



universität
wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

„Evolutionary origins and functional impact of novel genes on cephalopod developmental and morphological innovations“

verfasst von / submitted by

Elena Andrea Ritschard

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

Doctor of Philosophy (PhD)

Wien, 2022 / Vienna 2022

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on the student
record sheet:

UA 794 685 437

Dissertationsgebiet lt. Studienblatt /
field of study as it appears on the student record sheet:

Biologie

Betreut von / Supervisor:

Assoz. Prof. Dr. Oleg Simakov

Mitbetreut von / Co-Supervisor:

*This thesis is dedicated to my beautiful intercultural support network: mi familia and friends,
many who became like family abroad.*

Acknowledgements

Thanks to Oleg Simakov and Graziano Fiorito for their scientific supervision along this journey, for supporting me on important decisions during the PhD and motivating me to carry on. Special thanks, Oleg, for believing since the beginning in my potential and for encouraging my scientific growth through networking, conferences and experiences abroad. I would also like to thank many great minds that helped me with their scientific input: Bob Zimmerman, Nicolas Robert, Carrie Albertin, Pamela Imperadore and Giovanna Ponte. Big thanks to all the members of the Simakov Lab, Uli Technau's, Grigory Ghenikhovich's and Qi Zhou's groups for the insightful conversations and fun happy hours. A huge thank you to Hannah Schmidbaur, my dearest friend and colleague, with whom I shared one of the most beautiful personal growth experiences during this PhD journey. I am grateful to have had you on my side all this time! I thank also Christina Holzinger, for her motivation and the fun brain-storming sessions. I enjoyed a lot working together on the C2H2 paper! I am also very grateful to Sönke Johnsen and Bob Fitak for their guidance on the GPCR project and, together with Lori Schweikert, for their reassuring words in times of doubts.

Not enough words could express the gratitude I have towards the support network I built during these years. Friends who have made my days brighter with every sip of cafecito con chisme, game evenings with funny drawings, puppy cuddles, dancing nights, Yppenplatz summer evenings, transcendental conversations, new restaurant discoveries, trips around the world, long-distance calls, emotional reenounters en la tierrita, and I could just continue. Mi familia hermosa, que me ha apoyado en cada loca decisión y siempre vio lo mejor en mí aún cuando yo no podía verlo. Los amo. Ligia y Marcela, que me brindaron la ayuda profesional que mi salud mental requería: gracias por guiarme y ayudarme a hacer de esta montaña rusa algo invaluable en mi vida. La familia Robles-Mora-Schützenhofer, que me adoptó con tanto genuino cariño y enorme bondad. To every person that has been from the beginning, has arrived in the middle or only towards the end of this process: THANKS! You helped me overcome the difficult moments of the PhD and showed me, even if you didn't notice, how important it is to value the little, not-so-little, things of life. I'm here, living this very best moment thanks to your beautiful existence.

Finally, I'd like to thank the funding agencies that contributed to my PhD: the Austrian Science Fund (FWF) and the Stazione Zoologica Anton Dohrn in Naples, Italy.

Table of Contents

Abstract	6
Zusammenfassung	7
General introduction	9
Genome evolution driving organismal novelties: an overview	10
Towards a multi-scale approach to investigate organismal novelties	13
Cephalopods: model systems to study novelty evolution and its genomic background	14
Thesis aim and chapters overview	16
References	18
Chapter 1: Genomic signatures of G-protein-coupled receptor expansions reveal functional transitions in the evolution of cephalopod signal transduction	24
Abstract	25
Introduction	26
Material and methods	28
Results	31
Discussion	38
Conclusion	41
References	42
Chapter 2: Co-evolutionary signatures of gene family expansions and their contribution to cephalopod nervous system evolution	49
Abstract	50
Introduction	51
Material and methods	52
Results	56
Discussion	66
Closing remarks	69
References	71
Chapter 3: Combined contribution of orphan and expanded genes at the developmental and organismal level in cephalopods	75
Introduction	76
Materials and methods	77
Results and discussion	81
Conclusion	93
References	93

General discussion	97
Contribution of expanded adhesive proteins in neural wiring of the octopus axial nerve cord	97
Unravelling cephalopod neuronal network connectivity through contrasting C2H2 and GPCR expansion histories	99
Orphan genes and C2H2 zinc finger expansions in the evolution of octopus brain regulatory networks	100
Concluding remarks	101
References	101
Appendix	104
Supplementary material in Chapter 1	104
Supplementary material in Chapter 2	112
References	132

Abstract

Possessing the largest and most elaborated nervous system among invertebrates and lineage-specific organs associated with the adaptation to different marine environments, coleoid cephalopods (*i.e.*, octopuses, squids, cuttlefishes) provide exciting systems to investigate how developmental and morphological novelties evolve. With the increasing sequencing of cephalopod genomes, the genomic signatures underlying the evolution of such novelties have started to be unraveled but how exactly genomic innovations translate into organismal organization remains largely unresolved. In this thesis, the evolution of cephalopod genomic novelties at the coding sequence level (*i.e.*, two gene family expansions and orphan genes) was investigated, aiming to assess their collective functional impact at the organismal level. Using phylogenetics, 4 expansions of G protein-coupled receptors (GPCRs) and 19 of the Cys2His2- type (C2H2) zinc finger transcription factors were characterized. Gene expression and positive selection analyses revealed that the expanded GPCRs in cephalopods evolved - possibly by functional adaptation- new expression domains outside of the brain, in contrast with their conserved counterparts. This result differs to the expanded C2H2s, which maintained an ancestral expression domain in nervous tissues, mainly the brain. However, weighted gene co-expression network analysis (WGCNA) showed that genes from the same C2H2 expansions did not belong to the same co-expression modules suggesting putative functional divergence of the paralogues at the regulatory level within a same tissue. Furthermore, cephalopod-, octopus/squid- and species-specific orphan genes were characterized using orthologous gene family construction approaches. Although orphan genes were found to be expressed in a wide range of tissues, consistent expression of orthologous orphan genes and expanded C2H2s was only found in nervous tissues (both central and peripheral) of the 3 octopus' species studied. Lastly, I investigated the developmental expression of gene novelties that highlighted co-regulation properties of these novelties. Altogether, the genes here characterized may be of interest in further studies to understand the joint contribution of gene novelties and gene family expansions in the development of the large and uniquely structured cephalopod nervous system. This first approach to cephalopod gene novelties also offers a resource for future functional studies to understand their contribution to the evolution of these intriguing animals.

Zusammenfassung

Mit dem größten und ausgereiftesten Nervensystem unter den wirbellosen Tieren sowie stammesspezifischen Organen, die mit der Anpassung an verschiedene Meeresumgebungen verbunden sind, bieten Coleoiden Kopffüßer, (d. h. Oktopusse, Kalmare und Sepien) spannende Systeme, um zu untersuchen, wie sich entwicklungsbedingte und morphologische Neuerungen entwickeln. Mit der zunehmenden Sequenzierung der Genome von Kopffüßern wurden die genomischen Signaturen, die der Entwicklung solcher Neuerungen zugrunde liegen, allmählich entschlüsselt. Aber die Frage, wie genau sich genomische Innovationen in der Organisation des Organismus niederschlagen, bleibt weitgehend unbeantwortet. In dieser Arbeit wurde die Entwicklung von genomischen Neuerungen bei Kopffüßern auf der Ebene der kodierenden Sequenzen (d. h. zwei Expansionen von Genfamilien und orphan-genes) untersucht, um ihre kollektiven funktionellen Auswirkungen auf der Ebene des Organismus zu bewerten. Mithilfe der Phylogenetik wurden vier Expansionen von G-Protein-gekoppelten Rezeptoren (GPCRs) und 19 der Zinkfinger-Transkriptionsfaktoren vom Typ Cys2His2 (C2H2) charakterisiert. Genexpressions- und Positivselektionsanalysen ergaben, dass die erweiterten GPCRs in Kopffüßern - möglicherweise durch funktionelle Anpassung - neue Expressionsdomänen außerhalb des Gehirns entwickelt haben, im Gegensatz zu ihren konservierten Vertretern. Dieses Ergebnis unterscheidet sich von dem der erweiterten C2H2s, die eine angestammte Expressionsdomäne in Nervengewebe, hauptsächlich im Gehirn, beibehalten haben. Die Analyse des Gen-Koexpressions-netzwerks (WGCNA) zeigte jedoch, dass Gene aus denselben C2H2-Expansionen nicht zu denselben Koexpressionsmodulen gehören, was auf eine mögliche funktionelle Divergenz der paralogenen Gene auf regulatorischer Ebene innerhalb desselben Gewebes schließen lässt. Darüber hinaus wurden orphan-genes von Kopffüßern, Oktopusse und Kalmare mit Hilfe der Konstruktion orthologer Genfamilien charakterisiert. Obwohl orphan-genes in einer Vielzahl von Geweben eine Expression aufweisen, wurde eine konsistente Expression orthologer orphan-genes und expandierter C2H2s nur in Nervengewebe (sowohl zentral als auch periphar) der drei untersuchten Oktopusarten gefunden. Schließlich wurde die entwicklungsbedingte Ausprägung von Gen-Neuheiten untersucht, welche die Ko-Regulationseigenschaften dieser Neuheiten hervorhoben. Insgesamt könnten die hier charakterisierten Gene in weiteren Studien von Interesse sein, um den gemeinsamen Beitrag von Gen-Neuheiten und Gen-Familien-Expansionen bei der Entwicklung des großen und einzigartig strukturierten Nervensystems von Kopffüßern zu verstehen. Diese erste Annäherung an Gen-Neuheiten bei Kopffüßern bietet auch eine Ressource für zukünftige Funktionsstudien mit dem Ziel, deren Beitrag zur Evolution dieser faszinierenden Tiere zu verstehen.

General introduction

This introduction is mostly based on a review of my authorship published in the peer-reviewed journal BioEssays.

Citation:

E. A. Ritschard, B. Whitelaw, C. B. Albertin, I. R. Cooke, J. M. Strugnell, O. Simakov, Coupled Genomic Evolutionary Histories as Signatures of Organismal Innovations in Cephalopods: Co-evolutionary Signatures Across Levels of Genome Organization May Shed Light on Functional Linkage and Origin of Cephalopod Novelties. *BioEssays* 41(12) (2019), doi: 10.1002/bies.201900073.

Genome evolution driving organismal novelties: an overview

Complex transitions in body plan organization, like the origin of multicellularity, the appearance of bilateral symmetry and the evolution of centralized nervous systems, have defined animal evolution [1]. Understanding the mechanisms behind such transitions and how these have driven the diversity of life forms on earth is an ongoing quest in the evolutionary biology field. Since the development of so-called Next Generation Sequencing (NGS) methods (*i.e.*, platforms that perform massive parallel sequencing of genome fragments, which can be ultimately assembled by sequence overlap) and after more than twenty years of technological improvements, investigating big genomes has become more affordable and efficient [2]. This has resulted in an increasing availability of genomes covering most lineages across the tree of life and has opened the opportunity to approach that “big question” at the molecular level through comparative genomics. A first hypothesis on the genomic background of complex organismal transitions suggested that the latter were related to a “step-wise” increase in genome complexity. However, increasing evidence has shifted that postulate towards an already complex ur-metazoan ancestor (*e.g.*, refs. [3-8]). For instance, evidence for conserved gene order (also called synteny), protein domain architecture and gene structure (*e.g.*, exon–intron composition) has been found in early branching lineages, [3,5,9] suggesting these features were present already in the metazoan ancestor. Crucial cell signaling and developmental gene family complements, like Wnts, TGFbeta, Hedgehog, Hox, and Sox, are also highly conserved in most metazoans, including sponges [3] and cnidarians [5]. These have been proposed to be “true” animal gene novelties due to their absence outside Metazoa. [1,8] Altogether, these ancestral genomic features have largely contributed to the evolution of modern animals as building blocks that were modified in different lineages. Such modifications can be, but are not limited to, changes of ancestral synteny by **(a) genome re-arrangements** or expansions of preexisting gene repertoires via **(b) gene duplication** (*i.e.*, from independent gene families to whole genome duplications, as in the vertebrate lineage [10]). Novel genetic material can also arise by **(c) orphan gene formation** (*i.e.*, genes with no homologues outside of a taxon). These evolutionary events are what is referred here to as levels of genome organization, and each contributes differently to the evolution of individual metazoan lineages and their organismal novelties [11]. Gene duplications and orphan gene emergence are the two levels of genome organization deeply addressed in this thesis.

