universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

Developmental genes, proneuropeptides and peptide hormones in Mollusca: an *in-silico* approach

verfasst von / submitted by

André Luiz de Oliveira

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2018/ Vienna, 2018

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 794685437
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Biologie
Betreut von / Supervisor:	Univ.-Prof. DDr. Andreas Wanninger

Aos meus pais que, mesmo do outro lado do oceano Atlântico, tornaram possível essa realização. Esse título pertence também a vocês. Obrigado por tudo.

À Bruna, meu casal, pelo incondicional suporte e companheirismo. *“Qualquer amor já é um pouquinho de saúde, um descanso na loucura"* – João Guimarães Rosa.

“What characterizes the living world is both its diversity and its underlying unity.” - Jacob, 1977.

Acknowledgments

I would like to express my special appreciation and thanks …

to Andreas Wanninger for the guidance through the most intellectually stimulating years of my life so far, and to show me that mollusks are more than delicious source of food. Science is not an easy path for anyone, it requires sweat, tremendous amounts of coffee and a highly customisable Linux distro, however, your support and assistance made this whole journey enjoyable and enriching.

to Tim Wollesen for the suggestions and scientific input, especially during the first drafts of what would be my PhD proposal. It was never so easy for someone to dive in into the molluscan world with your help and patience.

to Thomas Schwaha for made me completely integrated into the Austrian/Viennese culture and way of life. My wife often says that I am the most Austrian of the Non-Austrian people, and she is probably right. Tausend dengschee oida.

to Andrew Calcino for the help with the establishment of a Bioinformatics niche in our research group, constructive criticism during my endless analyses and free “Australian slang phrases and terms” lectures.

to the present and past members of the Wanninger Lab, especially Maik Scherholz and Alen Kristof for promptly helping me during my settling in Vienna and the friendly counselling involving all the bureaucratic steps during my PhD and life issues.

to all colleagues at the Department of Integrative Zoology responsible for providing the perfect “*Gesellschaft*” and infrastructure to do high-quality research.

an die Deutsche Sprache, um mir zu zeigen, dass es kompliziertere Sachen als ein Doktorstudium gibt.

to the Brazilian scientific program “Science without Borders” (project number: 6090-13/3) for the full financial support during my doctoral studies.

Table of Contents

ABSTRACT

As one of the most diverse groups of invertebrate animals, with remarkable differences in their life cycles, developmental biology, body plans, and behavioural traits, mollusks represent powerful models for neurobiological and developmental studies. Significant progress has been made with respect to molluscan phylogeny, and recent phylogenomic studies show a deep dichotomous split dividing Mollusca into two lineages: the Aculifera (shell-less Aplacophora and eight-shelled Polyplacophora), and Conchifera (the remaining uni-shelled mollusks) Molecular studies within Mollusca are focused on the three most species-rich class-level taxa: Gastropoda, Bivalvia, and Cephalopoda. The remaining taxa, Polyplacophora, Neomeniomorpha (=Solenogastres), Chaetodermomorpha (=Caudofoveata), Monoplacophora and Scaphopoda, have been investigated to a much lesser extent. This PhD thesis uses in-house next-generation transcriptome and genome sequence data, combined with publicly available resources, to access the molecular landscape of representatives of all extant conchiferan and aculiferan class-level taxa and their putative lophotrochozoan allies. Sequence databases were screened for the presence of important developmental genes (e.g. Hox, ParaHox), and peptide signalling molecules (e.g. neuropeptides and hormones).

From the study focused on developmental genes, the Hox and ParaHox complement of the last common ancestor (LCA) of Mollusca could be inferred. It contained 11 Hox and three ParaHox genes, a situation that was retained in the LCA of Aculifera and Conchifera, and is in agreement with estimations of the complement in the lophotrochozoan LCA. Additionally, lineage-specific signatures in the molluscan *Hox5* and lophotrochozoan *Gsx* gene are presented, as well as the characterisation of other known important developmental gene families (e.g. Hedgehog, Wnt).

Regarding the neuropeptide and peptide hormone survey, a detailed overview of the molluscan proneuropeptide and peptide hormone toolkit is presented. The results expand the distribution of several peptide families within Lophotrochozoa (e.g., prokineticin, insulin-related peptides, prohormone-4, LFRFamide), and provide evidence for an early origin of others (e.g., GNXQN/prohormone-2). Furthermore, the presence of the Wnt antagonist *dickkopf1/2/4* ortholog in lophotrochozoans and nematodes is revealed. Phylogenetic analyses suggest that the egg-laying hormone family is a DH44 homolog restricted to gastropods. Finally, the data show that

numerous peptides evolved much earlier than previously assumed and that key signalling elements are extensively conserved among extant mollusks.

By reconstructing the phylogenetic history of the Euarthropoda ecdysis signalling pathway, this work shows that the key elements responsible for moulting have ancient origins at the deep nodes in the metazoan tree and are widespread among non-moulting animals, including all extant class-level taxa of Mollusca. Eclosion hormone and bursicon originated prior to the cnidarian-bilaterian split, whereas the crustacean cardioactive peptide traces back to the stem of Bilateria. The identification of the eclosion hormone, bursicon and crustacean cardioactive peptide in Onychophora, and eclosion hormone and crustacean cardioactive peptide in Tardigrada, strongly suggest a scenario in which the entire pathway was already functional in the last common ancestor of Panarthropoda. The finding of a trunk-like peptide in the comb jelly *Mnemiopsis* closely related to the arthropod prothoracicotropic hormone, a neurohormone that triggers the ecdysis behavioural cascade in insects, is of particular interest in the light of a proposed independent evolution of ctenophore and cnidarian-bilaterian nervous systems.

ZUSAMMENFASSUNG

Als eine der vielfältigsten Gruppen wirbelloser Tiere, mit bemerkenswerten Unterschieden in Morphologie, Lebenszyklen und Verhaltensmustern, stellen Mollusken leistungsfähige Modelle für die Neuro- und Entwicklungsbiologie dar. Neuere phylogenomische Studien zeigen eine tiefe dichotome Spaltung, die Mollusca in zwei Schwestergruppen aufteilt: die Aculifera (schalenlose Aplacophora und achtschalige Polyplaphophora) und Conchifera (die verbleibenden, primär einschaligen Mollusken). Genom- und Transkriptom-basierte Studien konzentrierten sich bisher auf die drei Klassen Gastropoda, Bivalvia und Cephalopoda. Die verbleibenden Taxa Polyplacophora, Neomeniomorpha (= Solenogastres), Chaetodermomorpha (= Caudofoveata), Monoplacophora und Scaphopoda sind auf molekularer Ebene weit weniger detailliert untersucht. Diese Dissertation verwendet in-house „next generation“ Transkriptom- und Genomsequenzanalysen, in Kombination mit öffentlich verfügbaren Sequenzdaten, um Teile der molekularen Ausstattung all rezenten Molluskenklassen sowie ihrer mutmaßlichen Lophotrochozoen-Verwandten zu rekonstruieren. Sequenzdatenbanken wurden hinsichtlich des Vorhandenseins von wichtigen Entwicklungsgenen (zum Beispiel Hox, ParaHox) und Peptidsignalmolekülen (zum Beispiel Neuropeptide und Hormone) untersucht.

Aus der Studie der Entwicklungsgene konnte das Hox- und ParaHox-Komplement des letzten gemeinsamen Vorfahrens der Mollusca abgeleitet werden. Es enthielt 11 Hox- und drei ParaHox-Gene, eine Situation, die in dem letzten gemeinsamen Vorfahren der Aculifera und Conchifera erhalten blieb und ebenfalls für den letzten gemeinsamen Vorfahren der Lophotrochozoa angenommen wird. Zusätzlich werden taxon-spezifische Signaturen von *Hox5* in Mollusken und *Gsx* innerhalb der Lophotrochozoen dargestellt sowie die Charakterisierung anderer bekannter wichtiger Entwicklungsgenfamilien präsentiert (zum Beispiel Hedgehog, Wnt).

In Bezug auf die Neuropeptid- und Peptidhormon-Studie wird ein detaillierter Überblick über das Proneuropeptid- und Peptidhormon-Arsenal der Mollusken gegeben. Die Ergebnisse erweitern das Vorhandensein etlicher Peptidfamilien innerhalb der Lophotrochozoa (zum Beispiel Prokineticin, Insulin-verwandte Peptide, Prohormone-4, LFRFamide) und liefern Hinweise auf einen frühen Ursprung anderer (z. B. GNXQN / Prohormone-2). Darüber hinaus wird das Vorhandensein des

Orthologs des Wnt-Antagonisten *dickkopf1/2/4* in Lophotrochozoen und Nematoden aufgedeckt. Phylogenetische Analysen legen nahe, dass die Egg-laying-hormone-Familie ein DH44-Homolog ist, das auf Gastropoden beschränkt ist. Schließlich zeigen die Daten, dass zahlreiche Peptide viel früher als bisher angenommen entstanden sind und dass wichtige Signalelemente unter den rezenten Mollusken weitgehend konserviert sind.

Eine in dieser Arbeit erstmalig präsentierte detaillierte Rekonstruktion der evolutionären Geschichte des Häutungs-Signalweges der Euarthropoda zeigt, dass die für die Häutung verantwortlichen Schlüsselelkomponenten tiefe Ursprünge innerhalb der Metazoen aufweisen und ebenfalls unter nicht-Ecdysozoen weit verbreitet sind. Dies schliesst auch die, Mollusken mit ein. Das Eclosionshormon sowie Bursicon entstanden vor der Aufspaltung der Bilateria und Cnidaria, wohingegen das „Crustacean cardioactive Peptid“ auf den Ursprung der Bilateria zurückführt. Die Identifizierung des Eclosionshormons, „Crustacean cardioactive Peptid“ und von Bursicon bei Onychophora sowie von Eclosionhormon und der „Crustacean cardioactive Peptid“ bei den Tardigrada deuten stark auf ein Szenario hin, in dem der gesamte Signalweg bereits im letzten gemeinsamen Vorfahren der Panarthropoda funktionell war. Das Auffinden eines „trunk-like Peptids“ in der Rippenqualle *Mnemiopsis*, dass eng verwandt mit dem Arthropoden-spezifischen „prothoracicotropic Hormon“ ist (ein Neurohormon, das die Häutungskaskade der Insekten auslöst), ist hierbei im Lichte einer vorgeschlagenen unabhängigen Evolution der Nervensysteme bei Ctenophora und Cnidaria/Bilateria von besonderem Interesse.

1. GENERAL INTRODUCTION

1.1 Mollusks

Having successfully conquered a wide range of niches in marine, freshwater and terrestrial environments, mollusks are virtually everywhere. With estimates of up to 200,000 extant species (Ponder & Lindberg, 2008), mollusks are not only speciose, but also morphologically variable and diverse in all aspects of life, including their behavioural traits. They can passively drift in the water column (planktonic) or actively swim (nektonic); they can be sessile or infaunal; they can be highly visual, agile predators, or practically immobile by cementing themselves to the substrate and feeding on suspended matter. Many mollusks are culturally and economically important as sources of food, jewellery (e.g., pearls, shell money), and biomedical applications (pain treatment, e.g. analgesic peptides of cone snails), while others, especially gastropods and bivalves, are a burden for society, destroying crops, being vectors for infectious diseases (e.g. schistosomiasis), and threats to native species as highly competitive neozoans (e.g. zebra and the quagga mussels).

Mollusca currently comprises eight distinct lineages (i.e. class-level taxa), including the well-known Gastropoda (snails, slugs), Cephalopoda (octopus and squids), and Bivalvia (mussels, oysters). Much less familiar groups include the flattened, eight shells plate-bearing chitons, i.e. Polyplacophora, the benthic elephant tooth-like shell mollusks, i.e. Scaphopoda, the limpet-like circular monoplacophorans, i.e. Monoplacophora, and the worm-like shell-less aplacophorans that include the Chaetodermomorpha (=Caudofoveata) and the Neomeniomorpha (=Solenogastres) (for review see Wanninger & Wollesen, 2018).

The striking variations of the molluscan body plan render Mollusca an interesting group for comparative studies, especially with respect to the evolutionary and developmental processes that underlie their rich phenotypic diversification. Moreover, the different lifestyles, developmental pathways (ranging from indirect development via various larval types to direct development), and their highly variable neuroanatomy make them ideal models to understand how evolution has brought about the great and distinct forms of behaviour that they display (Faller et al., 2012; Hochner & Glanzman, 2016; Shigeno et al., 2018).

Due to the extensive and exceptionally preserved fossil record (e.g., Parkhaev 2008, 2017; Sutton et al., 2012; Vinther et al., 2017), mollusks are equally important to the understanding of broader animal relationships and organismal evolution. However, the wealth of paleontological data and the morphological disparity among

mollusks have fostered numerous competing hypotheses over the evolutionary origins of the various molluscan lineages, and consequently about the archetype of the hypothetical ancestral mollusc (HAM or “urmollusk“), i.e. the last common ancestor of all mollusks (Haszprunar & Wanninger, 2012). The controversies about molluscan phylogeny, in accord with Telford & Budd, (2011), constitute one of the greatest challenges in invertebrate evolution, and part of the problem relies to great extent on unravelling molluscan interrelationships, which have always been contentious (Schödl & Stöger, 2014).

1.2 In the pursuit of a reliable mollusk phylogeny

The interrelatedness among the major molluscan lineages has been elusive for a long time (e.g., Haszprunar & Wanninger, 2012; Sigwart & Lindberg, 2014). Most traditional hypotheses of molluscan phylogeny are based on adult morphological characters and generally support the scenario in which worm-like aplacophoran mollusks, i.e. Neomeniomorpha and Chaetodermomorpha, are the earliest extant offshoots (Haszprunar, 2000). When considered paraphyletic, the placement of either of the two aplacophoran groups as sister to the remaining class-level taxa resulted in a clade termed Hepagastralia (Fig. 1A) and Adenopoda (Fig. 1B) (Salvani & Steiner, 1996; Haszprunar 2000). The Testaria concept proposed Polyplacophora as sister to all primarily uni-shelled mollusks, i.e. Conchifera (Waller, 1998) (Fig. 1C). These scenarios support the assumption that molluscan evolution involved a progressive increase in body plan complexity from simple worm-like neomeniomorphs and chaetodermomorphs via polyplacophorans to the more complex conchiferans.

The advance of molecular techniques and *in silico* frameworks to robustly infer homology from sequence datasets provided an independent and powerful source of data to infer organismal relationships (Dunn et al., 2008; Hejnol et al., 2009). Unfortunately, the first molecular investigations on deep molluscan phylogeny yielded inherently contradictory scenarios, mainly caused by the selection of inadequate molecular markers (Kocot, 2013). These studies heavily relied on the small (SSU or 18S) and large (LSU or 28S) subunits of ribosomal gene sequences, mitochondrial markers (*16S* and *COI* genes), and one of the main histone proteins in the structure of the eukaryotic cells, histone H3 (Winnepenninckx et al., 1996; Rosenberg et al.,

1997; Passamaneck et al., 2004; Giribet et al., 2006; Meyer et al., 2010; Wilson et al., 2010). Although these genes have been useful to clarify several questions regarding metazoan relationships (Halanych et al., 1995; Aguinaldo et al., 1997), they either failed to recover the monophyly of individual molluscan class-level taxa (e.g. paraphyly of Bivalvia in Passamaneck et al., 2004), or Mollusca as a whole (paraphyletic clade composed of Neomeniomorpha + Annelida + Sipunculida in Wilson et al., 2010). To further complicate matters, some studies recovered a monophyletic well-supported clade composed of Monoplacophora and Polyplacophora named Serialia (due the presence of serially repeated organs systems such as gills and dorsoventral muscles in these animals) (Giribet et al., 2006; Wilson et al., 2010).

During the vivid debates about molluscan phylogeny, a new method reliant on high-throughput sequencing technologies, phylogenomics, emerged as robust tool to infer phylogeny. This approach employs the comparison of thousands of homologous molecular characters obtained from genomic and transcriptomic data (Meusemann et al., 2010; Pick et al., 2010). The use of the entire complement of protein-coding sequences within species constitutes an appropriate way to circumvent the lack of resolution faced in targeted-gene approaches to solve molluscan phylogeny. In 2011, four phylogenomic studies finally brought some consensus in the understanding of molluscan phylogeny (Kocot et al., 2011; Meyer et al., 2011; Smith et al., 2011; Vinther et al., 2011).

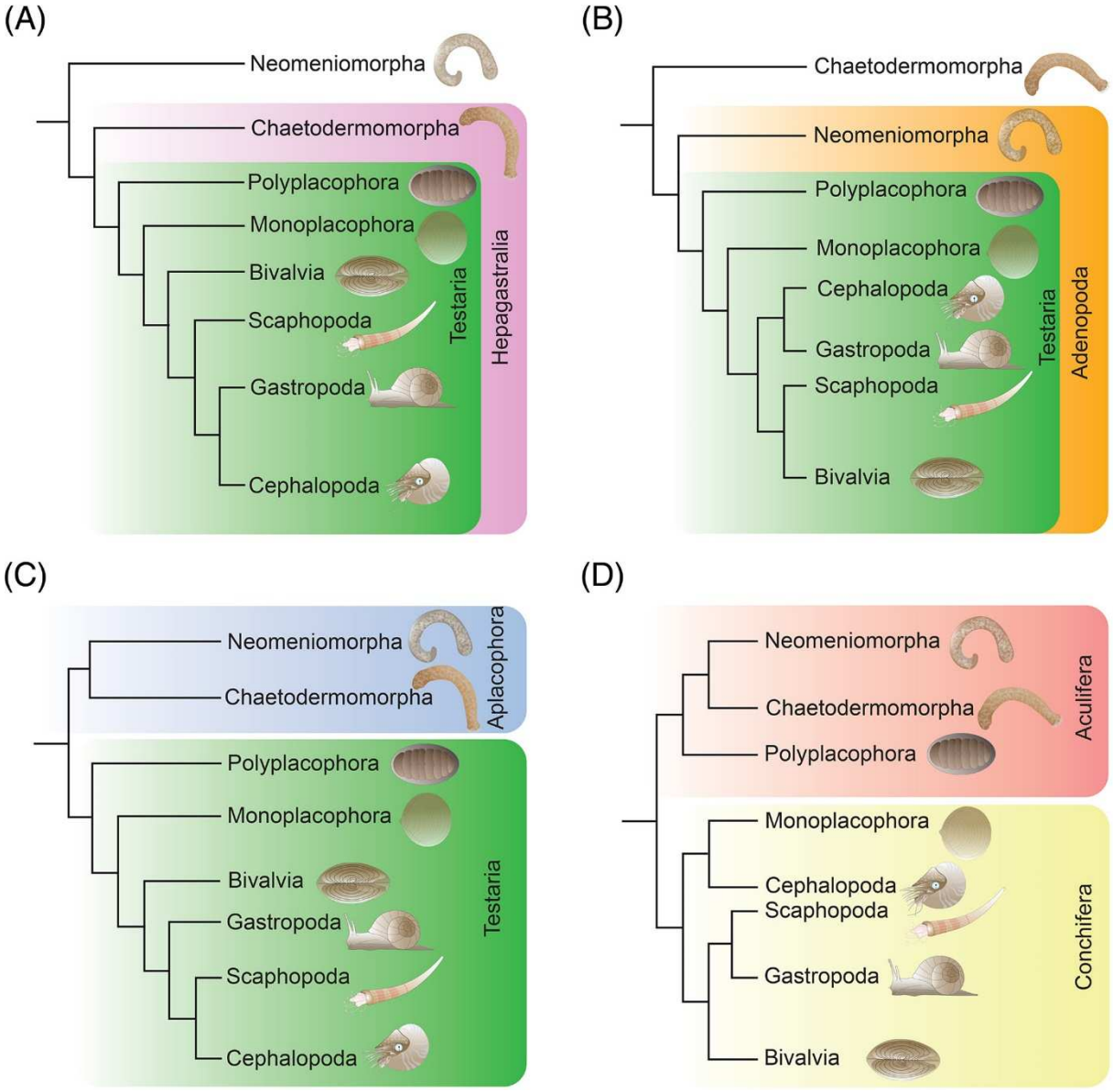


Figure 1 – Summary of major competing hypotheses of molluscan phylogeny (from Wanninger & Wollesen, 2018). (A) Hepagastralia-Testaria concept with Neomeniomorpha as the earliest molluscan branch. (B) Adenopoda – Testaria concept with Chaetodermomorpha as the earliest molluscan branch. (C) Aplacophora-Testaria concept with Aplacophora as sister to the monophyletic Testaria, i.e. Polyplacophora + primarily uni-shelled mollusks. (D) Aculifera-Conchifera concept, according to which Polyplacophora together with Aplacophora forms the clade Aculifera that is sister group to Conchifera. Note that despite the monophyly of Conchifera in all cases, its internal relationships are still contentious.

1.3 Bringing mollusks to the phylogenomics era

The first notable outcome of the 2011 studies was the recovery of the individual-class level taxa and Mollusca monophyly using sequence data, despite the

variable methodological frameworks to process and analyse the molecular data (Table 1). As a result, these four studies represent independent assessments with some congruent results on molluscan phylogeny. Three out of the four studies (Kocot et al., 2011; Vinther et al., 2011; Smith et al., 2011) recovered a monophyletic Aculifera with high support values (Neomeniomorpha + Chaetodermomorpha as monophyletic Aplacophora being sister to Polyplacophora). However, Kocot et al. (2011) and Smith et al. (2011) proposed a scenario in which Aculifera is the sister taxon to all remaining mollusks, implying a deep dichotomous split in the molluscan tree (Fig. 1D – the Aculifera-Conchifera concept; Scheltema, 1993, 1996). By contrast, Vinther et al. (2011) placed Cephalopoda as the closest relatives of aculiferan mollusks, rendering Conchifera paraphyletic. Although Meyer et al. (2011) recovered the monophyly of Mollusca, no support for the Aculifera and Conchifera clades was found.

Within Conchifera the internal relationships are still unclear with many competing scenarios. Of the four studies mentioned above, only one included all recent eight class-level taxa (Smith et al., 2011), with Monoplacophora missing in Vinther et al. (2011) and Kocot et al. (2011), and in Meyer et al. (2011), together with Scaphopoda and Neomeniomorpha. The two most complete phylogenomic studies (Kocot et al., 2011; Smith et al., 2011) presented a strong concordance regarding the Conchifera interrelationships, differing only in the ambiguous position of the Scaphopoda. Both studies recovered a close relationship between Gastropoda and Bivalvia, with Scaphopoda placed either as sister group to Gastropoda + Bivalvia (i.e. Pleistomollusca concept; Kocot et al., 2011), or to Gastropoda alone (Smith et al., 2011). Interestingly, the studies where the monophyly of Conchifera was not supported, the Pleistomollusca concept was also recovered (Meyer et al., 2011; Vinther et al., 2011).

The placement of Cephalopoda remains elusive, with hypotheses proposing a close sister relationship to aculiferan mollusks (Meyer et al., 2011; Vinther et al., 2011), or to the remaining Conchifera class-level taxa (Kocot et al., 2011). The more complete analysis (Smith et al., 2011) recovered an unconventional scenario placing Cephalopoda not the sister group of all other Conchifera, but to Monoplacophora.

Table 1 – Summary of four molecular studies about molluscan phylogeny published in 2011. Monoplacophora data was included only in Smith et al. (2011). Meyer et al. (2011) omitted Neomeniomorpha and Scaphopoda taxa, in addition to Monoplacophora, in their analyses.

	Kocot et al., 2011	Vinther et al., 2011	Smith et al., "big matrix", 2011	Meyer et al., "best 18", 2011
Type of data	Protein-coding sequences	Housekeeping genes	Protein-coding sequences	Ribosomal proteins
# of genes	308	7	301	18
# of taxa	49	51	46	79
# of mollusks	42	31	35	16
# of class-level taxa included	7	7	8	5
Generalised topology	Aculifera + Conchifera	Aculifera	Aculifera + Conchifera	Other

The Aculifera-Conchifera concept (Fig. 1D) is currently the favoured hypothesis, being endorsed by comparative ontogenetic analyses and paleontological data (Scherholz et al., 2013; Vinther at al., 2017), suggesting that worm-like aplacophorans are secondarily simplified in their vermiform body plan. However, as predicted by Telford & Budd (2011), "the fight over molluscan evolution may even now just be warming up", and many authors using different molecular markers, mitogenomics, and expanded taxon sampling have revived old concepts (e.g. Serialia) and the confusion surrounding molluscan phylogeny (Sigwart & Lindberg, 2014; Stöger et al., 2013; Schrödl & Stöger, 2014). The continuous progress in molecular techniques, evo-devo approaches, and bioinformatics will hopefully improve our understanding about molluscan evolutionary history (Kocot, 2013; Wanninger & Wollesen, 2018).

1.4 Molluscan genome sizes and genomics

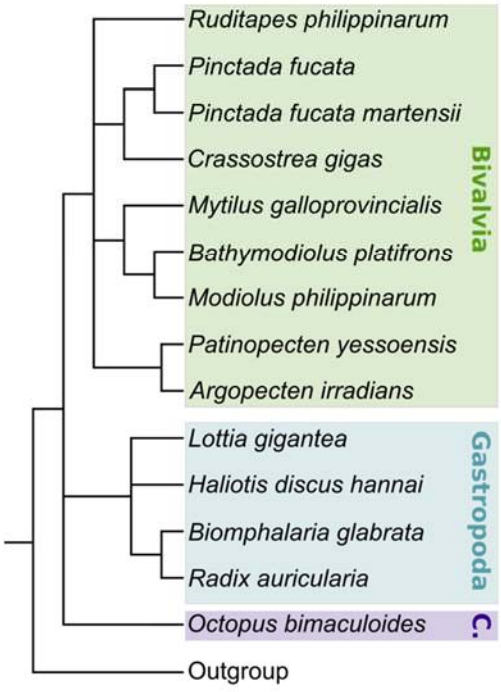
The rapidly changing field of DNA sequencing has proven fundamental for present day research within the biological sciences (Shendure et al., 2017). As the DNA sequencing technologies increased in speed and sequence quality, and

became dramatically cheaper, researches augmented the set of so-called model organisms (e.g. fruit fly, and humans) by the inclusion of poorly studied systems.

In 2012, the draft genomes of two commercially important bivalves, the pearl oyster *Pinctada fucata* and the Pacific oyster *Crassostrea gigas,* were published, rendering them the first molluscan genomic resources publicly available (Figure 2A; Takeuchi et al., 2012, 2016; Zhang et al., 2012). Thereafter, the molluscan research community rejoiced with the release of genomes of seven more bivalves (Du et al., 2017a, 2017b; Murgurella et al., 2016; Mun et al., 2017; Sun et al., 2017; Wang et al., 2017), four gastropods (Adema et al., 2017; Nam et al., 2017; Schell et al., 2017; Simakov et al., 2013), and one cephalopod (Albertin et al., 2015) (see Figure 2A). These studies have shown that many aspects of the biology in conchiferan mollusks (i.e. lifestyle, morphological novelties) are underlined by expansions of specific gene families. The expansion of heat shock protein 70 genes in the oysters *P. fucata* and *C. gigas* in response to the sessile life in the intertidal environment, and different genes involved in chemosensation, apoptosis and immune defense in the pulmonate gastropod *Biomphalaria glabrata*, the intermediate host of *Schistosoma mansoni*, are examples of the interplay between genes and environment (Takeuchi et al., 2012, 2016; Zhang et al., 2012; Adema et al., 2017). The molluscan genomes also provided new insights into the evolution of bilaterian lineages by the increase in the diversity of relatively small number of lophotrochozoan taxa characterised.

Mollusks are also extremely variable in terms of genome sizes, ranging from 0.29 gigabases (Gb), as in the neomeniomorph *Neomenia permagna*, to up to 7.85 Gb, as in the snail *Diplommatina kiiensis kiiensis* (Fig. 2B). This variation is explained, to some extent, by the large proportion of repetitive sequences that are widespread in the molluscan genomes making up more than 40% of the genomic landscape in some species (Fig. 2A). Transposable genetic elements constitute a major force driving genome expansion in mollusks, as observed in many other eukaryotic genomes (Feschotte & Pritham, 2007). Curiously, it has been suggested that in the cephalopod *Octopus bimaculoides* large scale genomic rearrangements closely associated with two bursts of transposon activity that occurred ~25-million and ~56-million years ago, might have played a role in boosting learning and memory.

A



Common name	Genome size (Gb)	# of genes	Repeats	References
Manila clam	1.37 a	108.034	26,38%	Mun et al. 2017
pearl oyster	1.15 b	30.619	37,46%	Takeuchi et al. 2012, 2016
pearl oyster	---	32.937	42,20%	Du et al. 2017a
Pacific oyster	0.637 b	28.027	36,00%	Zhang et al. 2012
Mediterranean mussel	1.60 a	10.891	36,13%	Murgarella et al. 2016
deep sea mussel	1.64 a	33.584	47,90%	Sun et al. 2017
mussel	2.38 a	36.549	62,00%	Sun et al. 2017
scallop	1.43 a	26.415	39,00%	Wang et al. 2017
bay scallop	0.99 a	72.187	---	Du et al. 2017b
own limpet	0.420 c	23.800	21,00%	Simakov et al. 2013
abalone	1.80 b	29.449	30,76%	Nam et al. 2017
fresh water snail	0.916 d	14.423	44,80%	Adema et al. 2017
fresh water snail	1.58 b	17.338	20,50%	Schell et al. 2017
octopus	2.68 b	33.638	45,00%	Albertin et al. 2015

a kmer frequency analysis **c** bulk fluorometric assay
b flow cytometry **d** Feulgen image analysis densitometry

B

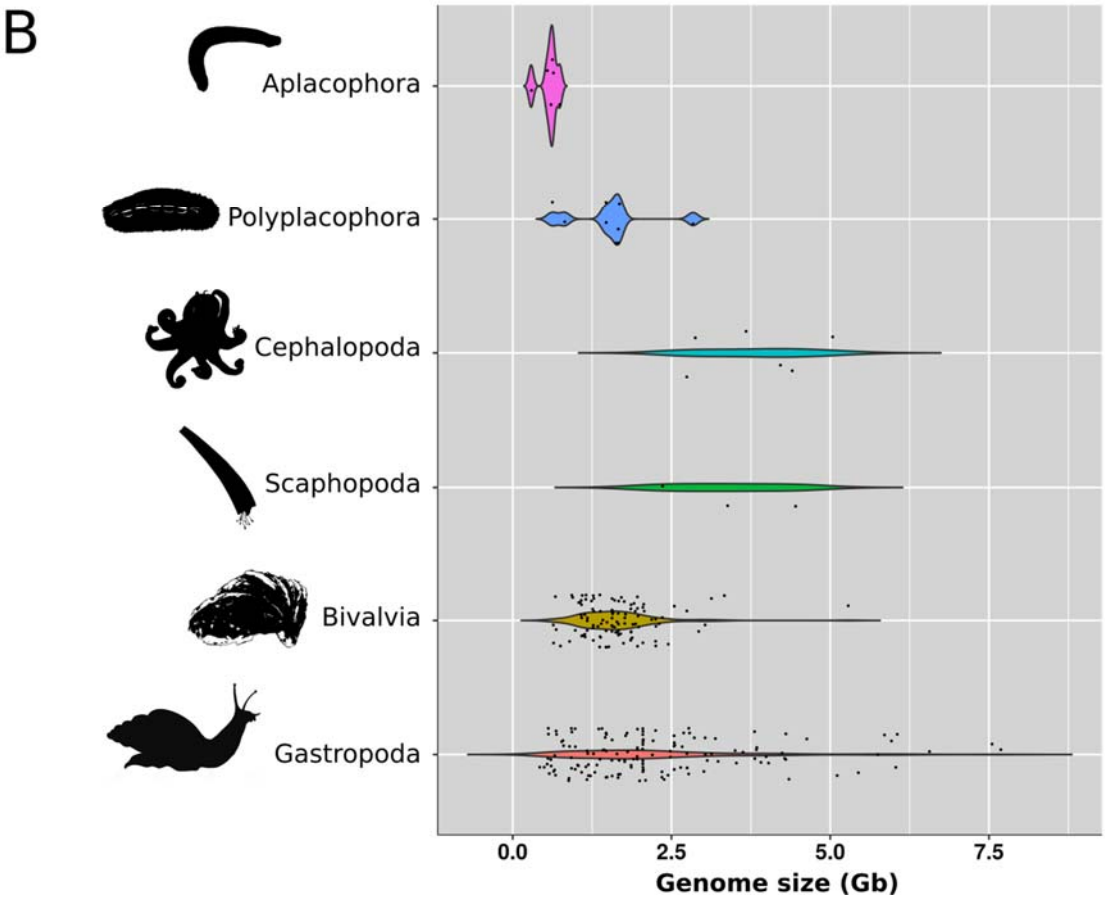


Figure 2 – Summary of publicly available molluscan nuclear genomes and estimates of molluscan genome sizes. (A) Table adapted from Takeuchi (2017) showing the estimated genome sizes from the 14 published molluscan genomes. Genome size estimations were

obtained with four different methodologies: a- kmer frequency analysis; b- flow cytometry; c- bulk fluorometric assay; d- Feulgen image analysis densitometry. The number of predicted genes and percentage of repetitive elements obtained from each molluscan genome are also shown. Phylogenetic relationships among Bivalvia, Gastropoda and octopus (Cephalopoda, abbreviated as C. in the figure) are depicted on the left. (B) Violin plots of 281 molluscan genomes obtained from the Genome Size Database (http://www.genomesize.com/). Haploid DNA contents (C-values, in picograms) were converted to base pairs following Dolezel et al. (2003) formula. Kernel probability density values for the different molluscan class-level taxa are shown (width of the plot). Individual data points, i.e. molluscan genome sizes, are represented by black dots randomly scattered over the graph. The violin plots depict the range of the genome sizes (horizontal axis) and their frequencies (vertical axis) in the different molluscan class-level taxa. Animal silhouettes were obtained from www.phylopic.org and are either licensed under the Creative Commons Attribution 3.0 Unported or available under public domain (credited images used Aplacophora and Polyplacophora: Noah Schlottman and Casey Dunn; Scaphopoda: Brockhaus and Efron; Bivalvia: Taro Maeda and David.Monniaux; Gastropoda: Fernando Carezzano).

1.5 Molluscan developmental genes and neuropeptides

"Evolution of form is very much a matter of teaching very old genes new tricks!"
— Sean B. Carroll

Evolutionary developmental biology (EvoDevo) is a discipline that aims to explore how developmental processes over time brought about the morphological diversity observed in recent and fossil organisms (Müller, 2007; Carrol, 2008). Two central findings in this field involve the recognition that principal regulatory genes are conserved across phyla, and the manner in which they interact with each other (e.g. gene regulatory networks) play a central role in morphogenesis. In other words, the formation and differentiation of many morphologically divergent structures are governed by homologous genes expressed in different spatiotemporal contexts (Raff, 2000; Davidson & Erwin, 2006; Peter & Davidson, 2011). Hox and ParaHox gene families are a classical example of how the very same battery of genetic elements gives rise to different morphological features, i.e. body plans, of animals.

Hox and ParaHox genes are evolutionary sister families that code for transcriptional regulators restricted to the animal kingdom. Among others, they play a

central role in axial patterning during embryonic development of most multicellular animals (Brooke et al., 1998). First identified in the fruit fly, Hox and ParaHox genes are present in the genomes of nearly all animals, and their phylogenetic origins are still under debate (Ramos et al., 2012; Fortunato et al., 2014; for review, Ferrier, 2015). Two remarkable features of Hox and ParaHox genes are their very distinctive organisation into clusters in the genomes, and the manner of how these genes are expressed during the ontogeny of the animals, often following a spatial (i.e. expression matches the genes´ relative order on the chromosome) and temporal (i.e anterior genes are expressed earlier during ontogeny than posterior ones) collinearity of expression (Monteiro & Ferrier, 2006).

Due to the considerable morphological diversity among the eight molluscan class-level taxa, mollusks are excellent candidates to investigate body plan evolution. Thus, many genomic studies have focused on the identification of the Hox and ParaHox complements and their chromosomal organisation in Mollusca (Fig. 3; Albertin et al., 2015; Simakov et al., 2013; Takeuchi et al., 2012, 2016; Wang et al., 2017; Zhang et al., 2012). These studies have shown that cluster alterations, i.e. physical splits of the Hox and ParaHox clusters, and gene losses are frequent within the conchiferan genomes (Fig. 3). In the cephalopod *Octopus bimaculoides*, for instance, the ParaHox gene *Xlox* and three Hox genes (*Hox2*, *Hox3*, and *Hox4*) were lost and the Hox and ParaHox clusters appear dissociated with the individual genes being distributed on different scaffolds.

It has been proposed that the diversity of molluscan morphology has resulted from the lack of a rigid exo- or internal skeleton constraints, as found in ecdysozoans and vertebrate deuterostomes, and the great plasticity in the developmental frameworks (reviewed by Hochner & Glanzman, 2016). The latter hypothesis is currently supported by evidence gathered from gene expression studies in four molluscan class-level taxa – Gastropoda, Cephalopoda, Scaphopoda, and Polyplacophora. In the gastropod *Gibbula varia* and in the Hawaiian bobtail squid, *Euprymna scolopes*, Hox genes are seemingly expressed in a non-collinear manner in both trochozoan- and molluscan-specific structures (Hinman et al., 2003; Lee et al., 2003; Samadi & Steiner, 2009, 2010), whereas in the polyplacophoran *Acanthochitona crinita*, a staggered expression pattern located in all germ layers along the antero-posterior axis is observed (Fritsch et al., 2015, 2016). Curiously, in the scaphopod *Antalis entalis* Hox genes are expressed in a near-to staggered

fashion during ontogeny, similar to *A. crinita*, however they also appear to be involved in the patterning of molluscan morphological traits, as in cephalopods and gastropods (Wollesen et al., 2018).

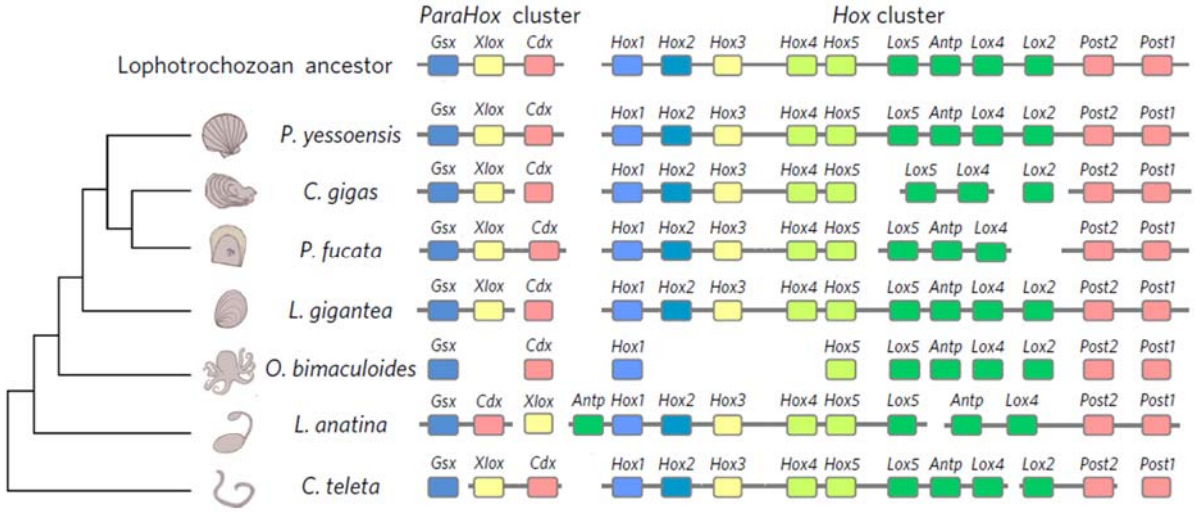


Figure 3 – Chromosomal organization of Hox and ParaHox genes in conchiferan mollusks and other lophotrochozoans (adapted from Wang et al., 2017). Hox and ParaHox clusters are usually fragmented in lophotrochozoans, with the exception of the scallop *Patinopecten yessoensis*, possibly representing the ancestral state of these clusters in the last common lophotrochozoan ancestor (top). Note that the *Lox2* gene of *Pinctada fucata* is missing in the figure, albeit with the release of the version 2.0 of the pearl oyster genome, all Hox and ParaHox genes were identified (Takeuchi et al., 2016).

Neuropeptides constitute a diverse class of neuronal signalling molecules crucial for homeostasis, behavioural patterns, and numerous physiological processes in animals, e.g. digestion, diuresis, pain perception, and food intake. They may act as neurotransmitters, neuromodulators, hormones, growth factors, and are important regulators of behavioural actions linked to courtship, sleep, learning, addiction, stress, and ecdysis (Schoofs et al., 2017; Elphick et al., 2018). Neuropeptides are evolutionary ancient signalling molecules widely distributed in metazoans, present in cnidarians (for review see Takahashi & Takeda, 2015), the nerveless placozoans (Smith et al., 2014; Varoqueaux et al., 2018), ctenophores (Moroz et al., 2014), and bilaterians (Jékely, 2013; Mirabeau & Joly, 2013). Despite their absence in sponges, the mining of the *Amphimedon queeslandica* genome identified many well-conserved processing enzymes (e.g. PAM, furins, preprotein convertases) required for cleavage and maturation of neuropeptides from their inactive precursors (Srivastava et al.,

2010). These findings indicate that the enzymatic toolkit necessary for neuropeptide signalling predates the origin of neuronal tissues.

Gastropod and bivalve mollusks have been successfully employed as model systems to understand how the interplay between the structural and functional diversity of neuropeptides generate and modulate the complex behaviours and physiology in these animals (Muneoka & Kobayashi, 1992; Bailey et al., 1996). For instance, aplysiid gastropods (sea hares) are paramount models in neuroscience, especially in the biology of learning and memory, due to their small number of giant neurons that facilitates physiological and biochemical studies (Glanzman, 2009; Moroz, 2011). Moreover, several studies on neuropeptide function and evolution have been triggered by discoveries from molluscan models. The FMRFamide peptide family, which has been shown to be widely distributed in metazoans (Walker et al., 2009; Jékely, 2013), was first identified in the bivalve *Macrocallista nimbosa* (Price and Greenberg, 1977). Despite the important role of mollusks in neuroscience and neuroendocrinology, broad comparative studies outside the Conchifera clade are still missing.

1.6 Thesis aims

Molecular studies within Mollusca are overwhelmingly focused on the three most species-rich class-level taxa: Gastropoda, Bivalvia, and Cephalopoda. The remaining taxa, the aplacophorans (Neomeniomorpha and Chaetodermomorpha), Polyplacophora, Monoplacophora and Scaphopoda have been investigated to a much lesser extent, hindering the understanding of how the highly variable forms of body plans, physiological processes, and behaviours evolved in this phylum. This thesis revolves around questions in molluscan evolution that benefit from a broad comparative approach, such as the identification of important molecular components involved in the body plan specification, and in the myriad of behaviours and physiological processes displayed by the eight extant class-level taxa of Mollusca. To fill this gap in knowledge, this thesis combines novel high-quality next-generation transcriptomic and genomic data with publicly available resources in order to:

i. Develop new *in silico* pipelines for molecular data pre-processing, assembling, and identification of the target genes and gene families (Manuscript 1, 2 and 3);

ii. Identify and compare the developmental gene toolkit, particularly Hox and ParaHox gene families, of the extant molluscan class-level taxa in the light of metazoan evolution (Manuscript 1);
iii. Identify and compare the neuropeptide toolkit of extant molluscan class-level taxa in the light of the metazoan evolution (Manuscript 2);
iv. Clarify the phyletic distribution of key components of the Euarthropoda ecdysis pathway in all major branches of the Metazoa (Manuscript 3).

By addressing the above specific aims, this thesis not only provides one of the first large-scale comparative genomic and transcriptomic studies on Mollusca, but also establishes an extensive framework for understanding some of the underlying molecular mechanisms involved in molluscan morphogenesis and many physiological processes. Additionally, the investigation about distribution of the ecdysis pathway components in mollusks, with a large comparison across metazoans, elucidates the evolutionary history of these elements within Metazoa and provides new insights into the evolution of moulting in Euarthropoda.

1.7 References

Adema, C. M., Hillier, L. W., Jones, C. S., Loker, E. S., Knight, M., Minx, P., et al. (2017). Whole genome analysis of a schistosomiasis-transmitting freshwater snail. *Nature communications*, *8*, 15451.

Aguinaldo, A. M. A., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A., et al. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, *387*(6632), 489-493.

Albertin, C. B., Simakov, O., Mitros, T., Wang, Z. Y., Pungor, J. R., Edsinger-Gonzales, E., et al. (2015). The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*, *524*(7564), 220.

Bailey, C. H., Bartsch, D., & Kandel, E. R. (1996). Toward a molecular definition of long-term memory storage. *Proceedings of the National Academy of Sciences*, *93*(24), 13445-13452.

Brooke, N. M., Garcia-Fernàndez, J., & Holland, P. W. (1998). The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature*, *392*(6679), 920.

Carroll, S. B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, *134*(1), 25-36.

Davidson, E. H., & Erwin, D. H. (2006). Gene regulatory networks and the evolution of animal body plans. *Science*, *311*(5762), 796-800.

Dolezel, J., Bartos, J., Voglmayr, H., & Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. *Cytometry*. *Part A: the journal of the International Society for Analytical Cytology*, *51*(2), 127-8.

Du, X., Fan, G., Jiao, Y., Zhang, H., Guo, X., Huang, R., et al. (2017a). The pearl oyster *Pinctada fucata martensii* genome and multi-omic analyses provide insights into biomineralization. *GigaScience*, *6*(8), gix059.

Du, X., Song, K., Wang, J., Cong, R., Li, L., Zhang, G. (2017b). Draft genome and SNPs associated with carotenoid accumulation in adductor muscles of bay scallop (*Argopecten irradians). Journal of genomics*, *5*, 83.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., et al. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, *452*(7188), 745.

Elphick, M. R., Mirabeau, O., & Larhammar, D. (2018). Evolution of neuropeptide signalling systems. *Journal of Experimental Biology*, *221*(3), jeb151092.

Faller, S., Rothe, B. H., Todt, C., Schmidt-Rhaesa, A., & Loesel, R. (2012). Comparative neuroanatomy of Caudofoveata, Solenogastres, Polyplacophora, and

Scaphopoda (Mollusca) and its phylogenetic implications. *Zoomorphology*, *131*(2), 149-170.

Ferrier, D. E. (2015). The origin of the Hox/ParaHox genes, the Ghost Locus hypothesis and the complexity of the first animal. *Briefings in functional genomics*, 15(5), 333-341.

