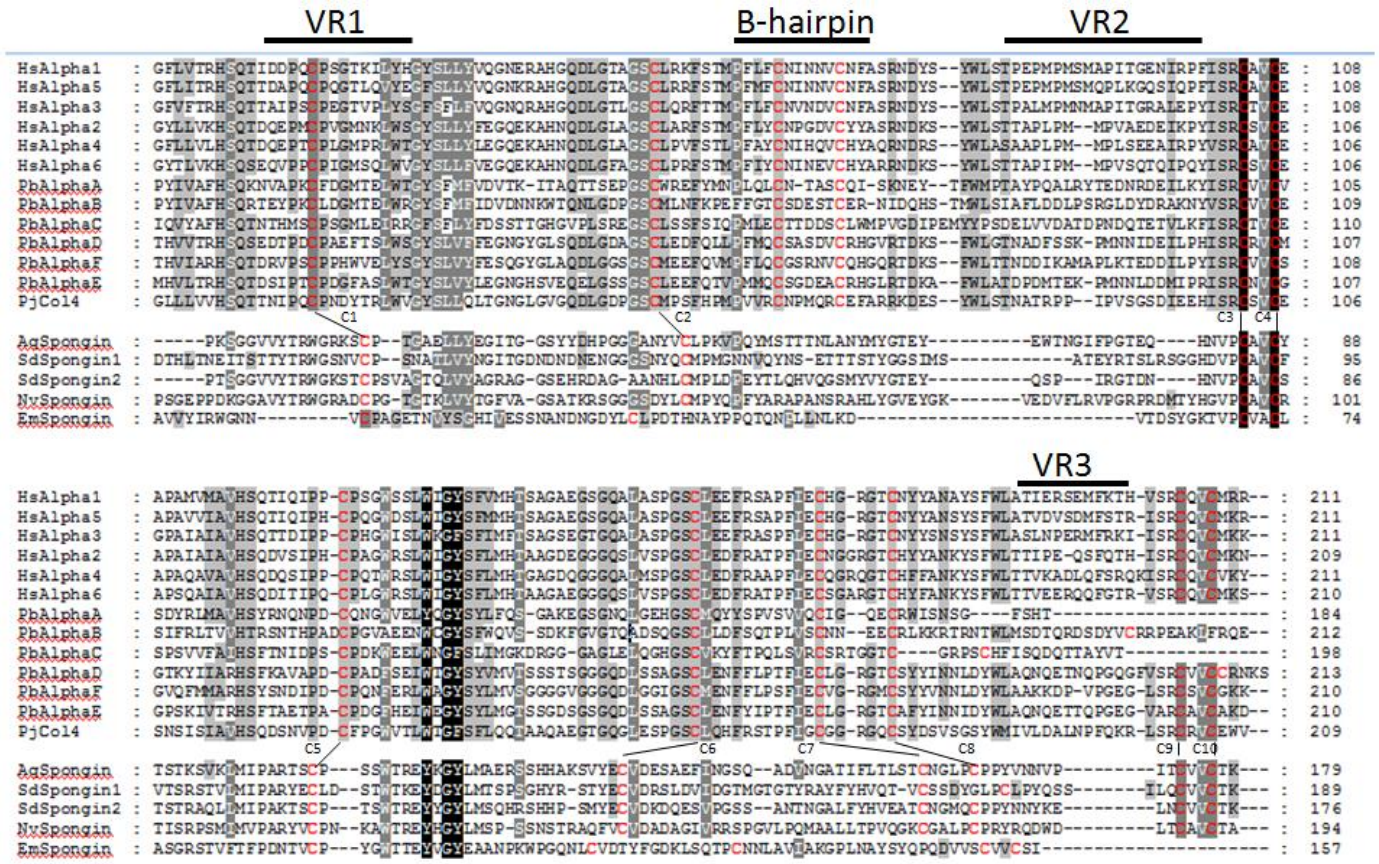


Figure 4: Multiple species alignments of Spongin Short-Chain Collagens (SSCC) and collagen Type IV alpha chain NC1 domains. Our alignments were produced using GeneDoc software. (Gene names and related organisms: Hsapha1-6: *Homo sapiens*, PbalphaA-E: *Pleurobrachia bachei*, PjCol4: *Pseudocorticum jarrei*, Agspongin: *Anopheles gambiae*, Sdspongin1-2: *Suberites domuncula*, Nvspongin: *Nematostella* vectensis, Emspongin: *Ephydatia mulleri*). Highlighted in red are Cystein domains, marked C1-10, conserved across SSCC and type IV alpha chains. Black shading indicates 100% conservation across phyla, while grey shading indicates partial homology. Noted above the alignments are NC1 specific domains VR1-3 (variable regions), and β-hairpin region. Homology between *Pleurobrachia bachei* and other type IV collagens is readily apparent.