(a) Genome rearrangements

From sponges to bilaterians, metazoan genomes have maintained regions with conserved local gene order and chromosomal-scale gene linkages over a wide evolutionary time span,

a phenomenon also termed “synteny” [12-14]. Synteny conservation has been suggested to be relevant in the regulatory context as it maintains interactions between regulatory elements and their target genes [15-17]. Re-arrangements of syntenic blocks impose a high impact on disrupting functional and regulatory genomic units and could have been thus constrained throughout the course of animal evolution. However, comparative genomic studies have found that novel syntenies generated by genome re-arrangements coincide with complex evolutionary transitions, like the appearance of bilateral symmetry in metazoans [13]. The extent to which genome re-arrangements contribute to evolutionary novelties by forming new regulatory interactions is just beginning to be understood (*e.g.*, [18]) and represents nowadays a field of study worth further exploration. These changes in regulatory landscape are also of relevance to this thesis, as these are suggested here to generate the “groundfloor” for other levels of genome organization (like gene duplications and orphan genes) to change over evolutionary time, interact with each other and consequently contribute altogether to organismal transitions and lineage-specific novelties [11].

(b) Gene duplications

Changes in the gene complement of genomes, from single genes to whole genome duplications, have been suggested to be major drivers of evolution and thus the focus of many studies [19,20]. For instance, gene duplication events driven by transposable elements have contributed to brain evolution within mammals [21]. Following duplication, copies may share at the beginning same regulatory properties and/or functions, and this redundancy will permit their subsequent divergence without altering the ancestral function. Two hypotheses have been posed on why redundant genes are kept and how these can diverge [22]. On one hand, sub-functionalization proposes that the two paralogues divide the work of the ancestral function by *e.g.*, dividing the repertoire of regulatory interactions between the two copies. On the other hand, neo-functionalization posits that by the acquirement of novel regulatory pathways the new copy can gain a new function. Thus, duplication of both regulatory sequences (*i.e.*, non-coding genes) and/or their target protein coding genes can generate new interactions in a regulatory network and contribute to the evolution of new functions [23].

Moreover, how paralogues diverge and the extent to which these play a role in regulatory network evolution also depends on the duplication dynamics. Duplications can appear tandemly (*i.e.*, adjacent to each other) or interspersed either within the same chromosome (*i.e.*, intrachromosomal) or in different ones (*i.e.*, interchromosomal) [24]. Tandem duplications can arise as errors of repair mechanisms of double-strand DNA breaks and the resulting maintenance of close proximity and thus ancestral interactions could favor sub- over neo-functionalization [11]. While interspersed duplications can be also a product of DNA strand

misrepair and reverse transcription of mRNAs [25], genome architecture modifications (*e.g.*, genome rearrangements) after a duplication event may result in translocation of paralogues to other genomic regions. Interspersed duplications can favor neo-functionalization, as novel regulatory interactions that were not present in the ancestral gene can be established in this process of new “territory colonization”.

(c) Orphan gene emergence

Genes lacking homologues outside of a taxon in question have been found to represent up to 20% of every genome sequenced [26,27]. These have been termed orphan or also taxonomically restricted genes (TRGs) and can be characterized using orthology inference methods based on sequence similarity searches and phylogenetics, as well as synteny-based approaches [28]. Their importance in the evolution of novelties has been increasingly investigated and various studies have evidenced that the function of orphan genes is tightly linked to lineage-specific adaptations to the environment (*e.g.*, an antifreeze glycoprotein orphan gene in Arctic codfishes, vital to survive in freezing sea temperatures [29]) and developmental processes (*e.g.*, TRGs genes of vertebrates involved in the growth of the axon-covering myelin sheaths structure [30]).

How orphan genes can emerge has been explained by two main hypotheses. The first hypothesis posits that after gene duplication, paralogues can undergo rapid divergence beyond a threshold of sequence similarity, thus losing the homology fingerprint of their gene family of origin [27]. The latter represents a plausible scenario for TRGs emergence, but requires extensive substitutions which, if not facilitated by *e.g.*, high transposons activity, can occur only rarely. The second mechanism, previously thought to be even less likely to occur, is *de novo* gene emergence from ancestral non-coding sequences [28]. Nowadays, increasing evidence supports this scenario of TRGs origination. In a *de novo* gene birth scenario, a non-genic DNA sequence containing equivocal splice -, polyadenylation – or transcription starting sites can be erroneously transcribed to RNA. At first, this molecule could function as non-coding antisense or structural RNA and acquire afterwards an open reading frame (ORF) that translate for a new protein [27]. Another plausible scenario, which has been proven *e.g.*, in the Arctic codfishes’ antifreeze glycoprotein gene AFGP [29], is the acquisition of an open reading frame before a promoter region, thus preceding transcription of the newly born orphan gene [28]. Regardless of the mechanism behind the origin of these orphan genes, it is important to note that their maintenance in the genome cannot occur without respective adaptations at the regulatory level. As in the case of duplicated genes evolving via neo-functionalization, the contribution of orphan genes to novel functions may be related to the generation of new regulatory interactions [11].

Towards a multi-scale approach to investigate organismal novelties

As mentioned above, the evolution of phenotypic novelties is associated with various levels of genome organization, where each plays a role in the modification and extension of the existing regulatory landscapes that define physiological systems, organs, tissues, and cell types [11,31-33]. However, it remains often difficult to directly associate genomic changes with their morphological and functional outcomes [11]. This endeavor becomes increasingly difficult for more “complex” systems, such as the genomic basis of vertebrate brain development and organization [34] or that of other highly elaborated nervous systems, such as those found in cephalopod molluscs. [35] While functional experiments (e.g., gene knock-down/out) are absolutely necessary in dissecting such systems, novel comparative genomic approaches are also required to quantify and reveal the key contributors among the various genomic characters to morphological and/or developmental innovations at the organismal level.

Given the increasing understanding of genome evolution across metazoans, it is becoming necessary to (re)consider the organismal context of genomic novelties [11]. In particular, genomic novelties affecting a single organ or cell type are likely to play a role in the evolution of functionally linked cells, tissues and/or structures [36,37]. In Ritschard *et al.*, 2019 [11], we propose that the evolution of functional linkages between structures in the organismal context can be understood under the framework of co-evolution. In the original definition, co-evolution is implicated when interacting populations diverge reciprocally (therefore the alternative term is co-divergence) as in, for example, host and symbiont taxa (e.g., refs. [38,39]). However, this term has been also applied to study protein interaction evolution. At the simplest level this involves, and can be measured as, co-evolution between genes of different gene families by, for example, phylogenetic profiling [40] or the mirrortree approach [41]. How this concept of co-evolution can be linked to the genomic basis of organismal novelties, remains unexplored. In our *BioEssays* review [11], we explore how different combinations of levels of genome organization have contributed to the origin of novelties in the organismal context. We ask whether, and how, the individual genomic evolutionary signatures across and within tissues and organs can show signs of co-evolution, reflecting functional linkage. To do so, a more complete approach across several levels of genomic organization is required. Such an approach should ideally take into account the contribution of all genomic characters (*i.e.*, levels of genome organization), such as the origin of orphan genes, gene duplications, non-coding elements and genome rearrangements [11]. The three chapters of this thesis represent some of the first studies that seek to find functional linkages between genomic novelties in a

group of animals with high potential of being models to study the genomic basis behind evolutionary novelty: cephalopods.

Cephalopods: model systems to study novelty evolution and its genomic background

Coleoid cephalopods (cuttlefishes, squids, and octopuses) are a class of mollusks with a highly divergent body plan from their molluscan relatives, in which the foot developed into flexible arms and the shell was almost completely lost. Cephalopods are also considered to possess the most elaborate invertebrate nervous system, a notion that is likely linked to their complex behavioral repertoire, their active predatory mode of life, and their ability to rapidly manipulate skin texture and color for camouflage and display [11]. Their brain size resembles that of vertebrates in number of cells and their brain to body ratio is even bigger than in vertebrates [31]. Cephalopods have also evolved convergently camera-like eyes with vertebrates, making them particularly interesting model systems to study the mechanisms behind convergent evolution. Additionally, their adaptation to a wide range of marine environments is associated with a suite of lineage-specific morphological novelties that have evolved across the subclass such as light and adhesive organs in *Euprymna scolopes* and *Idiosepius paradoxus*, respectively, accessory nidamental glands and sucker ring teeth in squids and cuttlefishes, toxin producing salivary glands in blue-ringed octopus' species of the *Hapalochlaena* genus, among others [31]. Cephalopods have thus been the target of several research areas [42,43] such as behavioral biology [44], phylogeography [45], development [46,47], neurobiology [35,48] and, more recently, genomics.

In 2015, the first cephalopod genome from the Californian two-spot octopus, *Octopus bimaculoides*, was sequenced [49]. Since then, coleoid cephalopods genomes of the two main lineages Decapodiformes (*i.e.*, squids and cuttlefishes) and Octopodiformes (*i.e.*, octopuses) have been published at an accelerating rate. Altogether, those studies revealed various levels of genome organization such as orphan genes, gene family expansions, RNA editing, and extensive genome rearrangements, as possible drivers for the evolution of cephalopod morphological novelties [11] (Figure 1). With the recent publication of one representative of the sister lineage of Coleoidea, namely *Nautilus pompilius*, the evidence hints to a coleoid origin of the aforementioned genomic novelties [50].

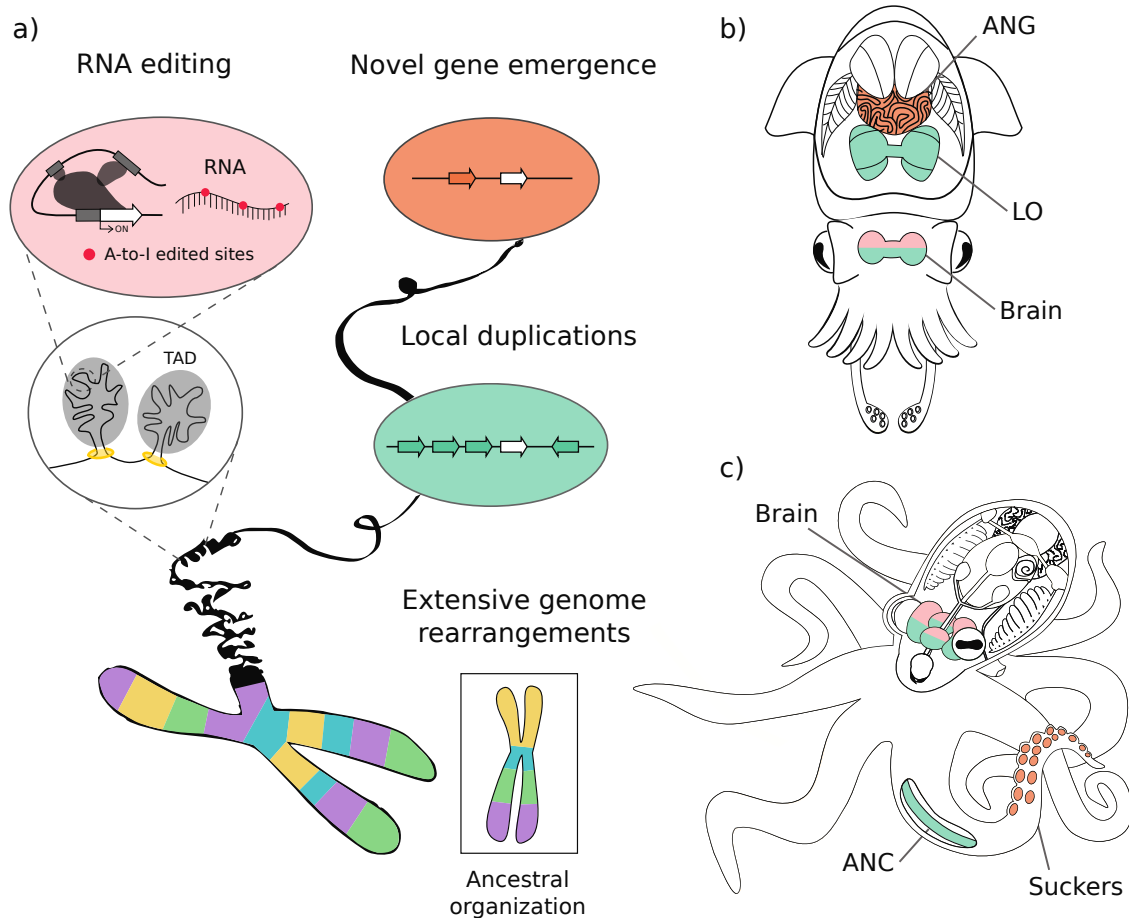


Figure 1. Different levels of genome organization and their contribution to morphological novelties in cephalopods. a) Individual levels of genome organization (*i.e.*, genome rearrangements, local duplications, novel gene emergence, and RNA editing) and their contribution at organismal level (see text) in b) *E. scolopes* (illustration by Hannah Schmidbaur) and c) *O. bimaculoides* are color-coded. ANG: accessory nidamental gland; LO: light organ; ANC: axial nerve cord. [11]