Feschotte, C., & Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.*, *41*, 331-368.

Fortunato, S. A., Adamski, M., Ramos, O. M., Leininger, S., Liu, J., Ferrier, D. E., et al. (2014). Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. *Nature*, 514(7524), 620.

Fritsch, M., Wollesen, T., De Oliveira, A. L., & Wanninger, A. (2015). Unexpected co-linearity of Hox gene expression in an aculiferan mollusk. *BMC evolutionary biology*, *15*(1), 151.

Fritsch, M., Wollesen, T., & Wanninger, A. (2016). Hox and ParaHox gene expression in early body plan patterning of polyplacophoran mollusks. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *326*(2), 89-104.

Giribet, G., Okusu, A., Lindgren, A. R., Huff, S. W., Schrödl, M., & Nishiguchi, M. K. (2006). Evidence for a clade composed of molluscs with serially repeated structures: monoplacophorans are related to chitons. *Proceedings of the National Academy of Sciences*, *103*(20), 7723-7728.

Glanzman, D. L. (2009). Habituation in *Aplysia*: the Cheshire cat of neurobiology. *Neurobiology of learning and memory*, *92*(2), 147-154.

Halanych, K. M., Bacheller, J. D., Aguinaldo, A. M., Liva, S. M., Hillis, D. M., & Lake, J. A. (1995). Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. *Science*, *267*(5204), 1641-1643.

Haszprunar, G. (2000). Is the Aplacophora monophyletic? A cladistic point of view. *American Malacological Bulletin*, *15*(2), 115-130.

Haszprunar, G., & Wanninger, A. (2012). Molluscs. *Current Biology*, *22*(13), R510-R514.

Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G. W., Edgecombe, G. D., et al. (2009). Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings of the Royal Society of London B: Biological Sciences*, *276*(1677), 4261-4270.

Hinman, V. F., O'Brien, E. K., Richards, G. S., & Degnan, B. M. (2003). Expression of anterior Hox genes during larval development of the gastropod *Haliotis asinina*. *Evolution & development*, *5*(5), 508-521.

Hochner, B., & Glanzman, D. L. (2016). Evolution of highly diverse forms of behavior in molluscs. *Current Biology*, *26*(20), R965-R971.

Jékely, G. (2013). Global view of the evolution and diversity of metazoan neuropeptide signaling. *Proceedings of the National Academy of Sciences*, *110*(21), 8702-8707.

Kocot, K. M., Cannon, J. T., Todt, C., Citarella, M. R., Kohn, A. B., Meyer, A. et al. (2011). Phylogenomics reveals deep molluscan relationships. *Nature*, *477*(7365), 452.

Kocot, K. M. (2013). Recent advances and unanswered questions in deep molluscan phylogenetics. *American Malacological Bulletin*, *31*(1), 195-208.

Lee, P. N., Callaerts, P., de Couet, H. G., & Martindale, M. Q. (2003). Cephalopod Hox genes and the origin of morphological novelties. *Nature*, *424*(6952), 1061.

Meusemann, K., von Reumont, B. M., Simon, S., Roeding, F., Strauss, S., Kück, P., et al. (2010). A phylogenomic approach to resolve the arthropod tree of life. *Molecular biology and Evolution*, *27*(11), 2451-2464.

Meyer, A., Todt, C., Mikkelsen, N. T., & Lieb, B. (2010). Fast evolving 18S rRNA sequences from Solenogastres (Mollusca) resist standard PCR amplification and give new insights into mollusk substitution rate heterogeneity. *BMC evolutionary biology*, *10*(1), 70.

Meyer, A., Witek, A., & Lieb, B. (2011). Selecting ribosomal protein genes for invertebrate phylogenetic inferences: how many genes to resolve the Mollusca? *Methods in Ecology and Evolution*, *2*(1), 34-42.

Mirabeau, O., & Joly, J. S. (2013). Molecular evolution of peptidergic signaling systems in bilaterians. *Proceedings of the national academy of sciences*, *110*(22), E2028-E2037.

Monteiro, A. S., & Ferrier, D. E. (2006). Hox genes are not always Colinear. *International journal of biological sciences*, *2*(3), 95.

Moroz, L. L. (2011). *Aplysia*. *Current biology*: *CB*, *21*(2), R60.

Moroz, L. L., Kocot, K. M., Citarella, M. R., Dosung, S., Norekian, T. P., Povolotskaya, I. S, et al. (2014). The ctenophore genome and the evolutionary origins of neural systems. *Nature*, *510*(7503), 109.

Müller, G. B. (2007). Evo–devo: extending the evolutionary synthesis. *Nature reviews genetics*, *8*(12), 943.

Mun, S., Kim, Y. J., Markkandan, K., Shin, W., Oh, S., Woo, J., et al. (2017). The whole-genome and transcriptome of the manila clam (*Ruditapes philippinarum*). *Genome biology and evolution*, *9*(6), 1487-1498.

Muneoka, Y., & Kobayashi, M. (1992). Comparative aspects of structure and action of molluscan neuropeptides. *Experientia*, *48*(5), 448-456.

Murgarella, M., Puiu, D., Novoa, B., Figueras, A., Posada, D., Canchaya, C. (2016). A first insight into the genome of the filter-feeder mussel *Mytilus galloprovincialis*. *PLoS One*, *11*(3), e0151561.

Nam, B. H., Kwak, W., Kim, Y. O., Kim, D. G., Kong, H. J., Kim, W. J., et al. (2017). Genome sequence of pacific abalone (*Haliotis discus hannai*): the first draft genome in family Haliotidae. *GigaScience*, *6*(5), 1-8.

Parkhaev, P. Y. (2008). The early Cambrian radiation of Mollusca. *Phylogeny and Evolution of the Mollusca, 33*.

Parkhaev, P. Y. (2017). Origin and the early evolution of the phylum Mollusca. *Paleontological Journal*, *51*(6), 663-686.

Passamaneck, Y. J., Schander, C., & Halanych, K. M. (2004). Investigation of molluscan phylogeny using large-subunit and small-subunit nuclear rRNA sequences. *Molecular phylogenetics and evolution*, *32*(1), 25-38.

Peter, I. S., & Davidson, E. H. (2011). Evolution of gene regulatory networks controlling body plan development. *Cell*, *144*(6), 970-985.

Pick, K. S., Philippe, H., Schreiber, F., Erpenbeck, D., Jackson, D. J., Wrede, P., et al. (2010). Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Molecular biology and evolution*, *27*(9), 1983-1987.

Ponder, W., & Lindberg, D. R. (Eds.). (2008). *Phylogeny and Evolution of the Mollusca*. Univ of California Press.

Price, D. A., & Greenberg, M. J. (1977). Structure of a molluscan cardioexcitatory neuropeptide. *Science*, *197*(4304), 670-671.

Raff, R. A. (2000). Evo-devo: the evolution of a new discipline. *Nature Reviews Genetics*, *1*(1), 74.

Ramos, O. M., Barker, D., & Ferrier, D. E. (2012). Ghost loci imply Hox and ParaHox existence in the last common ancestor of animals. *Current biology*, *22*(20), 1951-1956.

Rosenberg, G., Tillier, S., Tillier, A., Kuncio, G. S., Hanlon, R. T., Masselot, M., et al. (1997). Proceedings of an international meeting: ribosomal rna phylogeny of selected major clades in the mollusca. *Journal of Molluscan Studies*, *63*(3), 301-309.

Salvini-Plawen, L., & Steiner, G. Synapomorphies and plesiomorphies in higher classification of Mollusca. Origin and evolutionary radiation of the Mollusca. Edited by: Taylor J. 1996.

Samadi, L., & Steiner, G. (2009). Involvement of Hox genes in shell morphogenesis in the encapsulated development of a top shell gastropod (*Gibbula varia* L.). *Development genes and evolution*, *219*(9-10), 523-530.

Samadi, L., & Steiner, G. (2010). Expression of Hox genes during the larval development of the snail, *Gibbula varia* (L.)—further evidence of non-colinearity in molluscs. *Development genes and evolution*, *220*(5-6), 161-172.

Schell, T., Feldmeyer, B., Schmidt, H., Greshake, B., Tills, O., Truebano, M, et al. (2017). An annotated draft genome for *Radix auricularia* (Gastropoda, Mollusca). *Genome biology and evolution*, *9*(3), 00-00.

Scheltema, A. H. (1993). Aplacophora as progenetic aculiferans and the coelomate origin of mollusks as the sister taxon of Sipuncula. *The Biological Bulletin*, *184*(1), 57-78.

Scheltema, A. (1996). Phylogenetic position of Sipuncula, Mollusca and the progenetic Aplacophora. *Origin and evolutionary radiation of the Mollusca*.

Scherholz, M., Redl, E., Wollesen, T., Todt, C., & Wanninger, A. (2013). Aplacophoran mollusks evolved from ancestors with polyplacophoran-like features. *Current Biology*, *23*(21), 2130-2134.

Schoofs, L., De Loof, A., & Van Hiel, M. B. (2017). Neuropeptides as regulators of behavior in insects. *Annual review of entomology*, *62*, 35-52.

Schrödl, M., & Stöger, I. (2014). A review on deep molluscan phylogeny: old markers, integrative approaches, persistent problems. *Journal of natural history*, *48*(45-48), 2773-2804.

Sigwart, J. D., & Lindberg, D. R. (2014). Consensus and confusion in molluscan trees: evaluating morphological and molecular phylogenies. *Systematic biology*, *64*(3), 384-395.

Simakov, O., Marletaz, F., Cho, S. J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., et al. (2013). Insights into bilaterian evolution from three spiralian genomes. *Nature*, *493*(7433), 526.

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., et al. (2017). DNA sequencing at 40: past, present and future. *Nature*, *550*(7676), 345.

Smith, S. A., Wilson, N. G., Goetz, F. E., Feehery, C., Andrade, S. C., Rouse, G. W., et al. (2011). Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*, *480*(7377), 364.

Smith, C. L., Varoqueaux, F., Kittelmann, M., Azzam, R. N., Cooper, B., Winters, C. A., et al. (2014). Novel cell types, neurosecretory cells, and body plan of the early-diverging metazoan *Trichoplax adhaerens*. *Current Biology*, *24*(14), 1565-1572.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M. E., Mitros, T., et al. (2010). The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature*, *466*(7307), 720.

Stöger, I., Sigwart, J. D., Kano, Y., Knebelsberger, T., Marshall, B. A., Schwabe, E., et al. (2013). The Continuing Debate on Deep Molluscan Phylogeny: Evidence for Serialia (Mollusca, Monoplacophora). *BioMed Research International*, 2013.

Sun, J., Zhang, Y., Xu, T., Zhang, Y., Mu, H., Zhang, Y., et al. (2017). Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nature ecology & evolution*, *1*(5), 0121.

Sutton, M. D., Briggs, D. E., Siveter, D. J., Siveter, D. J., & Sigwart, J. D. (2012). A Silurian armoured aplacophoran and implications for molluscan phylogeny. *Nature*, *490*(7418), 94.

Takahashi, T., & Takeda, N. (2015). Insight into the molecular and functional diversity of cnidarian neuropeptides. *International journal of molecular sciences*, *16*(2), 2610-2625.

Takeuchi, T., Kawashima, T., Koyanagi, R., Gyoja, F., Tanaka, M., Ikuta, T., et al. (2012). Draft genome of the pearl oyster *Pinctada fucata*: a platform for understanding bivalve biology. *DNA research*, *19*(2), 117-130.

Takeuchi, T., Koyanagi, R., Gyoja, F., Kanda, M., Hisata, K., Fujie, M. (2016). Bivalve-specific gene expansion in the pearl oyster genome: implications of adaptation to a sessile lifestyle. *Zoological letters*, *2*(1), 3.

Takeuchi, T. (2017). Molluscan Genomics: Implications for Biology and Aquaculture. *Current Molecular Biology Reports*, *3*(4), 297-305.

Telford, M. J., & Budd, G. E. (2011). Invertebrate evolution: bringing order to the molluscan chaos. *Current Biology*, *21*(23), R964-R966.

Varoqueaux, F., Williams, E. A., Grandemange, S., Truscello, L., Kamm, K., Schierwater, B. (2018). High Cell Diversity and Complex Peptidergic Signaling Underlie Placozoan Behavior. *Current Biology*, *28*, 1-7.

Vinther, J., Sperling, E. A., Briggs, D. E., & Peterson, K. J. (2011). A molecular palaeobiological hypothesis for the origin of aplacophoran molluscs and their derivation from chiton-like ancestors. *Proceedings of the Royal Society of London B: Biological Sciences*, (*1732*), 1259-1268.

Vinther, J., Parry, L., Briggs, D. E., & Van Roy, P. (2017). Ancestral morphology of crown-group molluscs revealed by a new Ordovician stem aculiferan. *Nature*, *542*(7642), 471.

Walker, R. J., Papaioannou, S., & Holden-Dye, L. (2009). A review of FMRFamide- and RFamide-like peptides in metazoa. *Invertebrate neuroscience*, *9*(3-4), 111-153.

Waller, T. R. (1998). Origin of the molluscan class Bivalvia and a phylogeny of major groups. *Bivalves: An eon of evolution*, *1*(4), 5.

Wanninger, A., & Wollesen, T. (2018). The evolution of molluscs. *Biological Reviews*.

Wang, S., Zhang, J., Jiao, W., Li, J., Xun, X., Sun, Y., et al. (2017). Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nature ecology & evolution*, *1*(5), 0120.

Wilson, N. G., Rouse, G. W., & Giribet, G. (2010). Assessing the molluscan hypothesis Serialia (Monoplacophora+ Polyplacophora) using novel molecular data. *Molecular Phylogenetics and Evolution*, *54*(1), 187-193.

Winnepenninckx, B., Backeljau, T., & De Wachter, R. (1996). Investigation of molluscan phylogeny on the basis of 18S rRNA sequences. *Molecular Biology and Evolution*, *13*(10), 1306-1317.

Wollesen, T., Monje, S. V. R., de Oliveira, A. L., & Wanninger, A. (2018). Staggered Hox expression is more widespread among molluscs than previously appreciated. *Proc. R. Soc. B*, *285*(1888), 20181513.

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., et al. (2012). The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, *490*(7418), 49.

2. RESULTS

2.1 Manuscript 1 - Comparative transcriptomics enlarges the toolkit of known developmental genes in mollusks (published)

De Oliveira, AL; Wollesen, T; Kristof, A; Scherholz, M; Redl, E; Todt, C; Bleidorn, C; Wanninger, A. Comparative transcriptomics enlarges the toolkit of known developmental genes in mollusks. BMC Genomics, v.17, p.905, 2016.

Status: published in BMC Genomics

De Oliveira *et al. BMC Genomics* (2016) 17:905
DOI 10.1186/s12864-016-3080-9

BMC Genomics

RESEARCH ARTICLE Open Access

Comparative transcriptomics enlarges the toolkit of known developmental genes in mollusks

A. L. De Oliveira[1], T. Wollesen[1], A. Kristof[1], M. Scherholz[1], E. Redl[1], C. Todt[2], C. Bleidorn[3,4] and A. Wanninger[1*]

Abstract

Background: Mollusks display a striking morphological disparity, including, among others, worm-like animals (the aplacophorans), snails and slugs, bivalves, and cephalopods. This phenotypic diversity renders them ideal for studies into animal evolution. Despite being one of the most species-rich phyla, molecular and *in silico* studies concerning specific key developmental gene families are still scarce, thus hampering deeper insights into the molecular machinery that governs the development and evolution of the various molluscan class-level taxa.

Results: Next-generation sequencing was used to retrieve transcriptomes of representatives of seven out of the eight recent class-level taxa of mollusks. Similarity searches, phylogenetic inferences, and a detailed manual curation were used to identify and confirm the orthology of numerous molluscan Hox and ParaHox genes, which resulted in a comprehensive catalog that highlights the evolution of these genes in Mollusca and other metazoans. The identification of a specific molluscan motif in the Hox paralog group 5 and a lophotrochozoan ParaHox motif in the *Gsx* gene is described. Functional analyses using KEGG and GO tools enabled a detailed description of key developmental genes expressed in important pathways such as Hedgehog, Wnt, and Notch during development of the respective species. The KEGG analysis revealed *Wnt8*, *Wnt11*, and *Wnt16* as Wnt genes hitherto not reported for mollusks, thereby enlarging the known Wnt complement of the phylum. In addition, novel *Hedgehog* (*Hh*)-related genes were identified in the gastropod *Lottia* cf. *kogamogai*, demonstrating a more complex gene content in this species than in other mollusks.

Conclusions: The use of *de novo* transcriptome assembly and well-designed *in silico* protocols proved to be a robust approach for surveying and mining large sequence data in a wide range of non-model mollusks. The data presented herein constitute only a small fraction of the information retrieved from the analysed molluscan transcriptomes, which can be promptly employed in the identification of novel genes and gene families, phylogenetic inferences, and other studies using molecular tools. As such, our study provides an important framework for understanding some of the underlying molecular mechanisms involved in molluscan body plan diversification and hints towards functions of key developmental genes in molluscan morphogenesis.

Keywords: Bioinformatics, EvoDevo, Evolution, Evolutionary developmental biology, Genomics, Mollusca, Next-generation sequencing, NGS, Transcriptomics, RNA-seq

* Correspondence: andreas.wanninger@univie.ac.at
[1]Department of Integrative Zoology, Faculty of Life Sciences, University of Vienna, Althanstraße 14, Vienna 1090, Austria
Full list of author information is available at the end of the article

Background

Over the past decade, an ever increasing number of molecular data has become available for representatives of numerous animal phyla. It has been shown that many genes are evolutionary conserved, either sharing similar functions or being co-opted into various novel functions, thereby often displaying astounding functional plasticity during animal development (e.g., [1–3]). A substantial body of evidence suggests that evolutionary changes or variations in the regulation of highly conserved developmental genes, as well as divergence in gene sequences (e.g., duplications and mutations), have been responsible for major alterations in the evolution of animal body plans [4–6]. Within these conserved genes, two families of homeotic genes that encode transcription factors and are involved in bilaterian anterior-posterior axis and/or digestive tract patterning, the Hox and ParaHox genes, are among the best-investigated so far [7–9]. Therefore, understanding and reconstructing the evolutionary history of these gene families is crucial for inferring animal evolution and the relationships between genetic and morphological complexity [10, 11].

Comparisons between Hox and ParaHox gene clusters support the hypothesis that both families evolved from an early duplication of an ancient ProtoHox cluster [12–15]. Thereby, the Hox and ParaHox clusters underwent different evolutionary pathways, in which the Hox cluster expanded by several tandem duplications, whereas the ParaHox cluster, composed of *Gsx* (paralog of the anterior Hox genes), *Cdx* (paralog of the *Hox3* gene), and *Xlox* (paralog of the posterior Hox genes), did not. Within Lophotrochozoa, a major group of protostome animals that often show a spiral cleavage pattern and/or a ciliated larva in their life cycle, the Hox and ParaHox families are usually composed of 11 and three genes, respectively [16]. Although the majority of studies are restricted to two lophotrochozoan phyla (Mollusca and Annelida), these results suggest that the last common ancestor of all lophotrochozoan animals also harbored a toolkit that included 11 Hox and three ParaHox genes.

The phylum Mollusca comprises approximately 200,000 living species, ranking it the second-most speciose metazoan phylum [17]. Most mollusks, like numerous other lophotrochozoans, display a highly conserved pattern of spiral cleavage in the early embryo, resulting in the formation of four vegetal macromeres and four animal micromeres. In many basally branching clades, embryology is followed by indirect development via a free-swimming, ciliated trochophore-like larva which most likely constitutes the ancestral condition for Mollusca. This type of larva is commonly found in caudofoveates (= chaetodermomorphs) [18], polyplacophorans [19], gastropods [20], scaphopods [21–23], and bivalves (e.g., [24, 25]; see [26] for review). Many gastropods and bivalves develop a secondary, planktotrophic larva, the veliger, while solenogasters (= neomeniomorphs) and protobranch bivalves have independently evolved a secondary lecithotrophic larval type, the so-called pericalymma or test cell larva (see [26] for review; [27–29]).

In evo-devo research, mollusks occupy an important role in studies focused on the function and expression of regulatory genes during development, providing insights into the mechanisms that underlie the diversification of metazoan body plans [30]. To this end, several transcriptomic studies focusing on biomineralisation processes and their concordant genes have recently become available [31–34]. However, given the high morphological disparity, the complex life cycles, and the striking variation during the ontogeny among molluscan taxa, there is a considerable lack of molecular studies dealing with the expression of key developmental genes in this phylum. As such, only a few gene expression studies have been published, including Hox genes [35–41] and ParaHox genes [42–44]. These studies suggest a high plasticity and recruitment into novel functions of these genes at least in cephalopods and gastropods. Since these data stem from very few species only, the full complement of Hox and ParaHox gene expression domains (and hence their putative functions) in Mollusca is yet to be analysed. To this end, an improvement of the equally poor database of other molluscan developmental genes will significantly contribute to further insights into the molecular toolkit that governs key developmental processes of this important lophotrochozoan phylum [45, 46].

With the advent of next-generation sequencing technologies (e.g., [47, 48]), large-scale comparative genomic surveys of non-model species are now possible, allowing for deeper insights into ancestral versus novel features of the molecular machinery that underlies the ontogenetic establishment of animal body plans. Recently, four important molecular resources were established by sequencing and annotating complete genomes for mollusks using the bivalves *Crassostrea gigas* [49] and *Pinctada fucata* [50], the gastropod *Lottia gigantea* [16], and the cephalopod *Octopus bimaculoides* [51] as model organisms. Apart from useful insights into genome organisation and the structure of individual genes in these species, the studies identified the complete Hox and ParaHox complements, adding valuable knowledge about the diversity of these homeotic genes in mollusks.

To expand this database, we sequenced transcriptomes sampled from distinct developmental stages and provide in-depth analyses of the Hox and ParaHox gene families in representative species of seven out of the eight recent class-level taxa of mollusks. Furthermore, we screened our sequences for orthologs present in the Wnt, Notch, and Hedgehog signaling pathways. These highly conserved pathways contribute to orchestrating the broad display of morphology diversity found in bilaterians through

epigenetic interactions between cells and the entrainment of certain developmental programs (for review, see [52]). In addition, we provide a broad functional characterisation of the molluscan gene content using Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways.

Results

Pre-processing and *de novo* assembly of the transcriptomic libraries

The filtering pipeline discarded between 4.78 % (the bivalve *Nucula tumidula*) and 17.40 % (the neomeniomorph = solenogaster *Wirenia argentea*) of low-quality, adaptor contaminated, paired-end reads from the molluscan libraries (Table 1). The assembling process generated high-quality transcriptomes ranging from 34,794 (the gastropod *Lottia* cf. *kogamogai*) to 394,251 (*W. argentea*) sequences (Table 2). The difference in the number of reconstructed base pairs, transcripts, the values of the largest transcript, and the N50 (median transcript length) are obvious between 454 and Illumina libraries. The best 454 library (the cephalopod *Idiosepius notoides*) includes considerably less transcripts, base pairs, and N50 transcript length than any of the short-read Illumina libraries. To facilitate the downstream analysis, both assembled libraries derived from the cephalopod *I. notoides* were combined.

Identification of the coding sequence regions and clustering of the transcriptomes

This procedure generated high-quality redundant protein gene sets that contained between 17,163 (*Lottia* cf. *kogamogai*) and 216,221 (the polyplacophoran *Acanthochitona crinita*) sequences. The percentage of transcripts in each of the molluscan libraries that codes for a putative protein sequence ranges from 21 % (*Idiosepius notoides*) to 59 % (*A. crinita*) (Table 3). After the clustering and the elimination of protein sequence redundancy, the number of sequences lowered by more than 70 % in some protein gene sets (approx. 74 % in *Wirenia argentea*, approx. 72 % in *A. crinita*, and approx. 71 % in the scaphopod *Antalis entalis*). The 454 protein gene set derived from *L.* cf. *kogamogai* showed the lowest reduction in the number of protein sequences, in which just more than approx. 2 % of the sequences were clustered.

Assessment of the protein gene set completeness using BUSCO

The completeness in the molluscan protein gene sets, as approximated by the presence of universal single copy orthologs [53], showed an ample variability ranging from 68.21 % (*Lottia* cf. *kogamogai*) to 95.02 % (*Nucula tumidula*) (Table 4). A correlation of completeness and the sequencing technique is noticeable among the different molluscan protein gene sets. For instance, the most incomplete protein gene set using deep Illumina sequencing (the chaetodermomorph = caudofoveate *Scutopus ventrolineatus*: 79.83 % of completeness) is more complete than the one generated by the 454 pyrosequencing (*L.* cf. *kogamogai*: 68.21 % of completeness). Likewise, the number of fragmented BUSCOs in the *S. ventrolineatus* library is still lower than the number of fragmented BUSCOs in the *L.* cf. *kogamogai* 454 sequenced library. The statistics of pre-

Table 1 Summary of the pre-processing pipeline in the molluscan transcriptomic libraries

Organism	No. of reads[a] before pre-processing	No. of reads[a] after pre-processing	No. of reads[a] excluded
Gymnomenia pellucida (Neomeniomorpha)	53,751,440	50,292,634 (93.57 %)	3,458,806 (6.43 %)
Wirenia argentea (Neomeniomorpha)	50,456,889	41,678,466 (82.60 %)	8,778,423 (17.4 %)
Scutopus ventrolineatus (Chaetodermomorpha)	43,492,046	40,596,155 (93.34 %)	2,895,891 (6.66 %)
Acanthochitona crinita (Polyplacophora)	35,737,364	33,695,610 (94.29 %)	2,041,754 (5.71 %)
Idiosepius notoides[b] (Cephalopoda)	588,878	588,878 (100 %)	-
Idiosepius notoides (Cephalopoda)	38,267,214	35,131,600 (91.81 %)	3,135,614 (8.19 %)
Lottia cf. *kogamogai*[b] (Gastropoda)	402,814	402,814 (100 %)	-
Nucula tumidula (Bivalvia)	40,797,848	38,849,372 (95.22 %)	1,948,476 (4.78 %)
Antalis entalis (Scaphopoda)	24,194,021	22,881,795 (94.58 %)	1,312,226 (5.42 %)

[a]Read pairs for Illumina libraries
[b]Note that the 454 datasets were just trimmed and converted to fasta and fasta.qual files. The quality and length filtering was executed by the program MIRA4 during the assembling step

Table 2 Summary of assembly statistics from the nine molluscan transcriptomic libraries

Assembler	Organism	No. of transcripts	No. of transcripts > 1,000 bp	No. of reconstructed bases (bp)	No. of reconstructed bases in transcripts >1,000 bp	Length of the largest transcript reconstructed (bp)	N50
IDBA-tran	*Gymnomenia pellucida* (Neomeniomorpha)	228,678	136,889	408,484,174	355,797,467	26,833	2,616
	Wirenia argentea (Neomeniomorpha)	394,251	178,721	495,209,150	369,131,155	13,881	1,725
	Scutopus ventrolineatus (Chaetodermomorpha)	220,258	96,068	253,037,497	181,977,467	17,067	1,555
	Acanthochitona crinita (Polyplacophora)	364,800	234,607	689,247,497	614,059,419	17,023	2,737
	Idiosepius notoides (Cephalopoda)	285,863	93,114	297,178,066	189,330,826	19,705	1,399
	Nucula tumidula (Bivalvia)	273,272	126,403	378,309,195	296,427,272	20,605	2,100
	Antalis entalis (Scaphopoda)	351,943	125,869	369,111,329	241,658,022	28,825	1,399
MIRA4	*Idiosepius notoides* (Cephalopoda)	43,218	6,880	29,267,478	10,095,956	10,063	785
	Lottia cf. *kogamogai* (Gastropoda)	34,794	6,391	25,737,707	9,530,625	7,134	817

processing, assembly, and quality assessment pipelines are summarised in Table 5.

Identification of Hox and ParaHox sequences and phylogenetic analyses

A total of 64 Hox and eight ParaHox genes were identified and their orthology confirmed through Bayesian phylogenetic analysis (Fig. 1). Monophyly of paralog groups *Hox1, Hox2, Lox4, Post1, Post2* and the ParaHox groups *Gsx, Xlox*, and *Cdx* is well-supported (posterior probability > 0.9). Identity of other paralog groups was established by annotating them using information from well-characterised model metazoan and molluscan sequences they cluster with. Supposedly complete (11 genes) or almost complete (nine or more genes) sets of Hox genes were obtained from the polyplacophoran *Acanthochitona crinita*, the neomeniomorphs *Gymnomenia pellucida* and *Wirenia argentea*, as well as the scaphopod *Antalis entalis*. The putatively most incomplete set of Hox genes (three genes) was retrieved from the chaetodermomorph (caudofoveate) *Scutopus ventrolineatus* (Fig. 2).

The common paralog peptide signatures in the homeobox domain and in its flanking regions greatly differ between the different Hox and ParaHox paralog groups

Table 3 Summary of empirical homology-based prediction and clustering methodology in the molluscan transcriptomic libraries

Organism	No. of transcripts	No. of possible putative proteins	No. of selected putative proteins	No. of non-redundant putative proteins
Gymnomenia pellucida (Neomeniomorpha)	228,678	834,304	125,766	54,997
Wirenia argentea (Neomeniomorpha)	394,251	1,185,594	213,616	54,183
Scutopus ventrolineatus (Chaetodermomorpha)	220,258	499,165	87,291	39,631
Acanthochitona crinita (Polyplacophora)	364,800	1,663,283	216,221	59,271
Idiosepius notoides (Cephalopoda)	329,081	543,405	70,861	21,533
Lottia cf. *kogamogai* (Gastropoda)	34,794	47,120	17,163	16,781
Nucula tumidula (Bivalvia)	273,272	787,355	105,381	38,563
Antalis entalis (Scaphopoda)	351,943	739,709	124,738	35,443

Table 4 BUSCO summary of the molluscan protein gene sets

Organism	Complete Single-copy BUSCOs	Fragmented BUSCOs	Missing BUSCOs	Completeness (%)
Gymnomenia pellucida (Neomeniomorpha)	708	87	48	94.31
Wirenia argentea (Neomeniomorpha)	450	234	159	81.14
Scutopus ventrolineatus (Chaetodermomorpha)	506	167	170	79.83
Acanthochitona crinita (Polyplacophora)	660	136	47	94.42
Idiosepius notoides (Cephalopoda)	680	102	61	92.76
Lottia cf. *kogamogai* (Gastropoda)	286	289	268	68.21
Nucula tumidula (Bivalvia)	705	96	42	95.02
Antalis entalis (Scaphopoda)	697	100	46	94.54

(Fig. 3). The paralog group 1 (HPG-1) contains one conserved motif (positions 6-8) and two unique single amino acid signatures (positions 29 and 56) in the homeobox domain. Additionally, two non-basic amino acids in the N-terminal region inside of the homeobox at positions 2 and 3 (see [54]) and one conserved motif downstream of the homeobox (positions +1 and +2) in the C-terminal region provide unambiguous signatures for the paralog group 1.

The paralog groups 2 (HPG-2) and 3 (HPG-3) have a unique DNA-contacting residue that lies between two conserved basic amino acids at position 4 within the N-terminal region in the homeobox [54]. Furthermore, the paralog group 2 contains three unique single amino acid signatures at position 2, 24, and 58-59, whereas paralog group 3 contains one conserved bilaterian residue at position 14 and one specific lophotrochozoan "AL" motif in the positions 36-37.

The paralog groups 4 (HPG-4) and 5 (HPG-5) do not show any specific motifs or unique residues within the homeodomain. The unique signature of these two paralog groups is the motif "YPWM" located in the upstream N-terminal region outside the homeodomain. Moreover, the paralog group 4 contains a "LPNTK" diagnostic motif in the downstream C-terminal region of the homeodomain

Table 5 Summary of initial pre-processing and generation of the high-quality molluscan protein gene sets

Organism	No. of raw reads[a]	No. of reconstructed transcripts	No. of n-r[b] putative proteins	Gene set completeness (%)
Gymnomenia pellucida (Neomeniomorpha)	53,751,440	228,678	54,997	94.31
Wirenia argentea (Neomeniomorpha)	50,456,889	394,251	54,183	81.14
Scutopus ventrolineatus (Chaetodermomorpha)	43,492,046	220,258	39,631	79.83
Acanthochitona crinita (Polyplacophora)	35,737,364	364,800	59,271	94.42
Idiosepius notoides (Cephalopoda)	38,267,214	285,863	21,533[c]	92.76[c]
Idiosepius notoides (454) (Cephalopoda)	588,878	43,218	–	–
Lottia cf. *kogamogai* (Gastropoda)	402,814	34,794	16,781	68.21
Nucula tumidula (Bivalvia)	40,797,848	273,272	38,563	95.02
Antalis entalis (Scaphopoda)	24,194,021	351,943	35,443	94.54

[a]Read pairs for Illumina libraries
[b]Non-redundant
[c]After the assembly step the *Idiosepius notoides* libraries were combined together for the subsequent downstream analysis

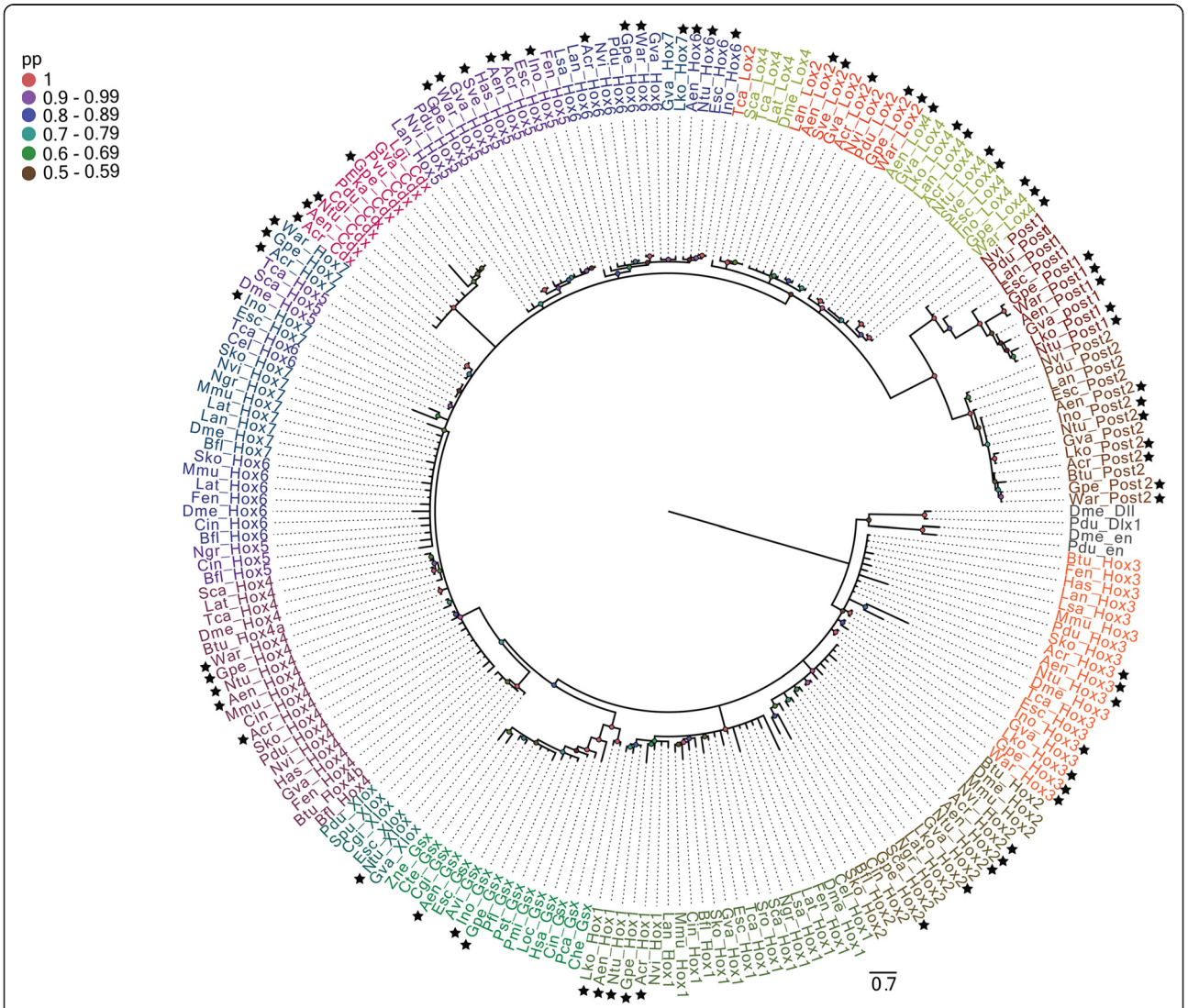


Fig. 1 Phylogeny of Hox and ParaHox genes from amino acid sequences containing homeodomain and flanking regions. The consensus tree was inferred by Bayesian phylogenetic analysis with MrBayes v3.2.2 discarding 25 % of the sampled trees as burn-in. The branch support values are posterior probability values. The new Hox and ParaHox sequences identified in this study are highlighted by black stars. Hox and ParaHox paralog groups are represented by different colors. The homeobox genes *distalless* and *engrailed* were used as outgroups

(see [55]). A unique specific molluscan motif with 6 residues ("HIAKNM") was discovered in the paralog group 5, located immediately after the last residue in the C-terminal region of the homeodomain. The paralog group 7 (HPG-7), albeit not possessing a unique signature within the homeodomain, is the only paralog group that contains the region that regulates the Hox-PBC interaction located close to the N-terminal region right before the start of the homeodomain (see [56]).

The Hox sequences belonging to the central class *Lox5* (HPG-6), *Lox2, Lox4* (HPG-8), and the posterior paralog groups *Post-2* and *Post-1* (HPG-9) were characterised by the presence of specific lophotrochozoan signature motifs and amino acid residues in the homeodomain and its surroundings. For example, the presence of a strongly conserved C-terminal parapeptide motif in the paralog genes *Lox5* (Lox5-parapetide), *Lox2,* and *Lox4* (Ubd-A-parapeptide), and the distinctive homeodomain residues in the paralog genes *Post-1* and *Post-2* (see [57]).

Alignment of the ParaHox genes *Xlox*, *Gbx*, and *Cdx* provides an overview of the conserved homeodomain peptides and the specific signature motifs located in the N- and C-terminal regions of these genes. Among these signature motifs, a specific lophotrochozoan pentapeptide motif ("LRTCD") in the C-terminal arm of the *Gsx* gene is present. Apart from the gastropod *Lottia* cf. *kogamogai*, the neomeniomorph *Wirenia argentea,* and the chaetodermomorph *Scutopus ventrolineatus*, at least

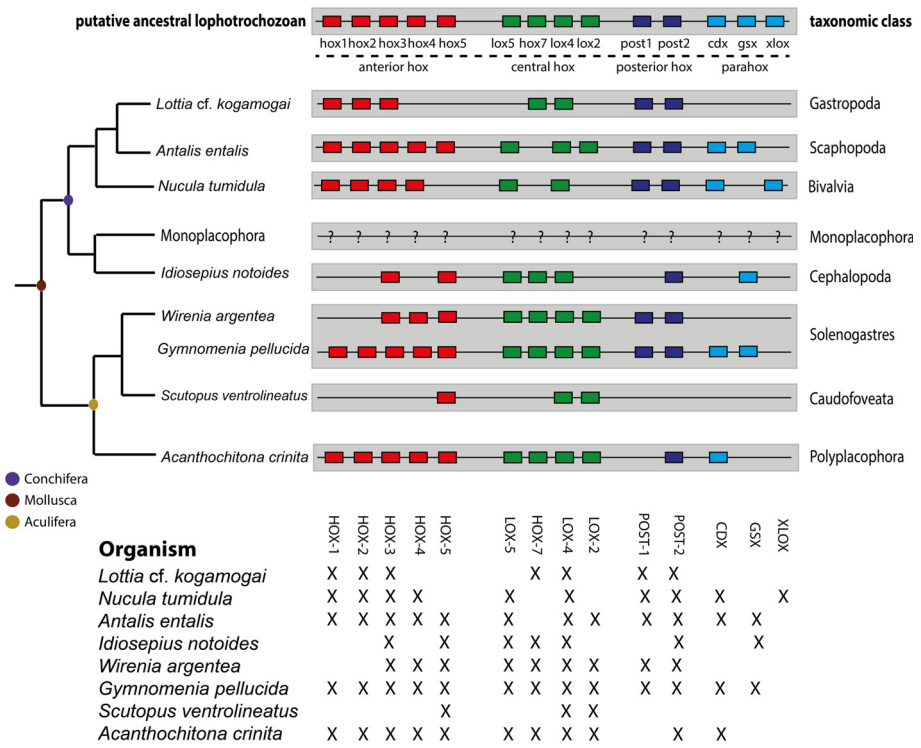


Fig. 2 Summary of the Hox and ParaHox genes identified in the eight molluscan species studied herein. For comparison, the putative ancestral lophotrochozoan gene toolkit is provided. Tree topology follows Smith et al. [61]. The colored boxes indicate the anterior, central, and posterior Hox as well as the ParaHox paralog groups. To date, no known Hox or ParaHox sequences (represented by question marks) belonging to a monoplacophoran mollusk have been identified. The colored circles in the branch nodes represent the last common ancestor of the monophyletic clades Conchifera, Aculifera, and Mollusca, respectively. It is important to note that the figure does not depict the chromosomal organisation of the Hox and ParaHox genes in the studied species. Thus, it is well possible (and in case of the cephalopod *Idiosepius* even likely) that the Hox complement is not organised in a distinct cluster in (some of) the species depicted herein

one ParaHox gene was identified in each of the species investigated.

Gene Ontology and KEGG annotation

The functional characterisation using the KEGG and Gene Ontology (GO) Slim terms revealed a similar relative percentage of genes distributed in the different functional categories among the molluscan protein gene sets, with a few exceptions (Table 6, Figs. 4 and 5). The percentage of classified proteins belonging to the different molluscan gene sets ranges from 29.59 % to 46.04 % in the KEGG analysis and from 52.0 % to 60.11 % in the GO. Despite the disparity in the number of protein sequences in the molluscan protein gene sets, the number of pathway maps, in which all KEGG Orthology (KO) groups were mapped, is very similar among the species (between 332 and 342).

As expected for transcriptomes sampled from different early developmental stages, the transcriptomes are enriched with proteins that bind and interact with DNA (and thus have, e.g., a putative role in the control of gene expression, chromatin regulation, etc.) and/or RNA (e.g., have a function in RNA processing and modification such as alternative splicing, editing, and polyadenylation). Biological processes involving transmembrane transport as well as carbohydrate and lipid metabolism are overrepresented in relation to other categories in both KEGG and GO analyses. The functional category "signal transduction" is overpopulated with a high relative percentage of proteins in both analyses (between 3-5 % in KEGG and 1–2 % in GO). A deeper look into the fine-grained functional categories inside "signal transduction" in KEGG shows that Notch, Hedgehog, and Wnt are common signaling pathways shared in all gene sets with a high percentage of genes.