In situ expression of PbalphaC and collagen type IV gene tree

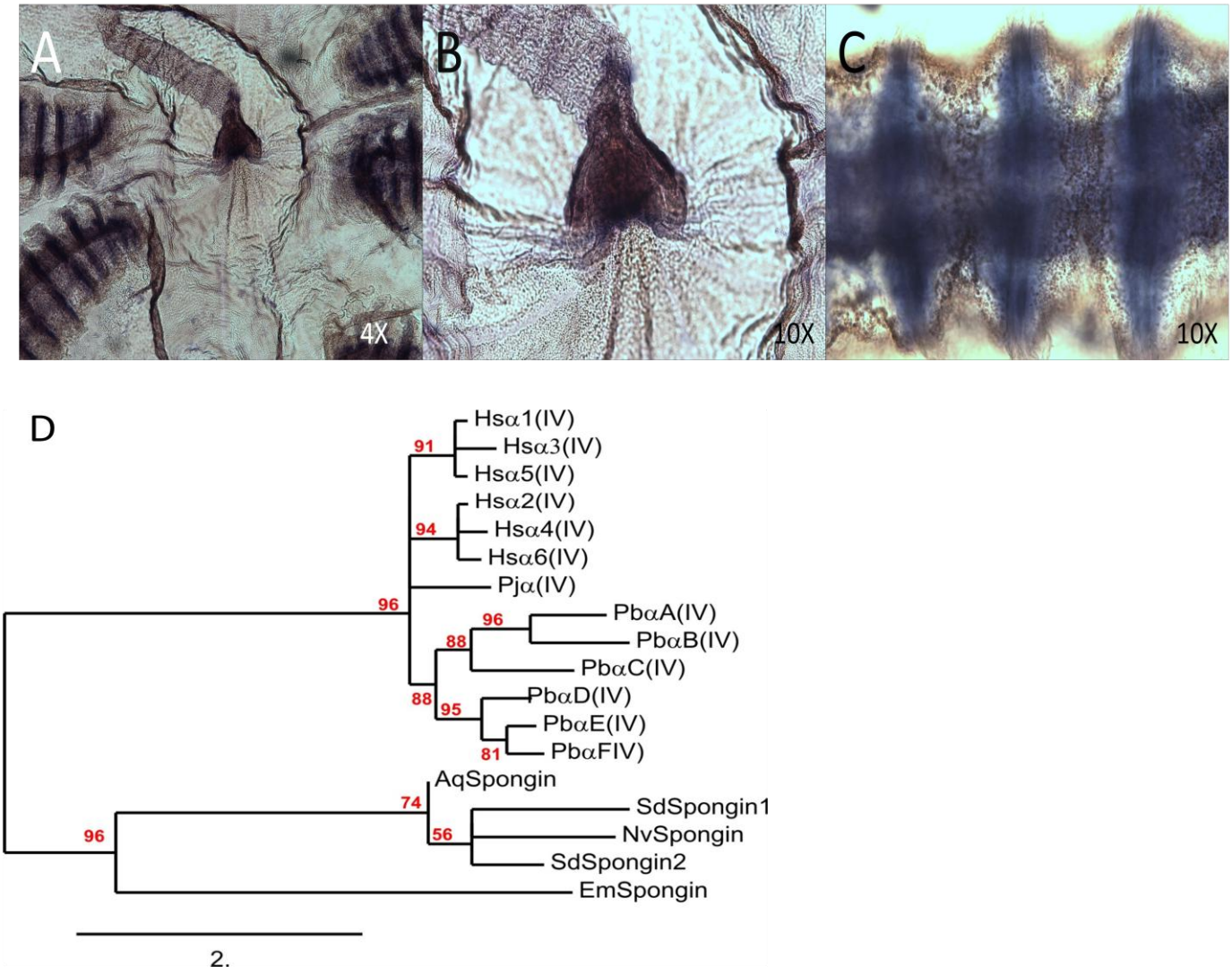
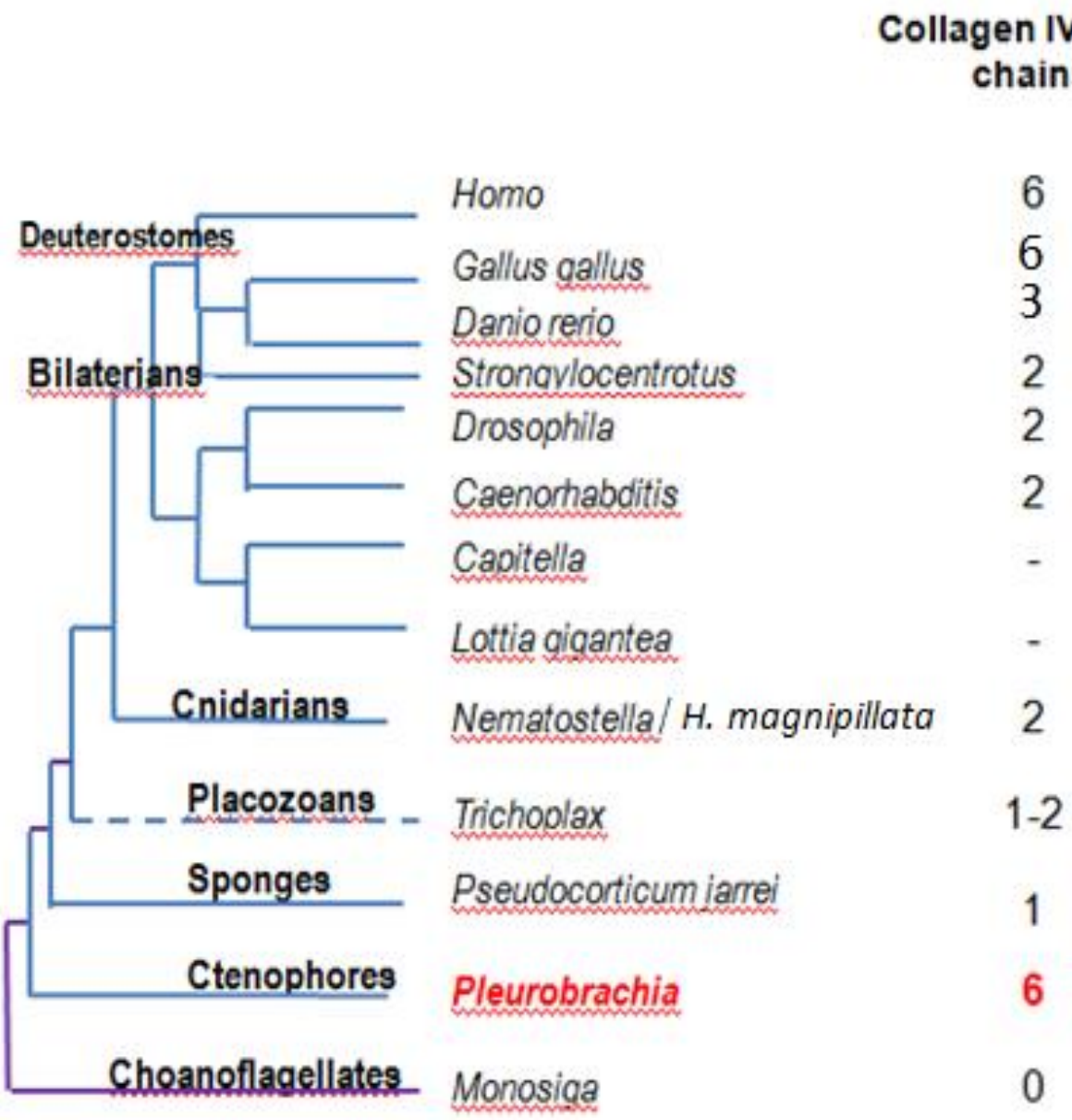