Three gene family expansions were firstly described in *O. bimaculoides* and further evidenced in other cephalopod genomes as well: C2H2 zinc finger transcription factors, protocadherins and G-protein coupled receptors. The expanded C2H2 zinc fingers and protocadherin gene families in both vertebrates and cephalopods have been argued to have facilitated their neuronal diversity [49,51-53], while the GPCR expansion has been suggested to contribute to sensory functions [49]. The Hawaiian bobtail squid (*Euprymna scolopes*) genome [54], has shown that its light organ (LO), a specialized symbiotic structure that houses luminescent bacteria [55], expresses tandemly duplicated genes (*i.e.*, reflectins). Another symbiotic organ present in many squid species, the accessory nidamental gland (ANG), was found to be enriched in taxonomically restricted genes [54]. Enrichment of orphan genes was found for other tissues in octopus' species as well, such as suckers in *Octopus bimaculoides* [49]. Furthermore, cephalopods employ epigenetic mechanisms such as extensive RNA editing to

diversify their transcripts [56]. RNA editing was found to underlie temperature adaptation in potassium channels in polar octopuses [57] and to be enriched in transcripts expressed in nervous tissues, similar to vertebrates [56,58]. Finally, chromosomal fusions and fissions shaped large-scale genome reorganization in coleoid cephalopods [53]. These genome rearrangements split conserved gene clusters found across different metazoan levels (e.g., mollusks, lophotrochozoan, bilaterians) providing novel cephalopod-specific gene regulatory domains [18].

Cephalopods provide a fruitful, yet almost entirely unexplored model system to study the role of functional linkages between genomic innovations shaping morphology, development, and behavior. Investigating genomic signatures behind those adaptations may help reveal fundamental insights into the co-evolution of novel features in the organismal context [11]. Following evidence on how independent gene duplication events of olfactory receptors and transcription factors have shaped the olfactory system architecture in vertebrates [58], in Ritschard *et al.*, 2019 [11], we propose that signatures of co-evolution between the C2H2, GPCRs and protocadherin expanded families can help us to address neural architecture and innovation in cephalopods. Moreover, whether cephalopod orphan genes contribute to the evolution of new or, rather, act as an addition to the already existing core gene regulatory networks remains unexplored. In addition to the warranted investigation of those genes' functions, their correlated emergence and expression with other types of novelty (e.g., gene duplications) may help establish functional links between and within tissues [11]. Altogether, emerging genomic data indicates the importance of looking simultaneously across different levels of genome organization, as many of them have been shown to be associated with cephalopod novelties, and co-evolutionary signatures may help us reveal their functional connectivity.

Thesis aim and chapters overview

In this thesis I explored how cephalopod novel genes (*i.e.*, the GPCRs and C2H2s gene family expansions and orphan genes) have contributed to the evolution of organismal novelties at the morphological and developmental levels in these fascinating mollusks. To do so I firstly used comparative genomic approaches to characterize the lineage-specificity of the genes. To designate expansion groups in the GPCRs and C2H2s and distinguish when in the cephalopod evolutionary timeline duplication events happened, I used phylogenetic construction of the gene families using the whole repertoire of a wide metazoan taxon sampling, including both octopus' and squid genomes. To define and characterize the

evolutionary origins of orphan genes I used a BLAST- and phylogenetic-based orthology inference method (*i.e.*, OrthoFinder) and further filtering steps. I then studied the putative function of these resulting novel genes by exploring their expression in adult tissues of different octopus and squid species and at different early developmental stages of the Hawaiian bobtail squid (*E. scolopes*). Finally, using co-expression analyses and the orthology results I also explored putative regulatory correlations between the different gene novelties. Altogether, my main aim was to start unravelling the synergically impact that gene family expansions and orphan genes have had in the evolution of cephalopod morphological and developmental innovations.

In chapter 1, I focused my analyses on the GPCR gene family and characterized 2 cephalopod-specific and 2 octopus-specific expansions. I discussed, by the hand of adult tissue transcriptomic data and positive selection analyses, the possible scenario of functional adaptation of the paralogues by acquiring new expression domains outside of the brain - the expression domain of the conserved GPCRs.

In chapter 2, I explored a second gene family expansion in cephalopods: the C2H2 zinc finger transcription factors. I found, in contrast to GPCRs, no transition of expression domains between conserved and expanded genes. I proposed, however, that the functional divergence of paralogues might be at the regulatory level within a same tissue, as genes originating from the same expansion events were assigned to different co-expression modules despite their shared expression in central nervous tissues.

Lastly, in chapter 3 I characterized cephalopod-, octopus- and squid-specific orphan genes and explored their contribution to both adult tissues of different species and during the early development of *E. scolopes*. I discussed how TRGs contributed mainly to late stages (pre-hatching embryos) and to a wide range of tissues in the adults. I highlighted the consistent nervous-system expression of orphan genes and C2H2s in the three octopus species analyzed and suggest further exploration of putative regulatory interactions between these groups of novel genes.

The resulting genomic and transcriptomic characterization of thousands of novel genes in this thesis provides a base for future functional experiments in cephalopods to further investigate if/how genomic novelties co-evolved. I believe that my PhD brings new perspectives while opens up more questions on how gene novelties have contributed to the development of fascinating cephalopod features, which make these animals interesting to study in the wide context of complex metazoan evolutionary transitions.

References

1. Simakov, O. and Kawashima, T. (2017). Independent evolution of genomic characters during major metazoan transitions. *Dev Biol.* 2017 Jul 15;427(2):179-192. 10.1016/j.ydbio.2016.11.012.
2. Giani, A.M., Gallo, G.R., Gianfranceschi, L. and Formenti, G. (2019). Long walk to genomics: history and current approaches to genome sequencing and assembly. *Comput Struct Biotechnol J*, 18, pp. 9-19. 10.1016/j.csbj.2019.11.002
3. Domazet-Loso, T., Brajković, J. and Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* 23(11):533-9. 10.1016/j.tig.2007.08.014.
4. Simakov, O. *et al.* (2013). Insights into bilaterian evolution from three spiralian genomes. *Nature.* 493(7433):526-31. 10.1038/nature11696.
5. Chénais, B., Caruso, A., Hiard, S. and Casse, N. (2012) The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene.* 509(1):7-15. 10.1016/j.gene.2012.07.042.
6. Srivastava M., *et al.* (2010). The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature.* 466(7307):720-6. 10.1038/nature09201.
7. Putnam, N.H., *et al.* (2007). Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science.* 317(5834):86-94. 10.1126/science.1139158.
8. Paps, J. and Holland, P.W.H. (2018). Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. *Nat Commun* 9, 1730. 10.1038/s41467-018-04136-5.
9. Raible F., *et al.* (2005). Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. *Science.* 310:1325–1326. 10.1126/science.1119089.
10. Kassahn, K. S., Dang, V. T., Wilkins, S. J., Perkins, A. C. and Ragan, M. A. (2009). Evolution of gene function and regulatory control after whole-genome duplication: Comparative analyses in vertebrates. *Genome Res.* 19: 1404–1418. 10.1101/gr.086827.108
11. Ritschard, E. A., Whitelaw, B., Albertin, C. B., Cooke, I. R., Strugnelli, J. M., and Simakov, O. (2019). Coupled Genomic Evolutionary Histories as Signatures of Organismal Innovations in Cephalopods: Co-evolutionary Signatures across Levels of Genome Organization May Shed Light on Functional Linkage and Origin of Cephalopod Novelty. *Bioessays* 41 (12), 1900073. 10.1002/bies.201900073

12. Simakov O, *et al.* (2022) Deeply conserved synteny and the evolution of metazoan chromosomes. *Sci Adv.* 8(5):eabi5884. 10.1126/sciadv.abi5884.
13. Robert, N.S.M., *et al.* (2022). Emergence of distinct syntenic density regimes is associated with early metazoan genomic transitions. *BMC Genomics* 23, 143. 10.1186/s12864-022-08304-2
14. Zimmermann, B., *et al.* (2019). Ancient animal genome architecture reflects cell type identities. *Nat Ecol Evol* 3, 1289–1293. 10.1038/s41559-019-0946-7
15. Irimia, M., *et a.* (2012). Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Res.* 22(12):2356-67. 10.1101/gr.139725.112.
16. Engström, P.G., Ho Sui, S.J., Drivenes, O., Becker, T.S. and Lenhard, B. (2007). Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res.* 17(12):1898-908. 10.1101/gr.6669607.
17. Kikuta, H., *et al.* (2007). Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* 17(5):545-55. 10.1101/gr.6086307.
18. Schmidbaur, H., *et al.* (2022). Emergence of novel cephalopod gene regulation and expression through large-scale genome reorganization. *Nat Commun* 13, 2172. 10.1038/s41467-022-29694-7
19. Carroll, S.B. (2005). Evolution at two levels: on genes and form. *PLoS Biol.* 3(7):e245. 10.1371/journal.pbio.0030245.
20. Moriyama, Y. and Koshida-Takeuchi, K. (2018). Significance of whole-genome duplications on the emergence of evolutionary novelties. *Briefings in Functional Genomics*, 17(5), 329–338. 10.1093/bfpg/ely007
21. Ferrari, R., Grandi, N., Tramontano, E. and Dieci, G. (2021). Retrotransposons as Drivers of Mammalian Brain Evolution. *Life* , 11, 376. 10.3390/life11050376
22. Innan, H. and Kondrashov, F. (2010). The evolution of gene duplications: Classifying and distinguishing between models. *Nat. Rev. Genet.* 11, pp. 97–108. 10.1038/nrg2689
23. Voordeckers, K., Pougach, K. and Verstrepen, K.J. (2015). How do regulatory networks evolve and expand throughout evolution? *Curr Opin Biotechnol.* 34:180-8. 10.1016/j.copbio.2015.02.001.
24. Mendivil Ramos, O. and Ferrier, D.E. (2012). Mechanisms of Gene Duplication and Translocation and Progress towards Understanding Their Relative Contributions to Animal Genome Evolution. *Int J Evol Biol.* 2012:846421. 10.1155/2012/846421.
25. Kaessmann, H., Vinckenbosch, N. and Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* 10, 19–31. 10.1038/nrg2487.

26. Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. and Bosch, T. C. G. (2009). More than just orphans: are taxonomically-restricted genes important in evolution? *Trends in Genetics*, 25, 9, 404-413. 10.1016/j.tig.2009.07.006.
27. Tautz, D. and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nat Rev Genet* 12, 692–702. 10.1038/nrg3053
28. Van Oss, S.B. and Carvunis, A-R. (2019). *De novo* gene birth. *PLoS Genet* 15(5): e1008160. 10.1371/journal.pgen.1008160.
29. Zhuang, X., Yang, C., Murphy, K.R. and Cheng, C.-H.C. (2019). Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. *Proc. Natl. Acad. Sci.* 116, 4400–4405. 10.1073/pnas.1817138116
30. Gould, R., Oakley, T., Goldstone, J., Dugas, J., Brady, S., and Gow, A. (2008). Myelin sheaths are formed with proteins that originated in vertebrate lineages. *Neuron Glia Biology*, 4(2), 137-152. 10.1017/S1740925X09990238
31. Wagner, G.P., Pavlicev, M. and Cheverud, J.M. (2007). The road to modularity. *Nat Rev Genet.* 8(12):921-31. 10.1038/nrg2267.
32. Davidson, E.H. (2010). Emerging properties of animal gene regulatory networks. *Nature.* 468(7326):911-20. 10.1038/nature09645.
33. Acemel, R.D., Maeso, I. and Gómez-Skarmeta, J.L. (2017). Topologically associated domains: a successful scaffold for the evolution of gene regulation in animals. *Wiley Interdiscip Rev Dev Biol.* 6(3). 10.1002/wdev.265.
34. Enard, W. (2016). The Molecular Basis of Human Brain Evolution. *Curr Biol.* 26(20):R1109-R1117. 10.1016/j.cub.2016.09.030.
35. Shigeno, S., Andrews, P.L.R., Ponte, G. and Fiorito, G. (2018). Cephalopod Brains: An Overview of Current Knowledge to Facilitate Comparison With Vertebrates. *Front. Physiol.* 9. 10.3389/fphys.2018.00952.
36. Rowe, T. (1996). Coevolution of the mammalian middle ear and neocortex. *Science.* 273(5275):651-4. 10.1126/science.273.5275.651.
37. Kaas, J.H. (2008). The evolution of the complex sensory and motor systems of the human brain. *Brain Res Bull.* 75(2-4):384-90. 10.1016/j.brainresbull.2007.10.009.
38. Clark, M.A., Moran, N.A., Baumann, P. and Wernegreen, J.J. (2000). Cospeciation between bacterial endosymbionts (*Buchnera*) and a recent radiation of aphids (*Uroleucon*) and pitfalls of testing for phylogenetic congruence. *Evolution.* 54(2):517-25. 10.1111/j.0014-3820.2000.tb00054.x.
39. Moeller, A.H., Caro-Quintero, A., Mjungu, D., Georgiev, A.V., Lonsdorf, E.V., Muller, MN, Pusey, A.E., Peeters, M., Hahn, B.H. and Ochman, H. (2016). Cospeciation of gut microbiota with hominids. *Science.* 353(6297):380-2. 10.1126/science.aaf3951.

40. Ruano-Rubio, V., Poch, O. and Thompson, J.D. (2009). Comparison of eukaryotic phylogenetic profiling approaches using species tree aware methods. *BMC Bioinformatics* 10, 383. 10.1186/1471-2105-10-383
41. Zhou, H. and Jakobsson, E. (2013). Predicting Protein-Protein Interaction by the Mirrortree Method: Possibilities and Limitations. *PLoS ONE* 8(12): e81100. 10.1371/journal.pone.0081100.
42. O'Brien, C.E., Roumbedakis, K. and Winkelmann, I.E. (2018). The Current State of Cephalopod Science and Perspectives on the Most Critical Challenges Ahead From Three Early-Career Researchers. *Front Physiol.* 9:700. 10.3389/fphys.2018.00700.
43. Kröger, B., Vinther, J. and Fuchs, D. (2011). Cephalopod origin and evolution: A congruent picture emerging from fossils, development and molecules: Extant cephalopods are younger than previously realised and were under major selection to become agile, shell-less predators. *Bioessays.* 33(8):602-13. 10.1002/bies.201100001.
44. Hanlon, R., and Messenger, J. (2018). *Cephalopod Behaviour* (2nd ed.). Cambridge: Cambridge University Press. 10.1017/9780511843600
45. Winkelmann, I., *et al.* (2013). Mitochondrial genome diversity and population structure of the giant squid *Architeuthis*: genetics sheds new light on one of the most enigmatic marine species. *Proceedings. Biological Sciences.* 280(1759):20130273. 10.1098/rspb.2013.0273.
46. Wollesen, T., *et al.* (2015). The ParaHox gene *Gsx* patterns the apical organ and central nervous system but not the foregut in scaphopod and cephalopod mollusks. *EvoDevo* 6, 41 (2015). 10.1186/s13227-015-0037-z.
47. Koenig, K.M, Sun, P., Meyer, E. and Gross, J.M. (2016). Eye development and photoreceptor differentiation in the cephalopod *Doryteuthis pealeii*. *Development.* 143(17):3168-81. 10.1242/dev.134254.
48. Shomrat, T., Turchetti-Maia, A., Stern-Mentch, N., Basil, J. and Hochner, B. (2015). The vertical lobe of cephalopods: an attractive brain structure for understanding the evolution of advanced learning and memory systems. *J. Comp. Physiol. A* 201 947–956. 10.1007/s00359-015-1023-6.
49. Albertin, C., *et al.* (2015). The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* 524, 220–224. 10.1038/nature14668.
50. Huang, Z., *et al.* (2022). Genomic insights into the adaptation and evolution of the nautilus, an ancient but evolving “living fossil”. *Molecular Ecology Resources*, 22, 15–27. 10.1111/1755-0998.13439

51. Styfhals, R., Seuntjens, E., Simakov, O., Sanges, R. and Fiorito, G. (2019). In silico Identification and Expression of Protocadherin Gene Family in *Octopus vulgaris*. *Front Physiol.* 9:1905. 10.3389/fphys.2018.01905.
52. Wang, Z.Y. and Ragsdale, C.W. (2017). Cadherin genes and evolutionary novelties in the octopus. *Semin Cell Dev Biol.* 69:151-157. 10.1016/j.semcdb.2017.06.007.
53. Albertin, C.B., *et al.* (2022). Genome and transcriptome mechanisms driving cephalopod evolution. *Nat Commun* 13, 2427. 10.1038/s41467-022-29748-w.
54. Belcaid M., *et al.* (2019). Symbiotic Organs Shaped by Distinct Modes of Genome Evolution in Cephalopods. *Proc. Natl. Acad. Sci. U. S. A.* 116, 3030–3035. 10.1073/pnas.1817322116
55. Peyer, S.M., Pankey, M.S., Oakley, T.H. and McFall-Ngai, M.J. (2014). Eye-specification genes in the bacterial light organ of the bobtail squid *Euprymna scolopes*, and their expression in response to symbiont cues. *Mech Dev.* 131:111-26. 10.1016/j.mod.2013.09.004.
56. Liscovitch-Brauer, N., Alon, S., Porath, H. T., Elstein, B., Unger, R., Ziv, T., Admon, A., Levanon, E. Y., Rosenthal, J. J. C. and Eisenberg, E. (2017). Trade-off between transcriptome plasticity and genome evolution in cephalopods. *Cell* 169, 191–202.e11. 10.1016/j.cell.2017.03.025.
57. Garrett, S. and Rosenthal, J.J. (2012). RNA editing underlies temperature adaptation in K⁺ channels from polar octopuses. *Science.* 335(6070):848-51. 10.1126/science.1212795.
58. Alon, S., Garrett, S. C., Levanon, E. Y., Olson, S., Graveley, B. R., Rosenthal, J.J.C. and Eisenberg, E. (2015) The majority of transcripts in the squid nervous system are extensively recoded by A-to-I RNA editing *eLife* 4:e05198. 10.7554/eLife.05198

Chapter 1: Genomic signatures of G-protein-coupled receptor expansions reveal functional transitions in the evolution of cephalopod signal transduction

This chapter has been published in the peer-reviewed journal *Proceedings of the Royal Society B*.

Citation:

E. A. Ritschard, R. R. Fitak, O. Simakov, S. Johnsen, Genomic signatures of G-protein-coupled receptor expansions reveal functional transitions in the evolution of cephalopod signal transduction. *Proc. R. Soc. B Biol. Sci.* 286 (2019), doi:10.1098/rspb.2018.2929.

Genomic signatures of G-protein-coupled receptor expansions reveal functional transitions in the evolution of cephalopod signal transduction

Elena A. Ritschard^{1,2}, Robert R. Fitak², Oleg Simakov¹, Sönke Johnsen²

¹Department of Molecular Evolution and Development, University of Vienna, Austria,

²Biology Department, Duke University, Durham, NC, USA

Corresponding authors:

Elena Ritschard elena.ritschard@univie.ac.at

Oleg Simakov oleg.simakov@univie.ac.at

Abstract

Coleoid cephalopods show unique morphological and neural novelties, such as arms with tactile and chemosensory suckers and a large complex nervous system. The evolution of such cephalopod novelties has been attributed at a genomic level to independent gene family expansions, yet the exact association and the evolutionary timing remain unclear. In the octopus genome, one such expansion occurred in the G-protein-coupled receptors (GPCRs) repertoire, a superfamily of proteins that mediate signal transduction. Here, we assessed the evolutionary history of this expansion and its relationship with cephalopod novelties. Using phylogenetic analyses, at least two cephalopod- and two octopus-specific GPCR expansions were identified. Signatures of positive selection were analysed within the four groups, and the locations of these sequences in the *Octopus bimaculoides* genome were inspected. Additionally, the expression profiles of cephalopod GPCRs across various tissues were extracted from available transcriptomic data. Our results reveal the evolutionary history of cephalopod GPCRs. Unexpanded cephalopod GPCRs shared with other bilaterians were found to be mainly nervous tissue specific. By contrast, duplications that are shared between octopus and the bobtail squid or specific to the octopus' lineage generated copies with divergent expression patterns devoted to tissues outside of the brain. The acquisition of novel expression domains was accompanied by gene order rearrangement through either translocation or duplication and gene loss. Lastly, expansions showed signs of positive selection and some were found to form tandem clusters with shared conserved expression

profiles in cephalopod innovations such as the axial nerve cord. Altogether, our results contribute to the understanding of the molecular and evolutionary history of signal transduction and provide insights into the role of this expansion during the emergence of cephalopod novelties and/or adaptations.

Keywords: G-protein coupled receptors, cephalopod evolution, gene family expansion, gene duplication, signal transduction

Introduction

G-protein-coupled receptors (GPCRs) comprise a large superfamily of membrane proteins that possess a characteristic 7-transmembrane domain (7TM) [1]. They trigger signal transduction pathways by sensing a variety of external and internal stimuli, including hydrogen and Ca^{2+} ions, small peptides (such as hormones and neurotransmitters), large proteins, and even photons as in the case of the light-sensitive opsins [2]. GPCRs thus play a crucial role in how organisms perceive and react to their environments, as well as in homeostatic regulation via endocrine and neuronal functions.

GPCRs have an ancient eukaryotic origin [3]. In bilaterians, the superfamily has evolved into five structural families: Glutamate, Frizzled, Adhesion, Secretin, and Rhodopsin [4]. Amino acid binding (e.g., glutamate and gamma-aminobutyric acid, GABA) is the predominant function in the members of the Glutamate family, but its expansion in vertebrates has resulted in pheromone-, taste, and calcium-sensing functions, as well as olfaction in fishes [5]. Receptors of the Frizzled family have a crucial role in animal development and therefore represent the most conserved family among GPCRs. Members of the Adhesion family are characterized by long N-termini that can contain several domains with adhesive properties (hence the 'adhesion' nomenclature), resembling or complementing the function of other molecules such as cadherins or integrins [6,7]. The proper processing and activation of the Adhesion GPCRs are thought to depend upon their conserved domain, the GPCR proteolytic site (GPS), but their exact physiological function remains unknown [8]. The Secretin family is closely related to the Adhesion GPCRs [5] and represents one of the largest groups of hormone and neuropeptide receptors involved in homeostatic functions [9]. Lastly, the Rhodopsin family is the largest group of GPCRs in most animals. Its members bind chemically diverse ligands and—in certain cases—sense light, thus mediating a variety of functions such as vision, neurotransmission, and immune responses.

Despite the conservation of the GPCR families during bilaterian evolution, the number of receptors in each family varies within individual lineages [3]. The enlargement of GPCR repertoires predominantly occurs by gene duplication and subsequent independent evolution of the copies [5], a powerful evolutionary mechanism that generates novel functions (via e.g. neo- or subfunctionalization) [10,11]. GPCR expansions are not surprising considering the diverse environments and stimuli these receptors evolved to process. Larger repertoires can increase the amount of available sensory information, thus facilitating adaptation to the environment [5] by enabling the evolution of sensory functions relevant to ecological contexts of species (e.g. gustatory receptors in disease-transmitting mosquitoes [12], chemoreceptors in nematodes [13,14], olfactory receptors in mammals [15]). Moreover, such expansions permit more complex homeostatic regulation [16]. However, relatively little is known about GPCR diversity and functionality outside vertebrates and model invertebrate species (e.g. *Drosophila melanogaster* [1]).