Regarding the Wnt gene family, at least one Wnt gene was found in each of the transcriptomes according to KEGG orthology assignments. The transcriptomes of the aculiferans *Acanthochitona crinita* and *Gymnomenia pellucida* are the most Wnt-rich transcriptomes with nine and eight Wnt representatives, respectively, whereas the gastropod *Lottia* cf. *kogamogai* and the chaetodermomorph *Scutopus ventrolineatus* harbor only the *Wnt5* gene. Additionally, most of the cardinal signaling components of the Notch and Hedgehog pathways were identified and characterised in all transcriptomes,

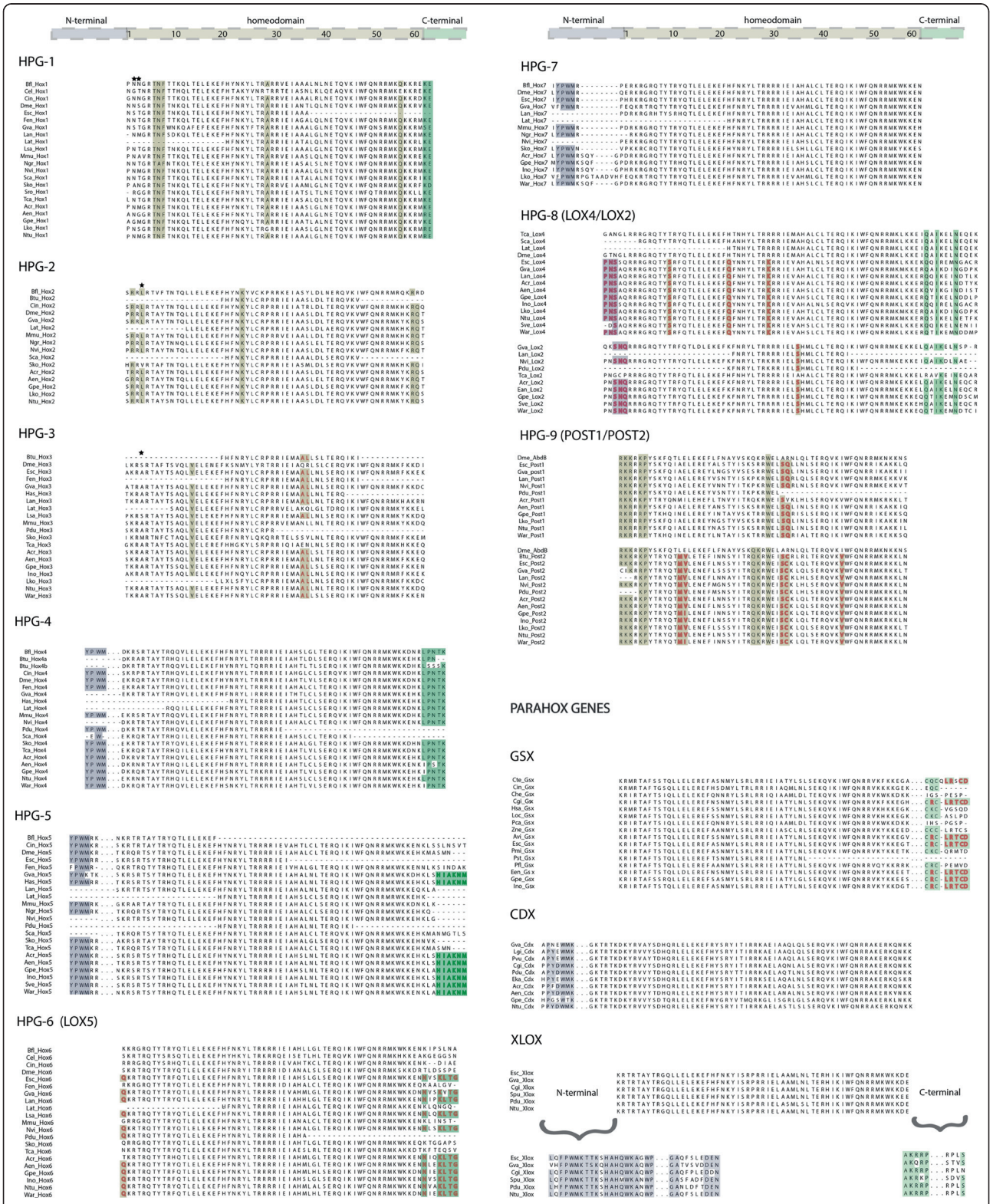


Fig. 3 Multiple sequence alignments of Hox and ParaHox sequences highlighting the conserved homeodomain and flanking regions. Bilaterian diagnostic peptides in the homeodomain and in the flanking regions are highlighted by colored boxes. Conserved lophotrochozoan and molluscan residues are highlighted by dark red and green colored letters, respectively. Black stars indicate DNA-contacting residues

Table 6 Functional annotation of the molluscan transcriptomes using KEGG analysis and Gene Ontology terms

Functional annotation	*Gymnomenia pellucida* (Neomeniomorpha)	*Wirenia argentea* (Neomeniomorpha)	*Scutopus ventrolineatus* (Chaetodermomorpha)	*Acanchothitona crinita* (Polyplacophora)	*Idiosepius notoides* (Cephalopoda)	*Lottia* cf. *kogamogai* (Gastropoda)	*Nucula tumidula* (Bivalvia)	*Antalis entalis* (Scaphopoda)
KEGG (total)	20,861	20,662	16,935	25,460	16,672	9,409	18,524	20,842
Pathways	341	338	340	342	336	332	337	338
Metabolism	3,720	4,324	4,371	5,806	2,951	1,910	3,984	4,958
Genetic Information	2,409	2,180	1,779	2,477	1,834	1,143	2,239	2,146
Enviromental Information	2,087	1,920	1,232	2,443	1,697	809	1,993	1,772
Cellular Processes	1,546	1,636	1,128	1,961	1,128	755	1,433	1,556
Organismal System	2,991	2,655	2,249	3,315	2,609	1,281	2,806	2,557
Human Diseases	3,915	3,619	2,634	3,803	3,127	1,828	4,386	3,367
Not classified	4,193	4,328	3,542	5,655	3,326	1,683	1,683	4,486
GO (total)	23,080	31,149	18,543	30,541	17,469	8,776	23,906	23,779
Biological Process	3,346	4,776	3,196	4,947	2,434	1,352	3,765	3,522
Molecular Function	12,989	16,957	9,683	15,688	9,186	4,662	12,999	12,630
Cellular component	6,745	9,416	5,664	9,906	5,846	2,762	7,142	7,627

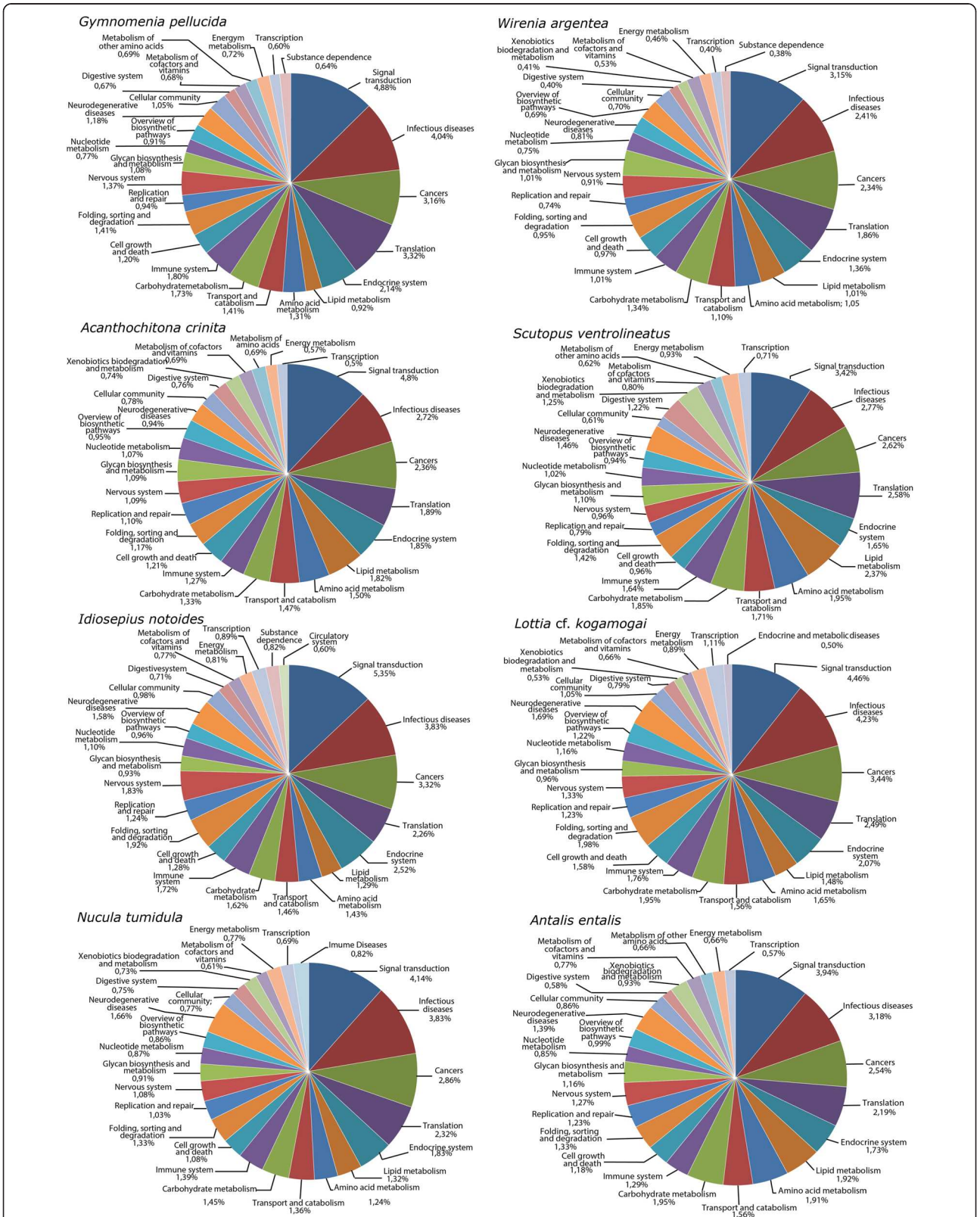


Fig. 4 Distribution of the 25 most represented KEGG functional categories in the eight molluscan transcriptomes. The numbers represent the relative percentage of mapped proteins in each category in regard to the total number of transcripts in the respective species

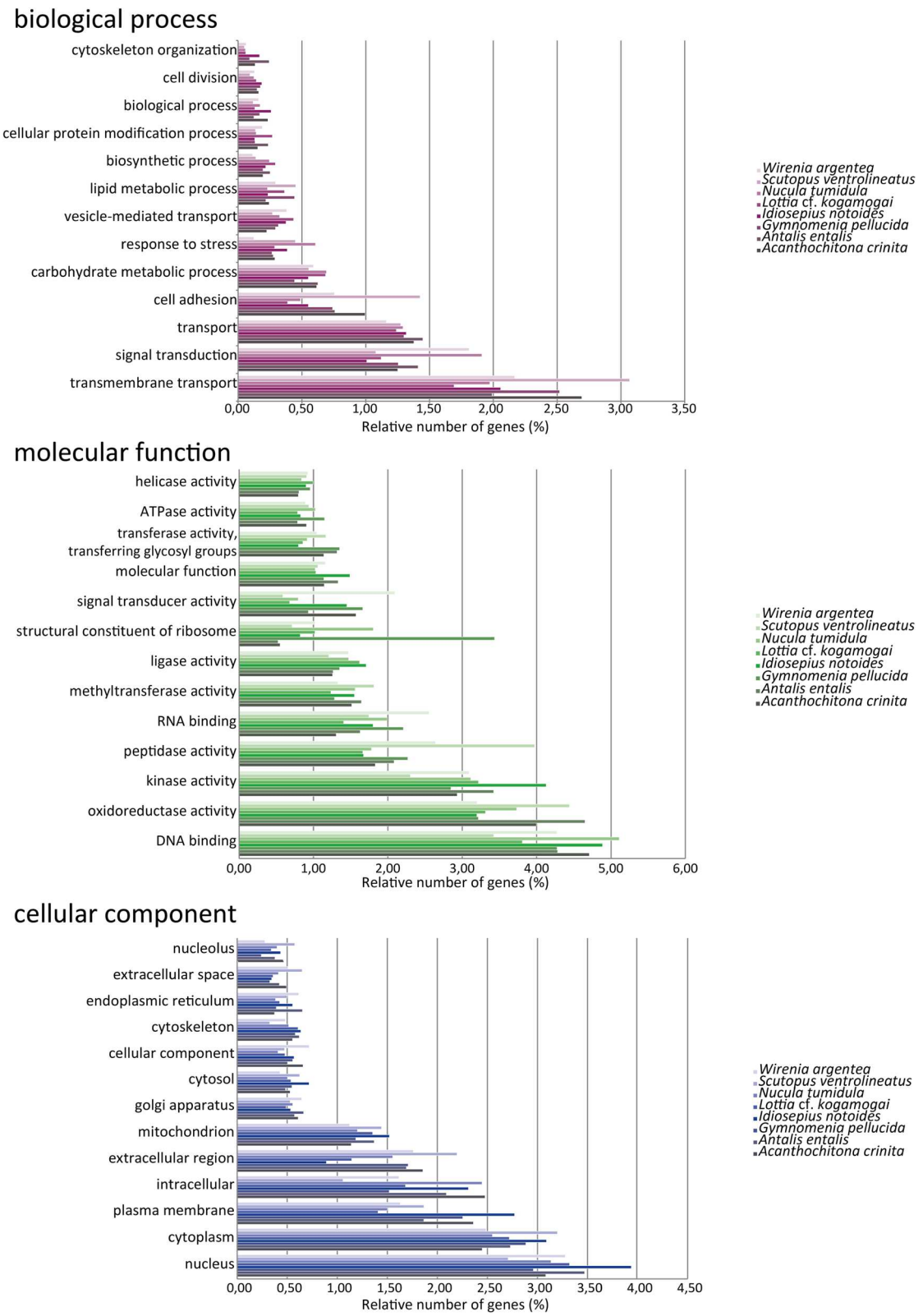


Fig. 5 Comparative functional classification using Gene Ontology-Slim terms. Only the 13 most expressed terms in each ontological domain are shown. The relative percentages represent the numbers of mapped GO terms in each category in reads to the total number of transcripts in the respective species

including the *Notch* and *Hh* orthologs (with the exception of *S. ventrolineatus*) (Fig. 6). Phylogenetic analysis with *Hh* genes confirmed the orthology of these molluscan genes and supports the monophyly of the three major clades of bilaterian animals (Deuterostomia, Lophotrochozoa, and Ecdysozoa) (Fig. 7a).

A detailed view at the sequence composition of *Hh* and *Notch* orthologs was performed, highlighting the organisation of the respective protein domains (i.e., N-terminal Hh and C-terminal Hint domain in *Hh* orthologs and EGF-like repeats, LNR, and ANK domains in *Notch* orthologs) and the characteristic conserved residues among these genes. The *Notch* sequences identified herein showed the absence of certain diagnostic motifs in all sequences, with the exception of the protobranch bivalve *Nucula tumidula*, implying that these genes, despite being classified as a *bona fide Notch*, are not represented in their totality (partial coding sequence region). Regarding the *Hh* genes, apart from *Lottia* cf. *kogamogai* in which the Hint domain is missing, all the *Hh* sequences harbor the full length diagnostic Hedgehog domains. Despite the partial *Hh* sequence of the limpet *Lottia* cf. *kogamogai*, the number of *Hh*-related genes retrieved from this transcriptome was the highest among all transcriptomes with 11 representatives. Phylogenetic analysis using the 11 *Hh*-related genes of the limpet, one polyplacophoran *Hh*-related gene obtained from *A. crinita*, and previously published sequences from other lophotrochozoans recovered the "Lophohog" clade, supporting the existence of a lophotrochozoan-specific *Hh*-related gene family ([58]; cf. Fig. 7b).

Overall, the diversity of different GO terms and KEGG functional categories in the protein gene sets show a high resolution picture of the molluscan transcriptomes.

Discussion

Feasibility of non-model mollusks for comparative transcriptomic studies

As next-generation sequencing costs have dramatically decreased during the last years, transcriptome shotgun sequencing has emerged as a powerful tool to investigate RNA dynamics of living organisms qualitatively (e.g., which genes are expressed during a given ontogenetic period) and quantitatively (e.g., expression levels of a specific gene) e.g., [59]. Accordingly, it is now feasible to obtain a full catalog of the transcriptome composition and complexity of organisms on a broader and comparative level, enabling to assess several questions in evolutionary biology with the assistance of genomic data [60–62]. Additionally, such a comparative approach is useful to discover shared and unique evolutionary events from different taxa, allowing plausible evolutionary inferences of specific biological questions. The 1KITE (1,000 Insect Transcriptome Evolution) project is a good example as to how next-generation transcriptome sequencing can form the base not only for phylogenetic analyses, but also for insights into genome and transcriptome evolution of species-rich animal clades [63].

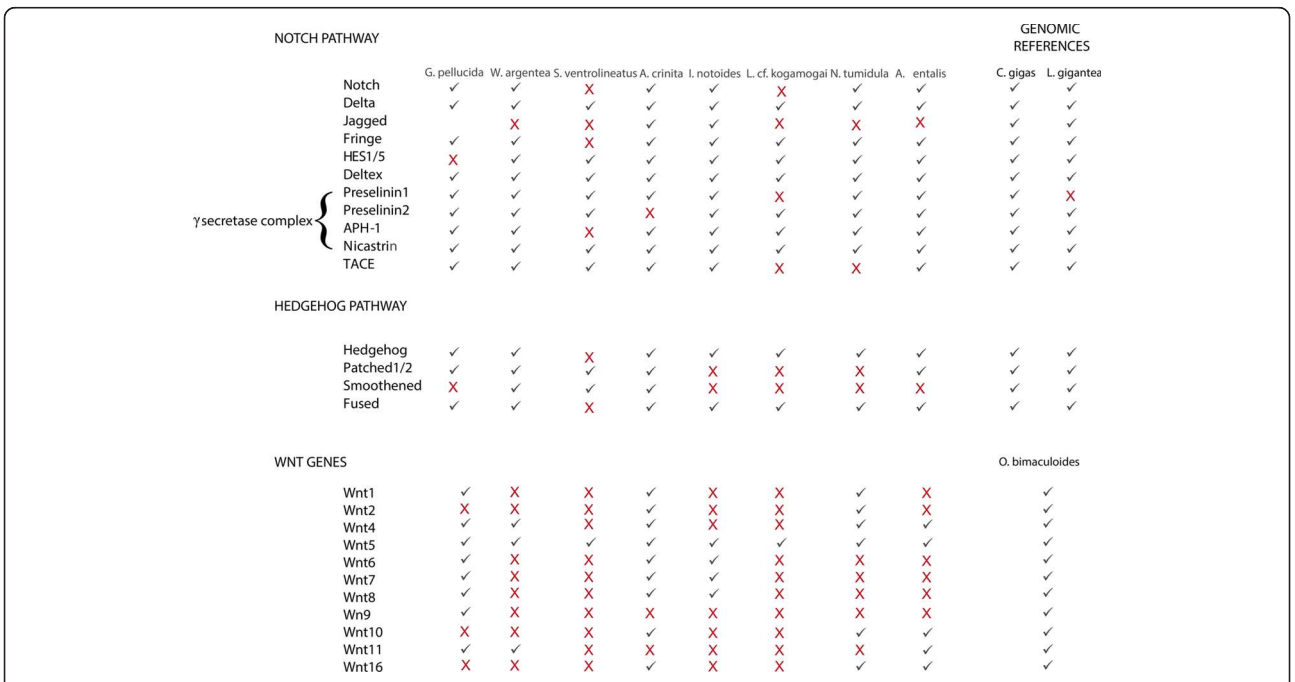


Fig. 6 Identification of cardinal gene components in Wnt, Hedgehog, and Notch signaling pathways in the transcriptomes. *Crassostrea gigas*, *Lottia gigantea*, and *Octopus bimaculoides* genes identified from genomic sequences were used for comparison. Not identified sequences are marked by the red "X"

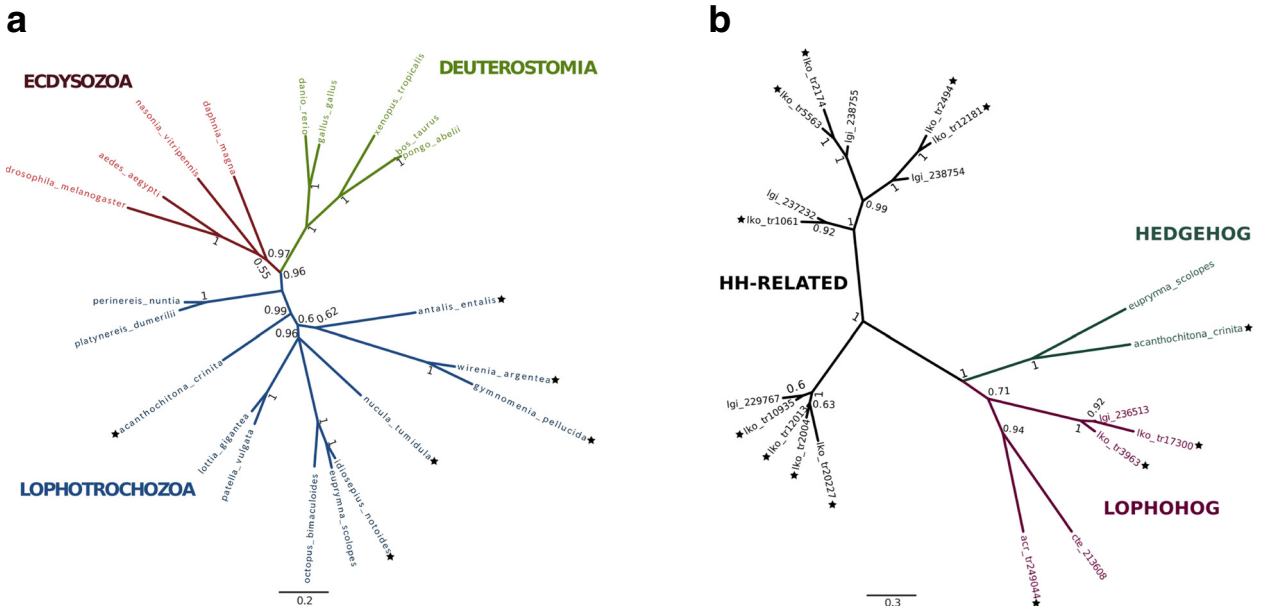


Fig. 7 Phylogenetic reconstruction of *Hh* (**a**) and *Hh*-related (**b**) genes from amino acid sequences. The consensus trees were inferred through Bayesian phylogenetic analysis with MrBayes v3.2.2 discarding 25 % of the sampled trees as burn-in. The branch support values are posterior probability values of Bayesian likelihood. Colored branches in A represent the three major superphyla Ecdysozoa, Deuterostomia, and Lophotrochozoa. The Lophohog *Hh*-related family was first described by Bürglin [58] and originally included two sequences retrieved from the lophotrochozoans *Lottia gigantea* (mollusk) and *Capitella teleta* (annelid). Names followed by black stars correspond to newly described sequences obtained in this work

In our study, nine new transcriptomes belonging to representatives of seven out of the eight recent class-level taxa of mollusks were deeply sequenced using next-generation sequencing (Illumina and 454). To generate reliable and good quality protein gene sets for downstream analyses (e.g., functional and phylogenetic analyses, sequence identification), various protocols for pre-processing, assembly, clustering, and coding sequence region prediction were established.

Despite many limitations in the *de novo* assembly and the scarce resources of molluscan genomic references (including fully annotated genomes), transcriptome sequencing offers a cost-effective method of characterizing the gene set of non-model species. One challenging aspect in every transcriptomic project is the comparison between assemblies using either common statistics (e.g., N50, number of reconstructed bases, and average length of the transcripts) or annotation-based metrics (e.g., number of single copy orthologs). As pointed out by O´ Neil & Emrich [64] and Mundry et al. [65], although many metrics have been used to evaluate and compare these assemblies, it is unclear how precise and accurate these metrics are. Despite these limitations, we assessed and evaluated common statistics in order to compare our assembly results with other recent transcriptome studies on lophotrochozoan organisms (e.g., N50 transcript and number of reconstructed bases). The assembly results obtained herein (excluding the 454 libraries) are at least comparable and in most cases outperform some recent transcriptome studies (cf. [46, 66, 67]). Regarding the completeness and integrity of the transcriptome (i.e. fragmentation of genes), the BUSCO analysis revealed a reasonable completeness in all molluscan libraries, corroborated by the great diversity of gene and gene families identified in the downstream analysis. The high proportion of fragmented genes in the transcriptome of the patellogastropod *Lottia* cf. *kogamogai*, as pointed out by the BUSCO analysis, reflects the high rates of insertions and deletions due to homopolymeric regions during the pyrosequencing process [68], creating frameshifts and disrupting the alignments between these sequences and their respective homologs. Indeed, the first phylogenetic analyses with *Lottia* cf. *kogamogai* Hox genes resulted in atypically long branches showing a great amount of genetic divergence between the patellogastropod sequences and their respective homologs in Mollusca and another bilaterians. Accordingly, even if it remains unclear as to how the aforementioned metrics most accurately reflect the assembly results, comparisons among our data as well as with those of different transcriptome studies clearly demonstrate the high quality of our results.

To date, there are only a few genetic or physical maps publicly available which describe genome organization, extrachromosomal DNA (mitochondrial genomes) [69–71], gene structure, or functional contents for lophotrochozoan animals and especially mollusks. However, three recent

studies based on a robust genome annotation in the patellogastropod limpet *Lottia gigantea* [16], the Pacific oyster *Crassostrea gigas* [49], and the octopus *Octopus bimaculoides* [51] have shown that the expected number of protein coding genes in these mollusks ranges from approx. 24,000 to 34,000. In our study, except for *Lottia* cf. *kogamogai* and *Idiosepius notoides*, all protein gene sets have an inflated number of putative proteins when compared to the patellogastropod, oyster, and octopus data. This elevated number of protein-coding genes does not necessarily represent the real complexity of the transcriptomic machinery in our study species; rather, it might be influenced by biases and limitations brought by the next-generation DNA platforms (i.e. fragmentation of genes, sequencing biases) (see [72]) and/or assembly artifacts. Considering the annotation of the coding sequence regions in the different molluscan libraries, a relatively small proportion of proteins (between 21–59 %; see Table 3) have shown sequence homology against well-curated public databases. This high proportion of non-annotated protein sequences is not unusual in transcriptome projects, and this feature is commonly observed in a vast diversity of taxa (cf. [73, 74]), including mollusks [75–77]. This lack of detectable sequence homology in the public databases may be due to several factors, including taxonomically restricted genes (e.g., orphan genes), novel isoform transcripts or protein-coding genes, non-functional coding sequence regions, and poor quality of the sequences themselves or the assembly procedures performed [78, 79]. Specifically in mollusks, various studies have described the emergence of numerous specific suites of genes and gene families, which are either present in different molluscan lineages or are restricted to a single one [51, 80]. The discovery of an independent large-scale expansion and evolution of the tyrosinase gene family in bivalves [81] is a good example of how comparative genomics and transcriptomics are useful to characterise novel lineage-specific genes and gene families.

Diversity of Hox and ParaHox genes in mollusks

To elucidate the utility of the molluscan transcriptomes for evo-devo studies, an extensive comparative survey was conducted focussing on Hox and ParaHox gene sequences. A total of 64 Hox and eight ParaHox genes were found and fully characterised. Prior to our study, complete (or near-complete) sets of Hox genes had only been identified in three bivalve species (*Pecten maximus, Crassostrea gigas*, and *Pinctada fucata*) [49, 82, 83], two marine gastropods (*Gibbula varia* and *Lottia gigantea*) [16, 38], and in two cephalopods (the squid *Euprymna scolopes* and the octopod *Octopus bimaculoides*) [51, 84]. We here report a complete Hox gene complement for the neomeniomorph *Gymnomenia pellucida*. Additionally, at least near-complete Hox gene complements were identified from the polyplacophoran *Acanthochitona crinita*, (10 genes), the scaphopod *Antalis entalis* (10 genes), and another neomeniomorph, *Wirenia argentea* (nine genes). Notably, only few ParaHox sequences were retrieved from our molluscan transcriptomes, considering that all three ParaHox genes had been found in various molluscan lineages prior to our analysis [16, 42, 49, 83, 85].

The publicly available genomic resources and the data presented here show that the molluscan Hox and ParaHox clusters share a similar composition in terms of gene content despite the great disparity of morphological features within the phylum [26]. This implies that the rich morphological diversity among different class-level taxa of mollusks lies in the regulation and subtle changes of the regulatory networks in the developmental program rather than in the physical organisation and composition of the Hox and ParaHox clusters. By comparing Hox sequences from a vertebrate, fly, and amphioxus, it was proposed earlier that many of the amino acid replacements used as diagnostic criteria for the different paralog groups are likely to be localised on the surface of the respective proteins and have a major functional impact on protein-protein interactions [86]. This fact, associated with the relaxed DNA-binding specificity of the homeodomain, provides the necessary toolbox to promptly originate new regulatory interactions between the Hox genes and their target genes [87], thereby forming an important prerequisite for the evolution of novel morphological features. Within Mollusca, a striking example as to how the possible relaxation of the regulatory constraints and the recruitment of novel regulatory genes are responsible for morphological changes has been reported for the cephalopod *Euprymna scolopes* [36]. Hox gene expression in this bobtail squid deviates from the proposed ancestral role of patterning the antero-posterior body axis; instead, the reported Hox genes are expressed during ontogeny of various taxon-specific morphological innovations such as the brachial crown, funnel, light organ, or the stellate ganglia. In addition to the striking plasticity of the Hox genes and their functional co-option during evolution, the study also proposed the possibility that the non-collinear mode of expression of these genes in cephalopods correlates with the disruption of the Hox cluster in the genome. This notion has recently been confirmed by detailed analyses of the genome of an octopod [51]. Concerning the ParaHox genes, it was shown that the expression of *Gsx* in the gastropod *Gibbula varia* coincides with the area that surrounds the radula anlage, indicating that the function of this homeobox gene was co-opted into the formation of this molluscan autapomorphy [42]. Studies on a scaphopod and the pygmy squid *Idiosepius*, however, revealed a different scenario, whereby *Gsx* is expressed in components of the developing larval and adult nervous system, respectively, but not in the digestive tract or

the developing radula, thus again demonstrating the plasticity of Hox and ParaHox expression domains across Mollusca [43].

It is difficult to determine whether the lack of specific Hox and ParaHox genes in the species of our study is due to gene loss, methodological biases, or low gene expression levels. However, loss of certain genes in both Hox and ParaHox clusters has been described as a recurrent event in the evolutionary history of metazoans [16, 88, 89] including mollusks [45, 49, 90]. Tunicates are a prime example as to how massive gene losses and disrupture of the cluster-like chromosomal organisation can occur in the Hox gene complement [91]. As such, disintegration of the Hox cluster and the loss of central class Hox genes have been reported for the tunicates *Oikopleura dioica* [92] and *Ciona intestinalis* [93]. Losses involving the anterior, central, and posterior Hox as well as the ParaHox genes have been shown by whole genome sequencing studies in cephalopods and bivalves [49, 51]. In addition, various molecular studies have failed to amplify and retrieve particular Hox and ParaHox gene fragments from a wide range of molluscan lineages [45, 84, 90, 94]. Taking into consideration these gene losses, the high degree of completeness of the scaphopod and polyplacophoran transcriptomes obtained from our BUSCO searches (94.54 % and 94.42 %, respectively) and the deep transcriptome sequencing, it seems reasonable to assume that both the polyplacophoran (*Acanthochitona crinita*) and the scaphopod (*Antalis entalis*) Hox set are made up of 10 genes and are represented in their totality in our analysis. Regardless of the Hox and ParaHox set completeness, it is important to notice that the Hox and ParaHox sequences identified in this study contain the full length protein-coding sequence and are long and informative enough for a great deal of molecular (e.g., *in situ* hybridisation) and bioinformatics applications (e.g., phylogenetic analysis).

The identification and characterisation of signature residues (i.e., residues that are shared at certain positions by orthologous proteins but not likely to be present in paralogous proteins) inside the homeodomain and in the surroundings of N-terminal and C-terminal regions provides a better understanding of the evolutionary history of Hox genes [7] and metazoan phylogeny. Herein, a hexapeptide molluscan motif in the paralog group 5 is described for the first time, together with a lophotrochozoan five residue motif in the C-terminal arm of the ParaHox gene *Gsx*. The molluscan-specific motif represents an important marker in distinguishing, from the same paralog, closely related species. To date, this is the first molluscan-specific motif related to a Hox paralog group. These findings show the suitability of the molluscan transcriptomes for the identification of target developmental genes and the specific fine-grained characterisation of these sequences in a phylogenetic context.

Recently, two phylogenomic studies have shed light on the evolutionary interrelationships between seven [60] or the entire eight recent class-level taxa of Mollusca ([61]; see also [95] for some corrections of their 2011 analysis). Remarkably, both analyses strongly support the Aculifera-Conchifera hypothesis, i.e., a basal split of Mollusca into a clade comprising all mollusks that derive from an ancestor with a single shell (Conchifera) and a taxon uniting both aplacophoran clades (Neomeniomorpha and Chaetodermomorpha) with the Polyplacophora as Aculifera. In the light of these results, the characterisation of the Hox and ParaHox gene sets described herein, which includes four aculiferan species, provides an important prerequisite for gene expression studies, and thus research into assessing the putative functional plasticity of these genes across Mollusca. As a matter of fact, expression patterns of ten Hox (all representatives except *Post-1*) and one ParaHox gene (*Cdx*), based on the transcriptome of the polyplacophoran *Acanthochitona crinita* analysed herein, have recently become available from our group [41, 44]. These studies show that the Hox genes in polyplacophorans are expressed in a conserved anterior-posterior pattern along the primary (i.e., longitudinal) body axis. Thereby, their expression was found to be staggered and not restricted to trochozoan- or molluscan-specific features such as the prototroch, the apical organ, or the anlagen of the shell (plates). Instead, the Hox genes are expressed in contiguous domains originating from different germ layers. This is in stark contrast to cephalopod and gastropod mollusks, where they are expressed in a non-staggered fashion in the foot, apical organ [35, 37, 38, 96] or in taxon-specific features of the squid *Euprymna* [36]. Thus, the polyplacophoran Hox gene expression pattern is more similar to annelids than to their molluscan allies. This has led to the conclusion that the Hox genes were co-opted into the patterning of morphological novelties in at least some conchiferans, a situation that most likely contributed to the evolutionary successes of its representatives (see [26]).

Functional characterisation and diversity of the gene repertoire in mollusks

The ability to correlate individual sequences and their respective molecular function is an important step to elucidate the biological background of large numbers of genes (e.g., a putative role in axis specification, neurogenesis, digestive tract formation, and the like). The categorisation of genes and gene products into well-constructed hierarchical classes and pathways aids in the understanding of both cell and organismal biology [97, 98]. This use of molecular information also aids in understanding genetic regulatory networks that control expression levels of mRNA and proteins. The GO as well as KEGG enrichment analyses showed a common overlap of functional categories, which are compatible with the biological

background where the transcriptomes were sampled. The functional GO terms "DNA binding", "nucleus", and "methyltransferase activity" are terms with a high relative percentage of proteins in all gene sets. This reflects the transcriptome background during the development of the species, composed by the presence of many proteins involved in the basal regulation of the transcription (e.g., general transcription factors), development (e.g., homeobox genes such as Hox and ParaHox genes), and protein methylation (e.g., regulation of the epigenetic levels that affect transcription).

Considerable differences were found between the KEGG categories and GO terms retrieved from the predatory sea snail *Rapana venosa* (larval and post-larval stages) [99] and the transcriptomes presented herein. The number of different metabolic pathways into which the proteins were mapped was also found to be higher in our study (between 332 and 342 pathways) than in that of Song et al. [99] (270 pathways). However, this discrepancy can be explained by the nature of the biological samples used to construct the RNA libraries. While the six *R. venosa* samples consisted of only larval and post-larval stages, our samples covered larval, post-larval, juvenile, and adult stages. Due to this broader sampling, one would expect a higher number of metabolic pathways in our analysis than in that of Song et al. [99]. The low percentage and absence of some developmental genes in the *Scutopus ventrolineatus* and *Lottia* cf. *kogamogai* transcriptomes, as revealed by the functional analysis with KEGG and GO, as well as the Hox and ParaHox survey, is a direct reflection of the use of adult specimens during the construction of the transcriptome library and the shallow depth of the 454 sequencing methodology, respectively.

The Wnt, Hedgehog, and Notch signaling pathways are related to the regulation of cell proliferation, transcription, translation, and the proper embryonic development of bilaterian animals, in which any interruption of their signaling activity has severe consequences on developmental outcomes [100]. Thirteen Wnt subfamilies have been characterised in metazoans, while lophotrochozoan representatives, such as the polychaete annelids *Capitella teleta* and *Platynereis dumerilli,* commonly possess only 12 subfamilies and the basal-branching gastropods *Patella vulgata* and *Lottia gigantea* only nine (*WntA*, *Wnt1*, *Wnt2*, *Wnt4*, *Wnt5*, *Wnt6*, *Wnt7*, *Wnt9,* and *Wnt10*) [101, 102]. We found three additional subfamilies in mollusks using KEGG orthology assignment, namely *Wnt8*, *Wnt11,* and *Wnt16*, suggesting that molluscan gene content in the Wnt subfamilies matches that of their lophotrochozoan relatives. Indeed, in a recent publication of the genome of the cephalopod *Octopus bimaculoides* [51], the presence of 12 Wnt genes was reported, corroborating our results and expanding the Wnt complement to at least 12 genes in Mollusca. The *Wnt3* gene is not present in any molluscan transcriptome analysed so far and is likewise absent in all other lophotrochozoans and ecdysozoans hitherto examined (but not in cnidarians) (see [103, 104]), reinforcing the idea that this gene was lost at the base of Protostomia.

Regarding the Hedgehog and Notch signaling pathways, no study focusing on the characterisation and phylogenetic relationships of these genes in mollusks is currently available. The limited knowledge about these important pathways is restricted to some gene expression studies in a few gastropod and cephalopod representatives [105, 106]. Comparisons with respect to the core components present in these two pathways between the transcriptomes described here and two molluscan reference genomes (the limpet *Lottia gigantea* and the oyster *Crassostrea gigas*) revealed a highly shared molecular framework. These results are not surprising, given that both signal transduction pathways play a fundamental role in animal development (e.g., patterning of body axes) and have been characterised in several metazoan animals, from sponges [107] to chordates including humans [108]. Our analysis of the domain organisation in *Notch* and *Hh* orthologs revealed different architectures and patterns of conservation within mollusks and other major groups of bilaterian animals (ecdysozoans and deuterostomes). The receptor Notch is a multidomain protein made by six different components: 30 to 40 amino acids EGF (epidermal growth factor) repeats containing six conserved cysteines; three LNR (lin-notch-repeat) or Notch domains; one NOD and NODP domain; a RAM 23 domain; a PEST domain; and, finally, several Ankyrin repeats [109]. Comparisons of the EGF domain content between the basally-branching bivalve *Nucula tumidula*, the polyplacophoran *Acanthochitona crinita*, and the gastropod *Lottia gigantea Notch* sequences revealed the presence of 34 to up to 36 repeats in these lophotrochozoan proteins. The presence of the NOD and NOPD domains has also been reported for the bivalve *N. tumidula* and is shared by the gastropod *L. gigantea*. The function of these domains is still obscure and remains to be elucidated, albeit they are present in almost all major metazoan lineages (with the exception of the Porifera) [110].

The *Hh* gene family is present throughout the Metazoa, being secondarily lost in some lineages. For example, the nematode *Caenorhabditis elegans* lacks an *Hh* ortholog, whereas *Drosophila melanogaster*, the sea anemone *Nematostella vectensis*, and mammals have one, two, and three *Hh* genes, respectively [111–114]. Herein, one single *Hh* gene was identified in each of the molluscan transcriptomes (apart from *Scutopus ventrolineatus*) through KEGG orthology assignments. Notably, a distinct *Hh*-related family named "Lophohog" was previously retrieved from the genomes of the annelid *Capitella sp. I* and the gastropod *Lottia gigantea* [58]. In this study, 12 *Hh*-related genes were

first identified and described for the basally branching gastropod *Lottia* cf. *kogamogai* (11 genes) and the polyplacophoran *Acanthochitona crinita* (one gene). No *Hh*-related genes were found in the remaining transcriptomes analysed in this study. Interestingly, three new *Hh*-related sequences (two from the limpet *L.* cf. *kogamogai* and one from the polyplacophoran *A. crinita*) showed a close relationship with Lophohog members, expanding the previously described Lophohog clade to five sequences. Accordingly, it seems that the genomes of the basally branching gastropods *L.* cf. *kogamogai* and *L. gigantea* are enriched with *Hh*-related genes, more than in any other molluscan representative investigated to date. The apparent lack of Lophohog representatives in the other mollusks investigated herein must be treated with care as it may not represent the real genetic background as fixed in the genome of these species due to the nature of the transcriptome sequencing; however, the available genomic and transcriptomic data so far support such a scenario. It is expected that the evolution of *Hh* and *Hh*-related sequences will become clearer as soon as additional molluscan genomes become available.

Conclusions

Mollusks show a striking diversity of body plans and are a key taxon for a better understanding of the underlying mechanisms that guide the evolution of developmental processes in multicellular animals. In this study, high-quality transcriptomes were generated from eight molluscan species, representing seven of the eight recent class-level taxa. Different pipelines were carefully designed and implemented, yielding results that are comparable with those generated from model organisms. Furthermore, an extensive catalog of annotated gene products was generated for application in a broad range of downstream analyses. The study focused on the identification and evolution of important developmental genes (Hox and ParaHox) and molecular pathways, nevertheless the results can be used in a broad range of *in silico* (e.g., phylogenomics and gene profiling) and molecular developmental and functional analyses (e.g., *in situ* localisation of mRNAs, expression and characterisation of cloned genes, gene silencing). The data presented herein increase the knowledge on the molecular toolkit of mollusks, especially of the understudied aplacophoran clades, and provides a valuable molecular resource, in particular for further research with a focus on comparative evolutionary developmental (i.e., evo-devo) studies.

Methods

Collection sites, animal cultures, RNA extraction, and fixation

Adults of the polyplacophoran *Acanthochitona crinita* were collected at the Station Biologique de Roscoff, (Roscoff, France) during the summers of 2013 and 2014. Embryos were cultured and staged as previously described [41, 115]. Several hundred individuals of early cleavage stages, blastulae, gastrulae, trochophore larvae, and metamorphic competent individuals as well as early juveniles were collected. Adults of the solenogasters (= neomeniomorphs) *Wirenia argentea* and *Gymnomenia pellucida*, the basally branching protobranch bivalve *Nucula tumidula*, and the caudofoveate (= chaetodermomorph) *Scutopus ventrolineatus* were collected from sediment that was sampled with a hyperbenthic sled at 180–220 meter depth on muddy seafloor in Hauglandsosen (Bergen, Norway) during the winters of 2012 and 2013. The solenogaster and bivalve embryos were cultured and staged as previously described [28, 115, 116]. Adults of *S. ventrolineatus* were kept at 6.5 °C in UV-treated millipore-filtered seawater (MFSW) at the marine living animal facilities at the Department of Biology, University of Bergen, and total RNA of two adult individuals was extracted. Several hundred individuals of early cleavage stages, blastulae, gastrulae, pericalymma (i.e., test cell) larvae, and metamorphic competent as well as juvenile individuals were collected from the solenogaster and bivalve species. Adults of the scaphopod *Antalis entalis* were collected from approx. 30 m depth by the staff of the research vessel *Neomys* off the coast of Roscoff (France). Embryos were cultured and staged as previously described [43]. A total of several hundred individuals of mixed developmental stages up to the early juveniles were collected. Adults of the pygmy squid *Idiosepius notoides* were dipnetted in the sea grass beds of Moreton Bay, Queensland, Australia. Adult squids were kept in closed aquaria facilities at the School of Biological Sciences of the University of Queensland and the RNA of seven nervous systems of adults was collected. Embryos of *I. notoides* were cultured and staged as previously described [117]. Several individuals (approx. 300) representing all stages from freshly laid zygotes to hatchlings were collected. Adults of the basally branching patellogastropod *Lottia* cf. *kogamogai* were collected from intertidal rocks and stones in the vicinity of the marine biological station Vostok (approx. 150 km north of Vladivostok, Russian Federation). Embryos and adults of *L.* cf *kogamogai* were cultured and staged as previously described [118, 119]. Several hundred *L.* cf. *kogamogai* embryos, larvae, and juveniles of key developmental stages (i.e. trochophore, veliger, metamorphic competent, early juvenile stages) were collected.

For RNA extraction, some individuals were stored in RNAlater (Lifetechnologies, Vienna, Austria) at −20 to −80 °C. The RNA of these specimens as well as freshly collected specimens was extracted with a Qiagen extraction kit (Roermond, Netherlands) and subsequently stored at −80 °C. Representatives of the cryptic monoplacophorans were not accessible to us for this study.

Next-generation sequencing, sequence pre-processing, and filtering

High-quality molluscan transcriptome libraries using next-generation sequencing were generated for each one of the aforementioned class-level taxa (Table 7). The short-read libraries were generated with Illumina HiSeq 2000, chemistry v3.0, 2 x 100 pb paired-end modules and the normalised random-primed cDNA. 454 libraries were generated with GS FLX+ with read length of up to 750 bp. The number of reads (or read pairs in Illumina libraries) generated per pooled transcriptomic library varied between 402,814 (*Lottia* cf. *kogamogai*) and 53,751,440 (*Gymnomenia pellucida*), depending on the sequencing technology used.

To remove low quality reads and avoid substandard results in the downstream analyses, different pre-processing bioinformatics pipelines were developed and empirically tested regarding the sequencing method used to obtain the transcriptomic libraries. The short-read libraries pre-processing (Illumina) was carried out using the multi-threaded command line tool trimmomatic v0.3.2 [120]. Known specific Illumina adapters were removed from the paired-end libraries with the parameter ILLUMINACLIP:adapters/TruSeq3-PE-2.fa:2:30. The filtering by quality and length was executed with the command line SLIDINGWINDOW:4:15 MINLEN:40 for all the transcriptomes except for the *Wirenia argentea* library, in which the parameters SLIDINGWINDOW:4:20 MINLEN:40 were defined. The long read libraries (454) were trimmed and converted from SFF (Standard Flowgram Format) to fasta and fasta.qual with the program sff_extract.py v0.3.0 included in the seq_crumbs package (http://bioinf.comav.upv.es/seq_crumbs/) with the default parameters as well as -min_left_clip = 30 parameter for *Lottia* cf. *kogamogai* and –min_left_clip = 32 for the *Idiosepius notoides* library. The quality of the filtered libraries was assessed with the software fastx_toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) taking into consideration the quality score of the bases, the GC-content, and the read length. The assemblies and all downstream analyses were conducted with high-quality and clean libraries.

Transcriptome assembly and quality assessment

The filtered short-read and long-read transcriptome libraries were reconstructed into contiguous cDNA sequences with IDBA_tran v1.1.1 [121] and MIRA4 [122] software, respectively. Information regarding the mRNA sources is summarised in Table 7. The short-read transcriptome assemblies with IDBA_tran were executed with the parameters –mink 20 –maxk 60 –step 5, except for the *Wirenia argentea* library for which the additional parameter –max_count 3 was used. All long-read transcriptome assembling was executed with the parameter mmhr = 2 and the default settings. The quantitative quality assessment of the reconstructed libraries were carried out using QUAST program v2.3 [123] regarding the number of transcripts, number of total bases reconstructed, N50 value, and GC content. The assembling results of the different *Idiosepius notoides* libraries (454 and Illumina) were combined and used in all posterior downstream analyses.

Table 7 Summary of the sequencing methods, organisms, and mRNA extraction sources

Organism	Class	mRNA source	Sequencing
Gymnomenia pellucida	Neomeniomorpha	1/5 total RNA from developmental stages (i.e. freshly hatched larvae until metamorphosis) – 4/5 mRNA from adults.	Illumina
Wirenia argentea	Neomeniomorpha	1/7 total RNA from developmental stages (i.e. freshly hatched larvae until metamorphosis) – 6/7 mRNA from adults.	Illumina
Scutopus ventrolineatus	Chaetodermomorpha	Total RNA from 2 adult individuals	Illumina
Acanthochitona crinita	Polyplacophora	Early cleavage stages – gastrula – early, midstage, and late trochophore larvae – metamorphic competent and settled (post metamorphic) individuals	Illumina
Idiosepius notoides	Cephalopoda	Central nervous system of 7 adult individuals	Illumina
Idiosepius notoides	Cephalopoda	2/3 total RNA from mixed developmental stages (i.e. stages collected after egg laying until the hatching stage) – 1/3 total RNA from adult central nervous system (brain), arm, and gonads tissue)	454
Lottia cf. *kogamogai*	Gastropoda	2/3 total RNA from mixed developmental stages (i.e. trochophore – veliger – pediveliger – metamorphic competent – first juvenile stages) – 1/3 total RNA from adult foot, and central nervous system (CNS)	454
Nucula tumidula	Bivalvia	Early cleavage stages – gastrula – early, midstage, and late pericalymma larvae – metamorphic competent and settled (post metamorphic) individuals	Illumina
Antalis entalis	Scaphopoda	Early cleavage stages – gastrula – early, midstage, and late trochophore larvae – metamorphic competent and settled (post metamorphic) individuals	Illumina

Identification of the coding sequence regions (CDS)

To predict the most probable coding sequence regions within the transcripts, an empirical homology-based methodology was designed using Novaes et al. [124] as a guide, rather than the use of gene prediction tools. The use of gene prediction tools requires the construction of a high-quality training dataset, an arduous task for understudied animals as those used herein. All the reconstructed sequences were translated into protein sequences (located between a start and a stop codon) greater than 50 amino acids in length with the program getorf from the EMBOSS package (http://emboss.sourceforge.net/). The libraries were then submitted to similarity searches with a defined e-value of 1e-06 against three well-curated reference libraries (Uniref90, Pfam and CDD) using the blastp [125], hmmsearch [126], and rps-blast [125] tools, respectively. An in-house Perl script was written in order to select the unique CDS in each transcript with the highest number of evidences (positive hits against the reference library). All the posterior downstream analyses were conducted with the protein gene set libraries created with the aforementioned procedure.