Figure 5: A-C show in situ preparations (combs oriented into the page) using *PbalphaC* . A: Note expression along the base of the combs, along the meridional canal (above the comb rows), and in the statolith (center). B: Close examination of the statolith reveals heavy expression in tissues composing the endoderm of the statolith, and specific single cell expression in the surrounding dermal layers. C: 10X photo of comb rows showing expression of *PbalphaC* in single cells surrounding comb rows, at the base of the comb rows, and along the meridional canals. D: Gene tree showing alignment of *P. bachei* collagen type IV alpha chains(Hsapha1-6: *Homo sapiens*, PbalphaA-E: *Pleurobrachia bachei*, PjCol4: *Pseudocorticum jarrei*, Agspongin: *Anopheles gambiae*, Sdspongin1-2: *Suberites domuncula*, Nvspongin: *Nematostella* vectensis, Emspongin: *Ephydatia mulleri*).. Note the two part grouping of *PbαA-C* and *PbαD-F*. Notice also that alpha chains in *P. bachei* show a more diverse intra-gene relationship than human alpha chains. Figure generated using MEGA tree software ML, bootstrap support of less than 70 omitted.
Probe Sequence on next page.

ATCTGGTATGTTGGAGATTAGGCGAGGTTTCTCCTTCCTCTACTTTGACTCTTCAACCGGTACTGGTCACGGTGTCC
CCCTGAGCAGAGAGGGAAGTTGCCTTTCATCTTTCTCCATCCAAC
CTATGCTGGAGTGTACTACCAAATACTGCGAGGTAAAGGGAGACGACAGCTCTCTCTGGATGCCAGTGGGAGATA
TTCCAGAGGGCAGTGCCGCTGGTATGTATTACCCTTCGGATGAACT
AGTGGTTGACGCGACAGATCCTAATGATCAAACTGAGACTGTCCTCAAGTTCATCTCTCGATGTACGGTCTGTGAA
TCACCATCAGTTGTGTTTGCCATTCACTCCTTCACCAACATTGACCC
CTCATGTCCTGATAAATGGGAGGAGCTATGGAATGGATTCTCTCTCATCATGGGTAAGGATCGAGGTGGTGGTGCT
GGGCTGGAACTACAGGGTCACGGATCTTGTGTCAAATACTTCACTCC
TCAGCTTTCCGTACGGTGCAGCCGAACCGGAGGAACAACTGGACGACCGAGAAGGGCGAATTCGTTTAAACCTGC
AGGACTAGTCCCTTTAGTGAGGGTTAATTCTGAGCTTGGCGTAATCA
TGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATAAAGT
GTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTG

Type IV Collagen alpha chains across Metazoa



PHYLA	CLASS	Species	COLLAGEN 4 COUNT	ACCESSION #
Porifera	Homoscleromorpha	*Pseudocorticum jarrei*	1	Q7JMZ8
Ctenophora	Tentacula	*Pleurobrachia bachei*	6	
Cnidaria	Hydrozoa	*H. magnipillata*	2	DN138061, DN636820
	Anthozoa	*Nematostella vectensis*	2	c438003055.Contig1, c415700716.Contig2
Arthrodpoda	Insecta	*D. melanogaster*	2	O18407, P08120
		Bombyx mori	2	BP125709, CK522520
		Anopheles gambiae	2	Q7PVR8, BM588632
Nematoda	Chromadorea	*C. elegans*	2	P1740, P17139
Echinodermata	Echinoidea	*S. pupuratus*	2	Q26640, Q07265
Placozoa	tricoplacia	*Trichoplax adhaerens*	1 or 2	XP_002116198.1, XP_002116198.1
Chordata	Ascidiacea	*C. intestinalis*	2	BW439169, BW229076
	Actinopterygii	*D. rerio*	3	
	Cephalochordata	*B. floridae*	2	BW844807, BW895761
	Aves	*Gallus gallus*	6	Q919K3, XP_416952, AAY43819, XP_422615, XP_420320, XP_420322
	Mammalia	*M Musculus*	6	NM_009931, NM_009932, NM_007734, NM_007735, NM_007736, NM_053185
	Mammalia	*H. sapiens*	6	P02462, P08572, Q01955, P53420, P29400, Q14031

Figure 5/Table 3: Tree shows collagen type IV alpha chain count across the metazoan tree. The most notable data in the tree is the fact that *P. bachei* has 6 distinct alpha chains, similar to the collagen profile of *H. sapiens.* This is again noticeable in table 3, which shows that *P. bachei* has a collagen count more similar to that of vertebrate chordates than other basal metazoans. This figure and table were adapted from Aouacheria et al., 2011.