The cephalopod (referring in this paper to coleoid cephalopods) body plan and nervous system are unique among molluscs. Morphological and neural novelties in these animals include flexible arms and a large and complex nervous system that may reach a total of 500 million neurons [17,18]. Rivalling vertebrate nervous systems, it is considered the largest among invertebrates. The cephalopod nervous system consists of central lobes surrounding the oesophagus and two optic lobes that together contain approximately a third of the neurons, with the remaining two-thirds distributed within the arms (e.g. in the axial nerve cord) [19]. This nervous system controls an outstanding behavioural repertoire [20], including their camouflaging abilities. Chromatophore cells contain pigment granules that are contracted or expanded through adjacent muscle action, allowing the animals to rapidly adjust colouration [21]. Moreover, cephalopod arms are considered a key innovation for their diversification as they might have enabled these animals to become agile predators [22]. The arms bear hundreds of tactile and chemosensory structures, known as suckers that interact with and provide information from the environment [23]. These structures, as well as the muscle coordination of the arms, are controlled by the axial nerve cords [19]. Additional perception of the external environment is achieved by the convergently evolved camera-type eyes harbouring one photoreceptor type and potentially by dermal photoreception [24].

Insights into the molecular basis of cephalopod innovations first arose with the sequencing of the California two-spot octopus' (*Octopus bimaculoides*) genome [25]. The genomic analyses revealed expansions in several key gene families involved in neuronal patterning, such as C2H2 zinc fingers and protocadherins, and additionally the GPCRs. The individual family composition, evolutionary dynamics, and patterns of gene expression of this octopus GPCR

expansion, however, have not been studied yet. Therefore, our aim was to assess the evolutionary history of this expansion and its relationship with morphological and neural cephalopod novelties. We studied genomic signatures of this gene superfamily in cephalopods, revealing potential evolutionary mechanisms behind their expansion and expression. Our results provide information about the duplication dynamics in this superfamily of proteins and the evolution of GPCR-mediated signal transduction in cephalopods.

Material and Methods

(a) Data collection and phylogenetic analyses

Pfam signatures for the Rhodopsin family (also classified and found in the Pfam database as class A: pf00001), Secretin and Adhesion families (both under class B: pf00002), Glutamate family (class C: pf00003), and Frizzled family (class F: pf01534) were used to compile a dataset of protein sequences from the UniProt database [26] for 14 species: *Anopheles gambiae* (African malaria mosquito), *Drosophila melanogaster* (fruit fly), *Caenorhabditis elegans* (round worm), *Capitella teleta* (polychaete worm), *Helobdella robusta* (Californian leech), *Schistosoma mansoni* (blood fluke), *Strongylocentrotus purpuratus* (purple sea urchin), *Saccoglossus kowalevskii* (acorn worm), *Branchiostoma floridae* (Florida lancelet), human, *Lottia gigantea* (giant owl limpet), *Mizuhopecten yessoensis* (Japanese scallop), *Crassostrea gigas* (Pacific oyster), and *O. bimaculoides* (California two-spot octopus). If available, only entries matching a reference UniProt proteome were downloaded (see electronic supplementary material, table S1). *Euprymna scolopes* (Hawaiian bobtail squid) and *Callistoctopus minor* (common long-arm octopus) sequences corresponding to the same GPCR families were included in the study using draft genome sequences (*E. scolopes*: Belcaid *et al.* [27]; *C. minor*: Kim *et al.* [28]). These 16 species were selected for our analysis to reconstruct the history of the cephalopod GPCR repertoire under a broad evolutionary context, having points of comparison between cephalopods (i.e. two Octopodiformes and one Decapodiformes species) and other species covering major bilaterian lineages (electronic supplementary material, figure S1). After removing protein sequences corresponding to equal gene entries (i.e. isoforms), the final dataset of all GPCRs comprised a total of 6194 sequences. Individual datasets for each GPCR class (A, B, C, and F) were also constructed for the phylogenetic analyses (electronic supplementary material, table S1). Sequences were aligned with MAFFT v7.312 [29] (default parameters) and alignments were cleaned with TrimAl v1.4 [30] using a 0.25 gap threshold, 0.25 residue overlap threshold, and 90% sequence overlap. The best-fit model of molecular evolution for each dataset was selected with ModelFinder, implemented in IQ-TREE v. 1.6.2 [31] using the

corrected Akaike Information Criterion (AICc). For the class A, class F, and all GPCR datasets, the LG model was assigned [32]. For class B and C, the WAG model was selected [33]. Maximum-likelihood trees were constructed with FastTree v2.1.10 [34] using four rounds of minimum-evolution nearest neighbour interchanges (NNI) and remaining parameters as default. Local support values were computed with FastTree by default with the Shimodaira–Hasegawa (S–H) test, a test that compares multiple topologies based on a non-parametric bootstrap [35,36]. To identify the expanded clades of GPCRs, we first identified nodes containing only cephalopod GPCRs (sequences of only one, two, or three cephalopod species) in the full GPCR tree and then ranked each node by the number of cephalopod GPCR sequences descending from the node. We only retained nodes that were outliers (10 or more paralogues) in the number of GPCRs present as expanded groups (electronic supplementary material, figure S2). Files with the sequences, alignments, and trees are available from the Dryad Digital Repository: <https://doi.org/10.5061/dryad.d3qh5c8> [37].

(b) Gene expression analyses

Normalized expression data, measured as transcripts per total expression counts, were collected across 12 tissue types for the genes encoding the *O. bimaculoides* GPCRs using a publicly available dataset [25]. Gene expression data, normalized as TPM counts (Transcripts per Million), for GPCRs of *C. minor* were collected across 18 tissue types from [28]. Expression data (TPM) for *E. scolopes* GPCRs were also collected across 7 tissues [27]. Tissue specificity of genes was measured calculating the parameter Tau [38], which varies between 0 (broad expression) and 1 (tissue specific). Genes with a Tau value higher than 0.8 were considered tissue specific [39]. Tau values between the non-expanded and the expanded groups were compared with a Kruskal–Wallis rank test [40]. A Dunn's test [41] was performed to report the results among the pairwise comparisons, using the Holm's *p*-value adjustment method [42]. Additionally, the proportion of expression of each gene in the different tissues was calculated by dividing its expression value in a tissue by the sum of its expression values in all tissues. The expression matrices for *E. scolopes*, *O. bimaculoides*, and *C. minor* used for the aforementioned analyses and calculations performed were deposited in Dryad Digital Repository: <https://doi.org/10.5061/dryad.d3qh5c8> [37].

(c) Annotation of expanded GPCR groups

Annotation of the nearest sequences to the expanded groups in the individual class trees was used to infer the putative function of the recently duplicated cephalopod GPCRs. Lastly, *O. bimaculoides* and *C. minor* GPCR sequences from the expanded groups were assigned putative functions by scanning for matches against protein signatures from the InterPro member databases [43] using InterProScan v5.7.48 [44].

(d) Positive selection analysis and GPCR genome location

Expanded groups were examined for positive selection by performing a maximum-likelihood branch test with codeml in PAML v4.9 [45]. Following the tree topology from the individual class phylogenies, subsets of the trees corresponding to the node containing each expanded group and its outgroup (S–H support greater than 0.70) were extracted using the drop.tip function in R (ape v5.0 package [46]). Protein sequences were re-aligned with MAFFT and resulting alignments were converted to their corresponding coding DNA sequence (CDS) alignments using the PAL2NAL program [47]. The CDS alignment was then cleaned with TrimAl, erasing sequences with not enough informative sites by using a 0.25 residue overlap threshold and 90% sequence overlap. In the case of *O. bimaculoides*, *C. gigas*, and *L. gigantea*, CDS sequences were obtained from the Ensembl Metazoa genomes browser (<https://metazoa.ensembl.org/index.html>) [48]. *C. minor* sequences were obtained from the supporting data of Kim *et al.* [28]. For the remaining species (i.e. *A. gambiae*, *D. melanogaster*, *M. yessoensis*, *S. purpuratus*, *B. floridae*, and human), corresponding coding DNA sequences were downloaded from NCBI (National Center for Biotechnology Information), searching by the gene ID entries. For each group, the CDS alignment and tree topology were used in codeml to compare likelihoods from two different models: (i) a null model assuming a single ratio of non-synonymous to synonymous substitutions (ω ; calculated by codeml from the dataset) across the entire tree, and (ii) an alternative model that estimated a different ω in the branches of interest (foreground), relative to the remaining (background) branches. For the alternative model, branches within groups 1–4 were labelled as foreground, whereas all other branches remained as background. The null and alternative models were compared with a likelihood ratio test, where the test statistic is chi-squared distributed with k degrees of freedom (k being the difference in the number of free parameters between models). Additionally, a sites test was carried out to identify specific amino acids under positive selection. In this test, the same null model was compared to an alternative model that allowed ω to vary among both branches and amino acid sites. Sites with a Bayes empirical Bayes (BEB) confidence greater than or equal to 95% and $\omega > 1$ were identified as under positive selection [49,50]. In the branch test, branches with $\omega > 100$ were discarded because they often overestimate ω as a result of poor alignment quality between deeply split or otherwise highly divergent sequences. Input files and codeml outputs, as well as likelihood ratio test calculations were deposited in Dryad Digital Repository: <https://doi.org/10.5061/dryad.d3qh5c8> [37].

Lastly, to search for evidence of tandem duplication, scaffold locations of *O. bimaculoides* genes from both non-expanded and expanded groups were extracted from

Albertin *et al.* [25]. Start and stop positions in the scaffolds, as well as the direction of these genes, were summarized and can be found in electronic supplementary material, table S2.

Results

(a) Phylogenetic analysis

The phylogenetic analyses resolved approximately 77% of branches at S–H support values greater than or equal to 0.70 in the tree containing all GPCR classes, 78% in class A, 83% in class C, and 79% in class B and F trees (see tree files deposited in Dryad Digital Repository). Using the ranking procedure of GPCR clades and a cut-off of at least 10 cephalopod sequences (Material and methods; electronic supplementary material, figure S2), two cephalopod-specific and two octopus-specific expansions were identified (labelled as groups 1–4 in figure 1) and distinguished from GPCRs shared with other bilaterians. Groups 1 and 2 belong to class A (Rhodopsin family) and contain, respectively, 63 and 21 GPCRs from the three cephalopod species, forming the cephalopod-specific expansions (S–H support greater than 0.70). Groups 3 and 4 belong to class B and contain 21 and 45 Secretin/Adhesion GPCR sequences from the two octopus species, respectively (i.e. *O. bimaculoides* and *C. minor*, with the exception of a single *E. scolopes* sequence in group 3; S–H support greater than 0.70). These constitute the putative octopus-specific expanded groups. We additionally estimated sequence divergence, based on total branch lengths, which resulted in a median of around 2.01 (first quartile 1.79, third quartile 2.5) substitutions per site between sequences in the expanded groups versus their closest non-cephalopod relatives. This indicates that, on average, around 1 substitution per site corresponds with the Cambrian radiation of those animal groups (approximately 500 million years ago) and 0.5 substitutions per site thus correspond with approximately 250 million years.