Generation of molluscan non-redundant gene sets

To decrease the computation resources required for the downstream analyses and prevent inflation of the results, the redundancy of the molluscan protein gene sets was reduced using the program UCLUST [127]. The protein sequences with 100 % identity were clustered together, in which the identity is a measure of the number of matches (identities) between two sequences divided by the number of alignment columns.

Assessment of completeness of protein gene sets

In addition to the statistical assessment of the assembled transcriptomes (e.g., N50 values, number of reconstructed base pairs), an analysis to assess the protein gene set completeness in terms of gene content was performed, in order to provide a better understanding and interpretation of the results obtained in the downstream analyses. The assessment of gene content and completeness of the protein gene sets was performed with the program BUSCO using the pre-defined metazoan Benchmarking set of Universal Single-Copy Orthologs with 843 evolutionary conserved orthologous groups [53]. The gene sets were classified into BUSCO metrics as follows: C: complete, D: duplicated, F: fragmented, M: missing.

Hox and Parahox sequence identification and phylogenetic analysis

The protein libraries from all transcriptomes were used in local similarity searches using the program blastp [125] against known and well-curated molluscan Hox and ParaHox sequences retrieved from GenBank non-redundant protein database. The top 3 blast hits of each similarity search were analysed and re-blasted against the entire GenBank non-redundant protein database to reconfirm the homology. Additionally, each putative Hox and ParaHox gene was independently aligned together with their representative homologs from several different metazoan phyla also retrieved from GenBank non-redundant database using the program mafft [128] with the parameters –max_iterate 1000 –localpair. The multiple sequence alignment containing the Hox and ParaHox sequences were searched for the presence of the diagnostic residues/motifs in the homeodomain as well as in the flanking regions. Frameshift errors in *Lottia* cf. *kogamogai Hox1*/*Hox2*/*Post-1*/*Post-2* sequences were corrected using the HMM-FRAME program [129]. All the sequences were then manually edited with the program aliview [130]. The phylogenetic analysis was carried out using MrBayes v3.2.6 [131] with Jones-Taylor-Thornton model of amino-acid substitution [132] as determined using Akaike information criterion (AIC) as implemented in prottest3 [133], 6 rates categories for the gamma distribution, and 30,000,000 generations. After the removal of the initial 25 % of the sampled trees as burn-in, the quality of the run was assessed using Tracer (http://beast.bio.ed.ac.uk/Tracer), regarding the convergence of the likelihood values. The final phylogenetic tree was created and edited with Figtree (http://tree.bio.ed.ac.uk/software/figtree/). The list of species and gene names, phyla, and GenBank accession numbers used in the phylogeny are available in Additional file 1: Table S1.

Identification of *Hh* and *Hh*-related genes and phylogenetic analysis

The *Hh* and *Hh*-related genes were retrieved from the molluscan transcriptomes based on the KEGG orthology assignments. All putative sequences were blasted against known protein databases (PFAM, CDD, and the non-redundant protein database from NCBI), in order to reconfirm the initial orthology assignments. The *Hh* and *Hh*-related sequences were aligned, edited, the phylogeny inferred, and the final tree generated as described above. Frameshift errors in *Lottia* cf. *kogamogai* lko_tr2004/lko_tr12013/lko_20227 were corrected using HMM-Frame program [129]. The substitution model, the number of generations, and sample frequency defined in MrBayes were WAG + G model of amino acid substitution [134], 30,000,000, and 1,000 respectively. The list of species and gene names, phyla, and GenBank accession numbers of the sequences used in the phylogeny are available in Additional file 2: Table S2 and Additional file 3: Table S3.

GO-Slim annotation and pathway mapping with KEGG

The Gene Ontology analyses (GO) were performed in two steps. First, all protein databases that originated from the

high-quality assembled transcriptomes were locally blasted against the UniProtKG database. In the second step all transcripts with positive GO-ids were categorised and quantified (with an in-house Perl script) into the generic 149 categories of the GO-Slim database (http://www.ebi.ac.uk), including the three main ontologies: biological process, cellular component, and molecular function.

The KEGG analysis was performed online through KAAS (KEGG Automatic Annotation Server) using the bi-directional best hit (BBH) methodology and the Gene database. First, all proteins were annotated using the KEGG GENES ortholog group database. This procedure assigned KO (Kegg Orthology) identifiers to the proteins, which were then mapped to BRITE hierarchies of functional classifications. The KEGG results were then categorised and quantified with the help of an in-house Perl script.

Additional files

Additional file 1: Table S1. Data used for the phylogenetic analysis of Hox and ParaHox genes, including the respective GenBank accession numbers. (DOC 31 kb)

Additional file 2: Table S2. Data used for the phylogenetic analysis of *Hedgehog* genes including the respective GenBank accession numbers. (DOC 31 kb)

Additional file 3: Table S3. Data used for the phylogenetic analysis of *Hh*-related genes including the respective GenBank accession numbers. (DOC 31 kb)

Abbreviations

AIC: Akaike information criterion; bp: Base pairs; BBH: Bi-directional best hit; BLAST: Basic local alignment search tool; BUSCO: Benchmarking set of Universal Single-Copy Orthologs; CDD: Conserved domain database; CDS: Coding sequence region; CNS: Central nervous system; DNA: Deoxyribonucleic acid; EGF: Epidermal growth factor; E-value: Expect value; GO: Gene ontology; Hh: Hedgehog; HMM: Hidden markov model; HPG: Hox paralog group; KAAS: KEGG automatic annotation server; KEGG: Kyoto encyclopedia of genes and genomes; KO: Kegg orthology; LNR: Lin-notch-repeat; NCBI: National center of biotechnology information; PEST domain: Proline-glutamic-acid-serine-threonine domain; RNA: Ribonucleic acid; SFF: Standard flowgram format

Acknowledgements

André Luiz de Oliveira acknowledges the financial support of the Brazilian programme "Science without Borders" (Ciência sem Fronteiras; Project Number 6090/13-3). CB is a "Ramon y Cajal" fellow supported by the Spanish MEC (RYC-2014-15615). TW was supported by ASSEMBLE (grant agreement no. 835 (SBR-1)) while collecting developmental stages of *Acanthochitona crinita* and *Antalis entalis*. TW thanks Bernard and Sandie Degnan (University of Queensland) for help in collecting and culturing *Idiosepius notoides*. Henrik Glenner (University of Bergen) is thanked for providing boat time and lab space for collection and culture of aplacophorans and bivalves. Furthermore, André Luiz de Oliveira thanks Martin Fritsch (University of Vienna) for the constructive suggestions in the initial phase of the project, especially concerning the polyplacophoran data, Sandy Richter (University of Leipzig) and Lars Hering (University of Kassel) for help with the bioinformatic analyses, Andrew Calcino (University of Vienna) for polishing the English, and Thomas Rattei (University of Vienna) for providing the high-end computational infrastructure needed in this work. Parts of this work were supported by grant no. P24276-B22 of the Austrian Science Foundation (FWF) on molluscan EvoDevo to AW.

Availability of data and materials

Sequences of the genes analyzed in this work are available through GenBank.

Authors' contributions

ALD and AW designed the project. ALD designed and executed the bioinformatics pipelines, performed the data analysis, and drafted the manuscript. CB contributed to the bioinformatics analysis and fine-tuning of the project. AK, ER, MS, TW, and CT sampled and expertly identified the specimens used in this analysis. ALD and AW jointly finalised the manuscript. All authors read, commented on, and approved the final version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

All authors agree to be named as co-author of this publication.

Ethics approval and consent to participate

Not applicable. No ethical approvals were required for the studies performed.

Data deposition

All sequences used in the phylogenetic analysis are available through the GenBank database. All the accession numbers are listed in Additional file 1: Table S1, Additional file 2: Table S2 and Additional file 3: Table S3.

Author details

[1]Department of Integrative Zoology, Faculty of Life Sciences, University of Vienna, Althanstraße 14, Vienna 1090, Austria. [2]University of Bergen, University Museum, The Natural History Collections, Allégaten 41, 5007 Bergen, Norway. [3]Museo Nacional de Ciencias Naturales, Spanish National Research Council (CSIC), José Gutiérrez Abascal 2, Madrid 28006, Spain. [4]Institute of Biology, University of Leipzig, Leipzig 04103, Germany.

Received: 2 March 2016 Accepted: 8 September 2016
Published online: 10 November 2016

References

1. Davidson EH, Erwin DH. Gene regulatory networks and the evolution of animal body plans. Science. 2006;311:796–800.
2. Carroll SB. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell. 2008;134:25–36.
3. Mallarino R, Abzhanov A. Paths less traveled. Evo-devo approaches to investigating animal morphological evolution. Annu Rev Cell Dev Biol. 2012;28:743–63.
4. Carroll SB. Endless forms: the evolution of gene regulation and morphological diversity. Cell. 2000;101:557–80.
5. Peter IS, Davidson EH. Evolution of gene regulatory networks controlling body plan development. Cell. 2011;144:970–85.
6. Leininger S, Adamski M, Bergum B, Guder C, Liu J, Laplante M, et al. Developmental gene expression provides clues to relationships between sponge and eumetazoan body plans. Nat Commun. 2014;5:3905.
7. Balavoine G, de Rosa R, Adoutte A. Hox clusters and bilaterian phylogeny. Mol Phylogenet Evol. 2002;24:366–73.
8. Ryan JF, Mazza ME, Pang K, Matus DQ, Baxevanis AD, Martindale MQ, et al. Pre-bilaterian origins of the Hox cluster and the Hox code: evidence from the sea anemone, *Nematostella vectensis*. PLoS One. 2007;2:e153.
9. Minelli A. EvoDevo and its significance for animal evolution and phylogeny. In: Wanninger A, editor. Evolutionary Developmental Biology of Invertebrates 1: Introduction, Non-Bilateria, Acoelomorpha, Xenoturbellida, Chaetognatha. Vienna: Springer; 2015. p. 1–23.
10. McGinnis W, Levine MS, Hafen E, Kuroiwa A, Gehring WJ. A conserved DNA sequence in homoeotic genes of the *Drosophila* Antennapedia and bithorax complexes. Nature. 1984;308:428–33.
11. De Rosa R, Grenier JK, Andreeva T, Cook CE, Adoutte A, Akam M, et al. Hox genes in brachiopods and priapulids and protostome evolution. Nature. 1999;399:772–6.
12. Brooke NM, Garcia-Fernàndez J, Holland PW. The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. Nature. 1998;392:920–2.

13. Ferrier DE, Holland PW. Ancient origin of the Hox gene cluster. Nat Rev Genet. 2001;2:33–8.
14. Garcia-Fernàndez J. Hox, ParaHox, ProtoHox: facts and guesses. Heredity (Edinb). 2005;94:145–52.
15. Ferrier DE. The origin of the Hox/ParaHox genes, the Ghost Locus hypothesis and the complexity of the first animal. Brief Funct Genomics. 2015; doi:10.1093/bfgp/elv05
16. Simakov O, Marletaz F, Cho SJ, Edsinger-Gonzales E, Havlak P, Hellsten U, et al. Insights into bilaterian evolution from three spiralian genomes. Nature. 2013;493:526–31.
17. Ponder WF, Linderg DR. Molluscan Evolution and Phylogeny: An introduction. In: Ponder WF, Lindberg DR (ed). Phylogeny and Evolution of the Mollusca. Berkeley (CA): Univ of California Press; 2008. p. 1-17.
18. Nielsen C, Haszprunar G, Ruthensteiner B, Wanninger A. Early development of the aplacophoran mollusc *Chaetoderma*. Acta Zool. 2007;88:231–47.
19. Henry JQ, Okusu A, Martindale MQ. The cell lineage of the polyplacophoran, *Chaetopleura apiculata*: variation in the spiralian program and implications for molluscan evolution. Dev Biol. 2004;272:145–60.
20. Kulikova VA, Kolbin KG, Kolotukhina NK. Reproduction and larval development of the gastropod *Cryptonatica janthostoma* (Gastropoda: Naticidae). Russ J Mar Biol. 2007;33:324–28.
21. Van Dongen CAM, Geilenkirchen WLM. The development of *Dentalium* with special reference to the significance of the polar lobe. I, II, III. Division chronology and development of the cell pattern in *Dentalium dentale* (Scaphopoda). Proc Kongl Ned Akad v Wet Ser C. 1974;77:57–100.
22. Van Dongen CAM, Geilenkirchen WLM. The development of *Dentalium* with special reference to the significance of the polar lobe. IV. Division chronology and development of the cell pattern in *Dentalium dentale* after removal of the polar lobe at first cleavage. Proc Kongl Ned Akad v Wet Ser C. 1975;78:358–75.
23. Van Dongen CAM, Geilenkirchen WLM. The development of *Dentalium* with special reference to the significance of the polar lobe. V and VI. Differentiation of the cell pattern in lobeless embryos of *Dentalium vulgare* (da Costa) during late larval development. Proc Kongl Ned Akad v Wet Ser C. 1976;79:245–66.
24. Cragg SM. The phylogenetic significance of some anatomical features of bivalve veliger larvae. In: Taylor JD, editor. Origin and evolutionary radiation of the Mollusca. USA: Oxford University Press; 1996. p. 371–80.
25. Morse MP, Zardus JD. Bivalvia. In: Harrison FW, Kohn AJ. Microscopic anatomy of invertebrates: Volume 6A Mollusca II. New York: Wiley-Liss Inc. 1997. p. 7-118.
26. Wanninger A, Wollesen T. Mollusca. In: Wanninger A, editor. Evolutionary Developmental Biology of Invertebrates 2: Lophotrochozoa (Spiralia). Vienna: Springer; 2015. p. 103–53.
27. Okusu A. Embryogenesis and development of *Epimenia babai* (Mollusca Neomeniomorpha). Biol Bull. 2002;203:87–103.
28. Todt C, Wanninger A. Of tests, trochs, shells, and spicules: Development of the basal mollusk *Wirenia argentea* (Solenogastres) and its bearing on the evolution of trochozoan larval key features. Front Zool. 2010;7:6.
29. Zardus JD, Morse MP. Embryogenesis, morphology and ultrastructure of the pericalymma larva of *Acila castrensis* (Bivalvia: Protobranchia: Nuculoida). Invertebr Biol. 1998;117:221–44.
30. Haszprunar G, Wanninger A. Molluscs. Curr Biol. 2012; doi: 10.1016/j.cub.2012.05.039
31. Jackson DJ, Degnan BM (in press). The importance of Evo-Devo to an integrated understanding of molluscan biomineralisation. J Struct Biol. 2016; doi: 10.1016/j.jsb.2016.01.005.
32. Kocot KM, Aguilera F, McDougall C, Jackson DJ, Degnan BM. Sea shell diversity and rapidly evolving secretomes: insights into the evolution of biomineralization. Front Zool. 2016;13:23.
33. Sleight VA, Thorne MA, Peck LS, Arivalagan J, Berland S, Marie A, et al. Characterisation of the mantle transcriptome and biomineralisation genes in the blunt-gaper clam, *Mya truncate*. Mar Genomics. 2016; doi:10.1016/j.margen.2016.01.003
34. Vendrami DL, Shah A, Telesca L, Hoffman JI. Mining the transcriptomes of four commercially important shellfish species for single nucleotide polymorphisms within biomineralization genes. Mar Genomics. 2016; doi:10.1016/j.margen.2015.12.009.3
35. Giusti AF, Hinman VF, Degnan SM, Degnan BM, Morse DE. Expression of a Scr/Hox5 gene in the larval central nervous system of the gastropod *Haliotis*, a non-segmented spiralian lophotrochozoan. Evol Dev. 2000;2:294–302.
36. Lee PN, Callaerts P, de Couet HG, Martindale MQ. Cephalopod Hox genes and the origin of morphological novelties. Nature. 2003;424:1061–5.
37. Samadi L, Steiner G. Involvement of Hox genes in shell morphogenesis in the encapsulated development of a top shell gastropod (*Gibbula varia* L.). Dev Genes Evol. 2009;219:523–30.
38. Samadi L, Steiner G. Expression of Hox genes during the larval development of the snail, *Gibbula varia* (L.)—further evidence of non-colinearity in molluscs. Dev Genes Evol. 2010;220:161–72.
39. Hashimoto N, Kurita Y, Wada H. Developmental role of dpp in the gastropod shell plate and co-option of the dpp signaling pathway in the evolution of the operculum. Dev Biol. 2012;366:367–73.
40. Focareta L, Sesso S, Cole AG. Characterization of homeobox genes reveals sophisticated regionalization of the central nervous system in the European cuttlefish *Sepia officinalis*. PLoS One. 2014;9:e109627.
41. Fritsch M, Wollesen T, de Oliveira AL, Wanninger A. Unexpected co-linearity of Hox gene expression in an aculiferan mollusk. BMC Evol Biol. 2015;15:151.
42. Samadi L, Steiner G. Conservation of ParaHox genes' function in patterning of the digestive tract of the marine gastropod *Gibbula varia*. BMC Dev Biol. 2010;10:74.
43. Wollesen T, Rodríguez Monje SV, McDougall C, Degnan BM, Wanninger A. The ParaHox gene *Gsx* patterns the apical organ and central nervous system but not the foregut in scaphopod and cephalopod molluscs. Evodevo. 2015;6:41.
44. Fritsch M, Wollesen T, Wanninger A. Hox and ParaHox gene expression in early body plan patterning of polyplacophoran mollusks. J Exp Zool B Mol Dev Evol. 2016;326:89–104.
45. Iijima M, Akiba N, Sarashina I, Kuratani S, Endo K. Evolution of Hox genes in molluscs: a comparison among seven morphologically diverse classes. J Molluscan Stud. 2006;72:259–66.
46. Riesgo A, Andrade SC, Sharma PP, Novo M, Pérez-Porro AR, Vahtera V, et al. Comparative description of ten transcriptomes of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa. Front Zool. 2012;9:33.
47. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11:31–46.
48. Bleidorn C. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. Syst Biodivers. 2016;14:1–8.
49. Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, et al. The oyster genome reveals stress adaptation and complexity of shell formation. Nature. 2012;490:49–54.
50. Takeuchi T, Kawashima T, Koyanagi R, Gyoja F, Tanaka M, Ikuta T, et al. Draft genome of the pearl oyster *Pinctada fucata*: a platform for understanding bivalve biology. DNA Res. 2012;19:117–30.
51. Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, et al. The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature. 2015;524:220–4.
52. Richards GS, Degnan BM. The dawn of developmental signaling in the metazoa. Cold Spring Harb Symp Quant Biol. 2009;74:81–90.
53. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
54. Merabet S, Hudry B, Saadaoui M, Graba Y. Classification of sequence signatures: a guide to Hox protein function. Bioessays. 2009;31:500–11.
55. Passamaneck YJ, Halanych KM. Evidence from Hox genes that bryozoans are lophotrochozoans. Evol Dev. 2004;6:275–81.
56. LaRonde-LeBlanc NA, Wolberger C. Structure of HoxA9 and Pbx1 bound to DNA: Hox hexapeptide and DNA recognition anterior to posterior. Genes Dev. 2003;17:2060–72.
57. Halanych KM. The new view of animal phylogeny. Annu Rev Ecol Evol Syst. 2004;35:229–56.
58. Bürglin TR. The Hedgehog protein family. Genome Biol. 2008;9:241.
59. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.
60. Kocot KM, Cannon JT, Todt C, Citarella MR, Kohn AB, Meyer A, et al. Phylogenomics reveals deep molluscan relationships. Nature. 2011;477:452–6.
61. Smith SA, Wilson NG, Goetz FE, Feehery C, Andrade SC, Rouse GW, et al. Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature. 2011;480:364–7.
62. Struck TH, Paul C, Hill N, Hartmann S, Hösel C, Kube M, et al. Phylogenomic analyses unravel annelid evolution. Nature. 2011;471:95–8.
63. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, et al. Phylogenomics resolves the timing and pattern of insect evolution. Science. 2014;346:763–7.
64. O'Neil ST, Emrich SJ. Assessing De Novo transcriptome assembly metrics for consistency and utility. BMC Genomics. 2013;14:465.

65. Mundry M, Bornberg-Bauer E, Sammeth M, Feulner PGD. Evaluating characteristics of de novo assembly software on 454 transcriptome data: A simulation approach. PLoS One. 2012;7:e31410.
66. Sadamoto H, Takahashi H, Okada T, Kenmoku H, Toyota M, Asakawa Y. De novo sequencing and transcriptome analysis of the central nervous system of mollusc *Lymnaea stagnalis* by deep RNA sequencing. PLoS One. 2012;7:e42546.
67. Mehr S, Verdes A, DeSalle R, Sparks J, Pieribone V, Gruber DF. Transcriptome sequencing and annotation of the polychaete *Hermodice carunculata* (Annelida, Amphinomidae). BMC Genomics. 2015;16:445.
68. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 2007;8:R143.
69. Meyer A, Todt C, Mikkelsen NT, Lieb B. Fast evolving 18S rRNA sequences from Solenogastres (Mollusca) resist standard PCR amplification and give new insights into mollusk substitution rate heterogeneity. BMC Evol Biol. 2010;10:70.
70. Osca D, Irisarri I, Todt C, Grande C, Zardoya R. The complete mitochondrial genome of *Scutopus ventrolineatus* (Mollusca: Chaetodermomorpha) supports the Aculifera hypothesis. BMC Evol Biol. 2014;14:197.
71. Senatore A, Edirisinghe N, Katz PS. Deep mRNA sequencing of the *Tritonia diomedea* brain transcriptome provides access to gene homologues for neuronal excitability, synaptic transmission and peptidergic signalling. PLoS One. 2015;10:e0123514.
72. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12:671–82.
73. Richardson MF, Sherman CDH. De Novo Assembly and Characterization of the Invasive Northern Pacific Seastar Transcriptome. PLoS One. 2015;10:e0142003.
74. Zhang D, Wang F, Dong S, Lu Y. De novo assembly and transcriptome analysis of osmoregulation in *Litopenaeus vannamei* under three cultivated conditions with different salinities. Gene. 2016;578:185–93.
75. Werner GD, Gemmell P, Grosser S, Hamer R, Shimeld SM. Analysis of a deep transcriptome from the mantle tissue of *Patella vulgata* Linnaeus (Mollusca: Gastropoda: Patellidae) reveals candidate biomineralising genes. Mar Biotechnol (NY). 2013;15:230–43.
76. Ding J, Zhao L, Chang Y, Zhao W, Du Z, Hao Z. Transcriptome sequencing and characterization of Japanese scallop *Patinopecten yessoensis* from different shell color lines. PLoS One. 2015;10:e0116406.
77. Harney E, Dubief B, Boudry P, Basuyaux O, Schilhabel MB, Huchette S, et al. De novo assembly and annotation of the European abalone *Haliotis tuberculata* transcriptome. Mar Genomics. 2016; doi: 10.1016/j.margen.2016.03.002
78. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 2009;25:404–13.
79. Gibson AK, Smith Z, Fuqua C, Clay K, Colbourne JK. Why so many unknown genes? Partitioning orphans from a representative transcriptome of the lone star tick *Amblyomma americanum*. BMC Genomics. 2013;14:135.
80. Vogeler S, Galloway TS, Lyons BP, Bean TP. The nuclear receptor gene family in the Pacific oyster, *Crassostrea gigas*, contains a novel subfamily group. BMC Genomics. 2014;15:369.
81. Aguilera F, McDougall C, Degnan BM. Evolution of the tyrosinase gene family in bivalve molluscs: independent expansion of the mantle gene repertoire. Acta Biomater. 2014;10:3855–65.
82. Canapa A, Biscotti MA, Olmo E, Barucca M. Isolation of Hox and ParaHox genes in the bivalve *Pecten maximus*. Gene. 2005;348:83–8.
83. Takeuchi T, Koyanagi R, Gyoja F, Kanda M, Hisata K, Fujie M, et al. Bivalve-specific gene expansion in the pearl oyster genome: implications of adaptation to a sessile lifestyle. Zoological Lett. 2016; doi: 10.1186/s40851-016-0039-2
84. Callaerts P, Lee PN, Hartmann B, Farfan C, Choy DW, Ikeo K, et al. HOX genes in the sepiolid squid *Euprymna scolopes*: implications for the evolution of complex body plans. Proc Natl Acad Sci U S A. 2002;99:2088–93.
85. Barucca M, Olmo E, Canapa A. Hox and paraHox genes in bivalve molluscs. Gene. 2003;317:97–102.
86. Sharkey M, Graba Y, Scott MP. Hox genes in evolution: protein surfaces and paralog groups. Trends Genet. 1997;13:145–51.
87. Carroll SB. Homeotic genes and the evolution of arthropods and chordates. Nature. 1995;376:479–85.
88. Ruvkun G, Hobert O. The taxonomy of developmental control in *Caenorhabditis elegans*. Science. 1998;282:2033–41.
89. Weiss JB, Von Ohlen T, Mellerick DM, Dressler G, Doe CQ, Scott MP. Dorsoventral patterning in the *Drosophila* central nervous system: the intermediate neuroblasts defective homeobox gene specifies intermediate column identity. Genes Dev. 1998;12:3591–602.
90. Degnan BM, Morse DE. Identification of eight homeobox-containing transcripts expressed during larval development and at metamorphosis in the gastropod mollusc *Haliotis rufescens*. Mol Mar Biol Biotechnol. 1993;2:1–9.
91. Stolfi A, Brown FD. Tunicata. In: Wanninger A, editor. Evolutionary Developmental Biology of Invertebrates 6: Deuterostomia. Vienna: Springer; 2015. p. 135–204.
92. Seo HC, Edvardsen RB, Maeland AD, Bjordal M, Jensen MF, Hansen A, et al. Hox cluster disintegration with persistent anteroposterior order of expression in *Oikopleura dioica*. Nature. 2004;431:67–71.
93. Spagnuolo A, Ristoratore F, Di Gregorio A, Aniello F, Branno M, Di Lauro R. Unusual number and genomic organization of Hox genes in the tunicate *Ciona intestinalis*. Gene. 2003;309:71–9.
94. Pernice M, Deutsch JS, Andouche A, Boucher-Rodoni R, Bonnaud L. Unexpected variation of Hox genes homeodomains in cephalopods. Mol Phylogenet Evol. 2006;40:872–9.
95. Smith SA, Wilson NG, Goetz FE, Feehery C, Andrade SCS, Rouse GW, et al. Corrigendum: Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature. 2013;493:708.
96. Hinman VF, O'Brien EK, Richards GS, Degnan BM. Expression of anterior Hox genes during larval development of the gastropod *Haliotis asinina*. Evol Dev. 2003;5:508–21.
97. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
98. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21:3674–6.
99. Song H, Yu ZL, Sun LN, Gao Y, Zhang T, Wang HY. De novo transcriptome sequencing and analysis of *Rapana venosa* from six different developmental stages using Hi-seq 2500. Comp Biochem Physiol Part D Genomics Proteomics. 2016;17:48–57.
100. Gerhart J. 1998 Warkany lecture: signaling pathways in development. Teratology. 1999;60:226–39.
101. Prud'homme B, Lartillot N, Balavoine G, Adoutte A, Vervoort M. Phylogenetic analysis of the Wnt gene family: Insights from lophotrochozoan members. Curr Biol. 2002;12:1395–400.
102. Cho SJ, Vallès Y, Giani Jr VC, Seaver EC, Weisblat DA. Evolutionary dynamics of the wnt gene family: a lophotrochozoan perspective. Mol Biol Evol. 2010;27:1645–58.
103. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science. 2007;317:86–94.
104. Chapman JA, Kirkness EF, Simakov O, Hampson SE, Mitros T, Weinmaier T, et al. The dynamic genome of Hydra. Nature. 2010;464:592–6.
105. Nederbragt AJ, van Loon AE, Dictus WJ. Evolutionary biology: hedgehog crosses the snail's midline. Nature. 2002;417:811–2.
106. Grimaldi A, Tettamanti G, Acquati F, Bossi E, Guidali ML, Banfi S, et al. A hedgehog homolog is involved in muscle formation and organization of *Sepia officinalis* (Mollusca) mantle. Dev Dyn. 2008;237:659–71.
107. Richards GS, Simionato E, Perron M, Adamska M, Vervoort M, Degnan BM. Sponge genes provide new insight into the evolutionary origin of the neurogenic circuit. Curr Biol. 2008;18:1156–61.
108. Lai E. Notch signaling: control of cell communication and cell fate. Development. 2004;131:965–73.
109. Artavanis-Tsakonas S, Rand MD, Lake RJ. Notch signaling: cell fate control and signal integration in development. Science. 1999;284:770–6.
110. Gazave E, Lapébie P, Richards GS, Brunet F, Ereskovsky AV, Degnan BM, et al. Origin and evolution of the Notch signalling pathway: an overview from eukaryotic genomes. BMC Evol Bio. 2009;9:249.
111. Mohler J, Vani K. Molecular organization and embryonic expression of the hedgehog gene involved in cell-cell communication in segmental patterning of *Drosophila*. Development. 1992;115:957–71.
112. Echelard Y, Epstein DJ, St-Jacques B, Shen L, Mohler J, McMahon JA, et al. Sonic hedgehog, a member of a family of putative signaling molecules, is implicated in the regulation of CNS polarity. Cell. 1993;75:1417–30.
113. Bürglin TR, Kuwabara PE. Homologs of the Hh signalling network in *C. elegans*. WormBook. 2006. p. 1-14. doi:10.1895/wormbook.1.76.1.
114. Matus DQ, Magie CR, Pang K, Martindale MQ, Thomsen GH. The Hedgehog gene family of the cnidarian, *Nematostella vectensis*, and implications for understanding metazoan Hedgehog pathway evolution. Dev Biol. 2008;313:501–18.
115. Wollesen T, Rodríguez Monje SV, Todt C, Degnan BM, Wanninger A. Ancestral role of *Pax2/5/8* in molluscan brain and multimodal sensory system development. BMC Evol Biol. 2015;15:231.

116. Redl E, Scherholz M, Todt C, Wollesen T, Wanninger A. Development of the nervous system in Solenogastres (Mollusca) reveals putative ancestral spiralian features. Evodevo. 2014;5:48.
117. Wollesen T, Cummins SF, Degnan BM, Wanninger A. FMRFamide gene and peptide expression during central nervous system development of the cephalopod mollusk, *Idiosepius notoides*. Evol Dev. 2010;12:113–30.
118. Kristof A, de Oliveira AL, Kolbin KG, Wanninger A. Neuromuscular development in Patellogastropoda (Mollusca: Gastropoda) and its importance for reconstructing ancestral gastropod bodyplan features. J Zoolog Syst Evol Res. 2016;54:22–39.
119. Kristof A, de Oliveira AL, Kolbin KG, Wanninger A. A putative species complex in the Sea of Japan revealed by DNA sequence data: a study on *Lottia* cf. *kogamogai* (Gastropoda: Patellogastropoda). J. Zoolog. Syst. Evol. Res. 2016; doi: 10.1111/jzs.12120.
120. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
121. Peng Y, Leung HC, Yiu SM, Lv MJ, Zhu XG, Chin FY. IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics. 2013;29:i326–34.
122. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, et al. Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. Genome Res. 2004;14:1147–59.
123. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5.
124. Novaes J, Rangel LT, Ferro M, Abe RY, Manha AP, de Mello JC, et al. A comparative transcriptome analysis reveals expression profiles conserved across three *Eimeria* spp. of domestic fowl and associated with multiple developmental stages. Int J Parasitol. 2012;42:39–48.
125. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
126. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009;23:205–11.
127. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26:2460–1.
128. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
129. Zhang Y, Sun Y. HMM-FRAME: accurate protein domain classification for metagenomic sequences containing frameshift errors. BMC Bioinformatics. 2011;12:198.
130. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 2014;30:3276–8.
131. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61:539–42.
132. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 1992;8:275–82.
133. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 2011;27:1164–5.
134. Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 2001;18:691–9.

2.1.1 Erratum

Erratum to the article "Comparative transcriptomics enlarges the toolkit of known developmental genes in mollusks." by De Oliveira AL, Wollesen T, Kristof A, Scherholz M, Redl E, Todt C, et al., published in BMC Genomics. 2016;17:905.

Following the publication of our recent article in BMC Genomics [1], it came to our attention that some values in the column named “No. of non-redundant putative proteins” in Tables 3 (page 4) and 5 (page 5); and the number of Wnt genes present in some lophotrochozoan representatives discussed in “Functional characterisation and diversity of the gene repertoire in mollusks” are incorrect.

1. Tables

Concerning the correct values of the “No. of non-redundant putative proteins”, they can be found below:

Organism	No. of non-redundant putative proteins
Gymnomenia pellucida (Neomeniomorpha)	38,400
Wirenia argentea (Neomeniomorpha)	55,194
Scutopus ventrolineatus (Chaetodermomorpha)	32,337
Acanthochitona crinita (Polyplacophora)	53,165
Idiosepius notoides (Cephalopoda)	29,064
Lottia cf. *kogamogai* (Gastropoda)	16,781
Nucula tumidula (Bivalvia)	41,781
Antalis entails (Scaphopoda)	40,347

Taking into consideration the aforementioned values the following statement in “Discussion: Feasibility of non-model mollusks for comparative transcriptomic studies”:

“In our study, except for *Lottia* cf. *kogamogai* and *Idiosepius notoides*, all protein gene sets have an inflated number of putative proteins when compared to the patellogastropod, oyster, and octopus data.”

Should be rewritten as follows:

“In our study, except for *Lottia* cf. k*ogamogai, Idiosepius notoides*, and *Scutopus ventrolineatus* all protein gene sets have an inflated number of putative proteins when compared to the patellogastropod, oyster, and octopus data.”

2. Wnt genes

Regarding the error in the number of Wnt genes in lophotrochozoan representatives, the following statement in “Abstract:Results”:

“The KEGG analysis revealed *Wnt8*, *Wnt11*, and *Wnt16* as Wnt genes hitherto not reported for mollusks, thereby enlarging the known Wnt complement of the phylum.”

and in the “Discussion: Functional characterisation and diversity of the gene repertoire in mollusks”:

“Thirteen Wnt subfamilies have been characterised in metazoans, while lophotrochozoan representatives, such as the polychaete annelids *Capitella teleta* and *Platynereis dumerilii*, commonly possess only 12 subfamilies and the basal-branching gastropods *Patella vulgata* and *Lottia gigantea* only nine (*WntA*, *Wnt1*, *Wnt2*, *Wnt4*, *Wnt5*, *Wnt6*, *Wnt7*, *Wnt9*, and *Wnt10*) [2,3]. We found three additional subfamilies in mollusks using KEGG orthology assignment, namely *Wnt8*, *Wnt11*, and *Wnt16*, suggesting that molluscan gene content in the Wnt subfamilies matches that of their lophotrochozoan relatives.

Should be rewritten, respectively, as follows:

“The KEGG analysis revealed the *Wnt8* gene family hitherto not reported for mollusks, thereby enlarging the known Wnt complement of the phylum.”

and:

“Thirteen Wnt subfamilies have been characterised in metazoans, while lophotrochozoan representatives, such as the polychaete annelids *Capitella teleta* [3] and *Platynereis dumerilii* [4] possess 12 subfamilies, and the basal-branching gastropods *Lottia gigantea* [3] and *Patella vulgata* [2] only 11 (*WntA*, *Wnt1*, *Wnt2*, *Wnt4*, *Wnt5*, *Wnt6*, *Wnt7*, *Wnt9*, *Wnt10*, *Wnt11*, *Wnt16*), and 4 (*WntA*, *Wnt1*, *Wnt2*, *Wnt10*) respectively. We found one additional subfamily in mollusks using KEGG orthology assignment, namely Wnt8, suggesting that molluscan gene content in the Wnt subfamilies matches that of their lophotrochozoan relatives.”

It is important to note that none of the downstream analyses and results in the paper were affected by these errors. We regret that these details were overlooked prior to publication, and we apologise for any inconvenience caused by them.

References

1. De Oliveira AL, Wollesen T, Kristof A, Scherholz M, Redl E, Todt C, et al. Comparative transcriptomics enlarges the toolkit of known developmental genes in mollusks. BMC Genomics. 2016;17:905.

2. Prud'homme B, Lartillot N, Balavoine G, Adoutte A, Vervoort M. Phylogenetic analysis of the Wnt gene family: Insights from lophotrochozoan members. Curr Biol. 2002;12:1395–400.

3. Cho SJ, Vallès Y, Giani Jr VC, Seaver EC, Weisblat DA. Evolutionary dynamics of the wnt gene family: a lophotrochozoan perspective. Mol Biol Evol. 2010;27:1645–58.

4. Janssen R, Le Gouar M, Pechmann M, Poulin F, Bolognesi R, Schwager EE, et al. Conservation, loss, and redeployment of Wnt ligands in protostomes: implications for understanding the evolution of segment formation. BMC Evol Biol. 2010;10:374.

2.2 Manuscript 2 - Extensive conservation of the proneuropeptide and peptide hormone complement in mollusks (in review)

De Oliveira[1], AL; Calcino[1], A and Wanninger[1] A.*

[1]Department of Integrative Zoology, Faculty of Life Sciences, University of Vienna, Althanstraße 14, Vienna, 1090, Austria

ORCID ID:

Andreas Wanninger: 0000-0002-3266-5838

André Luiz de Oliveira: 0000-0003-3542-4439

Andrew Calcino: 0000-0002-3956-1273

*author for correspondence: andreas.wanninger@univie.ac.at

Authors'email:

André Luiz de Oliveira: andre.luiz.de.oliveira@univie.ac.at

Andrew Calcino: andrew.calcino@univie.ac.at

Andreas Wanninger: andreas.wanninger@univie.ac.at

Status: submitted to Scientific Reports

2.2.1 Abstract

As one of the most diverse groups of invertebrate animals, mollusks represent powerful models for neurobiological and developmental studies. Neuropeptides and peptide hormones are a heterogeneous class of signalling molecules involved in chemical communication between neurons and in neuroendocrine regulation. Here we present a fine-grained view of the molluscan neuropeptide and peptide hormone toolkit. Our results expand the distribution of several peptide families (e.g., prokineticin, insulin-related peptides, prohormone-4, LFRFamide) within Lophotrochozoa and provide evidence for an early origin of others (e.g., GNXQN/prohormone-2, neuroparsin). We identified two new peptide families broadly distributed among mollusks, the protostomian fibrinogen-related family and the conchiferan PXRX family. We found the Wnt antagonist *dickkopf1/2/4* ortholog in lophotrochozoans and nematodes and reveal that the egg-laying hormone family is a DH44 homolog restricted to gastropods. Our data demonstrate that numerous peptides evolved much earlier than previously assumed and that key signaling elements are extensively conserved among extant mollusks.

2.2.2 Introduction

Neuropeptides and peptide hormones constitute a heterogeneous group of evolutionary related signaling protein molecules involved in neuro-modulation, neurotransduction, and hormonal functions (Liu et al., 2008), that commonly act via G protein-couple receptors (GPCRs). Two major differences between neuropeptides and peptide hormones concern the biological system in which they are functional as well as their signaling targets. Neuropeptides are secreted by neuronal cells and act on neighboring targets (cell-cell contact) whereas peptide hormones diffuse over long distances via haemolymph or blood, affecting targets far from the signaling source. The latter mechanism is controlled by the endocrine system (Hartenstein, 2006).
Neuropeptides and peptide hormones are synthesized in the form of large inactive precursor molecules known as proneuropeptides (pNPs) or prohormones. They are redirected to the secretory apparatus and are further cleaved and modified to regulate homeostatic processes and distinct behaviours in animals (Douglass et al., 1984). Structurally, pNPs and prohormones share common characteristics such as

the presence of an N-terminal signal peptide and one or more peptide sequences flanked by mono- or dibasic cleavage sites which are recognised by prohormone convertases. Each pNP and peptide prohormone may give rise to a single bioactive peptide, several copies of a single bioactive peptide, or more than one distinct bioactive peptide. Additional enzymatic processing steps, i.e. post-translational modifications (e.g. C-terminal alpha-amidation, N-terminal pyroglutamination) often occur before the generation of the active peptides (Eipper et al., 1992; Steiner, 1998; Hook et al., 2008).

The recent improvement of DNA sequencing technologies accompanied by the substantial reduction of costs has expanded the investigation of neuropeptide and hormonal signaling systems beyond the classical model organisms such as the nematode *Caenorhabditis elegans* (Nathoo et al., 2001), the fruit fly *Drosophila melanogaster* (Hewes & Taghert, 2001), and human (Fredriksson et al., 2003). Thus, comparative research into the evolution and diversity of metazoan neuropeptides, peptide hormones, and their molecular components today involves a range of previously neglected taxa from virtually all major metazoan lineages.

Proneuropeptides are widespread in eumetazoans (all animals except sponges) (Jékely, 2013; Mirabeau & Joly, 2013). The key components of the enzymatic toolkit essential for pNP and peptide prohormone processing, maturation, and secretion originated long before the emergence of Eumetazoa and are commonly recognized in organisms that lack a nervous system such as sponges and algae (Srivastava et al., 2010; Attenborough et al., 2012; Whalan et al.; 2012). Within Lophotrochozoa (a major clade of bilaterally symmetrical protostome animals that includes groups as diverse as platyhelminths, annelids, mollusks, or brachiopods), comprehensive investigations of the neuropeptide and peptide hormonal signaling systems have been conducted in the annelids *Capitella teleta* (Veenstra, 2011), *Helobdella robusta* (Veenstra, 2011), and *Platynereis dumerilii* (Conzelmann et al., 2013), as well as in two platyhelminths, the parasitic *Schistosoma mansoni* (Berriman et al., 2009*)* and the free-living *Schmidtea mediterranea* (Collins et al., 2010). The number of predicted peptide precursors (proneuropeptides and prohormones) in these species ranges from 13 in *S. mansoni* to 98 in *P. dumerilii.* These results show a tremendous variation in the composition of signaling peptides even in closely related organisms.

Mollusks comprise the most speciose and diverse lophotrochozoan phylum. They display highly variable behavioural and physiological repertoires, developmental

pathways (ranging from indirect development via various larval types to direct development), and neuroanatomical features. Molluscan nervous systems vary widely in their degree of complexity. They may exhibit little or no anterior centralization and may lack ganglia along their four longitudinal nerve cords (e.g., in aculiferans and monoplacophorans; Lemche & Wingstrand, 1959; Wingstrand, 1985; Todt et al., 2008; Faller et al., 2012; Sumner-Rooney & Sigwart 2018) or may have multiple (pairs of) ganglia (e.g., in the majority of the conchiferan clades). Neural complexity in mollusks peaks in the highly centralized, lobular brains of cephalopods (Hochner & Glanzman, 2016). Despite these considerable morphological differences, thorough assessments of the diversity of proneuropeptides and peptide prohormones in mollusks are only available for a few individual gastropod (Veenstra, 2010; Adamson et al., 2015; Ahn et al., 2017; Bose et al., 2017), bivalves (Stewart et al., 2014; Zhang et al., 2018), and cephalopod species (Zatylny-Gaudin et al., 2016*)* (Table 1). In the five remaining molluscan class-level taxa (Chaetodermomorpha, Neomeniomorpha, Polyplacophora, Scaphopoda and Monoplacophora) comprehensive and systematic investigations that are focused on peptidergic signaling systems are still lacking.

Previous studies have shown a high degree of conservation of the repertoire of neuropeptides and peptide hormones (e.g., achatin, allatotropin, elevenin and LFRFamide) between gastropods, cephalopods, bivalves, and other phyla, corroborating the notion that these molecules originated early in animal evolution (Jékely, 2013; Mirabeau & Joly, 2013). Screening molluscan databases for potential neuropeptides and peptide hormones resulted in the identification of hitherto unknown peptide families with representatives in other animal phyla such as annelids and insects (Zatylny-Gaudin et al., 2016). Numerous peptide families that are restricted to Mollusca or individual molluscan class-level taxa were also identified (Nagle et al., 1989; Bogdanov et al., 1998; Zhang et al., 2018).

In order to elucidate the evolutionary history of peptide signaling molecules and to assess whether the complexity of neural systems is reflected in the diversity of proneuropeptide and peptide hormone complements in mollusks, we analysed 62 publicly available datasets covering 35 molluscan and 19 other metazoan species. Sequence data from Mollusca and nine other lophotrochozoan phyla were included: Annelida, Brachiopoda, Ectoprocta, Entoprocta, Gastrotricha, Nemertea, Phoronida, Platyhelminthes, and Rotifera. We identified 67 peptide families with homologs in one

or more mollusk species. The homology of several other non-molluscan lophotrochozoan peptide sequences was confirmed and their relatedness with the molluscan pNP and peptide prohormones established (e.g., presence of shared conserved motifs, pattern of BLAST connections in the cluster maps). Our study represents the most complete and broad catalog of molluscan proneuropeptides and peptide hormones to date and constitutes an important resource for further investigations of molluscan and lophotrochozoan neural evolution, neurogenesis, and physiology.

2.2.3 Results

2.2.3.1 Prediction of molluscan and lophotrochozoan neuropeptidomes

Quality filtering of the molecular sequence databases followed by *de novo* assembly and identification of the coding sequence regions generated predicted protein datasets ranging from 12,808 (the shallow coverage of the chaetodermomorph *Chaetoderma* sp. data) to 606,184 sequences (combined ultra-deep *Dreissena rostriformis* libraries from different developmental stages). Assessments of completeness in the reconstructed protein datasets based on the presence of 978 benchmarking universal single copy metazoan orthologs (BUSCO; Simão et al., 2015) showed a great variation ranging from 4.73% completeness in the basally branching protobranch bivalve *Yoldia limatula* to up to more than 90.0% in the brachiopod *Lingula anatina*, the scaphopod *Gadila tolmiei*, and the annelid *Capitella teleta*. The 454 (Ronaghi et al., 1998; Margulies et al., 2005) sequenced libraries of the polyplacophoran *Chaetopleura apiculata*, the gastropods *Littorina littorea*, *Perotrochus lucaya,* and *Siphonaria pectinate,* the cephalopod *Nautilus pompilius*, and the bivalve *Yoldia limatula* present the highest number of missing BUSCOs. The BUSCO assessment results for the 54 lophotrochozoan proteomes are summarised in Additional file 1. The established non-redundant lophotrochozoan neuropeptidomes (set of proneuropeptides and peptide prohormones) have between 173 (in the gastropod *Biomphalaria glabrata*) and 14,195 (combined *Dreissena rostriformis* transcriptomes) secreted protein sequences with all hallmarks of either a bona fide proneuropeptide or a peptide hormone (e.g., signal peptide, non-folded protein domain, and repetitive motif sites). The 54 lophotrochozoan non-redundant

neuropeptidomes are mostly composed of full-length coding region protein sequences (Additional file 2).