FIBRILLAR COLLAGENS ACROSS METAZOA

Phylum/Gene	Species/NCBI accession	Phylum/Gene	Species/NCBI accession
Freshwater Sponge	E. mulleri	Abalone	(Haliotis discus)
Emu1α	P18856, Q06452	Hdcol1α	O97405
		Hdicol2α	O97406
Sponge	(A. queenslandia)	Sea Urchin	S. purpuratus
		Spu1α	Q26634
Amq1α	WGS	Spu2α	Q26639
Amq2α	WGS	Spu6α	XP_001192332
Amq3α	WGS		
Amq4α	WGS	Sea Urchin	Paracentrotus lividus
Amq5α	WGS	Pli5α	CAE53096
Amq6α	WGS		
Amq7α	WGS	Ascidian	(Ciona intestinalis)
		Cin759	ci0100150759
Hydra	(H. Magnipapillata)	Cin301	ci0100154301
HCol1	WGS	Cin606	ci0100131606
HCol2	WGS	Cin916	ci0100144916
HCol3	WGS		
HCol5	WGS	Human	Homo sapiens
		Hsaα1(I)	P02452
Sea anenome	(N. vectensis)	Hsaα2(I)	P08123
Nve1α	ID # 25350	Hsaα1(II)	PQ14027
Nve2α	ID #6496	Hsaα1(III)	P02461
Nve3α	ID #21796	Hsaα1(V)	P20908
Nve4α	ID #36007	Hsaα2(V)	P05997
Nve5α	ID #1737	Hsaα3(V)	P25940
Nve6α	ID #26644	Hsaα1(XI)	P12107
Nve7α	ID #166	Hsaα2(XI)	P13942
Nve8α	ID #40962	Hsaα1(XXIV)	Q7Z5L5
		Hsaα1(XXVII)	Q8IZC6
Mosquito	(Anopheles gambiae)		
Agaα1	ENSANGG00000016690		
Agaα2	ENSANGG00000018512		
Honeybee	Apis mellifera		
Ame1α	AADG02012535		
Ame2α	AADG02005865		

(From Exposito et al, 2008)

References

Aouacheria, A., C. Geourjon, et al. (2006). "Insights into early extracellular matrix evolution: spongin short chain collagen-related proteins are homologous to basement membrane type IV collagens and form a novel family widely distributed in invertebrates." Mol Biol Evol 23(12): 2288-2302.

Bruce Alberts, D. B., Julian Lewis, Martin Raff, Kieth Roberts, James D. Watson (1994). Molecular Biology of the Cell, Garland Publishing, Inc.

Bailey, W. J., J. Kim, et al. (1997). "Phylogenetic reconstruction of vertebrate Hox cluster duplications." Mol Biol Evol 14(8): 843-853.

Boot-Handford, R. P. and D. S. Tuckwell (2003). "Fibrillar collagen: the key to vertebrate evolution? A tale of molecular incest." Bioessays 25(2): 142-151.

Exposito, J. Y., C. Larroux, et al. (2008). "Demosponge and sea anemone fibrillar collagen diversity reveals the early emergence of A/C clades and the maintenance of the modular structure of type V/XI collagens from sponge to human." J Biol Chem 283(42): 28226-28235.

Exposito, J. Y., U. Valcourt, et al. (2010). "The fibrillar collagen family." Int J Mol Sci 11(2): 407-426.

Fowler, S. J., S. Jose, et al. (2000). "Characterization of hydra type IV collagen. Type IV collagen is essential for head regeneration and its expression is up-regulated upon exposure to glucose." J Biol Chem 275(50): 39589-39599.

Hudson, B. G., S. T. Reeders, et al. (1993). "Type IV collagen: structure, gene organization, and role in human diseases. Molecular basis of Goodpasture and Alport syndromes and diffuse leiomyomatosis." J Biol Chem 268(35): 26033-26036.

Kohn, A. B., M. R. Citarella, et al. (2012). "Rapid evolution of the compact and unusual mitochondrial genome in the ctenophore, Pleurobrachia bachei." Mol Phylogenet Evol 63(1): 203-207.

Leitinger, B. and E. Hohenester (2007). "Mammalian collagen receptors." Matrix Biol 26(3): 146-155.

Leys, S. P. and A. Riesgo (2011). "Epithelia, an evolutionary novelty of metazoans." J Exp Zool B Mol Dev Evol.

Ricard-Blum, S. and L. Ballut (2011). "Matricryptins derived from collagens and proteoglycans." Front Biosci 16: 674-697.

Shimizu, H., R. Aufschnaiter, et al. (2008). "The extracellular matrix of hydra is a porous sheet and contains type IV collagen." Zoology (Jena) 111(5): 410-418.

Tyler, S. (2003). "Epithelium--the primary building block for metazoan complexity." Integr Comp Biol 43(1): 55-63.