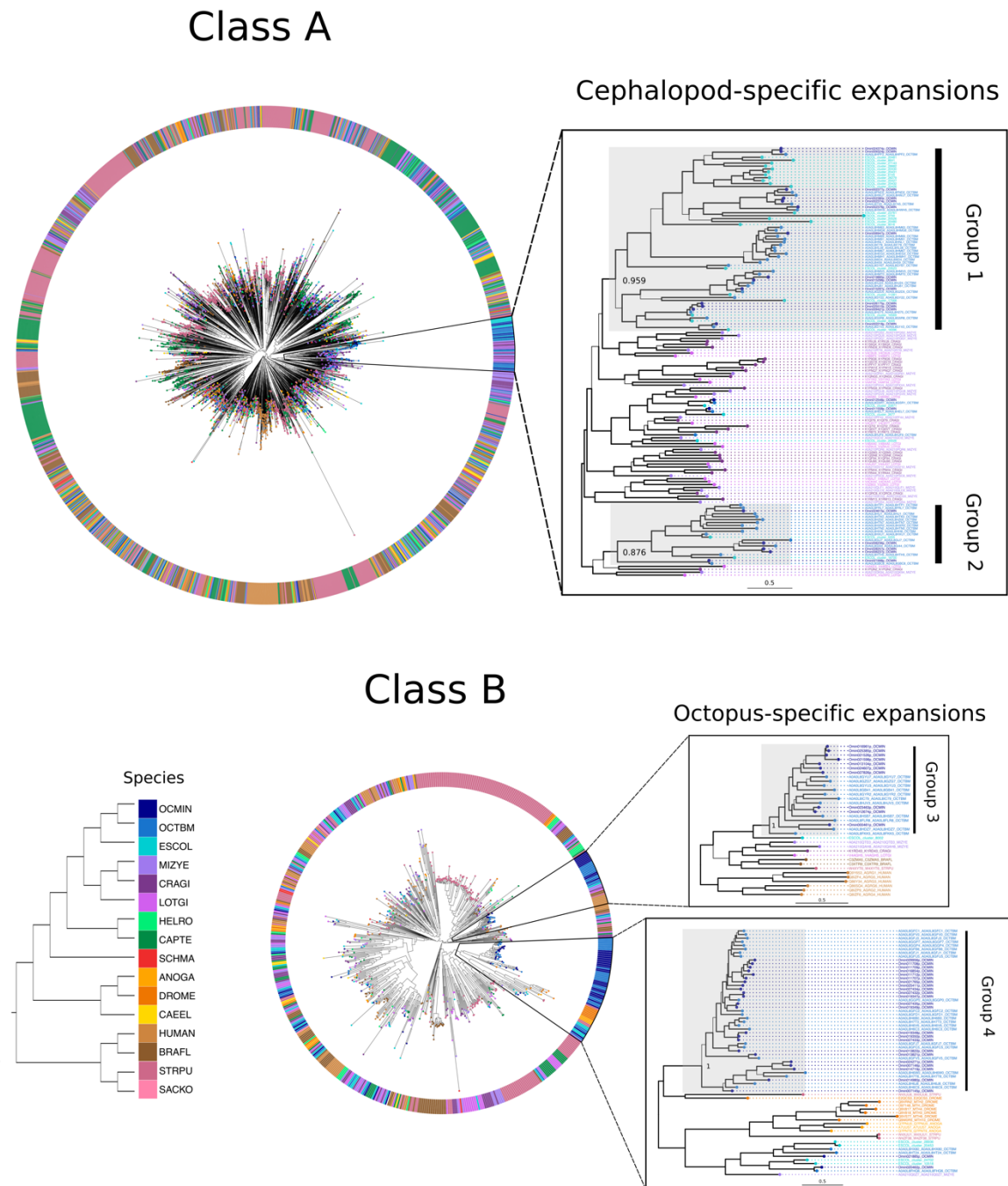


Figure 1. Phylogenetic trees for class A and B. Each colour corresponds to a species; abbreviations as follow. ANOGA: *Anopheles gambiae*, BRAFL: *Branchiostoma floridae*, CAEEL: *Caenorhabditis elegans*, CAPTE: *Capitella teleta*, CRAGI: *Crassostrea gigas*, DROME: *Drosophila melanogaster*, ESCOL: *Euprymna scolopes*, HELRO: *Helobdella robusta*, HUMAN, LOTGI: *Lottia gigantea*, MIZYE: *Mizuhopecten yessoensis*, OCMIN: *Callistocotopus minor*, OCTBM: *Octopus bimaculoides*, SACKO: *Saccoglossus kowalevskii*, SCHMA: *Schistosoma mansoni*, STRPU: *Strongylocentrotus purpuratus*. Zoom-ins show the expanded groups identified (1–4). Thicker branches represent significant S–H support (greater than 0.7). Bar length indicates 0.5 substitutions per site corresponding to approximately 250 Myr. Tree visualization was performed with the ggtree package [51] using R v3.4.2 [52].

(b) Gene expression analyses

To explore the possible functions of the cephalopod GPCRs and evolutionary transitions between the identified expansions, the expression was examined across various tissue types (figures 2–4; electronic supplementary material, figures S5–S9 and table S2). A majority of non-expanded GPCRs (i.e. genes not included within groups 1–4) were highly expressed in the nervous tissues of *O. bimaculoides*, such as the sub- and supraesophageal brain, optic lobe, and axial nerve cord, and the brain tissue in *C. minor* and *E. scolopes* (figures 2 and 4; electronic supplementary material, figures S5–S7). By contrast, expression of GPCRs in groups 1–4 was dominant outside of the brain (figures 2–4; electronic supplementary material, figures S5–S7). Expression in all groups was found to be mainly tissue specific. Mean Tau value and its 95% confidence interval in all but group 1 in *O. bimaculoides* and *E. scolopes*, and group 3 in *C. minor* was higher than 0.8 (electronic supplementary material, figure S4).

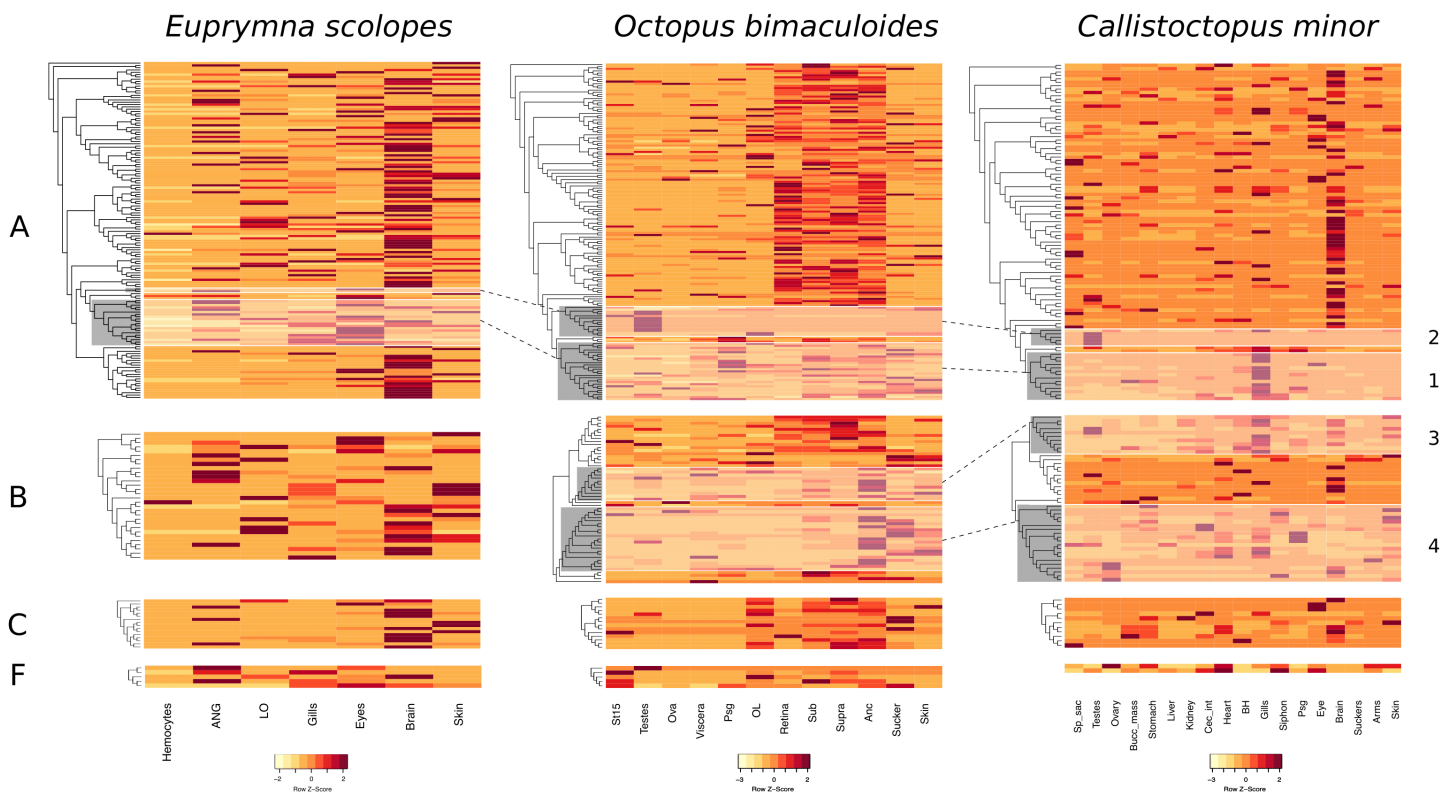


Figure 2. Heatmap of all *E. scolopes*, *O. bimaculoides*, and *C. minor* GPCRs. Genes (rows) are clustered following the trees resulting from the phylogenetic analyses performed for each class (a–f) independently. Clades highlighted in grey represent the four groups depicted in figure 1, group number is given to the right of the *C. minor* heatmap. Dotted lines connect corresponding expanded groups in both species. Tissues (columns) with transcriptomic data for *E. scolopes*: haemocytes, accessory nidamental gland, light organ (LO), gills, eyes, brain, and skin. Tissues (columns) with transcriptomic data for *O. bimaculoides*: sucker, testes, stage 15 (St15) embryo, and skin.

ova, skin, posterior salivary gland (Psg), viscera (heart, kidney, and hepatopancreas), subesophageal brain (sub), supraesophageal brain (supra), optic lobe (OL), axial nerve cord (Anc), and retina. Tissues (columns) with transcriptomic data for *C. minor*: liver, kidney, stomach, caecum intestine (Cec_int), posterior salivary gland (Psg), buccal mass (Bucc_mass), bronchial heart (BH), systemic heart (heart), suckers, arms, skin, gills, siphon, brain, eye, spermatophore sac (Sp_sac), testes, and ovary. Heatmap generated with the heatmap.2 function (gplots v3.0.1 package [53]) using R v3.4.2 [52].