2.2.3.2 The proneuropeptide/peptide prohormone complement of Mollusca

Using a bioinformatic pipeline for proneuropeptide and peptide prohormonal identification adapted from previous surveys (Jékely, 2013; Conzelmann et al., 2013) fine-grained 2D maps depicting the presence of major components of the molluscan neuropeptide/hormonal signaling systems were generated (Figure 1; Additional files 3 and 4). These depict hundreds of molluscan and lophotrochozoan homologs of known metazoan pNP/peptide prohormone families that were previously unknown from these clades. The deep molluscan taxonomic sampling identified 67 peptide families distributed in one or more molluscan taxa (Figure 2). The minimum pNP/peptide prohormone complement of the eight class-level taxa of Mollusca ranges from 28 families in monoplacophorans to 59 in bivalves and gastropods (Figure 3). The majority of the proneuropeptide and peptide prohormone families found in mollusks were also identified in other lophotrochozoans such as annelids (52 families in common) and nemerteans (41 families in common) (Figure 2). A full catalog of the mollusk/Lophotrochozoa-containing peptide families are provided in Additional file 5.

2.2.3.3 Eumetazoa-specific pNPs and peptide prohormones

Numerous peptide sequences retrieved from the molluscan and lophotrochozoan databases are also present in animals outside Lophotrochozoa, providing evidence that they were already present in the last common eumetazoan ancestor (Figure 2). These include the cysteine-knot glycoprotein hormones bursicon-alpha and bursicon-beta, insulin-related peptides (IRPs), orthologs of the insect eclosion-hormone (EH), and the extracellular signaling molecule trunk (related to the arthropod prothoracicotropic hormone, PTTH). A variety of mature short peptides encoded by FMRFamide and RYamide pNPs were found in mollusks and seven of the nine lophotrochozoan phyla under investigation (Annelida, Brachiopoda, Entoprocta, Gastrotricha, Nemertea, Phoronida, and Rotifera; Figure 2). Allatostatin-B or myoinhibitory peptides, characterised by the conserved N-terminal tryptophane

residue (W) and the C-terminal Wamide motif in the bioactive pNP, were identified in all eight molluscan classes, annelids, brachiopods, and nemerteans.

One surprising outcome was the identification of *dickkopf1/2/4* orthologs in ecdysozoan and lophotrochozoan representatives, expanding the phyletic distribution of this gene family to the entire Protostomia clade (Figure 4A and 4B). All newly identified protostomian dickkopf sequences (dkk) contain a signal peptide and two conserved cysteine-rich domains (CRD-1 and CRD-2) in which the N-terminal domain (CRD-1) corresponds to the dickkopf domain *per se* and the C-terminal domain (CRD-2) corresponds to the colipase fold (Figure 4A and 4C). Multiple sequence alignment revealed that the CRD-1 domains of the anthozoan cnidarian *Nematostella vectensis* and the protostomes all share eight cysteine residues (Additional file 5). *Hydra dkk1/2/4* orthologs lack the CRD-1 domain (Figure 4A, Additional file 5). Multiple sequence alignment of the colipase CRD-2 domains shows that all protostome, cnidarian, and deuterostome sequences possess ten highly conserved cysteine residues (Figure 4C). Outside the shared cysteine residues of the two CRDs, the dkk proteins show little sequence similarity. Bayesian phylogenetic inferences performed with CRD-2 domains recovered five distinct well-supported dkk clusters, two corresponding to the dkk-3 family (one belonging to deuterostomes and the other one to cnidarians) and the remaining three to the dkk1/2/4 family (Figure 4B). The parasitic nematode *Trichinella spirales*, the ectoproct *Membranipora membranacea*, and the nemertean *Lineus longissimus* sequences are closely related to the *Nematostella dkk/1/2/4* ortholog, while the remaining two lophotrochozoans, the bivalve mollusk *Ennucula tenuis* and the entoproct *Barentsia gracilis*, are more closely related to the hydrozoan *Hydra vulgaris*. Although in-cluster resolution was robust, a lack of resolution between clusters prevented a phylogenetic classification of two dkk-3 and the three dkk/1/2/4 groups relative to each other.

Another pNP family with a C-terminal colipase fold-related domain, named prokineticin, was identified in virtually all lophotrochozoan phyla sampled, with the exception of Platyhelminthes (Figure 2). Thirty-five transcripts belonging to the monoplacophoran *Laevipilina hyalina* with homology to other metazoan prokineticins were found (Additional file 4). Multiple sequence alignment and phylogenetic analyses show the presence of 4 groups of prokineticin-like peptides with high posterior probability support values in *Laevipilina* (Additional file 6).

2.2.3.4 Bilateria-specific pNPs and peptide prohormones

Numerous pNP/hormone representatives found in mollusks are present in the vast majority of other bilaterian clades, including 7B2, achatin, allatotropin, adipokinetic-hormone (AKH), allatostatin-C, crustacean cardio-active peptide (CCAP), elevenin (L11), glycoprotein-alpha and glycoprotein-beta, gonadotropin releasing hormone (GnRH), leucokinin, neuropeptide Y/F, proenkephalin, sulfakinin, tachykinin, and vasotocin/neurophysin (Figure 2). In many instances, peptide families were identified in all eight class-level taxa of Mollusca, such as EP, SIF/FFamide, allatostatin-A, luqin, pigment dispersing factor (PDF), pedal-peptide (ortholog of the ecdyszoan orcokinin), and small cardioactive peptide (sCAP). The insect single copy PDF pNPs formed a well-connected cluster with gastropod cerebrins and a number of other, previously uncharacterised, molluscan, annelid, and nemertean pNPs.

Despite orthology of calcitonin and diuretic hormone 31 (DH31), these two pNP families are split into two distinct clusters on the 2D cluster map (Figure 1A). The calcitonin pNP is present in aculiferan and conchiferan representatives, and, apart from the truncated *Ennucula tenuis* sequence, all sequences contain the two conserved cysteine residues in the mature neuropeptide. The polyplacophoran *Leptochiton rugatus* is the only investigated mollusk with both calcitonin and DH31 orthologs (Additional file 5). As in the *Platynereis dumerilii* and the insect DH31 orthologs, the molluscan DH31 sequence lacks cysteine residues in the bioactive peptide domain (Additional file 5).

As with calcitonin/DH31, the orthologous egg laying hormone (ELH) and DH44 families are split into two distinct clusters (Figure 1A). Identification of the conserved ELH/DH44 motif in annelids, nemerteans, mollusks, and arthropods showed different patterns of peptide repetition duplications, ranging from one motif in flies (*Drosophila melanogaster*), silkworm (*Bombyx mori*), and the nemertean *Tubulanus polymorphus*, to up to 16 in the polychaete annelid *Platynereis dumerilii* (Figure 5B, Additional file 7). Within Mollusca, dantaliid scaphopods (*Grapctacme eborea* and *Antalis entalis)* harbor three repetitions of the motif, whereas the gadiliid scaphopod *Gadila tolmiei,* the bivalves *Pinctada fucata, Crassostrea gigas* and *Patinopecten yessoensis*, and the polyplacophoran *Acanchochitona crinita* only have two (Figure 5B, Additional file 7). Multiple sequence alignments using DH44/ELH and corticotropin-releasing bioactive hormone domains showed higher conservation of amino acid positions

within the C- and N- terminal regions (Figure 5C). Bayesian phylogenetic inferences using molluscan ELH sequences and the protostomian diuretic hormone 44 (DH44) as well as the deuterostome corticotropin releasing factor (CRH) orthologs revealed the presence of three well-supported and distinct clades (Figure 5A). The first contains the ecdysozoan and deuterostome sequences, the second is exclusively composed of gastropod ELH sequences, and the third comprises the remaining non-gastropod mollusk, the annelid, and the nemertean sequences (Figure 5A). Thereby, the bivalve, scaphopod, and polyplacophoran peptide sequences show a higher degree of similarity to the *Platynereis* DH44 and the nemertean sequences than to their closest gastropod relatives (Figure 5A). These results are consistent with estimates of evolutionary divergence, which show that the sequences of the bivalves *C. gigas*, *P. fucata*, *P. yessoensis*, the polyplacophoran *A. crinita,* and the scaphopod *G. tolmiei* are less divergent from the annelid and nemertean sequences than from their gastropod counterparts (Additional file 8).

2.2.3.5 Protostomia-specific pNPs and peptide prohormones

Nine molluscan pNP/peptide prohormone families originated in the stem protostome (Figure 2), including prohormone-3 and prohormone-4, two myomodulin proneuropeptide precursors, neuroparsin, and PKYMDT/whitnin. Lophotrochozoan myomodulin pNPs generally yield multiple copies of small LRL- and LRMamide bioactive peptides, with the conserved motif located at the C-terminal end (although variations were observed in conchiferan and nemerteans representatives, e.g., VRL-, LRV-, and VRMamide) (Additional figure 9). Conversely, the sequence composition of the N-terminal region of the bioactive myomodulin neuropeptides is highly variable, resulting in the production of numerous distinct peptides from each precursor neuropeptide. In the case of the aplacophoran mollusks, each bioactive peptide produced from the myomodulin-2 pNPs is unique, while in gastropod and cephalopod myomodulin-1 pNPs multiple identical copies of the bioactive peptides are present (Additional file 9). In addition to LRL- and LRMamide peptides, molluscan and lophotrochozoan myomodulin pNPs (with the exception of gastropod myomodulin-1 and platyhelminth pNPs) encode a distinct class of PRXamide bioactive peptides (see Additional file 9).

The GNXQN family grouped together with the insect prohormone-2 peptides and forms a well-resolved cluster (Figure 1A). Motif searches revealed the presence of a highly conserved region (GN[QHR]QN) shared among all protostomians towards the N-terminal of all GNXQN/prohormone-2 pNPs (Additional file 5).

A cluster composed of lophotrochozoan representatives (mollusks, phoronids, brachiopods, and annelids) and one ecdysozoan, the scorpion *Mesobuthus gibbosus,* was identified in the analysis. All sequences, with the exception of the gastropod limpets *Lottia goshimai* and *Lottia gigantea*, share a conserved fibrinogen-related domain (FReD) towards the C-terminal end of the proteins (Additional file 5). This newly described FReD peptide was thus present in the last common protostomian ancestor.

2.2.3.6 Lophotrochozoa-specific pNPs and peptide prohormones

Fourteen lophotrochozoan-specific peptide families were identified (Figure 2). These include families previously restricted to individual molluscan classes (ie. Gastropoda and/or Bivalvia) such as the four repetitive peptide families LFRFamide, PRQFVamide, feeding circuit-activating peptide (FCAP), *Mytilus* inhibitory peptides (MIP), and the D-amino acid containing peptide family NdWFamide.

Precursors of the lophotrochozoan neuropeptide KY (NKY) form two distinct well-defined clusters of divergent proneuropeptide subgroups, NKY-1 and NKY-2. NKY family members are present in all eight class-level taxa of Mollusca, as well as in annelids and nemerteans. Multiple sequence alignments confirmed the presence of the conserved diagnostic lysine (Lys:K) and tyrosine (Tyr;Y) residues at the N- and C-terminal ends of these sequences. Conversely, the central region of the two NKY precursors, NKY-1 and NKY-2, differ considerably, being represented by FW[RQ]P[LM]G[YG] and G[YF]WIWMPAQG consensus peptide sequences, respectively.

The feeding circuit-activating peptide (FCAP) was identified in six of the eight molluscan class-level taxa and in all rotiferan taxa analysed here (*Rotaria tardigrada, Rotaria socialis,* and *Rotaria. sordida*) (Figure 2; Additional file 10). The molluscan pNPs contain multiple FCAP copies ranging from six in the pulmonate slug *Deroceras reticulatum* to up to 28 in the limpet *Lottia gigantea* (Additional file 10). The molluscan FCAP-bioactive peptides are usually 13 amino acids long; however,

differences in their length were observed (Additional file 10). The rotiferan FCAP-related bioactive peptides are shorter (with a fixed length of 11 amino acids) than their molluscan counterparts and are present in eight copies in the rotiferans *R. sordida* and *R. socialis* and in nine copies in *R. tardigrada* (Additional file 10). All lophotrochozoan FCAP-bioactive peptides are composed of related sequences that show species-specific variability towards the N-terminal region (Additional file 10).

2.2.3.7 Mollusca-specific pNPs and peptide prohormones

Seven peptide families with a distribution restricted to mollusks were recovered in the analysis (Figure 2). These include two well-known gastropod peptide families, abdominal ganglion (R3-14) and enterin, while the more widely distributed [A]PGWamides were found in all conchiferans except Monoplacophpora. Pleurins were recovered from bivalves, gastropods, and neomeniomorphs. Two pNP families composed of short potential bioactive peptides, referred to as LASGLI- and PSGYVRIamide, were identified in the bivalve *Dreissena rostriformis* and in the aplacophorans *Wirenia argentea* and *Gymnomenia pellucida*. A small peripheral group connected to the central cluster (Figure 1B) composed solely of conchiferan pNP sequences was recovered (Figure 2). The members of this peptide family showed no significant similarity against any known neuropeptide sequences available in the nr-database and thus likely represent an independent and divergent pNP family that evolved from sequences present in the central cluster (Figure 1A). All the sequences in this pNP family possess four conserved cysteine residues that are likely to give rise to two intramolecular disulfide bridges (Additional file 5). They also possess and a conserved P[FM]R[WY] protein motif, with the exception of two sequences belonging to the bivalve *Dreissena rostriformis.* According with conventions for pNP annotation, we name this conchiferan pNP family PXRX.

2.2.4 Discussion

2.2.4.1 Development of an in silico pipeline for proneuropeptide and peptide prohormone identification in Lophotrochozoa

No single best method has yet been established for the identification and retrieval of pNP and peptide prohormone sequences from genomic or transcriptomic databases. In 2013, two independent studies laid the framework for large-scale pNP and prohormone identification in metazoans (Jékely, 2013; Mirabeau & Joly, 2013) and subsequent studies have employed modified versions of these *in silico* pipelines (Conzelmann et al., 2013; Zatylny-Gaudin et al., 2016). Herein, we present an updated bioinformatic pipeline for pNP and peptide hormone identification and annotation (Figure 6) which has resulted in the identification and phylogenetic classification of hundreds of new pNPs and peptide hormones. Our greatly expanded but conservative new estimates of the pNP and peptide hormone complements of the eight molluscan class-level taxa are testament to the robustness of this pipeline.

It is difficult to state decisively that any particular peptide family is absent from the molecular databases analysed in our study. Methodological biases introduced during the data production and assembly steps, in addition to the meticulous avoidance of false positives, dictated by the stringency of the parameter settings used by the bioinformatics tools in the pipeline (e.g., signalP, hmmsearch, blastp), may have hindered the identification of some sequences. Some issues with the identification of molluscan and lophotrochozoan FMRFamide precursors may serve as example. FMRFamide peptides (Phe-Met-Arg-Phe-NH2) constitute one of the most well-known neuropeptide families studied in Mollusca since their discovery as a cardioacceleratory peptide (Price & Greenberg, 1977). They have since been identified in seven of the eight class-level taxa of mollusks (López-Vera et al., 2008; Faller et al., 2012; Redl et al., 2014), with the exception of Monoplacophora. However, the pipeline described herein failed to retrieve these sequences from the investigated databases. A careful inspection showed that FMRFamide pNP sequences belonging to all eight molluscan class-level taxa and other lophotrochozoan phyla (Brachiopoda and Nemertea), were later filtered out during the hmmsearch step in which those sequences with matches to any member of either the PfamA or PfamB database were removed (Figure 6B: “Removal of known folded protein domains”). This specific step was added in the pipeline in order to remove non-neuropeptide folded protein domain-containing sequences (with few exceptions, e.g., insulin-like domains). Manual curation of the resulting candidates revealed that matches against the pfamB model PF01581 (“FMRFamide-related peptide family”) had removed FMRFamide-related peptides from this list.

It is important to note that the aforementioned limitations are not solely restricted to this particularly work but are also present in other studies concerning pNP and peptide hormone identification. To elucidate the complete pNP and peptide prohormone repertoire of metazoans, rigorous manual inspection, and techniques such as mass-spectrometry, represent powerful tools to complement *in silico* automated bioinformatic screenings (Collins et al., 2010; Hauser et al., 2010; Xie et al., 2010; Diercksen et al., 2011; Conzelmann et al., 2013; Zatylny-Gaudin et al., 2016). Additionally, as suggested by Veenstra, 2010, the identification and characterisation of G protein-coupled receptors (GPCRs) is useful approach to fully understand the evolutionary history of a given peptide family, given the long-term coevolution of receptor-ligand pairs (Moyle et al., 1994; Park et al., 2002; Jékely, 2013; Mirabeau & Joly, 2013).

2.2.4.2 Mollusks as important models for clarifying the evolution and diversification of neuropeptide and peptide hormone families in metazoans

Studies focusing on the lophotrochozoan proneuropeptide and peptide hormone complement are still restricted to a few mollusks (Veenstra, 2010; Stewart et al., 2014; Adamson et al., 2015; Zatylny-Gaudin et al., 2016; Ahn et al., 2017; Zhang et al., 2018), flatworms (Collins et al., 2010), and annelids (Veenstra 2011; Conzelmann et al., 2013). To fill this gap of knowledge, molecular databases for the different eight class-level taxa of Mollusca as well as nine major additional lophotrochozoan phyla were mined for the presence of pNP and prohormone sequences.

The wide taxon sampling spanning the extant diversity of lophotrochozoan phyla showed that many peptide families that had previously only been known from annelids and mollusks (e.g. NKY, FXRI, LXRX, CLCCY) (Conzelmann et al., 2013) have orthologs in other lophotrochozoan phyla, rendering them *bona fide* lophotrochozoan families (i.e. peptide families that emerged at the base of Lophotrochozoa). Moreover, peptide families that were hitherto only known from mollusks are shown here to be widespread in other lophotrochozoans, such as the LFRFamide, PRQFVamide, NdWFamide, feeding circuit-activating peptide (FCAP), and *Mytilus* inhibitory peptide (MIP) families (Figure 2).

A few gene expression studies involving two of the aforementioned families, MIP and LFRFamide, have been performed in conchiferan mollusks (Hirata et al., 1988;

Kuroki et al., 1993; Fujisawa et al., 1999; Hoek et al., 2005; Zatylny-Gaudin et al., 2010; Bigot et al., 2014). Comparative physiological investigations involving MIPs in the bivalves *Mytilus edulis* and *Meretrix lusoria* (Hirata et al., 1988) as well as in the gastropods *Achatina fulica*, *Aplysia californica,* and *Aplysia kurodai* (Fujisawa et al., 1999) showed a strong inhibitory impact of these peptides on the contraction of different muscles in these animals. Regarding the LFRFamide peptides, a different scenario was revealed. In gastropods, LFRFamide peptides had an inhibitory activity on F2 neurons (Kuroki et al., 1993) as well as on the control of the feeding and reproduction behaviour in the snail *Lymnaea stagnalis* during schistosomiasis infections (Hoek et al., 2005). In the oyster *Crassostrea gigas* (Bigot et al., 2014) and in the squid *Sepia officinalis* (Zatylny-Gaudin et al., 2010) LFRFamide peptides are involved in energy metabolism and in the tonus and amplitude of rectal contraction.

Taken together, our results point to a pNP and peptide prohormone repertoire with evolutionary origins in Lophotrochozoa that consists of a minimum of 15 families, thus expanding the complement of ten previously identified families (Conzelmann et al., 2013). This finding, in combination with gene expression and functional studies, will enable testing of putative functions of these peptides in a broad range of lophotrochozoan taxa.

As a result of different evolutionary constraints (Martínez-Pérez et al., 2007; Wegener & Gorbashov 2008) and patterns of domain repetition and sequence divergence (i.e. little sequence similarity shared by related peptides from different phyla), the clustering approach is a robust method to elucidate and propose new evolutionary scenarios for pNPs and hormones. In addition, traditional phylogenetic reconstruction methods (e.g. maximum likelihood, Bayesian) become prohibitive and prone to errors when thousands of sequences are simultaneously analysed (Frickey & Lupas, 2004). The establishment of a close evolutionary relationship between insect prohormone-2 and the lophotrochozoan GNXQN family exemplifies how this method can expose the interconnectedness of peptide families that were previously unknown to be related. Prohormone-2 was fully characterised by Hummon et al. (2006) as a NVPIYQEPRF-containing neuropeptide in many insects, whereas GNXQN pNPs were first described in annelids, bivalves, and gastropods (Conzelmann et al., 2013). Our analysis not only expands the known phylogenetic distribution of the GNXQN pNPs to the remaining molluscan class-level taxa, with the exception of Scaphopoda, but also

indicates a homologous relationship between these two families. This points to a deeper origin of these families back to the last common protostomian ancestor.

A closer look into the prohormone-2/GNXQN cluster shows that insect prohormones are directly linked to some molluscan sequences. Likewise, the annelid sequences are also linked with the mollusks; however no direct link exists between the annelids and insects. Had such an analysis been conducted without the inclusion of the molluscan prohormone-2/GNXQN orthologs, no link would have been observed to indicate a relationship between prohormone-2 in insects and GNXQN in annelids (Figure 1A; Additional file 3). This result exemplifies the importance of broad taxonomic sampling when annotating pNPs and peptide hormones.

Our analysis revealed the presence of dickkopf (dkk) sequences in lophotrochozoan and ecdysozoan representatives, which had hitherto been considered lost in the protostome lineage (Niehrs, 2006). Dkks constitute a family that plays an important and ancient role in animal development by antagonising canonical Wnt signaling by competing with the Wnt-Frizzled complex for binding to the LRP receptors (Niehrs, 1999; Mao et al., 2001; Semënov et al., 2001, Augustin et al., 2006). Our analysis indicates the presence of two dkk genes in the last common ancestor of cnidarians and bilaterians, *dkk1/2/4* and *dkk3*, in which the first gave rise to the vertebrate *dkk1*, *dkk2*, and *dkk4* paralogs via gene duplication (Guder et al., 2006). *In silico* data mining of genomic and transcriptomic databases of model organisms, such as *Drosophila melanogaster* and *Caenorhabditis elegans*, have so far failed to recover any dkk orthologs within Protostomia (Niehrs, 2006). However, we found ecdysozoan and lophotrochozoan *dkk1/2/4* orthologs retrieved from nematodes, mollusks, ectoprocts, entoprocts, and nemerteans, which contain the two diagnostic cysteine-rich domains. These results demonstrate that the *dkk 1/2/4* ortholog was already present in the last common protostomian ancestor, while its paralog *dkk3* was secondarily lost in ecdysozoans and lophotrochozoans. Whether or not the Wnt-Dickkopf antagonism was functionally maintained in Ecdysozoa and Lophotrochozoa is yet to be demonstrated.

Since its discovery and isolation from the marine gastropod *Aplysia californica* (Strumwasser et al., 1969; Arch, 1976), the egg-laying hormone (ELH) has been subject to a number of studies focused on the molecular and neurophysiological mechanisms that dictate complex animal behaviour. When released into the hemocoel of a sexually mature gastropod, a series of behaviours are triggered (e.g.,

cessation of locomotion, inhibition of feeding, head movements), resulting in the extrusion of the egg mass (Arch & Smock, 1977). While ELH was initially only known from gastropod mollusks (Ebberink et al., 1985; Li et al., 1999; Veenstra, 2010), recent studies have confirmed its presence in many species of bivalves (Matsumoto et al., 2013; Stewart et al., 2014). Furthermore, ELH has been shown to be a homolog of the deuterostome corticotropin-releasing hormone (CRH) and the ecdysozoan and lophotrochozoan diuretic hormone 44 (Conzelmann et al., 2013; Mirabeau & Joly, 2013).

Phylogenetic analyses using the bioactive ELH domains showed that all molluscan sequences formed a unique clade (Stewart et al., 2014). Our results, using the bioactive ELH/DH44/CRH domain and its N-terminal flanking region, show that all gastropod ELH sequences form an independent and lineage-specific clade as sister group to the remaining molluscan and lophotrochozoan DH44 sequences. Interestingly, no ELH/DH44 sequences were retrieved from any cephalopod databases, including the predicted proteins from the *Octopus bimaculoides* genome (Albertin et al., 2015). These results are in agreement with another recent study that failed to retrieve any ELH/DH44 orthologs in transcriptomes built from the central nervous system of cuttlefish sampled during spawning (Zatylny-Gaudin et al., 2016). Furthermore, similarity searches using the genome of *Euprymna scolopes* confirmed this same scenario in the squid, pointing to a likely loss of the ELH/DH44 in the Cephalopoda lineage (H. Schmidbaur & O. Simakov, personal communication, May, 2018). It is difficult to assess whether this evolutionary scenario is underlain by changes in the role of these genes in annelids, nemerteans, and the different molluscan class-level taxa, since no comparative functional studies with DH44/ELH hormones have been reported outside of Gastropoda. In the ecdysozoan *Drosophila*, DH44 is involved in water regulation, excretion by the use of Malpighian tubules, and detection and consumption of nutritive sugars, but not in reproductive behaviour (Cabrero et al., 2002; Dus et al., 2015; Cannell et al., 2016). However, immunolocalization studies demonstrated that ELH-like peptides might play a role in the spawning processes of other ecdysozoans, e.g., decapod crustaceans (Liu et al., 2006; Ngernsoungnern et al., 2009).

2.2.4.3 Evolution of the neuropeptide and hormone complement within Mollusca

Mollusks show a huge diversity of body plans and nervous system complexity (Haszprunar & Wanninger; 2012; Hochner & Glanzman 2016). Consequently, neuropeptide and peptide prohormone toolkits in different mollusks constitute a valuable resource to elucidate the molecular mechanisms that control their development, growth, reproduction, and physiology. Comparative studies within Mollusca are still rare, and the few thorough analyses focusing on the pNP and prohormone complement are almost exclusively focused on individual gastropod and bivalve species (Veenstra 2010; Stewart et al., 2014; Adamson et al., 2015; Ahn et al., 2017; Bose et al., 2017; Zhang et al., 2018). Our work provides the repertoire of pNP and peptide prohormone signaling molecules for all the eight extant class-level taxa of Mollusca, including the understudied aculiferans, monoplacophorans, and scaphopods. The minimum class-level pNP/prohormone complement ranges from 28 to up to 59 in conchiferans and from 33 to 50 in aculiferans. The analyses revealed an unexpected conservation in the toolkit of pNPs and hormones within the phylum, regardless of the complexity of the nervous system and life styles of the respective protagonists (e.g., highly mobile predators versus slowly moving or sessile filter-feeders). FMRFamide, allatostatin-A and -B, NKY, pedal-peptides, and luqin are families all shared by all molluscan class-level taxa. Additionally, the peptide families retrieved from the molluscan databases show homology to virtually all described eumetazoan, bilaterian, protostomian, and lophotrochozoan families (Jékely, 2013; Mirabeau & Joly, 2013; Conzelmann et al., 2013) and only in a few cases lineage-specific innovations in the peptide complement was observed (Figure 2).

In some cases, peptide families were restricted a limited number of molluscan class-level taxa. This is the case for the dickkopf (Figure 4) and DH31 families. Previous studies claimed the secondary loss of DH31 in mollusks and the loss of dickkopf in protostomes (Conzelmann et al., 2013). Our results, however, show the presence of DH31 in a polyplacophoran (Figure 2; Additional file 3 and 5) and dickkopf in lophotrochozoans and at least one nematode (*Trichinella spiralis*). These findings demonstrate the importance of comparative analyses and broad taxon sampling in order to clarify the evolution of peptide families in metazoans.

Comparative studies suggest that regulatory gene families (i.e. protocadherins and C2H2s), post-transcriptional mechanisms (i.e. RNA-editing), genome rearrangements, and extensive transposable element activity are major forces behind the behavioural repertoire (e.g., camouflage displays, problem solving, and

observational learning) and the complex central nervous system (CNS) in cephalopods (Albertin et al., 2015; Liscovitch-Brauer et al., 2017). Our analysis suggests that the evolution of the complex CNS and the sophisticated behavioural repertoire of cephalopods was not paralleled by lineage-specific expansions of pNP or peptide hormone families. Although our homology-based approach for pNPs/peptide prohormone identification might have failed to identify particularly divergent homologs and lineage-specific peptide families, low number of peptide families (38 in total) identified using mass-spectrometry on the CNS of cuttlefish (*Sepia officinalis*) further corroborates our conclusions (Zatylny-Gaudin et al., 2016).
The neuropeptide/hormone complement described here shows considerable overlap with the results of previous works on gastropods, bivalves, and cephalopods. However, several peptide families (generally short amidated bioactive peptides) with either a broad (e.g., PXXXamide, Samide, and SPamide families; Zatylny-Gaudin et al., 2016) or a highly restricted phyletic distribution (e.g., CCFRamide; Conzelmann et al., 2013), even down to the species level (e.g., the scallop-specific GNamide, LRYamide, and Vamide families; Zhang et al., 2018) were not recovered in our study. It is therefore important to stress that the peptide families recovered in our study must not be regarded exhaustive, but rather as the minimum peptide complement present in the major class-level taxa of Mollusca.

2.2.5 Conclusions

The phylum Mollusca comprises more than 200,000 extant species and harbors a plethora of distinct body plans, neural architectures, and forms of behaviour. Through a comparative and integrative approach using *in silico* protocols and homology-based clustering, a detailed overview of the minimum proneuropeptide/hormone complement of all extant class-level taxa of Mollusca was obtained. Our study provides a high-quality, manually curated catalog containing multiple sequence alignments and peptide logos for 59 metazoan neuropeptide/peptide hormone families. We identified several new families such as PXRX and fibrinogen-related, expanded the phyletic distribution of others (e.g., neuroparsin, DH31), and established the homology of seemingly unrelated peptides (e.g., GNXQN and prohormone-2). We show for the first time the presence of a *dkk-1/2/4* ortholog gene in protostomes, whereby the lophotrochozoan and ecdysozoan sequences possess

the two diagnostic cysteine-rich dickkopf and colipase domains. ELH peptides are lineage-specific to gastropods but are closely related to their lophotrochozoan and non-gastropod molluscan orthologs, the diuretic hormone 44. Our results suggest that the complex nervous system and the extraordinary behavioural repertoire of cephalopods are not correlated with innovations of the downstream signaling elements (i.e. neuropeptides and hormones). Our pioneering study provides an important stepping stone towards a better understanding of the function and evolution of these conserved peptides not only in mollusks, but also in a wide range of other metazoans.

2.2.6 Material and Methods

2.2.6.1 Data collection, filtering, sequence reconstruction, proteome prediction, and completeness assessment

In order to identify as many molluscan and lophotrochozoan pNP groups as possible, several transcriptomes belonging to different class-level taxa of Mollusca and other lophotrochozoan phyla were downloaded from Sequence Read Archive database (www.ncbi.nlm.nih.gov/sra) and combined with molluscan transcriptomes generated by our group as described earlier (De Oliveira et al., 2016). Predicted coding sequence regions from genomic data were downloaded and included where available. The summary concerning the species, phyla, SRA accession numbers, and the file transfer protocol addresses (FTP) of the molecular data used in this study are available in Additional file 11.

The Illumina datasets retrieved from SRA were subject to a cleaning procedure (identification of adapters, poor quality regions) using trimmomatic (Bolger et al., 2014) and were reconstructed with IDBA-tran (Peng et al., 2013) using the parameters –max_isoforms and -step defined as 1 and 5, respectively. The 454 databases were reconstructed using successive rounds of assembly with MIRA4 (Chevreux et al., 1999) and CAP3 (Huang & Madan, 1999) programs using default parameters. The prediction of coding sequence regions from the reconstructed transcriptomes was performed with TransDecoder (http://transdecoder.github.io/) and only the longest coding sequence region of each reconstructed transcript was retained for the subsequent analyses (Figure 6A). The completeness of the individual

proteomes was assessed with BUSCO (Simão et al., 2015) with the default parameters using the pre-defined 978 metazoan Benchmarking set of Universal Single-Copy Orthologs. The proteomes were classified into BUSCO metrics as follows: complete, duplicated, fragmented, and missing.

2.2.6.2 Identification of molluscan and other lophotrochozoan pNPs

To date there is no publicly available program or script to perform a direct identification of pNPs in transcriptomic or proteomic datasets. To circumvent this limitation, a pipeline comprising several distinct bioinformatic strategies was implemented and executed based on the previous works of Jékely (2013) and Conzelmann et al. (2013). All the major steps are described in details below (Figure 6B).

2.2.6.3 Identification of signal peptide cleavage sites and establishment of the secretome databases

The initial identification of potential new lophotrochozoan pNPs was started with the identification of the signal peptide cleavage site using the program signalP 4.0 (Petersen et al., 2011). The program was executed under the following parameters: -m -n -u 0.45 -U 0.50, in which the parameters –m and –n control the output files (i.e. fasta file with the mature protein sequence and a gff annotation file, respectively) and the parameters –u and –U define the cut-off scores used to predict and identify the signal peptide cleavage site. The protein sequences that failed to present a signal peptide cleavage site were discarded. All subsequent analyses were carried out using the mature protein sequences (i.e. the protein sequence without the N-terminal signal peptide) (Figure 6B).

2.2.6.4 Removal of known folded protein domains, search for repetitive motifs, and establishment of the neuropeptidome databases

To avoid false positive results two distinct approaches were implemented: (1) identification and exclusion of sequences with known folded protein domains using the program hmmsearch (Eddy, 1998); (2) the identification of repetitive motifs

(cleavage sites) using a local Perl script. The similarity searches using hmmsearch were executed using the mature lophotrochozoan protein sequences as queries and the PFAM-A and B database under default parameters and a defined e-value of 1e-10. The protein sequences without matches to the PFAM-A or B database were screened for repetitive cleavage sites motifs using the following Perl requlars expressions: (R|K)*GKR(R|K)*, (R|K)*GRK(R|K)*, (R|K)*GRR(R|K)*, (R|K)*GKK(R|K)*, (R|K)*KR(R|K)*, (R|K)*RK(R|K)*, (R|K)*RR(R|K)*, (R|K)*KK(R|K)*, (R|K)*GR(R|K)*, (R|K)*GK(R|K)*. All mature lophotrochozoan proteins with a known folded protein domain and/or lacking any of the aforementioned repetitive motifs were discarded (Figure 6B). Redundancy was removed from the neuropeptidomes using cd-hit (Fu et al., 2012) with the parameter –c defined as 0.95 (sequence identity threshold).

2.2.6.5. Similarity searches against a curated non-redundant dataset of 6,692 pNPs

To avoid unrelated spurious sequences, to optimise the subsequent analyses, and to decrease computational burden in the phylogenetic steps, similarity searches were carried out using the blastp alignment tool (Camacho et al., 2009). The predicted neuropeptidomes were used as BLAST queries against a well-curated database composed of 6,692 metazoan pNPs (Jékely 2013; Conzelmann & Jékely 2013; Adamson et al., 2015; Ahn et al., 2017) using a loose e-value of 1e-03. The protein sequences without any similarity against the pNP database were removed from the next step of the pipeline (Figure 6B).

2.2.6.6 Clustering, multiple sequence alignment, motif identification, and illustration of the biological sequences

The remaining lophotrochozoan pNPs and peptide hormones (i.e. full length proteins including signal peptide) were used as input for the program CLANS (Frickey & Lupas, 2004), a Java application for visualising protein families based on pairwise similarity, together with the curated dataset of 6,692. The input dataset was clustered during approximately 20,000 rounds using local psi-blast using the following parameters: -evalue 1e-06 matrix BLOSUM62 -num_iterations 3 (Figure 1A).

Metazoan pNPs that failed to connect to any molluscan pNPs were excluded from the map. The large and strongly connected cluster composed by repetitive peptide sequences at the center of the map (Figure 1B) was re-analysed using CLANS and a non-iterative blastp similarity tool with an evalue of 1e-06. To help the identification of the peptide families, clusters were identified with the function “find cluster: convex clustering” under the default parameters. To improve and aid the overall classification and phyletic distribution of the pNP and hormone families in each cluster, motif searches using meme (Bailey et al., 2009), multiple sequence alignments using mafft (Katoh & Stanley, 2013), and additional phylogenetic inferences using mrbayes (Ronquist et al., 2012) were employed. The diagram of the proteins was drawn using IBS software (Liu et al., 2015). Any isolated pNP cluster smaller than 3 sequences and without any recognisable conserved domain(s) were excluded from the map. The peptide families identified in molluscan and lophotrochozoan representatives were classified according to their evolutionary origins following criteria established by Conzelmann et al. (2013) to distinguish pNP families present in the last common ancestor (LCA) of eumetazoans, bilaterians, protostomians, and lophotrochozoans. Additionally, peptides with their evolutionary origins tracing back to the LCA of Mollusca and different class-level taxa were also identified and classified. The final 3D maps were collapsed to 2D after the clustering for easier visualisation (Figure 6C, Additional Files 3, 4).

2.2.6.7 Phylogenetic analysis

Multiple sequence alignment files for each family were generated with the program mafft under the following parameters --maxiterate 1000 –localpair. The trimming of the poorly aligned regions in order to increase the accuracy of the subsequent phylogenetic inferences was performed with trimal or BMGE (Capella-Gutiérrez et al., 2009; Criscuolo & Gribaldo, 2010). Phylogenetic analyses were performed with mrbayes using the appropriate best-fit model of amino acid substitution as determined by Akaike information criterion (AIC) implemented in prottest3 (Darriba et al., 2011). The number of generations used in each phylogenetic run was determined using a convergence diagnostic (i.e. the standard deviation of split frequencies). All the runs were performed using the samplefreq parameter defined as 1000 and a

relative burn-in of 25%. The final phylogenetic consensus tree was edited with Figtree (http://tree.bio.ed.ac.uk/software/figtree).

2.2.7 Abbreviations

AIC: Akaike information criterion; AKH: adipokinetic hormone; BLAST: basic local alignment tool; BMGE: block mapping and gathering with entropy; BUSCO: benchmarking set of universal single-copy orthologs; CCAP: crustacean cardioactive peptide; CLANS: cluster analysis of sequences; CRH: corticotropin-releasing hormone; CNS: central nervous system; DH31: diuretic hormone 31; DH44: diuretic hormone 44; dkk: dickkopf; ELH: egg-laying hormone; EP: neuropeptide EP; FCAP: feeding circuit-activating peptide; gnRH: gonadotropin releasing hormone; GPCR: G protein-coupled receptor; IRP: insulin-related peptide; L11: elevenin; MIP: *Mytilus* inhibitory peptides; LCA: last common ancestor; NCBI: National Center for Biotechnology Information; NKY: neuropeptide KY; NPF: neuropeptide F; NS: nervous system; PDF: pigment dispersing factor hormone; PFAM: protein family database; pNP: proneuropeptide; PTTH: prothoracicotropic hormone; R3-14: gastropod neuropeptide R3-14; sCAP: small cardioactive peptide.

2.2.8 Competing interests

The authors declare that they have no competing interests.

2.2.9 Authors' contribution

André Luiz de Oliveira (ALDO) and Andreas Wanninger (AW) designed the project. ALDO designed, implemented and executed the bioinformatics pipelines, performed the data analysis, and drafted the manuscript with input from AW. Andrew Calcino performed the pre-processing, filtering, and transcriptome assembly of *Dreissena rostriformis*. ALDO and AW jointly finalised the manuscript. All authors read, commented on, and approved the final version of the manuscript.

2.2.10 Acknowledgments

We thank Gaspar Jékely (Exeter) and Christoph Bleidorn (Madrid and Göttingen) for the constructive suggestions and help during the initial stages of this work. We also thank Hannah Schmidbaur and Oleg Simakov (both Vienna) for the searches of ELH/DH44 proneuropeptide sequences in the genome of the squid *Euprymna scolopes*.

2.2.11 Funding

This work was supported by the Brazilian programme “Science without Borders” (Ciência sem Fronteiras; Project Number 6090/13-3) to ALDO and by a grant of the Austrian Science Fund (FWF; Project Number: P29455-B29) to AW.

2.2.12 Data availability

All data generated and/or analysed during this study are included in this published article (and its supplementary information files).

2.2.13 References

Adamson KJ, *et al*. Molecular insights into land snail neuropeptides through transcriptome and comparative gene analysis. *BMC Genomics* **16**, 308, doi:10.1186/s12864-015-1510-8 (2015).

Ahn SJ, Martin R, Rao S & Choi MY. Neuropeptides predicted from the transcriptome analysis of the gray garden slug *Deroceras reticulatum*. *Peptides*, **93**, 51-65, doi:10.1016/j.peptides.2017.05.005 (2017).

Albertin CB, *et al*. The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*, **524**, 220-224, doi:10.1038/nature14668 (2015).

Arch S. Neuroendocrine Regulation of Egg Laying in *Aplysia californica. Amer Zool*, **16**, 167-175 (1976).

Arch S & Smock T. Egg-laying behavior in *Aplysia californica*. *Behav Biol*, **19**, 45-54 (1977).

Attenborough RM, Hayward DC, Kitahara MV, Miller DJ & Ball EE. A “neural” enzyme in nonbilaterian animals and algae: preneural origins for peptidylglycine alpha-amidating monooxygenase. *Mol Biol Evol,* **29**, 3095-3109 (2012).

Augustin R, *et al*. Dickkopf related genes are components of the positional value gradient in *Hydra*. *Dev Biol,* **296**, 62-70 (2006).

Bailey TL, *et al*. MEME suite: Tools for motif discovery and searching. *Nucleic Acids Res*, **37**, W202-8, doi:10.1093/nar/gkp335 (2009).

Berriman M, *et al*. The genome of the blood fluke *Schistosoma mansoni*. *Nature*, **460**, 352-358, doi:10.1038/nature08160 (2009).

Bigot L, *et al*. Functional characterization of a short neuropeptide F-related receptor in a lophotrochozoan, the mollusk *Crassostrea gigas*. *J Exp Biol,* **217**, 2974-2982, doi:10.1242/jeb.104067 (2014).

Bogdanov YD, Balaban PM, Poteryaev DA, Zakharov IS & Belyavsky AV. Putative neuropeptides and an EF-hand motif region are encoded by a novel gene expressed in the four giant interneurons of the terrestrial snail. *Neuroscience*, **85**, 637-647 (1998).

Bolger AM, Lohse M & Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114-2120, doi: 10.1093/bioinformatics/btu170 (2014).

Bose U, *et al*. Neuropeptides encoded within a neural transcriptome of the giant triton snail *Charonia tritonis*, a Crown-of-Thorns Starfish predator. *Peptides,* **98**, 3-14, doi:10.1016/j.peptides.2017.01.004 (2017).

Cabrero P, *et al*. The *Dh* gene of *Drosophila melanogaster* encodes a diuretic peptide that acts through cyclic AMP. *J Exp Biol*, **205**, 3799-3807(2002).

Camacho C, *et al*. BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421, doi:10.1186/1471-2105-10-421 (2009).

Cannell E, *et al*. The corticotropin-releasing factor-like diuretic hormone 44 (DH44) and kinin neuropeptides modulate desiccation and starvation tolerance in *Drosophila melanogaster*. *Peptides,* **80**, 96-107, doi:10.1016/j.peptides.2016.02.004 (2016).

Capella-Gutiérrez S, Silla-Martínez JM & Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972-1973, doi: 10.1093/bioinformatics/btp348 (2009).

Chevreux B, *et al*. Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. *Genome Res,* **14**, 1147-1159 (2004).

Collins JJ 3rd, *et al*. Genome-wide analyses reveal a role for peptide hormones in planarian germline development. *PLoS Biol*, **8**, e1000509, doi: 10.1371/journal.pbio.1000509 (2010).

Conzelmann M, *et al*. The neuropeptide complement of the marine annelid *Platynereis dumerilii*. *BMC genomics*, **14**, 906, doi:10.1186/1471-2164-14-906 (2013).

Criscuolo A & Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol,* **10**, 210, doi:10.1186/1471-2148-10-210 (2010).

Darriba D, Taboada GL, Doallo R & Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, **27**, 1164-1165, doi:10.1093/bioinformatics/btr088 (2011).

De Oliveira AL, *et al*. Comparative transcriptomics enlarges the toolkit of known developmental genes in mollusks. *BMC genomics*, **17**, 905 (2016).

Dircksen H, *et al*. Genomics, transcriptomics, and peptidomics of *Daphnia pulex* neuropeptides and protein hormones. *J Proteome Res*, **10**, 4478-4504, doi:10.1021/pr200284e (2011).

Douglass J, Civelli O & Herbert E. Polyprotein gene expression: generation of diversity of neuroendocrine peptides. *Annu Rev Biochem*, **53**, 665-715 (1984).

Dus M, *et al*. Nutrient Sensor in the Brain Directs the Action of the Brain-Gut Axis in *Drosophila*. *Neuron*, **87**, 139-151, doi:10.1016/j.neuron.2015.05.032 (2015).

Ebberink RH, van Loenhout H, Geraerts WPM & Joosse J. Purification and amino acid sequence of the ovulation neurohormone of *Lymnaea stagnalis*. *Proc Nat. Acad Sci U S A*, **82**, 7767-7771 (1985).

Eddy SR. Profile Hidden Markov Models. *Bioinformatics*, **14**, 755-763 (1998).

Eipper BA, Stoffers DA, Mains RE. The biosynthesis of neuropeptides: peptide alpha-amidation. *Annu Rev Neurosci,* **15**, 57-85 (1992).

Faller S, Rothe BH, Todt C, Schmidt-Rhaesa A & Loesel R. Comparative neuroanatomy of Caudofoveata, Solenogastres, Polyplacophora, and Scaphopoda (Mollusca) and its phylogenetic implications. *Zoomorphology*, **131**, 149-170, doi.org/10.1007/s00435-012-0150-7 (2012).

Fredriksson R, Lagerstrom MC, Lundin LG & Schiöth HB. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol Pharmacol,* **63**, 256-272 (2003).

Frickey T & Lupas AN. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**, 3702-3704 (2004).

Fu L, Niu B, Zhu Z, Wu S & Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150-3152, doi: 10.1093/bioinformatics/bts565 (2012).

Fujisawa Y, *et al*. The *Aplysia mytilus* inhibitory peptide-related peptides: identification, cloning, processing, distribution, and action. *J Neurosci*, **19**, 9618-9634 (1999).

Guder C, *et al*. An ancient Wnt-Dickkopf antagonism in *Hydra*. *Development*, **133**, 901-911 (2006).