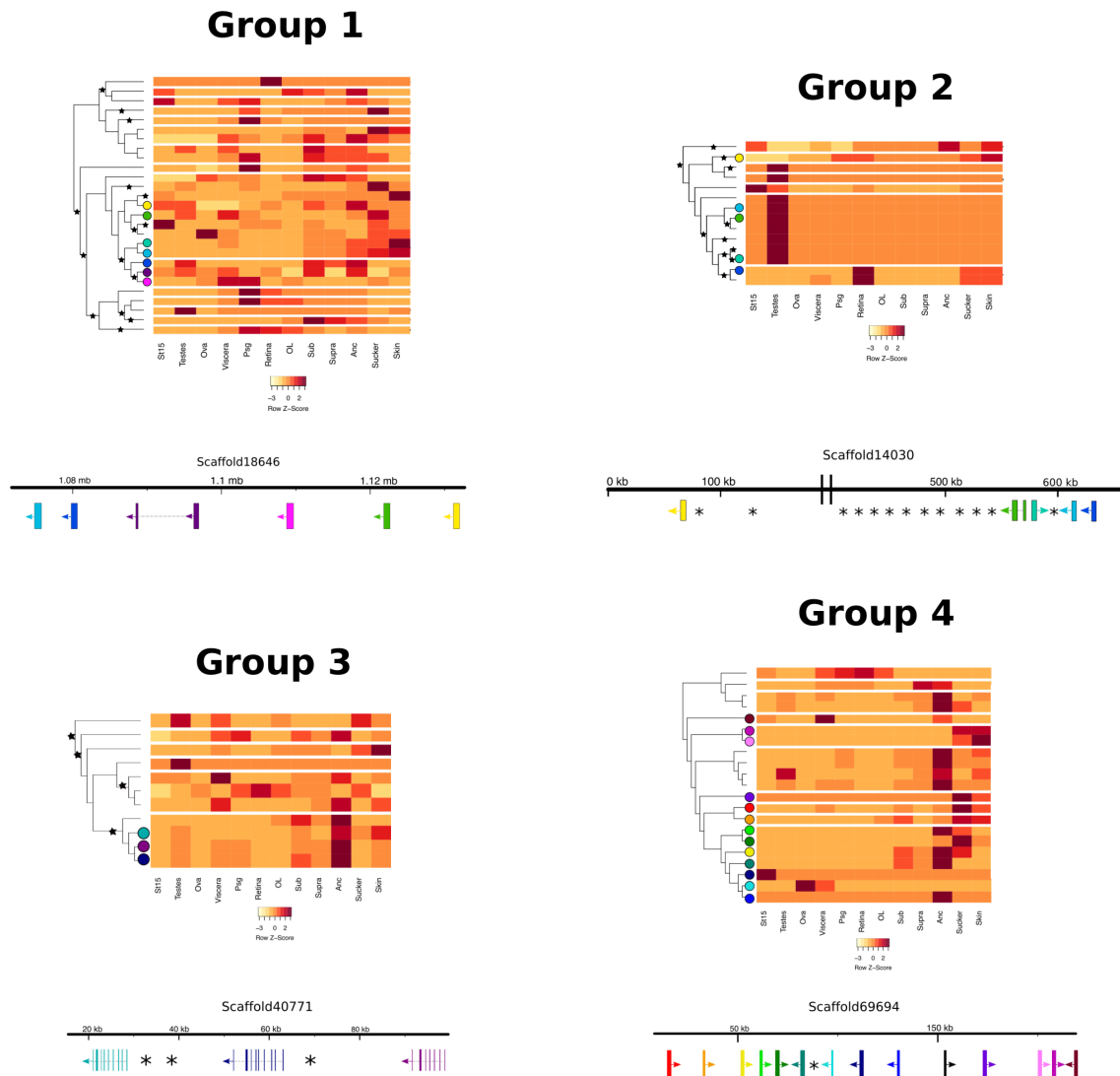


Figure 3. Expression profiles, positive selection results, and gene localization of cephalopod-specific expansions (groups 1 and 2) and octopus-specific expansions (groups 3 and 4) for *O. bimaculoides*. Genes (rows) in heatmaps are clustered following the trees resulting from the phylogenetic analyses. Spaces between rows indicate the presence of other sequences of cephalopod species as determined in the phylogenetic analyses. Black stars represent positive selection ($\omega > 1$). The scaffold with the most co-localized genes per group is represented below the heatmaps (see detailed information in the electronic supplementary material, table S2). Exon–intron composition of the genes is depicted as thick bars (exons) and grey lines connecting them (introns).

Direction of transcription is shown with an arrow. Asterisks (*) represent other genes (no GPCRs) found in the surrounding space of the co-localized GPCRs.

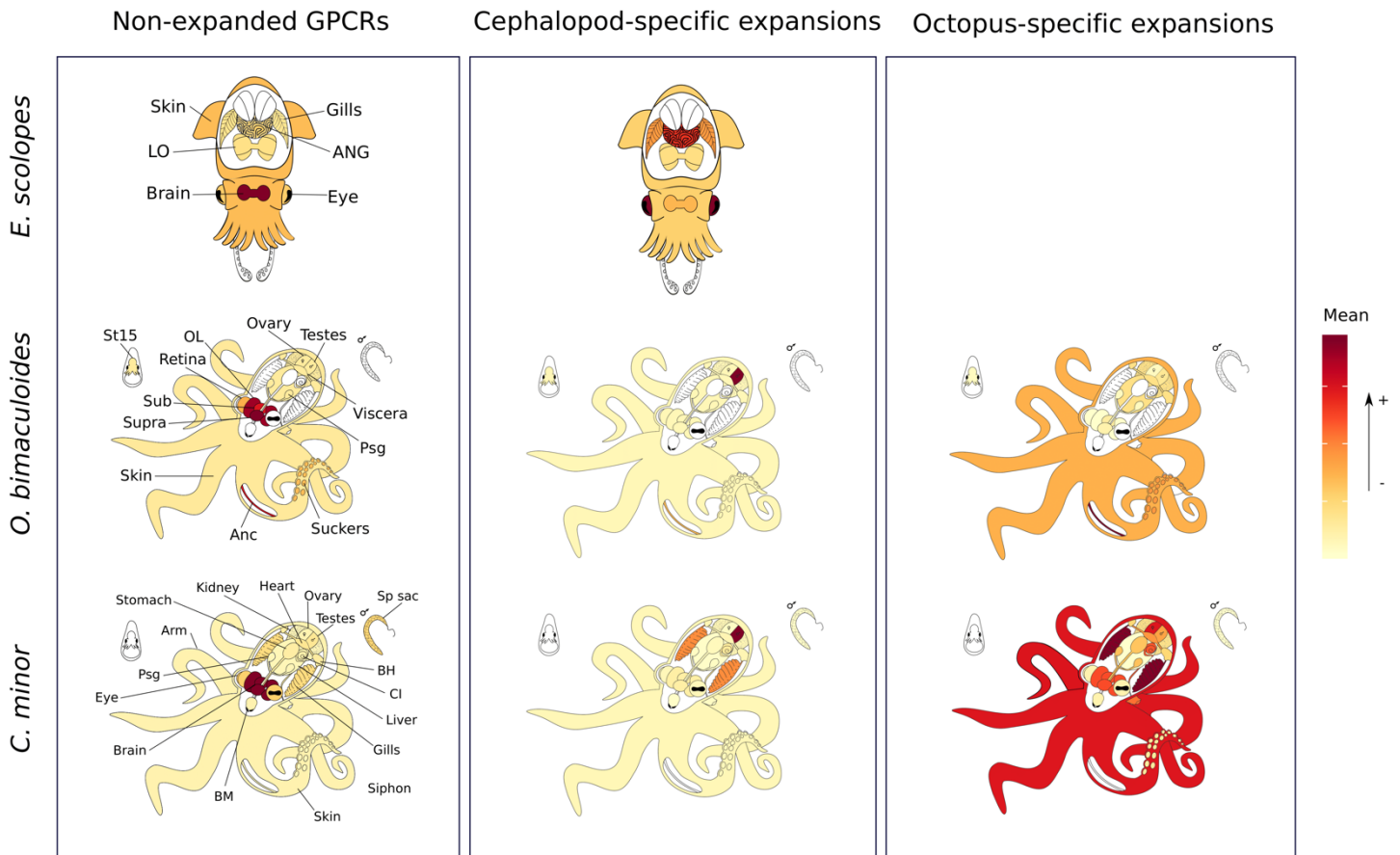


Figure 4. Expression proportions of *E. scolopes*, *O. bimaculoides*, and *C. minor* genes in each tissue in non-expanded GPCRs, cephalopod-specific expansions, and octopus-specific expansions. Tissue colouration gradient follows the mean values of gene expression proportion as shown in electronic supplementary material, figures S5–S7. Warmer colours represent the highest proportion of gene expression. *E. scolopes* (illustration designed by Hannah Schmidbaur) tissues: accessory nidamental gland (ANG), light organ (LO), gills, eyes, brain, and skin. *O. bimaculoides* tissues: sucker, testes, stage 15 (St15) embryo, ovary, skin, posterior salivary gland (Psg), viscera (heart, kidney, and hepatopancreas), subesophageal brain (sub), supraesophageal brain (supra), optic lobe (OL), axial nerve cord (Anc), and retina. *C. minor* tissues: liver, kidney, stomach, caecum intestine (CI), posterior salivary gland (Psg), buccal mass (BM), bronchial heart (BH), systemic heart (heart), suckers, arms, skin, gills, siphon, brain, eye, spermatophore sac (Sp sac), testes, and ovary.

GPCR genes from the cephalopod-specific expansion had a large expression component devoted to the testes in both *C. minor* and *O. bimaculoides* (group 2 in figure 3; electronic supplementary material, figures S6, S8, and S9). The genes mostly contributing to this high expression in testes formed a clade in group 2 and shared the same profile (figure 3; electronic supplementary material, figure S6). The highest proportion of expression in group 1

sequences of *C. minor* was in the gills (electronic supplementary material, figures S6 and S9). Expression profiles of *O. bimaculoides* genes in this cephalopod-specific group did not follow a phylogenetic pattern (i.e. closely related sequences did not show a similar profile, figure 3), and their highest expression was devoted to nervous and non-nervous tissues (electronic supplementary material, figure S8). In *E. scolopes*, the highest proportion of expression was found in the eyes, the ANG, and gills (figure 4; electronic supplementary material, figure S5). However, due to the low number of tissues sampled in the bobtail squid in comparison to both octopuses, no conclusions were drawn for this species.

The largest proportion of expression in the octopus-specific expansions in *O. bimaculoides* was in the axial nerve cord, followed by the suckers and the skin (figure 4; electronic supplementary material, figure S8). As in group 1, *O. bimaculoides* genes contributing to the highest expression (axial nerve cord) were found to be forming clades in groups 3 and 4 (figure 3). For the case of *C. minor*, the highest expression proportion values were found in the gills, the ovary, and the skin (figure 4; electronic supplementary material, figure S9). The highest expression in gills was mainly found in group 3 and these similar profiles were found again shared within clades (electronic supplementary material, figure S6). By contrast, *C. minor* genes in group 4 did not share similar expression profiles.

(c) Annotation of expanded GPCR groups

To further infer the putative function of the GPCR in groups 1–4, the annotation of the sister group sequences to the expansions were summarized (electronic supplementary material, table S3). Additionally, InterProScan annotation of the sequences from the four groups was performed to obtain information about putative functional domains. Groups 1 and 2 were found in a clade composed largely of molluscan sequences (i.e. *C. gigas*, *L. gigantea*, and *M. yessoensis*). These were found to be receptors of a variety of neuropeptides, such as Tachykinin-like peptides, Orexin (whose orthologue in deuterostomes is Allatotropin [54]), Cholecystokinin, Cephalotocin, Gonadotropin-releasing hormone (GnRH) II, Neuromedin-U, and others. Group 3 was found to be closely related to six human and two *M. yessoensis* adhesion GPCRs. All other related sequences (from *C. gigas*, *L. gigantea*, *S. purpuratus*, and *B. floridae*) were found to be uncharacterized proteins. The InterProScan searches confirmed the annotation of the cephalopod sequences to be adhesion GPCRs as it resulted in the identification of either the GPS motif and/or the adhesion 7TM domain in all cephalopod sequences in the group (with the exception of three *C. minor* sequences) (electronic supplementary material, table S3). Additional domains found in sequences of this expansion were the leucine-rich repeat (in 5 out of 22 sequences) and the Death domain (in one *O. bimaculoides* sequence). GPCRs from the octopus-specific expansion group 4 were

found to be closely related to Methuselah and Methuselah-like proteins (a subgroup of Secretin GPCRs) from *D. melanogaster*, *A. gambiae*, and *M. yessoensis*. All octopus sequences in this group were found to have matches with Methuselah-like 7TM signatures in our InterProScan searches (with the exception of three *C. minor* sequences).

(d) Positive selection analysis

For groups 1, 2, and 3, the likelihood ratio test rejected ($p < 0.05$) the null hypothesis of homogeneous evolution (equal ω) along the tree in favour of an alternative model of accelerated evolution (or positive selection) (electronic supplementary material, table S4). Signatures of positive selection ($\omega > 1$) could be found in branches of these three groups (figure 3; electronic supplementary material, figures S5 and S6). For group 4, the test did not reject ($p > 0.05$) the null hypothesis, favouring a model of homogeneous omega values of 0.306 (less than 1, purifying selection) along the tree. Moreover, no specific amino acid sites could be identified under positive selection (BEB less than 95%) using the sites test for any of the expanded groups. Positive selection was found, among others, on branches leading to genes sharing similar expression patterns, such as a high expression in *O. bimaculoides*' axial nerve cord and *C. minor*'s gills (figure 3; electronic supplementary material, figure S6).

(e) Genomic co-localization of GPCRs

To understand whether the more recently duplicated GPCRs tend to co-localize in the genome, we profiled their genomic locations in the draft genome assembly of *O. bimaculoides*. We found that only a single scaffold (out of 155) harboured three or more GPCRs from the non-expanded set, whereas groups 1–4 had the following distributions, respectively: 1 out of 17 scaffolds, 2 out of 6, 1 out of 9, and 2 out of 2 (electronic supplementary material, table S2). While impeded by the fragmented nature of the genome, comparison between the individual expansion groups and the non-expanded GPCRs shows almost complete re-distribution of the latter, supporting the scenario of local, tandem duplication and later dispersion.