Hartenstein V. The neuroendocrine system of invertebrates: A developmental and evolutionary perspective. *J Endocrinol*, **190**, 555-570 (2006).

Haszprunar G & Wanninger A. Molluscs. *Curr Biol*, **13**, R510-514, doi:10.1016/j.cub.2012.05.039 (2012).

Hauser F, *et al*. Genomics and peptidomics of neuropeptides and protein hormones present in the parasitic wasp *Nasonia vitripennis. J Proteome Res*, **9**, 5296-5310, doi:10.1021/pr100570j (2010).

Hewes RS & Taghert PH. Neuropeptides and neuropeptide receptors in the *Drosophila melanogaster* genome. *Genome Res*, **11**, 1126-1142 (2001).

Hirata T, *et al*. Structures and actions of *Mytilus* inhibitory peptides. *Biochem Biophys Res Commun*, **152**, 1376-1382 (1988).

Hochner B & Glanzman DL. Evolution of highly diverse forms of behavior in molluscs. *Curr Biol*, **26**, R965-971, doi:10.1016/j.cub.2016.08.047 (2016).

Hoek RM, *et al*. LFRFamides: a novel family of parasitation-induced –Rfamide neuropeptides that inhibit the activity of neuroendocrine cells in *Lymnaea stagnalis*. *J Neurochem*, **92**, 1073-1080 (2005).

Hook V, *et al*. Proteases for processing proneuropeptides into peptide neurotransmitters and hormones. *Annu Rev Pharmacol Toxicol*, **48**, 393-423, doi:10.1146/annurev.pharmtox.48.113006.094812 (2008).

Huang X & Madan A. CAP3: A DNA sequence assembly program. *Genome Res*, **9**, 868-877 (1999).

Hummon AB, *et al.* From the genome to the proteome: uncovering peptides in the *Apis* brain. *Science*, **314**, 647-649 (2006).

Jékely G. Global view of the evolution and diversity of metazoan neuropeptide signaling. *Proc Natl Acad Sci U S A*, **110**, 8702-8707, doi:10.1073/pnas.1221833110 (2013).

Katoh K & Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, **30**, 772-780, doi:10.1093/molbev/mst010 (2013).

Kuroki Y, *et al*. FMRFamide-related peptides isolated from the prosobranch mollusc *Fusinus ferrugineus*. *Acta Biol Hung*, **44**, 41-44 (1993).

Lemche H & Wingstrand KG. The anatomy of *Neopilina galatheae* Lemche, 1957. *Galathea Report*, **3**, 9-71 (1959).

Li L, *et al*. Egg-laying hormone peptides in the aplysiidae family. *J Exp Biol*, **202**, 2961-2973 (1999).

Liscovitch-Brauer N, *et al*. Trade-off between Transcriptome Plasticity and Genome Evolution in Cephalopods. *Cell*; **169**, 191-202, doi:10.1016/j.cell.2017.03.025 (2017).

Liu F, Baggerman G, Schoofs L & Wets G. The construction of a bioactive peptide database in Metazoa. *J Proteome Res*, **7**, 4119-4131, doi:10.1021/pr800037n (2008).

Liu W, *et al*. IBS: an illustrator for the presentation and visualization of biological sequences. *Bioinformatics*, **31**, 3359-3361, doi:10.1093/bioinformatics/btv362 (2015).

Liu Z, Sobhon P, Withyachumnarnkul B & Hanna P. Identification of a putative egg-laying hormone in neural and ovarian tissues of the black tiger shrimp, *Penaeus monodon*, using immunocytochemistry. *Invert Neurosci*, **6**, 41-46 (2006).

López-Vera E, Aguilar MB & Heimer de la Cotera EP. FMRFamide and related peptides in the phylum Mollusca. *Peptides*, **2**, 310-317, doi:10.1016/j.peptides.2007.09.025 (2008).

Margulies M, *et al*. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376-380 (2005).

Martínez-Pérez F, *et al*. Loss of DNA: a plausible molecular level explanation for crustacean neuropeptide gene evolution. *Peptides*; **28**, 76-82 (2007).

Matsumoto T, *et al*. Reproduction-related genes in the pearl oyster genome. *Zoolog Sci*, **30**, 826-850, doi:10.2108/zsj.30.826 (2013).

Mirabeau O & Joly JS. Molecular evolution of peptidergic signaling systems in bilaterians. *Proc Natl Acad Sci U S A*, **110**, e2028–2037, doi:10.1073/pnas.1219956110 (2013).

Mao B, *et al*. LDL-receptor-related protein 6 is a receptor for Dickkopf proteins. *Nature*, **411**, 321-325 (2001).

Moyle WR, *et al*. Co-evolution of ligand-receptor pairs. *Nature*, **368**, 251-255. (1994).

Nagle GT, *et al*. *Aplysia californica* neurons R3-R14: primary structure of the myoactive histidine-rich basic peptide and peptide I. *Peptides,* **10**, 849-857 (1989).

Nathoo AN, Moeller RA, Westlund BA & Hart AC. Identification of neuropeptide-like protein gene families in *Caenorhabditis elegans* and other species. *Proc Natl Acad Sci U S A*, **98**, 14000–14005 (2001).

Niehrs C. Head in the WNT: the molecular nature of Spemann's head organizer. *Trends Genet*, **15,** 314-319 (1999).

Niehrs C. Function and biological roles of the Dickkopf family of Wnt modulators. *Oncogene*, **25**, 7469-7481 (2006).

Ngernsoungnern P, *et al*. Abalone egg-laying hormone induces rapid ovarian maturation and early spawning of the giant freshwater prawn, *Macrobrachium rosenbergii*. *Aquaculture*, **296**, 143-149, doi:10.1016/j.aquaculture.2009.08.011 (2009).

Park Y, Kim YJ & Adams ME. Identification of G protein-coupled receptors for *Drosophila* PRXamide peptides, CCAP, corazonin, and AKH supports a theory of ligand-receptor coevolution. *Proc Natl Acad Sci U S A*; **99**, 11423-11428, (2002).

Peng Y, *et al*. IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics*, **29**, i326-334, doi:10.1093/bioinformatics/btt219 (2013).

Petersen TN, Brunak S, von Heijne G & Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods*, **8**, 785-786, doi:10.1038/nmeth.1701 (2011).

Price DA & Greenberg MJ. Structure of a molluscan cardioexcitatory neuropeptide. *Science*, **197**, 670-671 (1977).

Redl E, Scherholz M, Todt C, Wollesen T & Wanninger A. Development of the nervous system in Solenogastres (Mollusca) reveals putative ancestral spiralian features. *Evodevo*, **5**, 48, doi:10.1186/2041-9139-5-48 (2014).

Ronquist F, *et al*. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*, **61**, 539-542, doi:10.1093/sysbio/sys029 (2012).

Ronaghi M, Uhlén M & Nyrén P. A sequencing method based on real-time pyrophosphate. *Science*, **281**, 363-365 (1998).

Semënov MV, *et al*. Head inducer Dickkopf-1 is a ligand for Wnt coreceptor LRP6. *Curr Biol*, **11**, 951-961 (2001).

Shigeno S, Andrews PLR, Ponte G, Fiorito G. Cephalopod Brains: An Overview of Current Knowledge to Facilitate Comparison With Vertebrates. *Front Physiol*, 2018, **9**, 952, doi:10.3389/fphys.2018.00952 (2018).

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210-3212, doi:10.1093/bioinformatics/btv351 (2015).

Srivastava M, *et al*. The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature*, **466**, 720-726, doi:10.1038/nature09201 (2010).

Steiner DF. The proprotein convertases. *Curr Opin Chem Biol*, **2**, 31-39 (1998).

Stewart MJ, *et al*. Neuropeptides encoded by the genomes of the Akoya pearl oyster *Pinctata fucata* and Pacific oyster *Crassostrea gigas*: a bioinformatic and peptidomic survey. *BMC Genomics*, **15**, 840, doi: 10.1186/1471-2164-15-840 (2014).

Strumwasser F, Jacklet JW & Alvarez RB. A season rhythm in the neural extract induction of behavioural egg laying in *Aplysia*. Comp. *Biochem. Physiol,* **29**, 197-206 (1969).

Sumner-Rooney L & Sigwart JD. Do chitons have a brain? New evidence for diversity and complexity in the polyplacophoran central nervous system. *J. Morphol*, **279**, 936-949, doi: 10.1002/jmor.20823 (2018).

Todt C, Büchinger T & Wanninger A. The nervous system of the basal mollusk *Wirenia argentea* (Solenogastres): a study employing immunocytochemical and 3D reconstruction techniques. *Mar Biol Res*, **4**, 290-303 (2008).

Veenstra JA. Neurohormones and neuropeptides encoded by the genome of *Lottia gigantea*, with reference to other mollusks and insects. *Gen Comp Endocrinol*, **167**, 86-103, doi:10.1016/j.ygcen.2010.02.010 (2010).

Veenstra JA. Neuropeptide evolution: neurohormones and neuropeptides predicted from the genomes of *Capitella teleta* and *Helobdella robusta*. *Gen Comp Endocrinol*, **171**, 160-175, doi:10.1016/j.ygcen.2011.01.005 (2011)

Xie F, *et al*. The zebra finch neuropeptidome: prediction, detection and expression. *BMC Biol*, **8**, 28, doi:10.1186/1741-7007-8-28 (2010).

Wegener C & Gorbashov A. Molecular evolution of neuropeptides in the genus *Drosophila*. *Genome Biol*, **9**, R131, doi:10.1186/gb-2008-9-8-r131 (2008).

Whalan S, Webster NS & Negri AP. Crustose coralline algae and a cnidarian neuropeptide trigger larval settlement in two coral reef sponges. *PLoS One*, **7**, e30386, doi:10.1371/journal.pone.0030386 (2012).

Wingstrand KG. On the anatomy and relationships of Recent Monoplacophora. *Galathea Report*, **16**, 94 (1985).

Zatylny-Gaudin C, *et al*. Characterization of a novel LFRFamide neuropeptide in the cephalopod *Sepia officinalis*. *Peptides,* **31**, 207-214, doi:10.1016/j.peptides.2009.11.021 (2010).

Zatylny-Gaudin C, *et al*. Neuropeptidome of the cephalopod *Sepia officinalis*: Identification, tissue Mapping, and expression pattern of neuropeptides and neurohormones during egg laying. *J Proteome Res*, **15**, 48-67, doi:10.1021/acs.jproteome.5b00463 (2016).

Zhang M, *et al*. Identification and Characterization of Neuropeptides by Transcriptome and Proteome Analyses in a Bivalve Mollusc *Patinopecten yessoensis*. *Front Genet*, **9**, 197, doi:10.3389/fgene.2018.00197 (2018).

Table 1 – Summary of predicted peptide precursor genes dentified in gastropods, bivalves, and cephalopod mollusks.

Organism	Class-level taxa	Data source	# of peptide precursors	References
Lottia gigantea	Gastropoda	Genome	67	Veenstra, 2010
Theba pisana		Transcriptome	35	Adamson et al., 2015
Deroceras reticulatum		Transcriptome	65	Ahn et al., 2017
Charonia tritonis		Transcriptome	60	Bose et al., 2017
Pinctada fucata	Bivalvia	Genome and transcriptome	31	Stewart et al., 2014
Crassostrea gigas			44	
Patinopecten yessoensis			63	Zhang et al., 2018
Sepia officinalis	Cephalopoda	Transcriptome	55	Zatylny-Gaudin et al., 2016

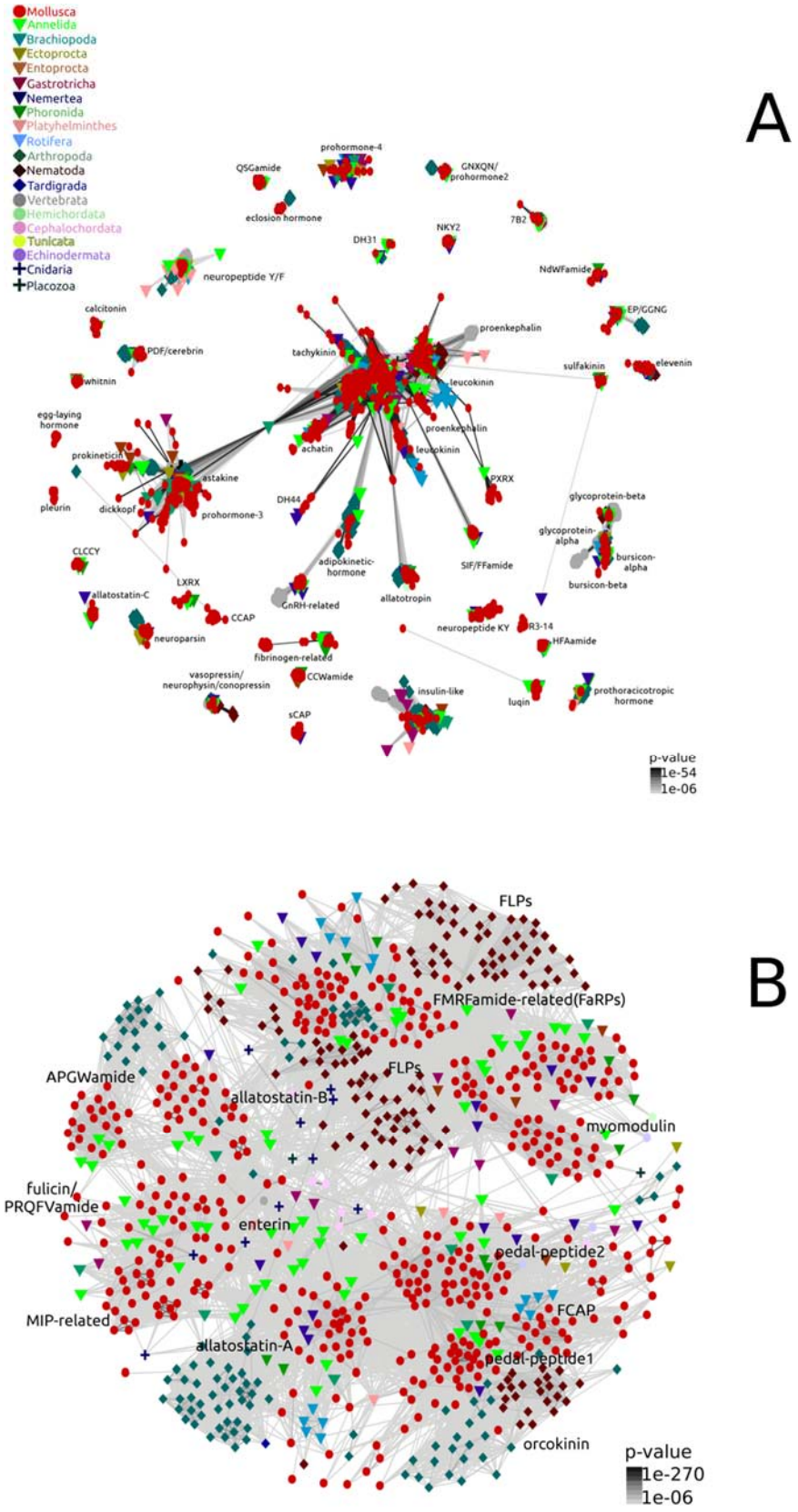


Figure 1: 2D cluster maps of molluscan and lophotrochozoan peptide families. Color and shape of nodes are based on the different phyla used in the analysis. (A) Psi-blast 2D cluster map (3 iterations) of molluscan and lophotrochozoan peptide

families. Edges correspond to psi-blast connections of P-value > 1e-06. (B) Non-iterative blastp 2D cluster map of repetitive peptide sequences. The central strongly connected cluster in the psi-blast 2D map (Fig. 1A) was reclustered with non-iterative blastp. The clusters were identified using convex-clustering, multiple sequence alignments, and motif identification. Edges correspond to blastp connections of P-value > 1e-06.

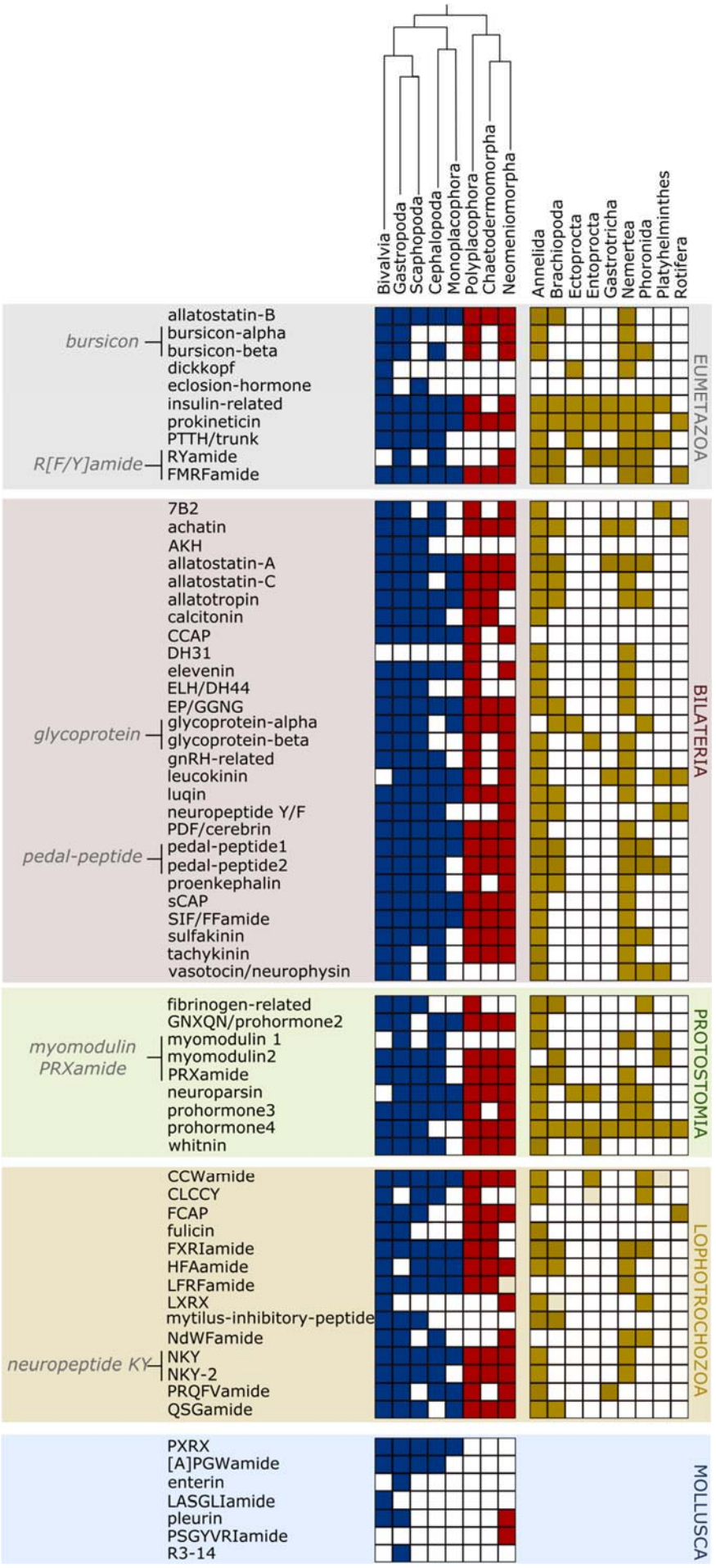

Bivalvia
Gastropoda
Scaphopoda
Cephalopoda
Monoplacophora
Polyplacophora
Chaetodermomorpha
Neomeniomorpha
Annelida
Brachiopoda
Ectoprocta
Entoprocta
Gastrotricha
Nemertea
Phoronida
Platyhelminthes
Rotifera
allatostatin-B
bursicon
bursicon-alpha
bursicon-beta
dickkopf
eclosion-hormone
insulin-related
prokineticin
PTTH/trunk
R[F/Y]amide
RYamide
FMRFamide
EUMETAZOA
7B2
achatin
AKH
allatostatin-A
allatostatin-C
allatotropin
calcitonin
CCAP
DH31
elevenin
ELH/DH44
EP/GGNG
glycoprotein
glycoprotein-alpha
glycoprotein-beta
gnRH-related
leucokinin
luqin
neuropeptide Y/F
PDF/cerebrin
pedal-peptide
pedal-peptide1
pedal-peptide2
proenkephalin
sCAP
SIF/FFamide
sulfakinin
tachykinin
vasotocin/neurophysin
BILATERIA
fibrinogen-related
GNXQN/prohormone2
myomodulin
PRXamide
myomodulin 1
myomodulin2
PRXamide
neuroparsin
prohormone3
prohormone4
whitnin
PROTOSTOMIA
CCWamide
CLCCY
FCAP
fulicin
FXRIamide
HFAamide
LFRFamide
LXRX
mytilus-inhibitory-peptide
NdWFamide
neuropeptide KY
NKY
NKY-2
PRQFVamide
QSGamide
LOPHOTROCHOZOA
PXRX
[A]PGWamide
enterin
LASGLIamide
pleurin
PSGYVRIamide
R3-14
MOLLUSCA

Figure 2: Minimum proneuropeptide/peptide prohormone complement of Mollusca. Peptide precursors were classified following criteria defined by Conzelmann et al. (2013), distinguishing peptide families present in the last common ancestor (LCA) of eumetazoans, bilaterians, protostomians, and lophotrochozoans. Peptide families present in the LCA of Mollusca and different molluscan class-level taxa are also displayed. The differently coloured boxes correspond to the presence of a given peptide family in conchiferan (blue), aculiferan (red), and lophotrochozoan (yellow) representatives.

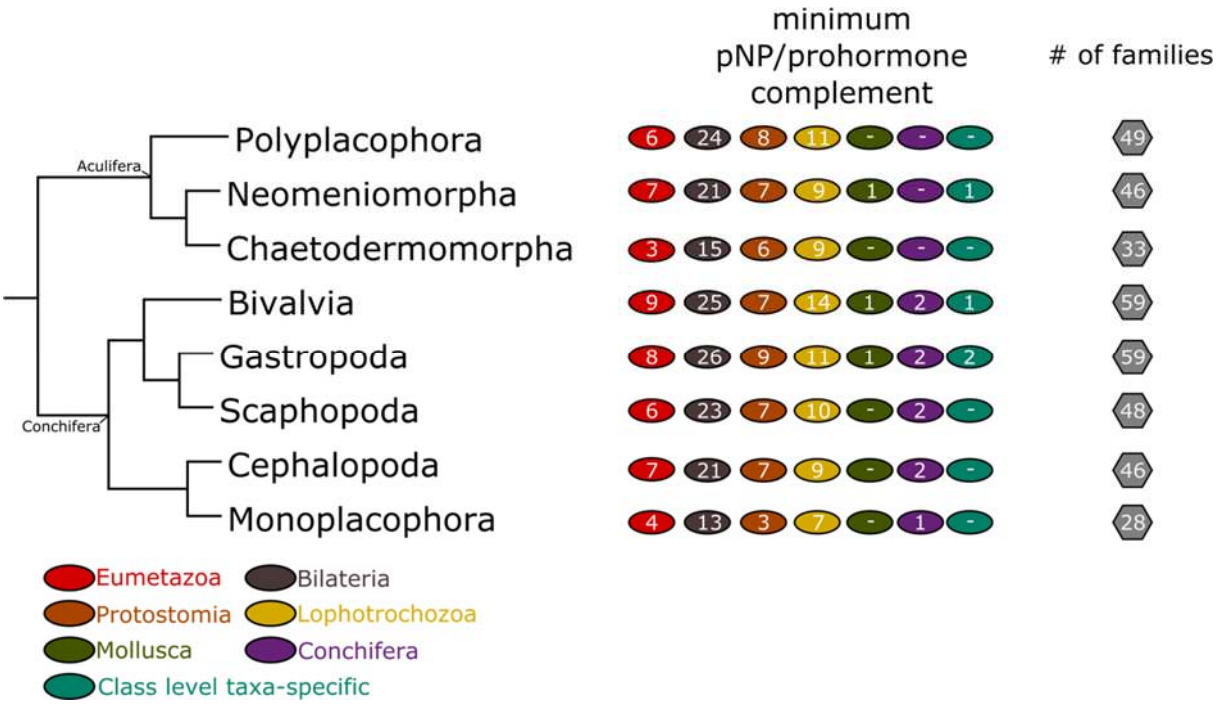


Figure 3: The distribution of the components of the pNP/peptide prohormone complement in molluscan class-level taxa using the currently widely accepted Conchifera/Aculifera hypothesis as a phylogenetic backbone. Coloured circles correspond to peptide families present in the last common ancestor of eumetazoans, bilaterians, protostomians, lophotrochozoans, mollusks, and, where present, conchiferans and specific class-level taxa, respectively. The numbers in the hexagons correspond to the minimum number of peptide families present in the last common ancestor of the various extant class-level taxa of Mollusca (in the right column).

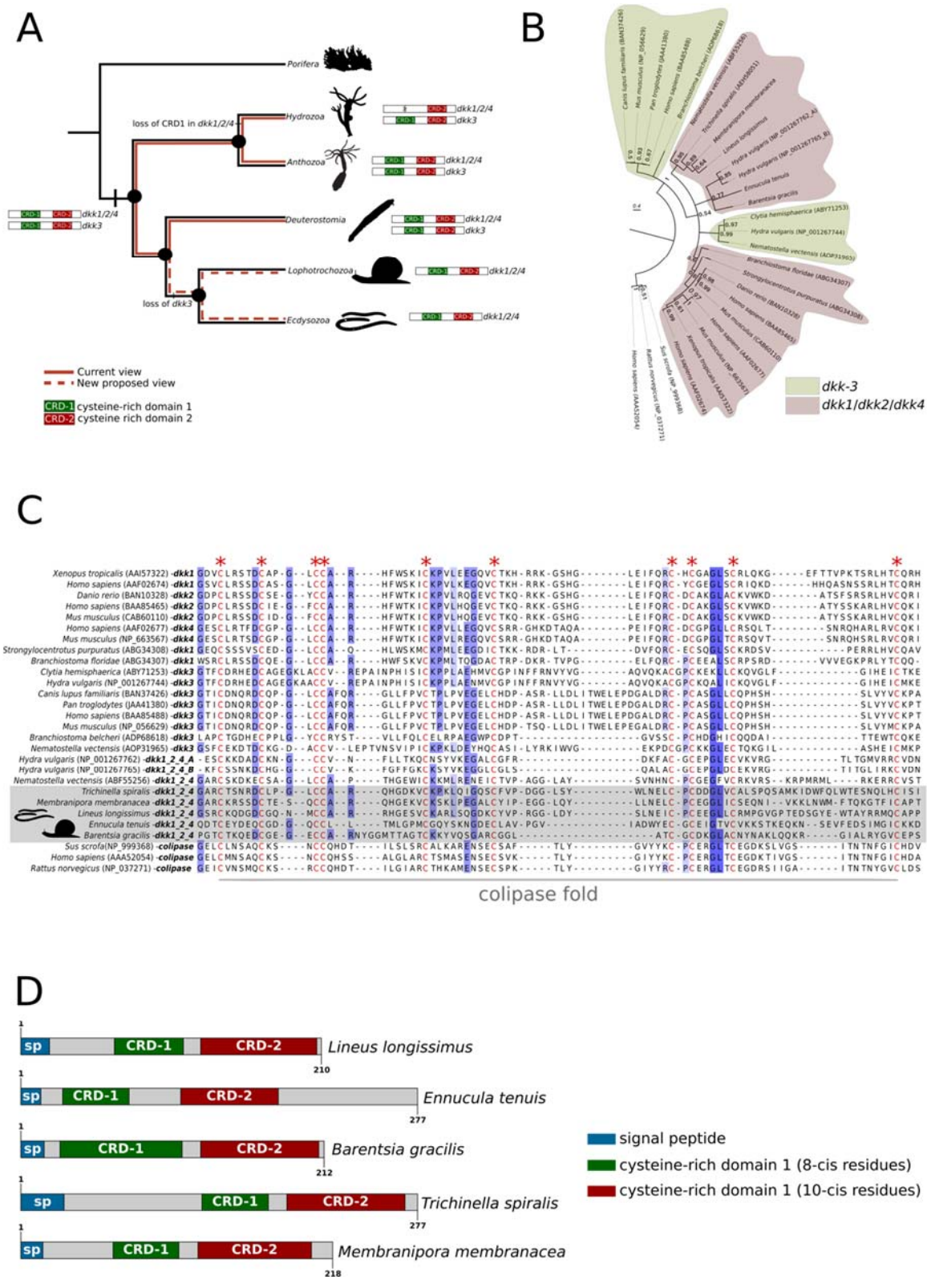


Figure 4: Evolution and distribution of dickkopf (dkk) proteins in Metazoa. (A) Traditional (red line) and novel (dotted red line) view of evolution and distribution of the *dkk 1/2/4* and *dkk 3* orthologs, highlighting the presence of *dkk-1/2/4* orthologs in protostomian animals. (B) Bayesian phylogenetic analysis of dkk proteins using the cysteine-rich domain-2 (colipase fold) found in lophotrochozoan and ecdysozoan

representatives. Sequences highlighted in bold correspond to the protostomian orthologs found in this study. NCBI accession numbers, when available, are displayed after the species names. The newly described dkk sequences are available in Additional file 3. Branch support values correspond to posterior probability values. Human, pig and rat colipases were used as outgroups. (C) Multiple sequencing alignment of the cysteine-rich domain-2 (colipase fold) showing the ten conserved cysteine residues as well as other conserved motifs in cnidarian, protostome, and deuterostome representatives. Lophotrochozoan and ecdysozoans orthologs are highlighted in the light gray box. (D) Domain structure of protostomian dkk sequences. Blue, green, and red boxes correspond to the signal peptide, cysteine-rich domain-1 (dkk domain), and cysteine-rich domain-2 (colipase fold), respectively. Animal silhouettes were obtained from www.phylopic.org and are either licensed under the Creative Commons Attribution 3.0 Unported or available under public domain (credited images used *Hydra*: Steven Traver; *Nematostella*: Jack Warner; *Branchiostoma*: Mali'o Kodis and Hans Hillewaert; Snail: Scott Hartman; *Trichinella*: Frank Förster).

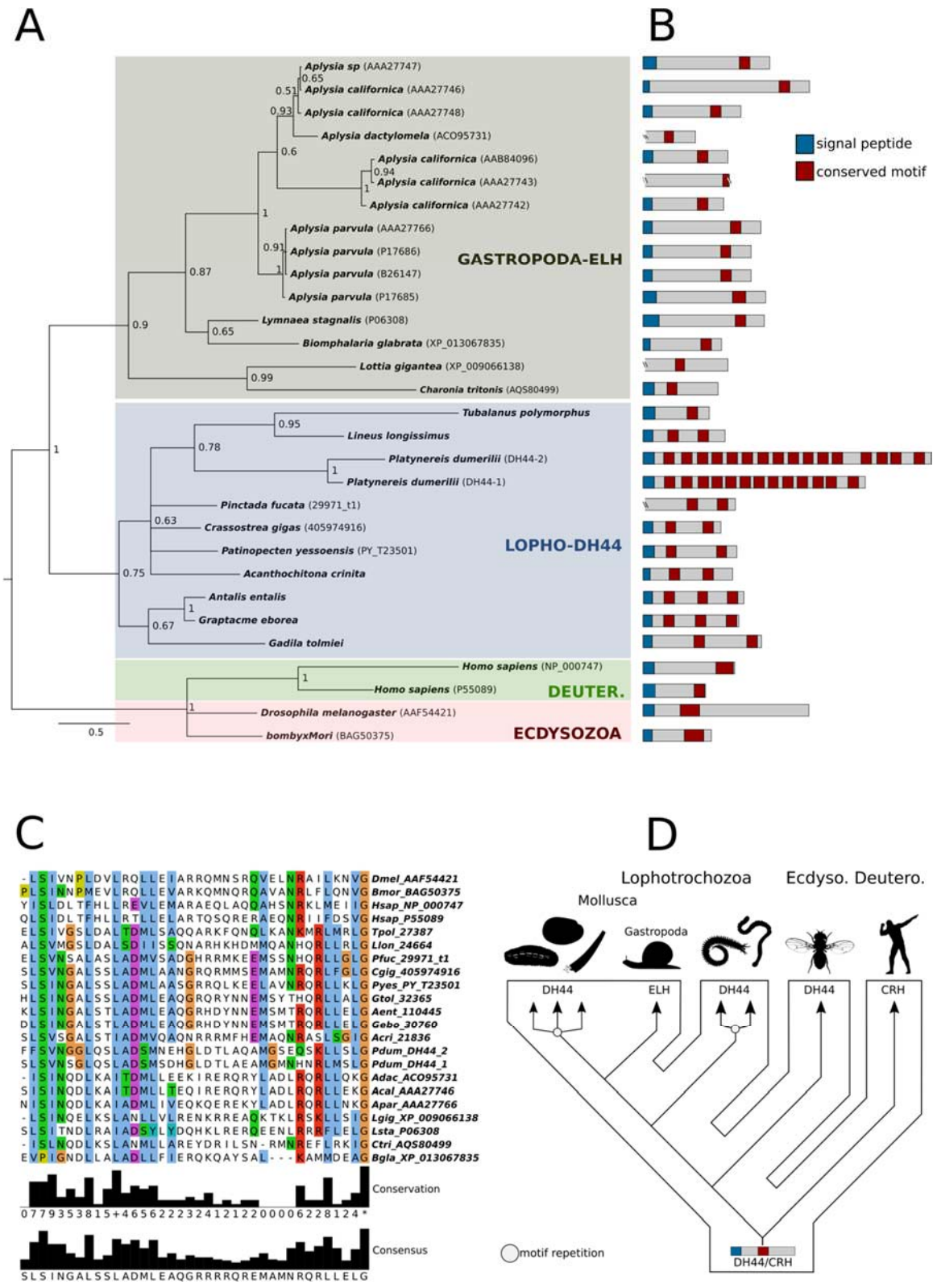


Figure 5: Evolution of DH44-ELH peptide hormone families in lophotrochozoans, ecdysozoans and deuterostomes. (A) Bayesian phylogenetic analysis using trimmed DH44/ELH protostomian sequences. Note the presence of three well-supported clusters: lophotrochozoan DH44, gastropod ELH, and ecdysozoan/deuterostome

DH44/CRH. NCBI accession numbers, where available, are displayed after the species names. The newly described dkk sequences are available in Additional file 3. Branch support values correspond to posterior probability values. (B) Domain structure of DH44 and ELH sequences showing the signal peptide (blue box) and shared conserved motifs corresponding to the predicted amidated peptides (red boxes). (C) Multiple sequence alignment of ELH, DH44, and CRH bioactive domains in metazoans. Species names are abbreviated for convenience (e.g. *Drosophila melanogaster* = Dmel; *Bombyx mori* = Bmor). The conservation histogram corresponds to the number of conserved amino acid physico-chemical properties for each column of the alignment. The consensus displayed below the alignment is the percentage of the modal residue per column including gaps. (D) New evolutionary scenario of ELH/DH44 prohormone sequences within Mollusca. White circles correspond to the presence of motif repetitions within the precursor sequences. Animal silhouettes were obtained from www.phylopic.org and are either licensed under the Creative Commons Attribution 3.0 Unported or available under public domain (credited images used Scaphopoda: Brockhaus and Efron; Polyplacophora: Noah Schlottman and Casey Dunn; Bivalvia and Gastropoda: Scott Hartman; *Platynereis*: B. Duygu Özpolat; *Tubulanus*: Mali'o Kodis and Rebecca Ritger; *Drosophila*: Thomas Hegna and Nicolas Gompel; *Homo sapiens*: David Orr).

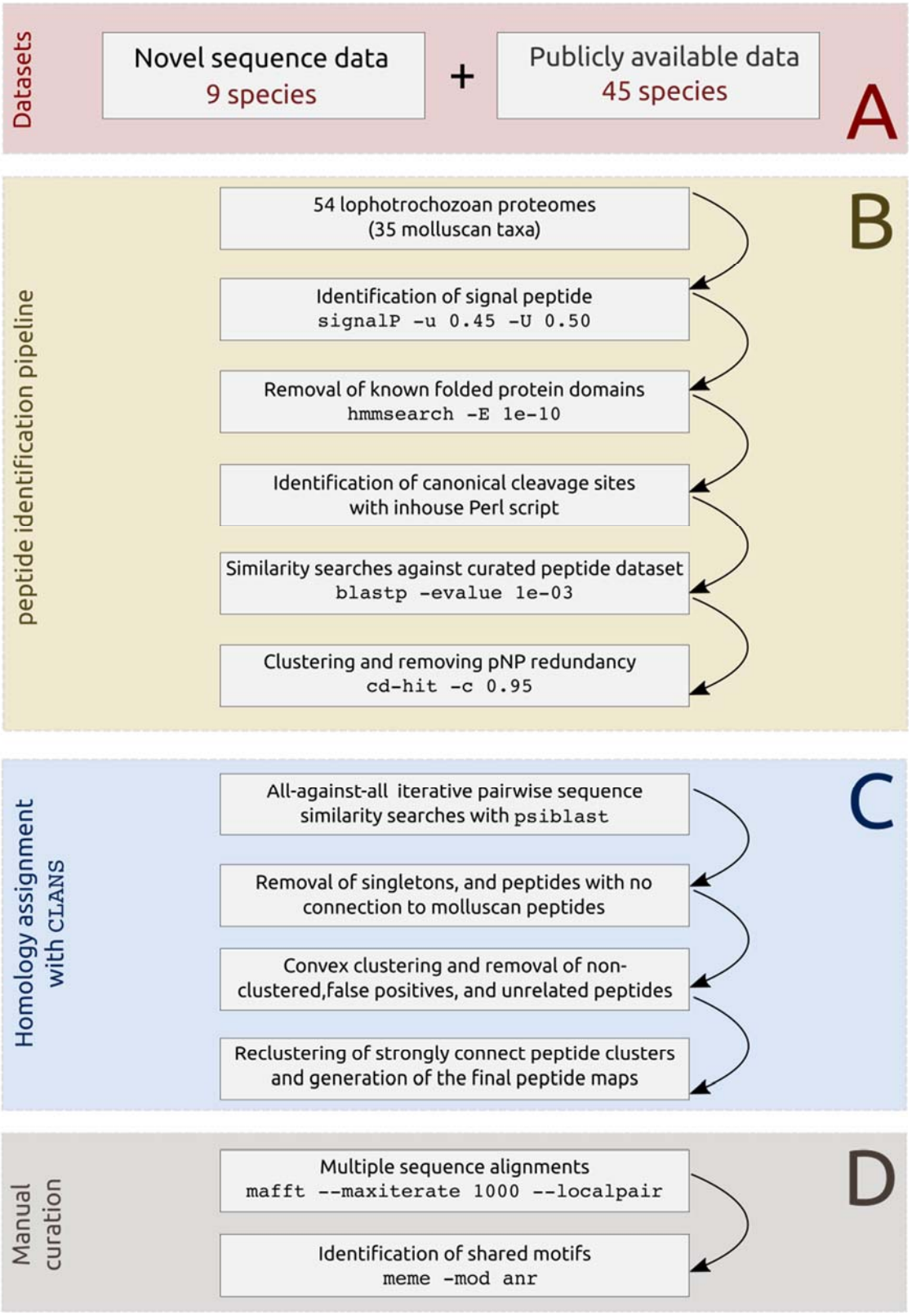


Figure 6: Bioinformatics pipeline developed for the identification and classification of molluscan and lophotrochozoan pNP and prohormone sequences. (A) The lophotrochozoan and molluscan databases were downloaded from Sequence read archive (https://www.ncbi.nlm.nih.gov/sra), pre-processed, and assembled locally

(not shown). (B) Predicted coding sequence regions from genomic data were downloaded and used where available. Identification of the *sine-qua-non* prerequisites present in the peptide sequences, clustering, and removal of false positive sequences were performed with several bioinformatics tools (e.g., signal, blast, hmmsearch) and *in-house* Perl scripts. (C) Orthology assignments and phylogenetic analysis were performed using all-against-all comparisons with psi-blast and blastp as implemented in the CLANS software. (D) Downstream annotation of the identified peptide sequences were aided through multiple sequence alignments, motif identification, and manual inspection.

Additional files

Additional file 1: BUSCO assessment of gene content and completeness of the lophotrochozoan protein sets. Gray stars indicate species in which genomic data were available. Datasets were classified into BUSCO metrics as follows: C: complete (C) and single copy (S); complete (C) and duplicated (D); fragmented (F); and missing (M). Black stars correspond to predicted protein sets obtained from available genomic data.

Additional file 2: Compressed zip file of all lophotrochozoan non-redundant neuropeptidomes in protein fasta file format. All steps and different programs used to generate the datasets are explained in “Material and methods”, and summarised in the Figure 6.

Additional file 3: Compressed zip file of the 2D final map in CLANS format depicting the phylogenetic relationships of eumetazoan proneuropeptides and peptide hormones. To read the CLANS format file, uncompress the file, install CLANS and run the command: java –jar clans.jar CLANS-FILE.rtf. CLANS can be downloaded from the following address: ftp://ftp.tuebingen.mpg.de/pub/protevo/CLANS/.

Additional file 4: Compressed zip file of the 2D final map for the central cluster in CLANS format. To read the CLANS format file, uncompress the file, install CLANS and run the command: java –jar clans.jar CLANS-FILE.rtf. CLANS can be downloaded from the following address: ftp://ftp.tuebingen.mpg.de/pub/protevo/CLANS/.

Additional file 5: Manually curated catalog of molluscan/lophotrochozoan peptide families containing trimmed multiple sequence alignments and peptide logos. Multiple sequence alignments were generated using the program mafft, and subsequently trimmed with trimal program. Peptide logos were generated using meme software. To access the high-quality pdf files please copy in your browser the following address: https://zoology.univie.ac.at/index.php?id=210716

Additional file 6: Trimmed multiple sequence alignments and phylogenetic analysis of four groups of prokineticin-like peptides in the monoplacophoran *Laevipilina hyalina*. The branch support values are posterior probabilities values. Multiple sequence alignments were generated using the program mafft, and subsequently trimmed with trimal program. The phylogenetic inferences were performed with mrbayes. For detailed information see “Material and methods – “Phylogenetic analysis”.

Additional file 7: Structure of molluscan and lophotrochozoan ELH/DH44 peptides, highlighting the identified repetitive peptide motifs. Motif identification was performed with meme software, and the presence of signal peptide was revealed with signalP program.

Additional file 8: Estimates of the evolutionary divergence of molluscan and lophotrochozoan ELH/DH44. The number of amino acid substitutions per site from between sequences is shown. Analyses were conducted using the JTT matrix-based model. The analysis involved 21 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 42 positions in the final dataset. Evolutionary analyses were conducted in MEGA7.

Additional file 9: Structure of molluscan and lophotrochozoan myomodulin and PRXamide peptides and identified repetitive peptide motifs. Motif identification was performed with meme software, and the presence of signal peptide was revealed with signalP program.

Additional file 10: Structure of molluscan and lophotrochozoan FCAP peptides and identified repetitive peptide motifs. Motif identification was performed with meme

software, and the presence of signal peptide was revealed with signalP program. The tables indicate the number of different copies identified in each proneuropeptide sequences.

Additional file 11: Table containing the species, phyla, SRA accession numbers, and the file transfer protocol addresses (FTP) of the molecular databases used in this study.

2.3 Manuscript 3 - Evolution and phylogenetic distribution of the euarthropod ecdysis pathway components (in preparation)

De Oliveira, AL; Calcino, A and Wanninger A.

Status: In preparation

2.3.1 Introduction

Ecdysis or moulting is the process of casting off an outer integument, the cuticle, for the purpose of growth. It is often considered an autapomorphy of Ecdysozoa, i.e. euarthropods, tardigrades, onychophorans, nematodes and related phyla (Aguinaldo et al., 1997; Schmidt-Rhaesa et al., 1998; De Rosa et al., 1999; Telford et al., 2008; Dunn et al., 2008). Despite the universality of the “moulting behaviour” within Ecdysozoa, the neuroendocrine components underlying this process remain elusive for the majority of the ecdysozoans outside of Euarthropoda. For instance, the molecular components responsible for ecdysis in the well-established nematode model organism *Caenorhabditis elegans* are still being investigated (reviewed by Page et al., 2014, and Lažetić & Fay, 2017).

In Arthropoda, ecdysis can be divided into three distinct stages, pre-ecdysis, ecdysis, and post-ecdysis, which correlates with major behavioural, molecular and cellular changes in the animal body. Each stage encompasses a series of specific skeletal muscular contractions controlled by a cascade of hormones and neuropeptides (Truman, 2005). Among the components present in this signalling pathway, prothoracicotropic hormone (PTTH), eclosion hormone (EH), crustacean cardioactive peptide (CCAP), and bursicon have been identified as key peptides involved in the ecdysis process. These four signaling molecules and their respective receptors are briefly introduced in the following.

Prothoracicotropic hormone (PTTH) and its ortholog Trunk: Prothoracicotropic hormone (PTTH) was the first insect brain peptide to be discovered (Kopeć, 1922; Wigglesworth, 1934). In holometabolous insects, the process of metamorphosis is initiated by the production of PTTH, which stimulates the prothoracic glands, the primary insect endocrine organ, to synthesize and release ecdysone. Ecdysone, an active steroid moulting hormone, in turn is required for insect growth, moulting and metamorphosis. The insect *ptth* is arthropod-specific paralog of *trunk* (Rewitz et al., 2009), an ancient bilaterian family known from lophotrochozoans and the cephalochordate *Branchiostoma floridae* (Jékely, 2013). Both paralogs present cystine knot-type structures similar to TGF-beta inhibitors and noggin factors found in non-bilaterians such as placozoans and sponges, as well as in bilaterian animals (Noguti el al., 1995; Jékely, 2013). PTTH and Trunk are ligands for the receptor

tyrosine kinase (RTKs) Torso, which is expressed in the prothoracic gland (Rewitz et al., 2009).

Eclosion hormone (EH): This hormone was first identified as a blood-borne factor by Truman & Riddiford (1970) in three lepidopteran species, *Hyalophora cecropia*, *Antheraea polyphemus*, and *Antheraea pernyi*, and was shown to act on the nervous system of these silkworms triggering a species-specific behaviour during the pre-ecydsis stage (Baker et al., 1999). Molecular characterisation showed that EH is a 62 amino acid peptide that contains six conserved cysteine residues that form three disulfide bridges (Truman, 1992; Žitňan et al., 2007). EH binds to and activates a guanylyl cyclase receptor that is expressed in epitracheal Inka cells (Chang et al., 2009). Traditionally considered to be confined to arthropods, a recent study showed the presence of EH and its receptor guanylyl cyclase in echinoderms (Zandawala et al., 2017).

Crustacean cardioactive peptide (CCAP): First isolated from the shore crab *Carcinus maenas*, the crustacean cardioactive peptide is a highly conserved amidated neuropeptide that increases the heartbeat in crustaceans and insects (Stangler et al., 1987; Cheung et al., 1992; Lehman et al., 1993). CCAP has multiple functions in addition to its cardioaccelatory activity stimulating oviduct contractions in *Locusta migratoria* (Donini et al., 2001) and the release of digestive proteins in *Periplaneta americana* (Sakai et al., 2006). Studies show that CCAP also plays an important role in ecdysis in crustaceans and insects by initiating the ecdysis motor program that marks the end of the pre-ecdysis stage (Gammie & Truman, 1997a, 1997b; Philippen et al., 2000; Lee et al., 2013). CCAP receptor is a G protein-couple receptor (GPCR) first screened from the *Drosophila* genome, and subsequently identified in many other insects (Cazzamali et al., 2003; Arakane et al., 2008; Vogel et al., 2013).

Bursicon: Bursicon was identified in the mid-sixties as a neurohormone responsible for the sclerotization and melanisation (tanning) of the insect cuticle right after the shedding of the old one during the last steps of the ecdysis process, i.e. post-ecdysis (Cottrell, 1962a,b; Fraenkel & Hsiao, 1965). Additionally, bursicon also plays a role in the development and expansion of the wings and other integumentary structures (Bai

& Palli, 2010). Comparative studies on the evolution and diversity of metazoan peptide families showed that the phyletic distribution of bursicon is broader than hitherto assumed, being found, in addition to ecdysozoans, in cnidarians (*Nematostella vectensis*), echinoderms (*Strongylocentrotus purpuratus)*, and annelids (Conzelmann et al., 2013; Jékely 2013). The gene coding for bursicon receptor, *rickets*, is a class-A rhodopsin-like GPCR (Baker & Truman, 2002; Arakane et al., 2008; Vogel et al., 2013).

Recent findings showed for the first time the presence of these previously mentioned peptide ligands in many lophotrochozoan phyla hinting to a deeper phylogenetic origin and wider distribution of these genes in the metazoan tree (De Oliveira et al., *in review*). To obtain a fine-grained overview of the evolutionary history of these ligand-receptor pairs, metazoan molecular databases were screened for the presence PTTH/Trunk, EH, CCAP, and bursicon ligands and receptors using sensitive iterative similarity search methods based on profile hidden Markov models (pHMMs).

2.3.2 Results

Prothoracicotropic hormone and its ortholog trunk: The homolog of the arthropod *ptth* gene, *trunk*, was found in lophotrochozoans (Annelida, Mollusca, Ectoprocta Nemertea, Brachiopoda, Phoronida, Gastrotricha), deuterostomes (Hemichordata, Cephalochordata), and in the ecdysozoan phyla Tardigrada and Onychophora (Fig. 4). A putative *trunk* ortholog was also recovered from the genome of the warty comb-jelly *Mnemiopsis leidyi* (Fig. 4A). The ctenophore trunk-like peptide is connected with significant sequence similarity (p-value < 1e-05) to lophotrochozoan, deuterostome and ecdysozoan *trunk* ortholog sequences. Although no *trunk* ortholog ligand was retrieved from the genome of the sea anemone *Nematostella vectensis*, similarity searches against the NCBI protein database identified this gene in other anthozoans, *Stylophora pistillata* (PFX31008.1) and *Orbicella faveolata* (XP_020630744.1 and XP_020630745.1).The receptor Torso was identified in non-bilaterians, ecdysozoans, lophotrochozoans and deuterostomes (Fig. 5). The receptor-ligand pair was found in cephalochordate, mollusk, onychophoran and arthropod representatives. Within Mollusca, Trunk was retrieved

from polyplachophorans and all conchiferans but not from aplacophorans. Torso was identified in aculiferans, i.e. Neomeniormorpha, Polyplacophora (but not Chaetodermomorpha), and conchiferan mollusks, i.e. Monoplacophora, Gastropoda, and Bivalvia, being absent in Cephalopoda and Scaphopoda. The ligand-receptor pair indicates a deeper origin of Trunk-PTTH-Torso neuropeptide signalling than previously assumed, tracing back to the last common ancestor of Cnidaria/Ctenophora and Bilateria.

Eclosion hormone: *eh* ligand orthologs were found in Cnidaria, Xenacoelomorpha, Ambulacraria, Lophotrochozoa and Panarthropoda (Fig. 4B). EH receptor guanylyl cyclase is similarly widespread amongst Ambulacraria, Lophotrochozoa and Panarthropoda (Fig. 6). Although *eh* ortholog was not found in cephalochordates, ectoprocts, and brachiopods, its guanylyl cyclase receptor was retrieved from these phyla (Fig. 6). Within Mollusca, all the extant classes possess the *eh* ligand ortholog, while its receptor was retrieved only from Neomeniomorpha, Monoplacophora, Gastropoda and Bivalvia (Fig. 4A and 6) All *eh* ligand orthologs harbour the six cysteine diagnostic residues, apart from Cnidaria in which only five are present. The ligand-receptor ancestry of these signalling molecules dates back at least to the stem urbilaterian and may have already been present in the last common ancestor of cnidarians and bilaterians.

Crustacean cardioactive peptide: The CCAP ligand was retrieved from numerous protostomian animals such as annelids, mollusks, nemerteans, platyhelminthes, rofiterans and several panarthropods (Tardigrada, Onychophora and Arthropoda; Fig. 4C). The receptor was found in all deuterostome phyla analysed, with the exception of Tunicata, the basally branching bilaterian lineage Xenacoelomorpha, several lophotrochozoans (annelids, mollusks, phoronids, rotiferans), and ecdyszoans (tardigrades and arthropods) (Fig. 7). Within Mollusca, the CCAP ligand was identified in all class-level taxa apart from Chaetodermomorpha, and its receptor in Polyplacophora, Monoplacophora, Gastropoda and Bivalvia. The ligand-receptor pair indicates a bilaterian origin of these peptides, with loss of the ligand in the Deuterostomia lineage.

Bursicon: The *bursicon* gene was found in non-bilaterian animals, the sea anemome *Nematostella vectensis*, ambulacrarians, (Hemichordata and Echinodermata), annelids, mollusks, nemerteans, phoronids, rotifers, and in the ecdysozoan phyla Arthropoda and Onychophora (Fig. 4D). The receptor Rickets was found only in mollusks, annelids, nemerteans, and arthropods (Fig. 8). Within Mollusca, bursicon was found in neomeniomorphs, polyplacophorans, bivalves, gastropods, and cephalopods. The receptor is present in Neomeniomorpha and all the conchiferan class-level taxa, with the exception of Monoplacophora. The bursicon-rickets pair indicates its presence already in the last common ancestor of Cnidaria and Bilateria, and the subsequent retention of either the ligand or receptor in the three major superphyla, i.e. Ecdysozoa, Lophotrochozoa and Deuterostomia.

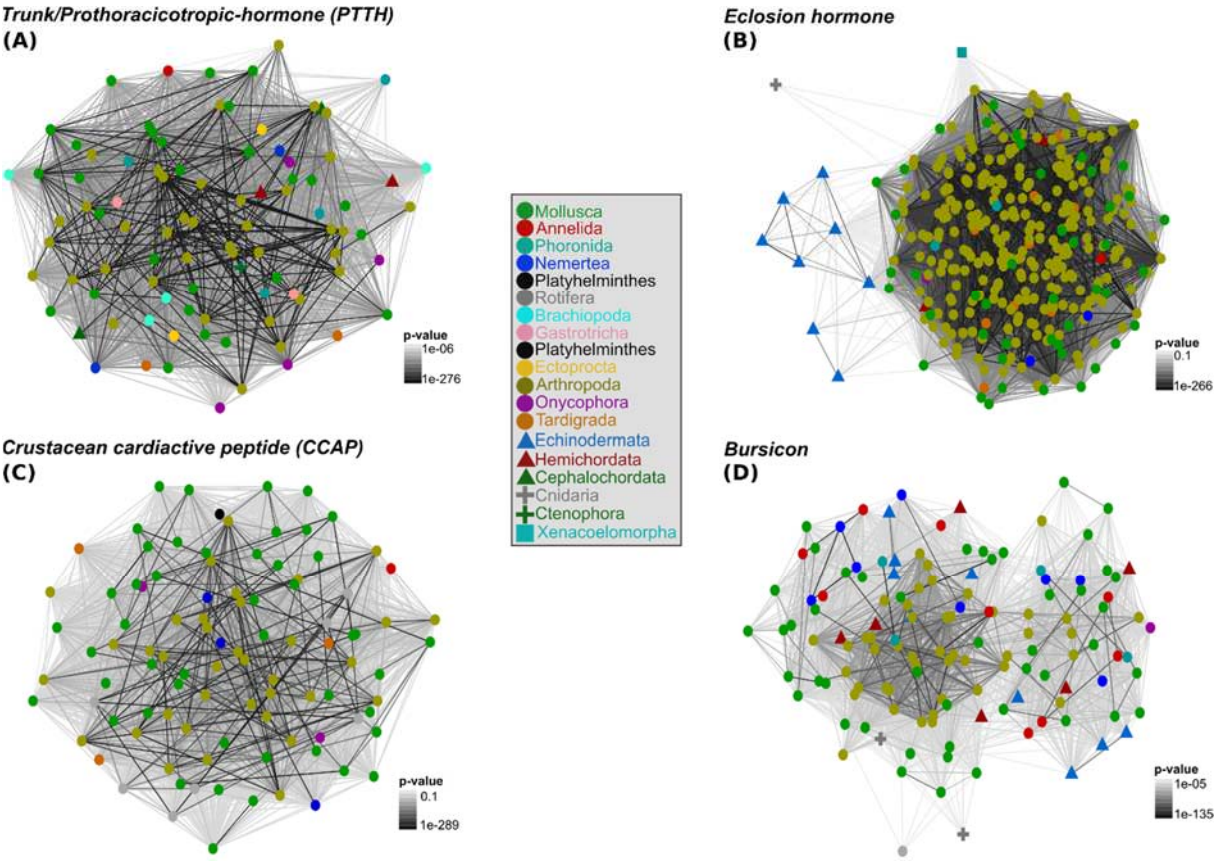


Figure 4 – 2D cluster maps of PTTH/Trunk, EH, CCAP and bursicon ligands. Colour shapes and nodes are based on the different phyla investigated. Edges correspond to BLAST connections.

2.3.3 Discussion

With the boom of genome and transcriptome sequence data from a plethora of animals and the development of sophisticated bioinformatics tools, significant progress was made over the past 10 years in reconstructing the peptidergic signalling systems in the animal kingdom (Jékely, 2013; Mirabeau & Joly, 2013). Here, by interrogating metazoan molecular databases, we were able to reconstruct the evolutionary history of key components of the Euarthropoda ecdysis signalling pathway. Importantly, owing to the high sequence divergence of some peptide ligands, e.g. eclosion hormone and trunk, many remotely conserved homologs were only identified due to the use of more sophisticated similarity search tools based on profile hidden Markov models.

The finding of a peptide in the comb jelly *Mnemiopsis leidyi* with a close relationship to insect neuropeptides represents, to the best of our knowledge, the first peptide identified in a ctenophore that is homologous to a metazoan neuropeptide precursor (i.e. prothoracicotropic hormone). This result is particularly interesting in the light of the proposed independent evolution of ctenophore and cnidarian-bilaterian nervous systems (Moroz et al., 2014; Jékely et al., 2015).

The combined analysis of neuropeptide/hormone precursors described herein show that genes with a conserved role in moulting in arthropods are widespread in non-bilaterians, deuterostomes, lophotrochozoans, and xenacoelomorphs, suggesting their presence already in the bilaterian and/or metazoan last common ancestors. Additionally, the data obtained from the investigated ecdysozoans show the conservation of PTTH/Trunk, EH, CCAP and Bursicon in panarthropods, with a possible secondary loss of Bursicon peptidergic signalling in tardigrades. The latter finding is corroborated by independent proneuropeptide and peptide hormone surveys in the tardigrade genomic and EST data (Christie et al., 2011; Koziol, 2018).

No homologs of the herein investigated peptides have been identified in the nematode *Caenorhabditis elegans* in accordance with previous studies (Page et al., 2014; Lažetić & Fay, 2017). Moreover, other key moulting hormones, the ecdysteroids ecdysone (E), 20-hydroxyecdysone (20E), and Halloween genes have been reported missing from the *C. elegans* genome as well (Frand et al., 2005, Schumann et al., 2018). Curiously, E and 20E have been identified in parasitic nematodes (Cleator et al., 1987; Shea et al., 2004) and, outside the ecdysozoans clade, in the platyhelminth *Monieza expansa*, the gastropod mollusks *Lymnaea*

stagnalis and *Helix pomatia*, and the annelid *Hirudo medicinalis* (Mendis et al., 1984; Nolte et al., 1986; Garcia et al., 1989; Barker et al., 1990; Paxton, 2005; Pilato et al., 2005). These results strongly suggest a different signalling pathway underlying this behaviour in the free-living nematode, notion that is currently supported by recent studies (Russel et al., 2011; Lažetić & Fay, 2017).

Our findings show that key players responsible for ecdysis in Euarthropoda evolved early in animal evolution and were most likely involved in different (hitherto unknown) biological functions. The identification of the euarthropod ecdysis pathway components in the closely related Onychophora and Tardigrada points towards a conserved functional pathway involved in ecdysis in the last common ancestor of Panarthropoda. However, the involvement of these genes and their function in the closest euarthropod allies remain unknown. Additionally, independent recruitments of novel components into the insect ecdysis signalling cascade have been reported (Kim et al., 2004, 2006). Corazonin and FMRFamide peptides, highly conserved metazoan neuropeptides (Jékely, 2013; Mirabeau & Joly, 2013), were identified as important players in ecdysis behaviour in the moth *Manduca* and the fruit fly *Drosophila* (Kim et al., 2004, 2006). These findings illustrate the variability of components underlying the ecdysis pathway in different lineages, and indicate that more detailed functional investigations into the euarthropod ecdysis pathway are required.

To conclude, further investigations into the closely related Nematoda allies (priapulids, kinorhynchs, lorificerans, and nematomorphs) are essential to fully appreciate the molecular mechanisms underlying the ecdysis behaviour in the eight ecdysozoan phyla. Our results suggest that ancient metazoan genes were involved in the moulting behaviour in the last common ancestor of Panarthropoda and may have acquired this function already at the base of Ecdysozoa.

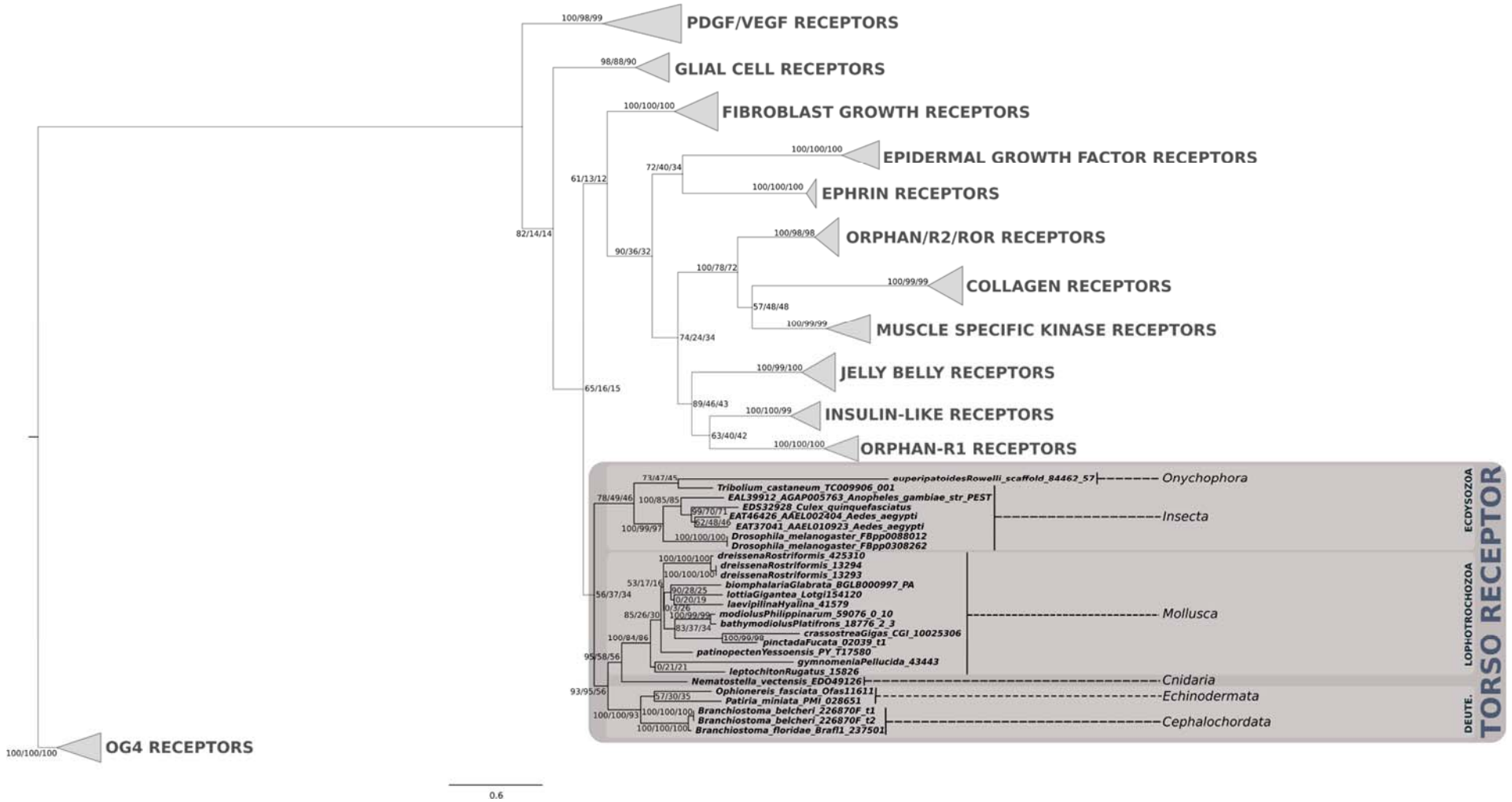


Figure 5 – Phylogenetic inference of PTTH/trunk receptor tyrosine kinase torso showing its distribution in eumetazoans. Support values for the tree nodes obtained from mrbayes, RAxML, and PhyML are shown as percentage. Tree topology obtained from RAxML was used as a backbone, and conflicting topology branches from mrbayes and PhyML inferred trees are marked by brackets ([ ]) around the support values.

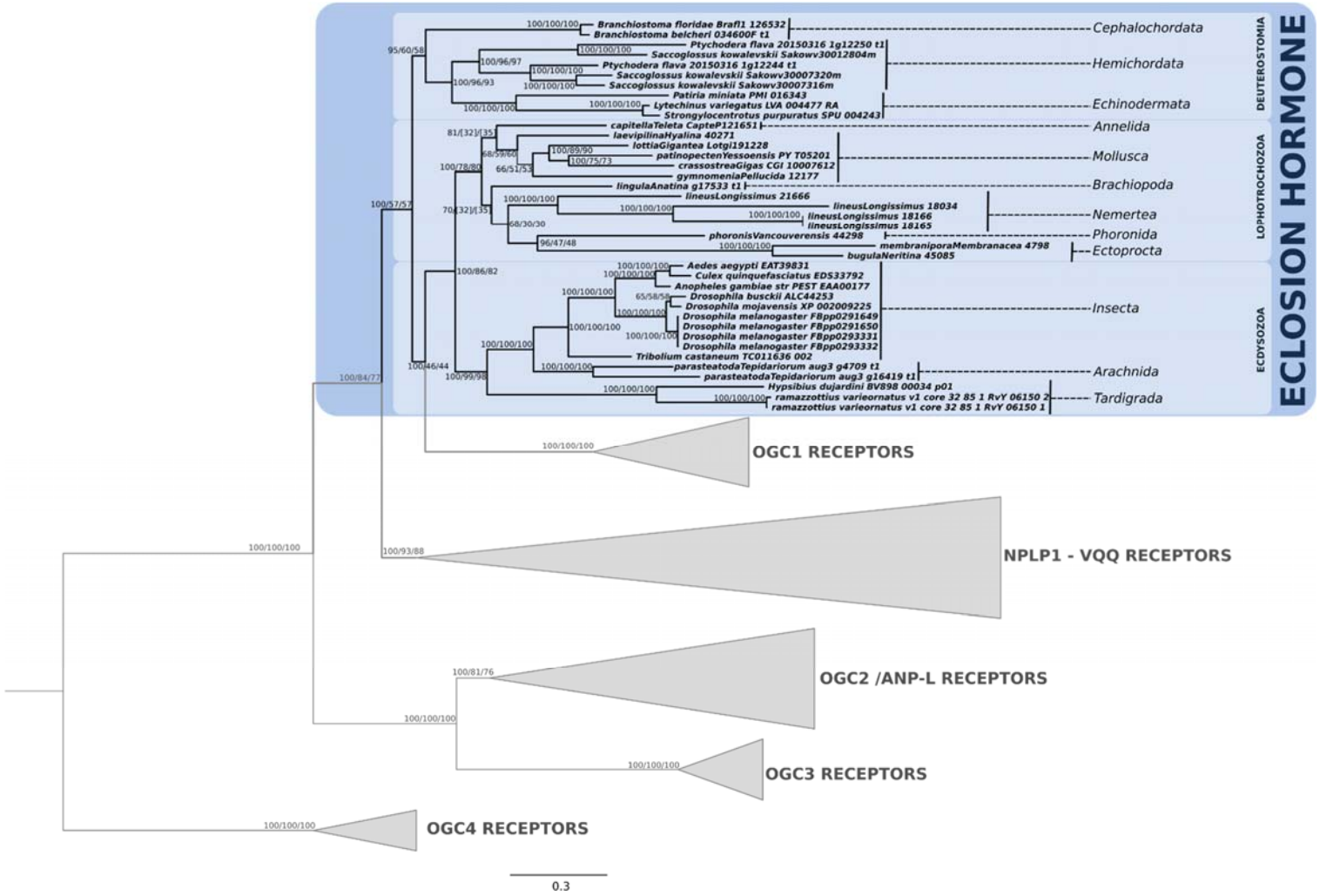


Figure 6 –Phylogenetic inference of guanylyl cyclase eclosion hormone receptor showing its distribution in nephrozoans, i.e. all bilaterians except xenacoelomorphs. Support values for the tree nodes obtained from mrbayes, RAxML, and PhyML are shown as percentage. Tree topology obtained from mrbayes was used as a backbone, and conflicting topology branches from RAxML and PhyML inferred trees are marked by brackets ([ ]) around the support values.

Figure 7 – Phylogenetic inference of G-protein-coupled CCAP receptor showing its distribution in bilaterians. Support values for the tree nodes obtained from mrbayes, RAxML, and PhyML are shown as percentage. Tree topology obtained from RAxML was used as a backbone, and conflicting topology branches from mrbayes and PhyML inferred trees are marked by brackets ([ ]) around the support values.

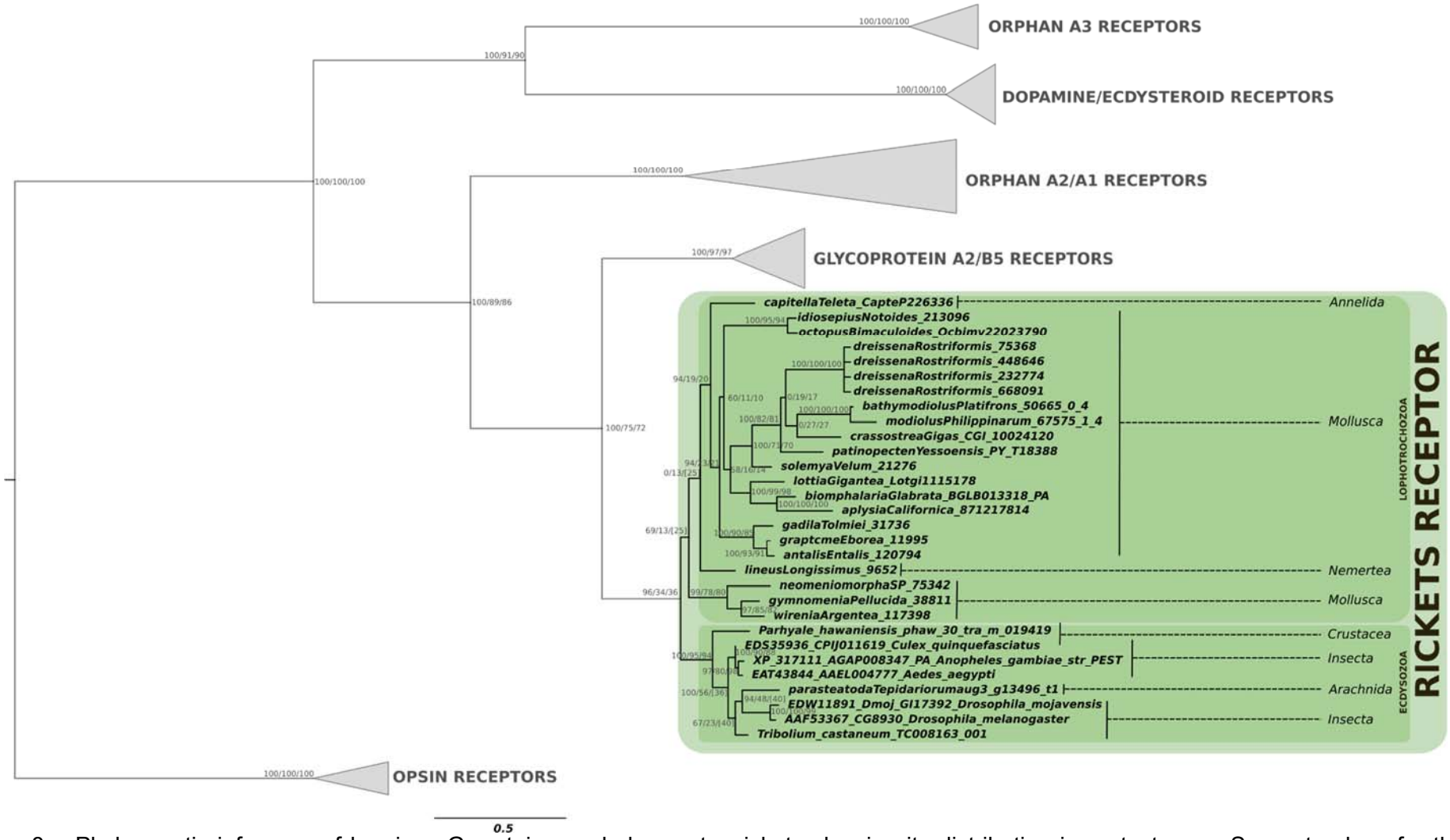


Figure 8 – Phylogenetic inference of bursicon G-protein-coupled receptor rickets showing its distribution in protostomes. Support values for the tree nodes obtained from mrbayes, RAxML, and PhyML are shown as percentage. Tree topology obtained from RAxML was used as a backbone, and conflicting topology branches from mrbayes and PhyML inferred trees are marked by brackets ([ ]) around the support values.

2.3.4 Material and methods

2.3.4.1 Data collection, filtering, sequence reconstruction, and proteome prediction

To obtain a comprehensive phylogenetic sampling of the metazoan tree, ecdysozoan, deuterostome, and non-Bilateria protein coding sequence (CDS) databases were downloaded from publicly available sites and combined with previous lophotrochozoan transcriptomes (De Oliveira et al., *in review*). The xenacoelomorph transcriptomic data were pre-processed and assembled as described in Manuscript 2. The databases include Porifera, Ctenophora, Cnidaria, Placozoa, Echinodermata, Hemichordata, Chordata, Annelida, Brachiopoda, Ectoprocta, Entoprocta, Gastrotricha, Nemertea, Phoronida, Platyhelminthes, Rotifera, Arthropoda, Tardigrada and Nematoda representatives. The choanoflagellate *Monosiga brevicollis* was sampled as outgroup. Table 2 summarises the recently incorporated databases, and the publicly available repository from which they were obtained. Sequence read archive (SRA) accession numbers for the basal branching xenacoelomorphs are also displayed.

2.3.4.2 Sensitive similarity searches with jackhmmer

Sensitive probabilistic iterative similarity searches based on profile hidden Markov models (HMMs) were performed with jackhmmer (Johnson et al., 2010) against the mentioned metazoan and choanoflagellate databases. Insect *eh*, *ccap*, *ptth*, *bursicon* orthologs were retrieved from NCBI (National Center for Biotechnology Information), and their respective receptors from Vogel et al. (2013). These sequences were used as queries in the similarity searches. The searches were performed under the default parameters using varying e-value thresholds (1 to 1e-06) controlled by the options –E and –domE, as defined in jackhmmer. The best hits found in the metazoan and choanoflagellate databases were stored in fasta format and used in the subsequent analyses.

2.3.4.3 Clustering and phylogenetic analysis

EH, CCAP, PTTH and bursicon ligand candidates retrieved from the metazoan and choanoflagellate databases were used as input, together with their respective insect orthologs, in the program clans (Frickey & Lupas, 2004) under different e-value thresholds (0.1 to 1e-06) and blast programs, i.e. blastp or psiblast (Camacho et al., 2009). Singleton sequences, i.e. isolated unconnected sequences, were excluded from the map. To further improve the orthology assessment, multiple sequence alignments were performed with mafft (Katoh & Standley, 2013), and the presence of shared conserved amino acid regions and residues were investigated with aliview (Larsson, 2014). The final 3D maps were collapsed into 2D after the clustering for easier visualisation.

Putative EH, CCAP, PTTH and bursicon receptor candidates retrieved from the metazoan and choanoflagellate databases were aligned with mafft together with their respective orthologs, when found, and subsequently trimmed with BMGE software under the following parameters: –h 1 –b 1 –m BLOSUM30 –t AA (Criscuolo & Gribaldo, 2010). Outgroups for the phylogenetic analyses were defined in conformity with Vogel et al. (2013).

Phylogenetic analyses were performed using RAxML (Stamatakis, 2014), PhyML (Guindon et al., 2010) and mrbayes (Ronquist et al., 2012) softwares using the appropriate best-fit model of amino acid substitution as described in Manuscript 2. RaxML was executed under default parameters and rapid bootstrap. PhyML was executed under the default parameters and an optimised starting tree (-o tlr option). The number of bootstrap values was set to 1.000 in RaxML and PhyML, and the number of generations used in the mrbayes was determined using a convergence diagnostic. All runs in mrbayes were performed with the samplefreq and relative burn-in defined as 1000 and 25%, respectively. The three final phylogenetic trees obtained for each of the four different receptors were visualised and combined with TreeGraph2 (Stöver & Müller, 2010).

Table 2 – Metazoan CDS databases and the publicly available repositories from which they were obtained. SRA accession numbers are also displayed.

Organism	Superphylum or phylum	Biological sequence data repositories
Australostichopus mollis	Deuterostomia	http://ryanlab.whitney.ufl.edu/genomes/Amol/

Branchiostoma belcheri	Deuterostomia	http://genome.bucm.edu.cn/lancelet/download_data.php
Branchiostoma floridae	Deuterostomia	http://genome.jgi.doe.gov/Brafl1/Brafl1.download.html
Ciona intestinalis	Deuterostomia	http://genome.jgi.doe.gov/Cioin2/Cioin2.download.ftp.html
Danio rerio	Deuterostomia	http://mar2015.archive.ensembl.org/Danio_rerio/Info/Index
Homo sapiens	Deuterostomia	http://grch37.ensembl.org/Homo_sapiens/Info/Index
Lytechinus variegatus	Deuterostomia	http://www.echinobase.org/Echinobase/LvDownloads
Mus musculus	Deuterostomia	http://www.ensembl.org/Mus_musculus/Info/Index
Patiriella regularis	Deuterostomia	http://ryanlab.whitney.ufl.edu/genomes/Preg/
Ophionereis fasciata	Deuterostomia	http://ryanlab.whitney.ufl.edu/genomes/Ofas/
Patiria miniata	Deuterostomia	http://www.echinobase.org/Echinobase/PmDownload
Ptychodera flava	Deuterostomia	https://groups.oist.jp/molgenu/hemichordate-genomes
Saccoglossus kowalevskii	Deuterostomia	https://groups.oist.jp/molgenu/hemichordate-genomes
Strongylocentrotus purpuratus	Deuterostomia	http://www.echinobase.org/Echinobase/SpDownloads
Caenorhabditis_elegans	Ecdysozoa	http://www.ensembl.org/Caenorhabditis_elegans/Info/Index
Drosophila melanogaster	Ecdysozoa	http://www.ensembl.org/Drosophila_melanogaster/Info/Index
Euperipatoides rowelli	Ecdysozoa	https://www.ncbi.nlm.nih.gov/Traces/wgs/?val=PXIH01#scaffolds
Tribolium castaneum	Ecdysozoa	http://metazoa.ensembl.org/Tribolium_castaneum/Info/Index

Parhyale hawaniensis	Ecdysozoa	https://figshare.com/articles/supplemental_data_for_Parhyale_hawaniensis_genome/3498104
Hypsibius dujardini	Ecdysozoa	http://ensembl.tardigrades.org/Hypsibius_dujardini_nhd315/Info/Index
Ramazzottius varieornatus	Ecdysozoa	http://download.tardigrades.org/v1/sequence/
Parasteatoda tepidariorum	Ecdysozoa	https://i5k.nal.usda.gov/content/data-downloads
Peripatopsis capensis	Ecdysozoa	https://www.ncbi.nlm.nih.gov/Traces/wgs/?val=PXIH01#scaffolds
Nematostella vectensis	Cnidaria	http://metazoa.ensembl.org/Nematostella_vectensis/Info/Index
Mnemiopsis leidyi	Ctenophora	https://research.nhgri.nih.gov/mnemiopsis/
Trichoplax adhaerens	Placozoa	http://metazoa.ensembl.org/Trichoplax_adhaerens/Info/Index
Amphimedon queenslandica	Porifera	http://amphimedon.qcloud.qcif.edu.au/
Isodiametra pulchra	Xenacoelomorpha	SRR2681926
Convolutriloba macropyga	Xenacoelomorpha	SRR2681679
Eumecynostomum macrobursalium	Xenacoelomorpha	SRR3105705
Diopisthoporus longitubus	Xenacoelomorpha	SRR3105704
Diopisthoporus gymnopharyngeus	Xenacoelomorpha	SRR3105703
Childia submaculatum	Xenacoelomorpha	SRR3105702
Monosiga brevicollis	Choanoflagellata	http://genome.jgi.doe.gov/Monbr1/Monbr1.download.ftp.html

2.3.5 References

Aguinaldo, A. M. A., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, *387*(6632), 489-493.

Arakane, Y., Li, B., Muthukrishnan, S., Beeman, R. W., Kramer, K. J., & Park, Y. (2008). Functional analysis of four neuropeptides, EH, ETH, CCAP and bursicon, and their receptors in adult ecdysis behavior of the red flour beetle, *Tribolium castaneum*. *Mechanisms of development*, *125*(11-12), 984-995

Bai, H., & Palli, S. R. (2010). Functional characterization of bursicon receptor and genome-wide analysis for identification of genes affected by bursicon receptor RNAi. *Developmental biology*, *344*(1), 248-258.

Baker, J. D., McNabb, S. L., & Truman, J. W. (1999). The hormonal coordination of behavior and physiology at adult ecdysis in *Drosophila melanogaster*. Journal of *Experimental Biology*, *202*(21), 3037-3048.

Baker, J. D., & Truman, J. W. (2002). Mutations in the *Drosophila* glycoprotein hormone receptor, rickets, eliminate neuropeptide-induced tanning and selectively block a stereotyped behavioral program. *Journal of Experimental Biology*, *205*(17), 2555-2565.

Barker, G. C., Chitwood, D. J., & Rees, H. H. (1990). Ecdysteroids in helminths and annelids. *Invertebrate Reproduction & Development*, *18*(1-2), 1-11.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, *10*(1), 421.

Chang, J. C., Yang, R. B., Adams, M. E., & Lu, K. H. (2009). Receptor guanylyl cyclases in Inka cells targeted by eclosion hormone. *Proceedings of the National Academy of Sciences*, *106*(32), 13371-13376.

Cazzamali, G., Hauser, F., Kobberup, S., Williamson, M., & Grimmelikhuijzen, C. J. (2003). Molecular identification of a *Drosophila* G protein-coupled receptor specific for crustacean cardioactive peptide. *Biochemical and biophysical research communications*, *303*(1), 146-152.

Cheung, C. C., Loi, P. K., Sylwester, A. W., Lee, T. D., Tublitz, N. J. (1992). Primary structure of a cardioactive neuropeptide from the tobacco hawkmoth, *Manduca sexta*. *FEBS letters*, *313*(2), 165-168.

Christie, A. E., Nolan, D. H., Garcia, Z. A., McCoole, M. D., Harmon, S. M., Congdon-Jones, B. (2011). Bioinformatic prediction of arthropod/nematode-like peptides in non-arthropod, non-nematode members of the Ecdysozoa. *General and comparative endocrinology*, *170*(3), 480-486.

Cleator, M., Delves, C. J., Howells, R. E., & Rees, H. H. (1987). Identity and tissue localization of free and conjugated ecdysteroids in adults of *Dirofilaria immitis* and *Ascaris suum*. *Molecular and biochemical parasitology*, *25*(1), 93-105.

Conzelmann, M., Williams, E. A., Krug, K., Franz-Wachtel, M., Macek, B., & Jékely, G. (2013). The neuropeptide complement of the marine annelid *Platynereis dumerilii*. *BMC genomics*, *14*(1), 906.

Cottrell, C. B. (1962a). The Imaginal Ecdysis of Blowflies. The Control of Cuticuar Hardening and Darkening. *Journal of Experimental Biology*, *39*(3), 395-412.

Cottrell, C. B. (1962b). The imaginal ecdysis of blowflies. Detection of the blood-borne darkening factor and determination of some of its properties. *Journal of Experimental Biology*, *39*(3), 413-430.

Criscuolo, A., & Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC evolutionary biology*, *10*(1), 210.

De Oliveira, A.L., Calcino, A. Wanninger, A. Extensive conservation of the proneuropeptide and peptide hormone complement in mollusks. Manuscript submitted for publication.

De Rosa, R., Grenier, J. K., Andreeva, T., Cook, C. E., Adoutte, A., Akam, M. (1999). Hox genes in brachiopods and priapulids and protostome evolution. *Nature*, *399*(6738), 772.

Donini, A., Agricola, H. J., & Lange, A. B. (2001). Crustacean cardioactive peptide is a modulator of oviduct contractions in *Locusta migratoria*. *Journal of insect physiology*, *47*(3), 277-285.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, *452*(7188), 745.

Fraenkel, G., & Hsiao, C. (1965). Bursicon, a hormone which mediates tanning of the cuticle in the adult fly and other insects. *Journal of Insect Physiology*, *11*(5), 513-556.

Frand, A. R., Russel, S., & Ruvkun, G. (2005). Functional genomic analysis of *C. elegans* molting. *PLoS biology*, *3*(10), e312.

Frickey, T., & Lupas, A. (2004). CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, *20*(18), 3702-3704.

Gammie, S. C., & Truman, J. W. (1997a). An endogenous elevation of cGMP increases the excitability of identified insect neurosecretory cells. *Journal of Comparative Physiology A*, *180*(4), 329-337.

Gammie, S. C., & Truman, J. W. (1997b). Neuropeptide hierarchies and the activation of sequential motor behaviors in the hawkmoth, *Manduca sexta*. *Journal of Neuroscience*, *17*(11), 4389-4397.

Garcia, M., Gharbi, J., Girault, J. P., Hetru, C., & Lafont, R. (1989). Ecdysteroid metabolism in leeches. *Invertebrate Reproduction & Development*, *15*(1), 57-68.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, *59*(3), 307-321.

Jékely, G. (2013). Global view of the evolution and diversity of metazoan neuropeptide signaling. *Proceedings of the National Academy of Sciences*,*110*(21), 8702-8707.

Jékely, G., Paps, J., & Nielsen, C. (2015). The phylogenetic position of ctenophores and the origin (s) of nervous systems. *Evodevo*, *6*(1), 1.

Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC bioinformatics*, *11*(1), 431.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), 772-780.

Kim, Y. J., Spalovská-Valachová, I., Cho, K. H., Zitnanova, I., Park, Y., Adams, M. E., et al. (2004). Corazonin receptor signaling in ecdysis initiation. *Proceedings of the National Academy of Sciences*, *101*(17), 6704-6709.

Kim, Y. J., Žitňan, D., Galizia, C. G., Cho, K. H., & Adams, M. E. (2006). A command chemical triggers an innate behavior by sequential activation of multiple peptidergic ensembles. *Current Biology*, *16*(14), 1395-1407.

Kopec, S. (1922). Studies on the necessity of the brain for the inception of insect metamorphosis. *The Biological Bulletin*, *42*(6), 323-342.

Koziol, U. (2018). Precursors of neuropeptides and peptide hormones in the genomes of tardigrades. *General and Comparative Endocrinology*.

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, *30*(22), 3276-3278.

Lažetić, V., & Fay, D. S. (2017, January). Molting in *C. elegans*. In Worm (Vol. 6, No. 1, p. e1330246). Taylor & Francis.

Lee, D. H., Orchard, I., & Lange, A. B. (2013). Evidence for a conserved CCAP-signaling pathway controlling ecdysis in a hemimetabolous insect, *Rhodnius prolixus*. *Frontiers in neuroscience*, *7*, 207.

Lehman, H. K., Murgiuc, C. M., Miller, T. A., Lee, T. D., Hildebrand, J. G. (1993). Crustacean cardioactive peptide in the sphinx moth, *Manduca sexta*. *Peptides*,*14*(4), 735-741

Mendis, A. H., Rees, H. H., & Goodwin, T. W. (1984). The occurrence of ecdysteroids in the cestode, *Moniezia expansa*. Molecular and biochemical parasitology, *10*(2), 123-138.

Mirabeau, O., & Joly, J. S. (2013). Molecular evolution of peptidergic signaling systems in bilaterians. *Proceedings of the national academy of sciences*, *110*(22), E2028-E2037.

Moroz, L. L., Kocot, K. M., Citarella, M. R., Dosung, S., Norekian, T. P., Povolotskaya, I. S. (2014). The ctenophore genome and the evolutionary origins of neural systems. *Nature*, *510*(7503), 109.

Noguti, T., Adachi-Yamada, T., Katagiri, T., Kawakami, A., Iwami, M., Ishibashi, J., et al. (1995). Insect prothoracicotropic hormone: a new member of the vertebrate growth factor superfamily. *FEBS letters*, *376*(3), 251-256.

Nolte, A., Koolman, J., Dorlöchter, M., & Straub, H. (1986). Ecdysteroids in the dorsal bodies of pulmonates (Gastropoda): synthesis and release of ecdysone. *Comparative Biochemistry and Physiology Part A: Physiology*, *84*(4), 777-782.

Page, A. P., Stepek, G., Winter, A. D., & Pertab, D. (2014). Enzymology of the nematode cuticle: a potential drug target? International Journal for Parasitology: *Drugs and Drug Resistance*, *4*(2), 133-141.

Paxton, H. (2005). Molting polychaete jaws - ecdysozoans are not the only molting animals. *Evolution & development*, *7*(4), 337-340.

Pilato, G., Binda, M. G., Biondi, O., D'Urso, V., Lisi, O., Marletta, A. (2005). The clade Ecdysozoa, perplexities and questions. *Zoologischer Anzeiger-A Journal of Comparative Zoology*, *244*(1), 43-50.

Phlippen, M. K., Webster, S. G., Chung, J. S., Dircksen, H. (2000). Ecdysis of decapod crustaceans is associated with a dramatic release of crustacean cardioactive peptide into the haemolymph. *Journal of Experimental Biology*, *203*(3), 521-536.

Rewitz, K. F., Yamanaka, N., Gilbert, L. I., & O'Connor, M. B. (2009). The insect neuropeptide PTTH activates receptor tyrosine kinase torso to initiate metamorphosis. *Science*, *326*(5958), 1403-1405.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, *61*(3), 539-542.

Russel, S., Frand, A. R., & Ruvkun, G. (2011). Regulation of the *C. elegans* molt by *pqn-47*. *Developmental biology*, *360*(2), 297-309.

Sakai, T., Satake, H., & Takeda, M. (2006). Nutrient-induced α-amylase and protease activity is regulated by crustacean cardioactive peptide (CCAP) in the cockroach midgut. *Peptides*, *27*(9), 2157-2164.

Schmidt-Rhaesa, A., Bartolomaeus, T., Lemburg, C., Ehlers, U., & Garey, J. R. (1998). The position of the Arthropoda in the phylogenetic system. *Journal of Morphology*, *238*(3), 263-285.

Schumann, I., Kenny, N., Hui, J., Hering, L., & Mayer, G. (2018). Halloween genes in panarthropods and the evolution of the early moulting pathway in Ecdysozoa. *Open Science*, *5*(9), 180888.

Shea, C., Hough, D., Xiao, J., Tzertzinis, G., & Maina, C. V. (2004). *An rxr/usp* homolog from the parasitic nematode, *Dirofilaria immitis*. *Gene*, *324*, 171-182.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313.

Stangier, J., Hilbich, C., Beyreuther, K., & Keller, R. (1987). Unusual cardioactive peptide (CCAP) from pericardial organs of the shore crab *Carcinus maenas*. *Proceedings of the National Academy of Sciences*, *84*(2), 575-579.

Stöver, B. C., & Müller, K. F. (2010). TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. *BMC bioinformatics*, *11*(1), 7.

Telford, M. J., Bourlat, S. J., Economou, A., Papillon, D., & Rota-Stabelli, O. (2008). The evolution of the Ecdysozoa. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *363*(1496), 1529-1537.

Truman J.W., Riddiford L.M. (1970). Neuroendocrine control of ecdysis in silkmoths. *Science*;*167*:1624–1626.

Truman J.W. The eclosion hormone system of insects (1992). *Progress in Brain Research*;*92*:361-74.

Truman, J. W. (2005). Hormonal control of insect ecdysis: endocrine cascades for coordinating behavior with physiology. *Vitamins & Hormones*, *73*, 1-30.

Vogel, K. J., Brown, M. R., & Strand, M. R. (2013). Phylogenetic investigation of peptide hormone and growth factor receptors in five dipteran genomes. *Frontiers in endocrinology*, *4*, 193.

Wigglesworth, V. B. (1934). Memoirs: The physiology of ecdysis in *Rhodnius prolixus* (Hemiptera). II. Factors controlling moulting and 'metamorphosis'. *Journal of Cell Science*, *2*(306), 191-222.

Zandawala, M., Moghul, I., Guerra, L. A. Y., Delroisse, J., Abylkassimova, N., Hugall, A. F., et al. (2017). Discovery of novel representatives of bilaterian neuropeptide families and reconstruction of neuropeptide precursor evolution in ophiuroid echinoderms. *Open biology*, *7*(9), 170129.

Žitňan, D., Kim, Y. J., Žitňanová, I., Roller, L., & Adams, M. E. (2007). Complex steroid–peptide–receptor cascade controls insect ecdysis. *General and comparative endocrinology*;*53*(1), 88-96.

3. GENERAL DISCUSSION

3.1 Next-generation sequencing

On 15 September of 2005, a landmark paper (Margulies et al., 2005) resulted in a revolution in the DNA sequencing fied by proposing a novel sequencing technology. As they wrote in their paper:

> "Here we describe a scalable, highly parallel sequencing system with raw throughput significantly greater than that of state-of-the-art capillary electrophoresis instruments. The apparatus uses a novel fibre-optic slide of individual wells and is able to sequence 25 million bases, at 99% or better accuracy, in one four-hour run. To achieve an approximately 100-fold increase in throughput over current Sanger sequencing technology, we have developed an emulsion method for DNA amplification and an instrument for sequencing by synthesis using a pyrosequencing protocol optimized for solid support and picolitre-scale volumes (Margulies et al., 2005)."

Since then many different technologies based on highly parallel or high-throughput sequencing methods were developed, e.g. Illumina, PacBio, or Nanopore (for review see Metzker, 2010; Levi & Myers, 2016; Goodwin et al., 2016), that rival traditional Sanger sequencing. These methods, referred to as next-generation sequencing, were closely followed by a rich inventory of applications that has been changing our understanding of biology. Amongst these applications, the identification and quantification of the complete set of transcripts within a biological sample, i.e. transcriptome, has been the most requested methodology by scientists (Figure 9; Reuter et al., 2015). Its popularity stems from the dramatic reduction of the DNA sequencing costs during the last years and the ability to investigate the functional elements of the genome without an *a priori* knowledge of the sequences being interrogated or the genome itself. The latter factor is a major advantage for research labs working on non-model species with limited or unavailable genomic information.

In the thesis presented here, by the use of next-generation transcriptome and genome sequence data, partly generated in-house and partly downloaded from online repositories, the molecular landscape of many mollusks belonging to all extant

conchiferan and aculiferan class-level taxa and their putative lophotrochozoan allies were screened for the presence of important developmental genes and signalling molecules. Furthermore, the evolution and distribution of key components of the Euarthropoda ecdysis pathway within Mollusca and different metazoans is also presented.

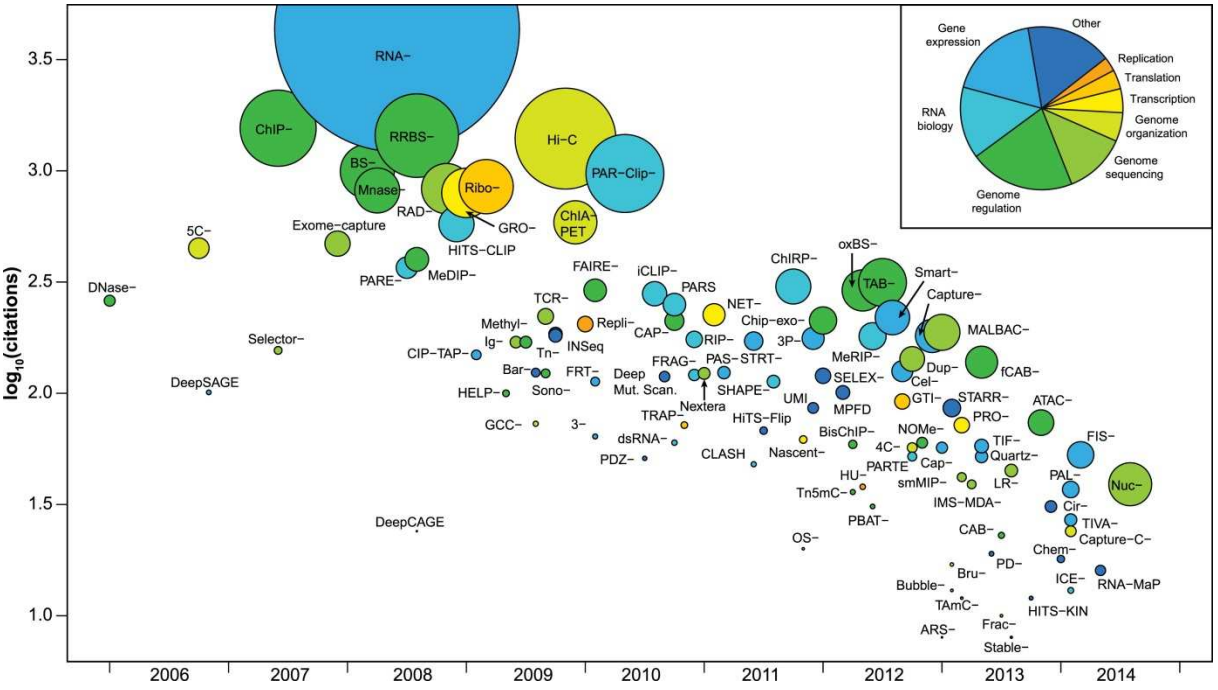


Figure 9 – Overview of the inventory of applications based on next-generation sequencing (from Reuter et al., 2015). The horizontal axis corresponds to the publication´s year of a representative article describing a new method based on next-generation sequencing technology (e.g. RNA-seq, ChIP-seq, Hi-C). The vertical axis shows the number of citations that a specific article received until 2014. The size of the circles is proportional to the publication rate (citation/month), and the colour indicates the different categories (e.g. gene expression, genome regulation, translation). The pie-chart at top right corner shows the colour key as well as the proportion of methods in each category.

3.2 Molluscan developmental gene studies (cf. Manuscript 1)

Manuscript 1 (*"Comparative transcriptomics enlarges the toolkit of known developmental genes in mollusks*") provides a survey of the Hox and ParaHox complement in Mollusca, especially on the understudied Aplacophora (Neomeniomorpha + Chaetodermomorpha). Until recently, whole genome sequencing projects and transcriptomic studies have largely focused on conchiferan

representatives (e.g. Simakov et al., 2013; Albertin et al., 2015; Takeuchi et al., 2016; Sun et al., 2017; Wang et al., 2017; see section 1.4 Molluscan genome sizes and genomics). By inclusion of three aplacophoran species, the neomeniomorphs *Wirenia argentea* and *Gymnomenia pellucida* as well as the chaetodermomorph *Scutopus ventrolineatus*, a solid hypothesis on the evolution of Hox and ParaHox families in Mollusca was proposed.

By comparing the Hox and ParaHox gene distribution in the monophyletic Aculifera and Conchifera clades (Iijima et al., 2006; Biscotti et al., 2014; De Oliveira et al., 2016) it can be inferred that the last common ancestor (LCA) of Mollusca had at least 11 Hox (*Hox1*, *Hox2*, *Hox3*, *Hox4*, *Hox5*, *Lox5*, *Hox7*, *Lox4*, *Lox2*, *Post-1* and *Post-2*) and three ParaHox genes (*Cdx*, *Gsx*, *Xlox*), respectively. These results corroborate estimations of the complement of the lophotrochozoan LCA. (De Rosa et al., 1999; Fröbius et al., 2008; Simakov et al., 2013; Luo et al., 2015).

Duplications and losses of the Hox and ParaHox genes are a common feature in lophotrochozoans (for review Barucca et al., 2016). A recent genomic study has shown the duplication of six Hox genes in the nemertean *Notospermus geniculatus* (*Hox1*, *Hox2*, *Hox3*, *Hox4*, *Hox5* and *Post-2*), and the loss of *Hox5*, *Hox7* and *Post-1* in the phoronid *Phoronis australis* (Luo et al., 2018). In contrast, the ParaHox gene *Xlox* is absent from the nemertean, whereas the phoronid retained the entire set, i.e. *Cdx*, *Gsx*, and *Xlox*. Within Mollusca, no duplicated Hox or ParaHox genes or gene sets have been reported so far; however Iijima et al. (2006), suggested the presence of *Hox5* and *Hox6* duplicates in many aculiferan and conchiferan representatives. Molluscan lineage-specific Hox losses have been identified in the Pacific oyster *Crassostrea gigas*, and the cephalopod *Octopus vulgaris*, in which the Hox complement is composed of ten (loss of *Hox7*) and eight genes (loss of *Hox2*, *Hox3*, *Hox4*), respectively (Zhang et al., 2012; Albertin et al., 2015). Interestingly, apart from *C. gigas*, the full Hox and ParaHox complement have been identified in the remaining pteriomorph bivalve genomes investigated (Takeuchi et al., 2016; Sun et al., 2017; Wang et al., 2017) as well as in the fresh-water mussel *Dreissena rostriformis* (A. Calcino – personal communication, October 2018). The ParaHox gene *Xlox* is absent from the octopus genome (Albertin et al., 2015).

Owing to the highly conserved homeodomain, Hox and ParaHox genes are not ideal targets to infer or reconstruct phylogenies. Nonetheless, the presence of specific amino acid residues, peptide motifs, and different paralog groups provide

useful hints towards the phylogenetic affinities of major animal taxa. This is the case for the Hox genes *Lox5*, *Lox2*, *Lox4*, *Post-1* and *Post-2*, and *Ubx*, *Adb-a* and *Adb-b*, in which their presence support the monophyly of Lophotrochozoa and Ecdysozoa, respectively, and provide several distinct lineage-specific diagnostic signatures on Hox and ParaHox genes (De Rosa et al., 1999; Kobayashi et al., 1999; Telford, 2000; Balavoine et al., 2002). From Manuscript 1, additional lineage-specific signatures were identified in the molluscan *Hox5* (Figure 10) and in the lophotrochozoan ParaHox *Gsx* gene orthologs (De Oliveira et al., 2016). The *Hox5* hexapeptide motif represents the first molluscan-specific signature identified in Hox genes.

The Hox and ParaHox gene sets described in Manuscript 1 provide an important resource for gene expression studies into the function of these genes across Mollusca. Three recent publications from our lab benefited directly from these data (Fritsch et al., 2015, 2016; Wollesen et al., 2018), including the first gene expression study of Hox and ParaHox genes in a series of developmental stages belonging to an aculiferan mollusk, the polyplacophoran *Acanthochitona crinita*. By the combination of the data generated in the three aforementioned studies with gene expression investigations of Hox and ParaHox genes available in other mollusks (Hinman et al., 2003; Lee et al., 2013; Samadi & Steiner, 2009, 2010; Wang et al., 2017), many unknown aspects of the evolution (e.g. manner of expression – spatial and temporal collinearity) and putative function of these genes in the molluscan developmental biology were discovered (see section 1.5 “Molluscan developmental genes and neuropeptides”). Furthermore, the transcriptome databases described in Manuscript 1 were employed as a starting point for other three gene expression studies that focused on *twist*, *otx*, *six3*, and *pax6* developmental genes in neomeniomorphs mollusks (Redl et al., 2016, 2018; Scherholz et al., 2017). These studies constitute the first investigations in the organisation, regulation and putative function of developmental genes in Aplacophora.

3.3 Molluscan proneuropeptide and peptide hormone families (cf. Manuscript 2)

For a long time the “classical” animal model organisms used in neurobiology, evolutionary biology and medical research had been restricted to two major branches

of the metazoan tree: Deuterostomia (e.g. vertebrates, ascidians, and sea urchins), and Ecdysozoa (e.g. arthropods, nematodes). However, this scenario has been slowly changing, and several organisms from the hitherto neglected Lophotrochozoa branch have been established as powerful model systems for molecular studies (Alvarado, 2003; Tessmar-Raible & Arendt, 2003; Fischer & Dorresteijn, 2004; Weisblat & Kuo, 2009). Among them, mollusks, especially gastropods and cephalopods, stood out as models in neurobiology being employed in Nobel Prize winning studies focused on the underlying mechanisms responsible for the propagation of electrical impulse in nerve cells and memory formation (Hodgkin et al., 1952; Kandel, 2001).

Recently, with the understanding of the role of neuropeptides and peptide hormones in a plethora of physiological processes and behaviours in animals, this huge class of signalling molecules has received the attention of many research groups. While there are few recent landmark studies on several animal groups (Jékely, 2013; Mirebeau & Joly, 2013; Zandawala et al., 2017; Thiel et al., 2018), no comparable study exists across the eight class-level taxa of mollusks. Manuscript 2 (*Extensive conservation of the proneuropeptide and peptide hormone complement in mollusks*) fills this gap by presenting a large suite of hitherto unknown peptide families for this important invertebrate phylum.

By looking at all major lineages of mollusks and numerous of their putative lophotrochozoan allies, the minimum proneuropeptide and peptide hormone complement of all individual class-level taxa of mollusks were reconstructed. This study represents the first thorough neuropeptidome survey performed outside of Gastropoda (Veenstra, 2010; Adamson et al., 2015; Ahn et al., 2017; Bose et al., 2017), Bivalvia (Stewart et al., 2014; Zhang et al., 2018) and Cephalopoda (Zatylny-Gaudin et al., 2016) and includes all Aculifera lineages, Monoplacophora and Scaphopoda.

The results point to a widely diverse molluscan complement consisting of a set of more than 60 peptide families with distinct phylogenetic origins, ranging from ancient neuropeptide and hormone precursors (e.g. FMRFamide, cysteine-knot protein hormones, insulin-related peptides) to more recent lophotrochozoan (e.g. neuropeptide KY, fulicin) and lineage-specific families (e.g. PXRXamide, pleurin). One important finding was that the proneuropeptide and hormone complement in the eight investigated class-level taxa is extensively conserved. In other words, the

various degrees of nervous system complexity displayed by mollusks do not correlate with the proneuropeptide and hormone complement found in these animals. Furthermore, the wide taxon sampling of extant lophotrochozoan phyla and the comparative approach showed that many peptide families that had previously been only known from a limited number of taxa are widespread throughout the metazoan tree and are also present in the different molluscan class-level taxa. Together, the results strengthen the notion that the neuropeptidergic components (neuropeptides and hormones) are deeply conserved between very distinct eumetazoan phyla (Jékely, 2013; Mirabeau & Joly, 2013).

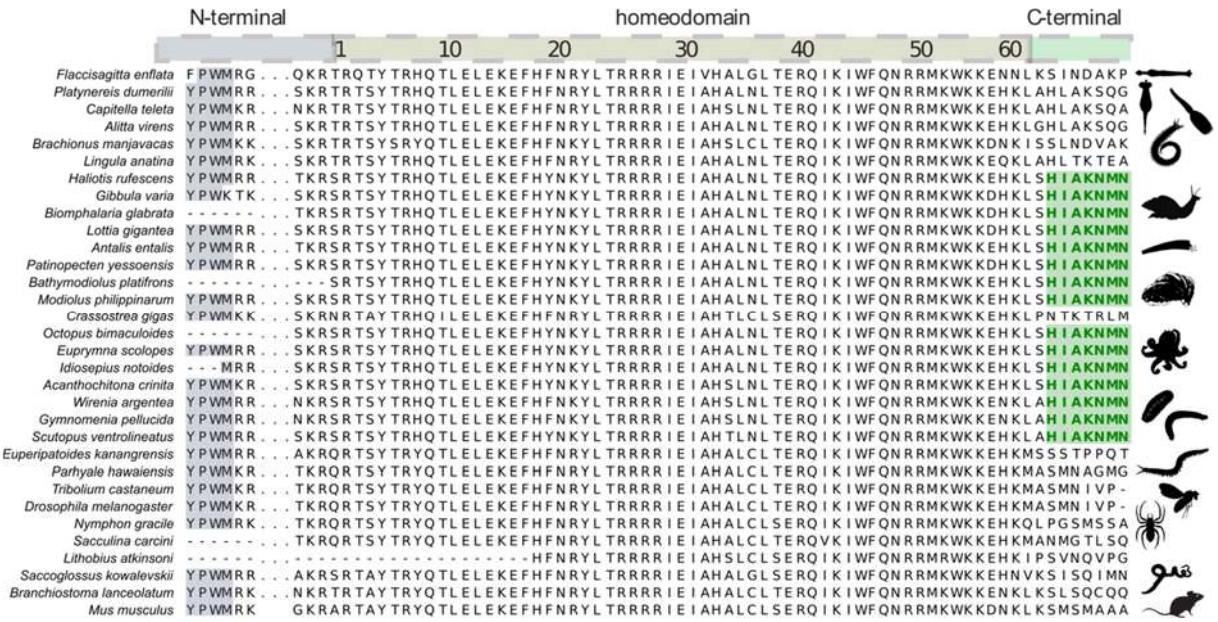


Figure 10 – Updated multiple sequence alignment of eumetazoan *Hox5* orthologs showing the conserved homeodomain, C- and N-terminal regions. Additional *Hox5* orthologs retrieved from publicly available molluscan genomes were added to the original alignment published by De Oliveira et al. (2016). The bilaterian diagnostic peptide motif in the N-terminal region is highlighted in light purple. The conserved molluscan-specific region is highlighted by dark green coloured letters. Note that the *Hox5* molluscan-specific signature is absent in *Crassostrea gigas*. Animal silhouettes were obtained from www.phylopic.org and are either licensed under the Creative Commons Attribution 3.0 Unported or available under public domain (credited images used Chaetognatha: Michelle Site; Rotifera: Diego Fontaneto, Elisabeth A. Herniou, Chiara Boschetti, Manuela Caprioli, Giulio Melone, Claudia Ricci, Timothy G. Barraclough T. Michael Keesey; *Nereis*: B. Duygu Özpolat; Gastropoda: Fernando Carezzano; Scaphopoda: Brockhaus and Efron; Bivalvia: Taro Maeda and David.Monniaux; Aplacophora and Polyplacophora: Noah Schlottman and Casey Dunn;

Onychophora and Hemichordata: Yan Wong; *Drosophila*: Ramiro Morales-Hojas; Mouse: Daniel Jaron).

3.4 Evolution and phylogenetic distribution of the euarthropod ecdysis pathway components (cf. Manuscript 3)

During the investigations of the proneuropeptides and peptide hormones described in Manuscript 2, one peptide cluster composed of bivalve, scaphopod, and arthropod representatives stood out as a particularly interesting result. Thorough analysis of the sequences for the presence of conserved motifs and sequence similarity searches showed that these constitute *eh*, i.e. eclosion hormone, orthologs. Eclosion hormone is a peptide part of a metabolic pathway responsible to trigger a species-specific behavioural pattern in insects that culminates in moulting. Furthermore, other three peptides shown in recent studies to be key components of the "ecdysis behaviour" in insects and crustaceans (prothoracicotropic hormone, crustacean cardioactive peptide and bursicon) were identified in Manuscript 2 to be homologous to many lophotrochozoan sequences. The discovery in Lophotrochozoa of many important peptides involved in the euarthropod ecdysis signalling pathway indicates a deeper evolutionary origin for these neuropeptidergic components, and raises interesting questions about the distribution and role of these signalling molecules outside of Euarthropoda.

The genetic basis underlying the evolution of novel traits, including ecdysis, is one of the major topics in modern EvoDevo, however it is still poorly understood (Moczek, 2008). Genomic and transcriptomic approaches have become the standard state-of-the-art method to integrate molecular information with the origin and diversification of novel features and behavioural traits (Anholt & Mackay, 2004; Smith et al., 2015; Conith et al., 2018). By the employment of sensitive similarity searches against a wide range of metazoan molecular databases the distribution of the major gene components of the Euarthropoda ecdysis pathway was explored. The results show the presence of prothoracicotropic hormone (and its paralog *trunk*), eclosion hormone, crustacean cardioactive, and bursicon peptide ligand-receptors in many protostomian, deuterostome, and non-bilaterian animals. At least in Panarthropoda (Euarthropoda + Onychophora + Tardigrada), the molecular framework underpinning ecdysis seems to be conserved, and supported by already preexisting genes

widespread in many “non-moulting” animals. Eclosion hormone and bursicon trace back to the last common ancestor of Cnidaria and Bilateria, whereas crustacean cardioactive peptide has its origins in the bilaterian stem. The prothoracicotropic hormone, an important trigger for the initiation of ecdysis in insects, is a recent paralog of the ancient extracellular molecule Trunk, which is present in cnidarians, ctenophores, and bilaterians (see section 2.3 Manuscript 3 - Evolution and phylogenetic distribution of the insect ecdysis pathway components). The finding of a trunk-like peptide in the comb jelly *Mnemiopsis* that bears close sequence similarity to the insect prothoracicotropic hormone, represents the first report of a metazoan neuropeptide homologous to a ctenophore signalling peptide and may thus hint towards a common origin of metazoan neural systems.

Taken together, the results indicate a recruitment and co-option of preexisting genes into euarthropod ecdysozoan moulting. This scenario corroborates a general trend in the evolution of novel traits, in which major phenotypic innovations arise by the tinkering of ancient molecular components deployed in different contexts. The employment of ancient gene sets into new complex phenotypes has been identified in many animals, and it is recognised as a major force promoting evolutionary changes in development and metabolic systems (True & Carrol, 2002; Moczek & Rose, 2009; Aguilera et al., 2017; Hilgers et al., 2018).

3.5 References

Adamson, K. J., Wang, T., Zhao, M., Bell, F., Kuballa, A. V., Storey, K. B., & Cummins, S. F. (2015). Molecular insights into land snail neuropeptides through transcriptome and comparative gene analysis. *BMC genomics*, *16*(1), 308.

Aguilera, F., McDougall, C., & Degnan, B. M. (2017). Co-option and de novo gene evolution underlie molluscan shell diversity. *Molecular biology and evolution*, *34*(4), 779-792.

Albertin, C. B., Simakov, O., Mitros, T., Wang, Z. Y., Pungor, J. R., Edsinger-Gonzales, E., et al. (2015). The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*, *524*(7564), 220.

Alvarado, A. S. (2003). The freshwater planarian *Schmidtea mediterranea*: embryogenesis, stem cells and regeneration. *Current opinion in genetics & development*, *13*(4), 438-444.

Anholt, R. R., & Mackay, T. F. (2004). Quantitative genetic analyses of complex behaviours in *Drosophila*. *Nature Reviews Genetics*, *5*(11), 838.

Ahn, S. J., Martin, R., Rao, S., & Choi, M. Y. (2017). Neuropeptides predicted from the transcriptome analysis of the gray garden slug *Deroceras reticulatum*. *Peptides*, *93*, 51-65.

Balavoine, G., de Rosa, R., & Adoutte, A. (2002). Hox clusters and bilaterian phylogeny. *Molecular phylogenetics and evolution*, *24*(3), 366-373.

Barucca, M., Canapa, A., & Biscotti, M. A. (2016). An overview of Hox genes in Lophotrochozoa: Evolution and functionality. *Journal of developmental biology*, *4*(1), 12.

Biscotti, M. A., Canapa, A., Forconi, M., & Barucca, M. (2014). Hox and ParaHox genes: a review on molluscs. *genesis*, *52*(12), 935-945.

Bose, U., Suwansa-ard, S., Maikaeo, L., Motti, C. A., Hall, M. R., & Cummins, S. F. (2017). Neuropeptides encoded within a neural transcriptome of the giant triton snail *Charonia tritonis*, a Crown-of-Thorns Starfish predator. *Peptides*, *98*, 3-14.

Conith, M. R., Hu, Y., Conith, A. J., Maginnis, M. A., Webb, J. F., & Albertson, R. C. (2018). Genetic and developmental origins of a unique foraging adaptation in a Lake Malawi cichlid genus. *Proceedings of the National Academy of Sciences*, 201719798.

De Oliveira, A. L., Wollesen, T., Kristof, A., Scherholz, M., Redl, E., Todt, C. (2016). Comparative transcriptomics enlarges the toolkit of known developmental genes in mollusks. *BMC genomics*, *17*(1), 905.

De Rosa, R., Grenier, J. K., Andreeva, T., Cook, C. E., Adoutte, A., Akam, M. (1999). Hox genes in brachiopods and priapulids and protostome evolution. *Nature*, *399*(6738), 772.

Fischer, A., & Dorresteijn, A. (2004). The polychaete *Platynereis dumerilii* (Annelida): a laboratory animal with spiralian cleavage, lifelong segment proliferation and a mixed benthic/pelagic life cycle. *Bioessays*, *26*(3), 314-325.

Fritsch, M., Wollesen, T., De Oliveira, A. L., & Wanninger, A. (2015). Unexpected co-linearity of Hox gene expression in an aculiferan mollusk. *BMC evolutionary biology*, *15*(1), 151.

Fritsch, M., Wollesen, T., & Wanninger, A. (2016). Hox and ParaHox gene expression in early body plan patterning of polyplacophoran mollusks. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *326*(2), 89-104.

Fröbius, A. C., Matus, D. Q., & Seaver, E. C. (2008). Genomic organization and expression demonstrate spatial and temporal Hox gene colinearity in the lophotrochozoan *Capitella* sp. I. *PLoS One*, *3*(12), e4004.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333.

Hilgers, L., Hartmann, S., Hofreiter, M., & von Rintelen, T. (2018). Novel genes, ancient genes, and gene co-option contributed to the genetic basis of the radula, a molluscan innovation. *Molecular biology and evolution*, *35*(7), 1638-1652.

Hinman, V. F., O'Brien, E. K., Richards, G. S., & Degnan, B. M. (2003). Expression of anterior Hox genes during larval development of the gastropod *Haliotis asinina*. *Evolution & development*, *5*(5), 508-521.

Hodgkin, A. L., Huxley, A. F., & Katz, B. (1952). Measurement of current-voltage relations in the membrane of the giant axon of *Loligo*. *The Journal of physiology*, *116*(4), 424-448.

Iijima, M., Akiba, N., Sarashina, I., Kuratani, S., & Endo, K. (2006). Evolution of Hox genes in molluscs: a comparison among seven morphologically diverse classes. *Journal of molluscan studies*, *72*(3), 259-266.

Jékely, G. (2013). Global view of the evolution and diversity of metazoan neuropeptide signaling. *Proceedings of the National Academy of Sciences*, *110*(21), 8702-8707.

Kandel, E. R. (2001). The molecular biology of memory storage: a dialogue between genes and synapses. *Science*, *294*(5544), 1030-1038.

Kobayashi, M., Furuya, H., & Holland, P. W. (1999). Evolution: Dicyemids are higher animals. *Nature*, *401*(6755), 762.

Lee, P. N., Callaerts, P., de Couet, H. G., & Martindale, M. Q. (2003). Cephalopod Hox genes and the origin of morphological novelties. *Nature*, *424*(6952), 1061.

Levy, S. E., & Myers, R. M. (2016). Advancements in next-generation sequencing. *Annual review of genomics and human genetics*, *17*, 95-115.

Luo, Y. J., Takeuchi, T., Koyanagi, R., Yamada, L., Kanda, M., Khalturina, M., et al. (2015). The *Lingula* genome provides insights into brachiopod evolution and the origin of phosphate biomineralization. *Nature communications*, *6*, 8301.

Luo, Y. J., Kanda, M., Koyanagi, R., Hisata, K., Akiyama, T., Sakamoto, H. (2018). Nemertean and phoronid genomes reveal lophotrochozoan evolution and the origin of bilaterian heads. *Nature ecology & evolution*, *2*(1), 141.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376.

Metzker, M. L. (2010). Sequencing technologies- the next generation. *Nature reviews genetics*, *11*(1), 31.

Mirabeau, O., & Joly, J. S. (2013). Molecular evolution of peptidergic signaling systems in bilaterians. *Proceedings of the national academy of sciences*, *110*(22), E2028-E2037.

Moczek, A. P. (2008). On the origins of novelty in development and evolution. *BioEssays*, *30*(5), 432-447.

Moczek, A. P., & Rose, D. J. (2009). Differential recruitment of limb patterning genes during development and diversification of beetle horns. *Proceedings of the National Academy of Sciences*, pnas-0809668106.

Redl, E., Scherholz, M., Wollesen, T., Todt, C., & Wanninger, A. (2016). Cell proliferation pattern and *twist* expression in an aplacophoran mollusk argue against segmented ancestry of Mollusca. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *326*(7), 422-436.

Redl, E., Scherholz, M., Wollesen, T., Todt, C., & Wanninger, A. (2018). Expression of *six3* and *otx* in Solenogastres (Mollusca) supports an ancestral role in bilaterian anterior-posterior axis patterning. *Evolution & development*, *20*(1), 17-28.

Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular cell*, *58*(4), 586-597.

Samadi, L., & Steiner, G. (2009). Involvement of Hox genes in shell morphogenesis in the encapsulated development of a top shell gastropod (*Gibbula varia* L.). *Development genes and evolution*, *219*(9-10), 523-530.

Samadi, L., & Steiner, G. (2010). Expression of Hox genes during the larval development of the snail, *Gibbula varia* (L.)—further evidence of non-colinearity in molluscs. *Development genes and evolution*, *220*(5-6), 161-172.

Scherholz, M., Redl, E., Wollesen, T., de Oliveira, A. L., Todt, C., & Wanninger, A. (2017). Ancestral and novel roles of Pax family genes in mollusks. *BMC evolutionary biology*, *17*(1), 81.

Simakov, O., Marletaz, F., Cho, S. J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., et al. (2013). Insights into bilaterian evolution from three spiralian genomes. *Nature*, *493*(7433), 526.

Smith, C. R., Helms Cahan, S., Kemena, C., Brady, S. G., Yang, W., Bornberg-Bauer, E. (2015). How do genomes create novel phenotypes? Insights from the loss of the worker caste in ant social parasites. *Molecular biology and evolution*, *32*(11), 2919-2931.

Stewart, M. J., Favrel, P., Rotgans, B. A., Wang, T., Zhao, M., Sohail, M. (2014). Neuropeptides encoded by the genomes of the Akoya pearl oyster *Pinctata fucata* and Pacific oyster *Crassostrea gigas*: a bioinformatic and peptidomic survey. *BMC genomics*, *15*(1), 840.

Sun, J., Zhang, Y., Xu, T., Zhang, Y., Mu, H., Zhang, Y., et al. (2017). Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nature ecology & evolution*, *1*(5), 0121.

Takeuchi, T., Koyanagi, R., Gyoja, F., Kanda, M., Hisata, K., Fujie, M., et al. (2016). Bivalve-specific gene expansion in the pearl oyster genome: implications of adaptation to a sessile lifestyle. Zoological letters, 2(1), 3.

Telford, M. J. (2000). Turning Hox “signatures” into synapomorphies. *Evolution & development*, *2*(6), 360-364.

Tessmar-Raible, K., & Arendt, D. (2003). Emerging systems: between vertebrates and arthropods, the Lophotrochozoa. *Current opinion in genetics & development*, *13*(4), 331-340.

Thiel, D., Franz-Wachtel, M., Aguilera, F., Hejnol, A., & Wray, G. (2018). Xenacoelomorph neuropeptidomes reveal a major expansion of neuropeptide systems during early bilaterian evolution. *Molecular Biology and Evolution*.

True, J. R., & Carroll, S. B. (2002). Gene co-option in physiological and morphological evolution. *Annual review of cell and developmental biology*, *18*(1), 53-80.

Veenstra, J. A. (2010). Neurohormones and neuropeptides encoded by the genome of *Lottia gigantea*, with reference to other mollusks and insects. *General and comparative endocrinology*, *167*(1), 86-103.

Wang, S., Zhang, J., Jiao, W., Li, J., Xun, X., Sun, Y., et al. (2017). Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nature ecology & evolution*, *1*(5), 0120.

Weisblat, D. A., & Kuo, D. H. (2009). *Helobdella* (Leech): a model for developmental studies. *Cold Spring Harbor Protocols*, *2009*(4), pdb-emo121.

Wollesen, T.; Rodríguez, M.S.; de Oliveira, A.L.; Wanninger, A. Staggered Hox expression is more widespread among mollusks than previously anticipated. *Proceedings of the Royal Society B*, *285*: 20181513.

Zandawala, M., Moghul, I., Guerra, L. A. Y., Delroisse, J., Abylkassimova, N., Hugall, A. F. (2017). Discovery of novel representatives of bilaterian neuropeptide families and reconstruction of neuropeptide precursor evolution in ophiuroid echinoderms. *Open biology*, *7*(9), 170129.

Zatylny-Gaudin, C., Cornet, V., Leduc, A., Zanuttini, B., Corre, E., Le Corguillé. (2015). Neuropeptidome of the Cephalopod *Sepia officinalis*: identification, tissue

mapping, and expression pattern of neuropeptides and neurohormones during egg laying. *Journal of proteome research*, *15*(1), 48-67.

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., et al. (2012). The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, *490*(7418), 49.

Zhang, M., Wang, Y., Li, Y., Li, W., Li, R., Xie, X., et al. (2018). Identification and characterization of neuropeptides by transcriptome and proteome analyses in a bivalve mollusc *Patinopecten yessoensis*. *Frontiers in Genetics*, 9.

4. CONCLUSION AND KEY FINDINGS

The scientific contributions presented in this thesis constitute the first *in silico* large-scale comparative work focused on developmental gene, proneuropeptide and peptide hormone complement in all extant conchiferan and aculiferan class-level taxa of mollusks. Furthermore, the evolutionary trajectories of key molecular components involved in the ecdysis pathway in Euarthropoda were reconstructed. The key findings presented in this thesis are the following:

- The repertoire of developmental genes (e.g. Hox, ParaHox, Wnt, Hedgehog, Notch), proneuropeptides and peptide hormones in the extant class-level taxa of Mollusca is comparable to many metazoan model organisms.

- The identification of a near-complete set of Hox genes in a scaphopod, in aplacophorans, and in a polyplacophoran mollusk indicates that the last common ancestor (LCA) of Mollusca had 11 Hox genes.

- There are lineage-specific signatures in the molluscan *Hox5* and lophotrochozoan *Gsx* gene.

- The repertoire of proneuropeptide and peptide hormones in Mollusca is composed of ~60 peptide families with distinct origins in the metazoan tree.

- The proneuropeptide and peptide hormone complement in the eight extant class-level taxa of Mollusca is extremely conserved, irrespective of the various degrees of complexity in the nervous system.

- Important components of the Euarthropoda ecdysis pathway are broadly distributed throughout the metazoan tree (i.e., including taxa that do not moult). The results suggest that ancient metazoan genes were involved in the moulting behaviour in the last common ancestor of Panarthropoda and may have acquired this function already at the base of Ecdysozoa.

- The presence of a *trunk-like* homolog in the genome of *Mnemiopsis leidyi* represents the first ctenophore peptide with homology to a known metazoan neuropeptide (i.e. the arthropod prothoracicotropic hormone).

In summary, this work describes a large catalog of hitherto unknown gene and gene families in Mollusca, providing an important framework to understand some of the underlying molecular mechanisms revolving around body plan diversification, neuroendocrine regulation, neurobiology, and homeostasis maintenance in Mollusca. The strong evolutionary focus of the study gave rise to novel insights on molluscan, lophotrochozoan, and metazoan evolution. These include the evolution of various peptide (e.g. egg-laying hormone, diuretic hormone 44), and developmental gene families (e.g. Hox, ParaHox, Lophohog) in Mollusca and distinct lophotrochozoans, as well as the reconstruction of the developmental gene, neuropeptide and hormone complement in the LCA of Mollusca and the extant class-level taxa.

5. APPENDIX

5.1 Relevant coauthorships

5.1.1 Published

1. Fritsch, M., Wollesen, T., De Oliveira, A. L., & Wanninger, A. (2015). Unexpected co-linearity of Hox gene expression in an aculiferan mollusk. BMC evolutionary biology, 15(1), 151.

2. Kristof, A., de Oliveira, A. L., Kolbin, K. G., & Wanninger, A. (2016a). Neuromuscular development in Patellogastropoda (Mollusca: Gastropoda) and its importance for reconstructing ancestral gastropod bodyplan features. Journal of Zoological Systematics and Evolutionary Research, 54(1), 22-39.

3. Kristof, A., de Oliveira, A. L., Kolbin, K. G., & Wanninger, A. (2016b). A putative species complex in the Sea of Japan revealed by DNA sequence data: a study on *Lottia* cf. *kogamogai* (Gastropoda: Patellogastropoda). Journal of Zoological Systematics and Evolutionary Research, 54(3), 177-181.

4. Scherholz, M., Redl, E., Wollesen, T., de Oliveira, A. L., Todt, C., & Wanninger, A. (2017). Ancestral and novel roles of Pax family genes in mollusks. BMC evolutionary biology, 17(1), 81.

5. Wollesen, T.; Rodríguez, M.S.; de Oliveira, A.L., Wanninger, A. Staggered Hox expression is more widespread among mollusks than previously anticipated. Proceedings of the Royal Society B, 285: 20181513

5.1.2 In preparation or under review:

1. Calcino, A; Oleg, S; de Oliveira, AL, Schwaha, T, Wanninger, A. The quagga mussel genome and the evolution of fresh water tolerance. In preparation.

5.2 Curriculum Vitae

1. Personal Data

Name:	André Luiz de Oliveira
Date of birth:	07 August 1986
City/State:	Jaboticabal/São Paulo
Country:	Brazil
Marital status:	Married
Parents:	Marcia Roberta Costa Claro de Oliveira (Mother) Oswaldo Luiz de Oliveira (Father)
Address:	Kurzbauergasse, 3 Top 18, 1020 Vienna - Austria
Phone number:	+43 660 4682369
Email	andre.luiz.de.oliveira@univie.ac.at

2. Education

Since Dez-2013	PhD Candidate in Biology at University of Vienna Department of Integrative Zoology Advisor: Prof. Andreas Wanninger Work Title: Developmental genes, proneuropeptides and peptide hormones: an *in-silico* approach
2010 – 2012	Master's degree in Science from University of Sao Paulo (USP, Brazil) Department of the Parasitology - Program of Host-Pathogen Interactions Advisor: Arthur Gruber Work Title: GenSeed-HMM: development of a platform for sequence reconstruction and application on next-generation sequencing data Date of approval: 14th of August 2012.
2005 – 2009	Bachelor's degree in Biological Science from Federal University of Ouro Preto (UFOP, Brazil) Date of approval: 14th of August 2009.

3. Research projects

Since Dez-2013 **Comparative transcriptomics in Mollusca**

Description: The main goal of the project is to perform an integrated and comparative transcriptomic analysis with novel and unpublished data obtained by massive parallel sequencing (i.e. 454 and Illumina sequencing), derived from different developmental stages in almost all class-level taxa of mollusks, in order to elucidate and characterise the transcriptome machinery of these species.

Keywords: Evo-devo; Bioinformatics; Transcriptomics; Next-generation Sequencing; Phylogenetic Analysis; Hox; Para-Hox; Transcription-factors; Neuropeptide; Hormone.

2010-2012 **GenSeed-HMM: development of a platform for sequence reconstruction and application on next-generation sequencing data**

Description: GenSeed-HMM is a program that implements a seed-driven progressive assembly, an approach to reconstruct specific unassembled data, starting from short nucleotide or protein seed sequences or profile Hidden Markov Models (HMM). The program can use any one of a number of sequence assemblers (Newbler, CAP3, Velvet, AbySS, SOAPdenovo). Assembly is performed in multiple steps and relatively few reads are used in each cycle, consequently the program demands low computational resources. The GenSeed-HMM tool was validated using real life data originated from prokariotic, eukaryotic and metagenomic samples, produced by different next-generation sequencing platforms.
Keywords: Evolution; Metagenomics; Bioinformatics; Viruses; Next-generation Sequencing; Sequence Assembly; Software development; PERL language.

4. Important Courses and Training

2016 to present Bioinformatics for non-bioinformaticians: how to build and analyse a Transcriptome – 300191UE (Two weeks theorical and practical course)
University of Vienna (Organiser/Teacher)

2015 Introduction to Animal Gene Expression (Two weeks course)
University of Vienna, Vienna, Austria (Student)

Molecular phylogenetics (Semestral Course)
University of Vienna, Vienna, Austria (Student)

2011 Evolutionary Genomics (Semestral Course)
University of Sao Paulo, Sao Paulo, Brazil (Student)

2010 Development of Applications for Bioinformatics (Semestral course)
University of Sao Paulo, Sao Paulo, Brazil (Student)

Fundamentals of Molecular Evolution of Microorganisms and Phylogenetic Reconstruction (Semestral Course)
University of Sao Paulo, Sao Paulo, Brazil (Student)

2009 Summer Bioinformatics course at University of Sao Paulo (45 hours course)
University of Sao Paulo, Sao Paulo, Brazil (Student)

2008 Introduction to Bioinformatics and its tools (8 hours course)
First Brazilian School of Bioinformatics (Student)

RNA Bioinformatics (8 hours course)
First Brazilian School of Bioinformatics (Student)

5. Bibliography

1. Calcino, A; Oleg, S; de Oliveira, AL, Schwaha, T, Wanninger, A. **The quagga mussel genome and the evolution of fresh water tolerance**. *In preparation*.
2. De Oliveira, A.L., Calcino, A., Wanninger A. **Evolution and phylogenetic distribution of the euarthropod ecdysis pathway components**. *In preparation*:*
3. De Oliveira, AL; Calcino A; Wanninger, A. **Extensive conservation of the proneuropeptide peptide hormone complement in mollusks**. *In review* in Scientific Reports.*
4. Wollesen, T; Rodríguez, MS; de Oliveira, AL; Wanninger, A. **Staggered Hox expression is more widespread among mollusks than previously anticipated**. *In press* in Proceedings of the Royal Society B. 285: 20181513
5. Scherholz, M; Redl, E; Wollesen, T; De Oliveira, AL; Todt, C; Wanninger, A. **Ancestral and novel roles of Pax family genes in mollusks**. BMC Evolutionary Biology, v. 17, p. 81, 2017.
6. De Oliveira, AL; Wollesen, T; Kristof, A; Scherholz, M; Redl, E; Todt, C; Bleidorn, C; Wanninger, A. **Comparative transcriptomics enlarges the toolkit of known developmental genes in mollusks**. BMC Genomics, v.17, p.905, 2016.*
7. Alves, JMP; De Oliveira, AL; Sandberg, TOM; Gallego, JLM; Toledo, MAF; de Moura; EMM; Menhert, DU; Oliveira, LS; Durham, AM; Zanotto; Reyes, A; Gruber A. GenSeed-HMM: **A tool for progressive assembly using profile HMMs as seeds and its application in *Alpavirinae* viral discovery from metagenomic data**. Frontiers in Microbiology, v.4, 2016.*
8. Kristof, A; De Oliveira, AL; Kolbin, KG; Wanninger, A. **A putative species complex in the Sea of Japan revealed by DNA sequence data: a study on *Lottia* cf. *kogamogai* (Gastropoda: Patellogastropoda)**. Journal of Zoological Systematics and Evolutionary Research, v. 1439, 2016.
9. Fritsch, M; Wollesen, T; De Oliveira, AL; Wanninger, A. **Unexpected co-linearity of Hox gene expression in an aculiferan mollusk**. BMC Evolutionary Biology, v. 15, p.151, 2015.
10. Kristof, A; De Oliveira, AL; Kolbin KG; Wanninger, A. **Neuromuscular development in Patellogastropoda (Mollusca: Gastropoda) and its importance for reconstructing ancestral gastropod bodyplan features**. Journal of Zoological Systematics and Evolutionary Research, v. 24, 2015.

*First authorship

6. Scholarships and Grants

Dez 2013 – Nov 2018 Scholarship for Doctoral Studies at University of Vienna
Ciência sem Fronteiras (Science without Borders)

project # 6090-13/3

Aug 2010 – Aug 2012 Scholarship for Master Studies at University of Sao Paulo
Fundação de Amparo à Pesquisa de São Paulo (FAPESP)
project # 2010/04609-1

2008 – 2009 Scholarship for Undergraduate Studies University Student at Federal University of Ouro Preto (UFOP)
Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG)

7. Conferences, symposiums and meetings

1. De Oliveira, AL; Calcino, A.; Wanninger, A. **The evolutionary history of key components of the eclosion signaling pathway**. 2018. 7th meeting of the European Society for Evolutionary Developmental Biology (EED). Galway, Ireland (Poster presentation).
2. Salamanca, D.; de Oliveira, AL; Calcino, A; Wanninger, A. **Hox Genes & Body Plan Evolution on a Bivalve Mollusk**. 2018. 7th meeting of the European Society for Evolutionary Developmental Biology (EED). Galway, Ireland (Poster presentation).
3. De Oliveira, AL; Wanninger, A. **Deciphering molluscan developmental genes by comparative transcriptomics**. 2016. Neptune Conference – Lisbon, Portugal (Speaker – 10 minutes talk).
4. De Oliveira, AL, Wanninger, A. **investigating the developmental genes toolkit of molluscs by next-generation sequencing**. 2016. Talk at Naturhistorisches Museum Wien – Vienna, Austria. (Speaker – 20 minutes talk).
5. De Oliveira, AL, Wanninger, A. **Comparative transcriptome analyses in molluscs reveal novel genes and gene families.** 2015. COSB PhD Seminar – Vienna, Austria (Speaker – 30 minutes talk).
6. De Oliveira, AL; Sobreira, TJP; Toledo, MAF; Zanotto, PMA; Gruber, A. **Sequence Reconstruction Based on Profile HMM Seeds: Implementation and Proof of Principle For The Diagnosis of Novel Viruses**. 2010. X-Meeting 2010 – Ouro Preto, Brasil (Poster presentation).

8 Awards

2018 Best speaker Award - Science Day Biology 2018 with the talk “Key components of the neuropeptide signaling pathway involving the eclosion hormone are found throughout the Metazoa - Faculty of Life Sciences of the University of Vienna

2018 Runner-up for best poster entittled "The evolutionary history of key components of the eclosion signaling pathway" at the 7th Meeting of the European Society for Evolutionary Developmental Biology

2015	Honorable Mention on the session Software and Development during the “X-Meeting 2015 – 11th International Conference of the AB3C + Symposium of Bioinformatics” held in Sao Paulo – Brazil, November 3 to 6 2015 due the work entitled: GenSeed-HMM: a tool for progressive assembly using profile HMMs as seeds – application in virus discovery of *Alpavirinae* from metagenomic data.

9. Languages

	Portuguese	Spanish	English	German
Writing	Native	Basic	Good	Intermediate
Speaking	Native	Basic	Good	Intermediate
Listening	Native	Good	Good	Good
Reading	Native	Good	Good	Good