The expression profiles of GPCRs co-localized in the genome of the cephalopod-specific GPCR expansions did not follow the same expression pattern, except for some of the genes in group 2 with the shared highest expression in testes (figure 3). By contrast, most of the genes in the octopus-specific expansion groups localized on the same scaffolds were highly expressed in the axial nerve cord (figure 3; electronic supplementary material, table S2). Moreover, evidence of local duplication under positive selection was found in some of these genes with shared expression patterns (i.e. group 4, Scaffold40771; figure 3).

Discussion

(a) Non-expanded GPCRs are predominantly expressed in cephalopod nervous tissues

GPCRs are essential components of animal nervous systems as they mediate signal transduction by binding neurotransmitters [55]. Here, we found that the non-expanded GPCRs (i.e. those outside of the expanded groups 1–4) in the three cephalopod species are mainly expressed in neural tissues, such as the sub- and supraoesophageal brains, the optic lobes, and the axial nerve cords in *O. bimaculoides* and the brain in *C. minor* and *E. scolopes* (figure 2; electronic supplementary material, figures S7–S9). This suggests that a large proportion of the GPCR repertoire related to neural functions in octopuses derives from evolutionarily conserved families of GPCRs, most showing one-to-one orthology with other bilaterian species. These are possibly correlated with shared and highly conserved neuronal signalling functions [56] and not with cephalopod neuronal innovations. Rather, those innovations seem to be related to other large gene families such as C2H2 and protocadherins [25,57].

(b) Expression divergence in the cephalopod-specific expansions

In contrast with the non-expanded GPCRs, we found a shift in expression patterns in the cephalopod-specific expansions comprising different nervous and non-nervous tissues, mainly outside the brain (figures 2–4). Expression of duplicates in group 2 was most prominent in testes (electronic supplementary material, figures S8 and S9), contributed by closely related genes (figure 3; electronic supplementary material, figure S6). Both *O. bimaculoides* and *C. minor* showed similar patterns. By contrast, sequences of group 1 showed discordant expression patterns between the three species analysed. Divergent expression patterns were found between *O. bimaculoides* duplicates, with the highest expression covering a wide range of tissues (figure 3; electronic supplementary material, figure S8), whereas *C. minor* showed higher expression in the gills (figure 3; electronic supplementary material, figures S6 and S9). This inconsistency in expression patterns could be an artefact resulting from differences in tissue sampling (e.g. *O. bimaculoides* lacks expression data for the gills) or a result of divergence time as most of these duplications have a coleoid cephalopod origin. Moreover, expression of these cephalopod-specific sequences in Decapodiformes (e.g. *Euprymna scolopes*) needs to be further investigated.

Both cephalopod-specific expansions were found inside a molluscan clade of neuropeptide receptors (figure 1; electronic supplementary material, table S3). These receptors and their

ligands comprise families with a bilaterian origin and are known to control a variety of physiological processes like reproduction and sexual behaviour (e.g. GnRH), heart activity (e.g. cholecystokinin [58]), and food intake (e.g. Neuromedin-U, orexin/allatotropin) [54]. Here, we found evidence that part of this expanded repertoire of neuropeptide receptors (i.e. group 2) has converged in Octopodiformes in functions devoted to mainly two tissues. On the one hand, most sequences of group 2 were related to male reproduction through their testes-specific expression; however, their exact function in this tissue remains unknown. On the other hand, the majority of group 1 sequences in *C. minor* were expressed in the gills. These are potentially related to specific functions in this tissue (e.g. respiration, circulation, excretion), as they were found to be tissue specific (electronic supplementary material, figure S4). Additionally, we found signatures of positive selection in some gill-specific receptors (figure 3), suggesting that the maintenance of the copies and their expression domains was potentially advantageous for these animals. This GPCR expression could be related to circulatory or respiratory adaptations in cephalopods as highly active marine predators [59]. However, expression data of gills in *O. bimaculoides* and other cephalopod species would be needed to confirm this trend.

(c) Evolution of octopus GPCR paralogues

Similar to cephalopod-specific GPCRs, the expression of the octopus-specific expansions was not related to the brain (figures 3 and 4). In *O. bimaculoides*, the highest proportion of expression was predominantly in the axial nerve cord, followed by the suckers and the skin (figure 4; electronic supplementary material, figure S8), whereas in *C. minor*, it was found to be again in gills (mostly in group 3), followed by the skin and ovaries (figure 4; electronic supplementary material, figure S9).

Proteins of octopus-specific group 3 were identified as adhesion GPCRs, as we found evidence for the presence of its characteristic domains: the GPS and the adhesion 7TM domains. Additional functional domains found here, like the leucine-rich repeat, have been also reported in the N-termini of human adhesion GPCRs [6]. Despite the small amount of information available on this GPCR family, expression data in vertebrates have suggested an important role in the central and peripheral nervous system [6]. As these proteins resemble or complement other adhesive molecules, such as cadherins, they could be involved in neuronal plasticity and axon guidance. The octopus genome revealed an expansion of protocadherins, highly expressed in nervous tissues and exceptionally enriched in the axial nerve cord and the optic lobes [25,57]. Consistently, some of these receptors showed higher expression in the axial nerve cord of *O. bimaculoides* (figure 3). These constitute a potential functional cluster, as they were co-localized and showed signatures of positive selection. Enrichment of these

adhesion GPCRs in *O. bimaculoides* was also found in the suckers and skin, structures directly associated with the peripheral nervous system of cephalopods. Thus, this expanded subgroup of GPCRs could be complementing the function of the enlarged protocadherin repertoire reported in genomes of octopuses in these tissues. An axial nerve cord sample is missing from *C. minor* transcriptome sampling. However, the skin showed high expression of sequences from that group, whereas total arm and sucker tissues did not (figure 4; electronic supplementary material, figure S9). Most of the sequences were also found to be highly expressed in the gills in *C. minor* (absent in the available *O. bimaculoides* tissue sampling). Thus, a consistent tissue sampling for transcriptomic data would help elucidate whether GPCRs expanded in the Octopodiformes lineage have diverged in expression in individual species or followed a shared pattern related to octopus-specific novelties.

Our annotation results suggest that the second octopus-specific expansion (group 4) comprises Methuselah-like GPCRs. The *methuselah* gene and its 15 paralogues (*methuselah-like 1–15*) were first described in *Drosophila* [60]. These constitute a gene family with an early metazoan origin that has gone through various events of extinctions (e.g. in vertebrates) and expansions (e.g. in insects) [61]. Little is known about the function of the 15 paralogues, but the *methuselah* gene has been widely studied in *Drosophila* and has been found to be related with stress response, lifespan, and embryonic development [62]. Here, we found that the octopus-specific expansion derives from the Methuselah-like 15 GPCR, one of the most ancient paralogues in *D. melanogaster*. The evidence of high co-localization found in this group suggests that this expansion arose from various events of tandem duplication (figure 3). As duplicates show divergent expression patterns in both species, duplication potentially followed subfunctionalization of the copies under homogeneous purifying selection (see Results).

(d) GPCR expansions and evolution of novel expression domains

In all expanded GPCR groups, we found evidence of tandem co-localization of genes in *O. bimaculoides* yet with diverging expression profiles (e.g. Scaffold 69 694 in group 4, figure 3; electronic supplementary material, table S2). Some of these were also observed to be in monophyletic groups with other non-co-localized genes (e.g. clade containing the genes of Scaffold 18 646 in group 1 or Scaffold 14 030 in group 2, figure 3). This indicates that the detected expression divergence could be a result of both accumulated mutations in the regulatory regions of clusters (for the case of co-localized genes, e.g. octopus-specific group 4) as well as the acquirement of new regulatory domains by rearrangement of gene order via translocation, gene conversion, or duplication and subsequent loss (e.g. cephalopod-specific groups 1 and 2) [63]. We found, additionally, significant signatures for positive selection on

divergently expressed duplicates (e.g. groups 1 and 2, figure 3), suggesting an adaptive scenario for their retention and neo- or subfunctionalization.

Taken together, our phylogenomic analyses suggest a possible scenario for GPCR evolution. First, the large, anciently expanded (in the metazoan or bilaterian ancestors) GPCR complement shows a largely neuronal expression domain in cephalopods (figure 4; electronic supplementary material, figures S7–S9), indicating a high selective pressure to maintain nervous system-related functions. Genes duplicated in the cephalopod lineage (i.e. cephalopod-specific expansions) diverged their expression patterns to tissues mainly outside of the brain (figure 4). Translocation or segmental duplication and subsequent loss likely rearranged the original order of these cephalopod-specific duplicates and facilitated the evolution of novel expression domains. Some duplicates remained spatially linked, in which case accumulated mutations in their regulatory regions possibly coincided with the expression divergence between them. Finally, local gene duplications occurred more recently in the Octopodiformes lineage and resulted in the emergence of expression domains devoted to the axial nerve cord and associated structures (i.e. skin and suckers) in *O. bimaculoides* (figure 4; electronic supplementary material, figure S8), and to the gills, skin, and ovary in *C. minor* (figure 4; electronic supplementary material, figure S9). Here, in the case of physically co-localized genes with similar expression profiles, their expression patterns suggest a general regulatory property of those clusters (i.e. shared regulatory elements due to proximity [63] or tandem duplication of promoters and/or regulatory elements). Additionally, the GPCR expansions were accompanied by positive selection, suggesting functional adaptation during their evolution.

Conclusion

Duplication and divergence are powerful evolutionary mechanisms that can generate novel functions [10]. Lineage-specific GPCR expansions have predominantly occurred through this mechanism of gene duplication and subsequent independent evolution of the copies [5]. Here, we presented an analysis of the evolutionary history of the cephalopod GPCR repertoire. Our phylogenetic analysis distinguished cephalopod- and octopus-specific expansions from evolutionary older GPCR families (figure 1). Transcriptome data combined with genomic location of genes helped elucidate the likely evolutionary transitions between these expansions (figure 4). Our results reveal functional transitions in the evolution of cephalopod signal transduction, starting with non-expanded receptors having nervous system-related functions. Duplications shared between cephalopod lineages followed, which developed

diverging expression across different tissues, in particular testes and gills. These were identified as neuropeptide receptors. Finally, more recent octopus-specific expansions with co-localized genes showed expression related to highly adapted octopod tissues and organs, such as the suckers and the axial nerve cord of the arms of octopuses (figures 3 and 4). These recent expansions were identified as Adhesion GPCRs and *methuselah/methuselah-like* GPCRs. These results help to reconstruct the evolutionary history of this superfamily of proteins and contribute to our understanding of the molecular mechanisms underlying the evolution of unique features in cephalopods.

Data accessibility

All supporting data (i.e. sequences, alignments, phylogenetic trees, expression matrices, Tau and expression proportion calculations, codeml files and likelihood ratio test calculations) are available from the Dryad Digital Repository: <https://doi.org/10.5061/dryad.d3qh5c8> [37].

Author's contributions

E.A.R. conceived the project, under the guidance of S.J. and O.S. Data analyses were performed by E.A.R, with contributions from R.R.F. and O.S. Manuscript was written by E.A.R. with input from all co-authors.

Funding

E.A.R and O.S. were supported by the grant from the Austrian Science Fund (FWF): P30686-B29.

References

1. Hanlon CD, Andrew DJ. 2015. Outside-in signaling – a brief review of GPCR signaling with a focus on the Drosophila GPCR family. J. Cell Sci. 128, 3533-3542. (doi:10.1242/jcs.175158)
2. Kobilka BK. 2007. G protein coupled receptor structure and activation. Biochim. Biophys. Acta 1768, 794-807. (doi:10.1016/j.bbamem.2006.10.021)

3. Fredriksson R, Schiöth HB. 2005. The repertoire of G-protein–coupled receptors in fully sequenced genomes. *Mol. Pharmacol.* 67, 1414-1425.
(doi:10.1124/mol.104.009001.sequenced)
4. Schiöth HB, Fredriksson R. 2005. The GRAFS classification system of G-protein coupled receptors in comparative perspective. *Gen. Comp. Endocrinol.* 142, 94-101.
(doi:10.1016/j.ygcen.2004.12.018)
5. Strotmann R, Schröck K, Bösselt I, Stäubert C, Russ A, Schöneberg T. 2010. Evolution of GPCR: change and continuity. *Mol. Cell. Endocrinol.* 331, 170-188.
(doi:10.1016/j.mce.2010.07.012)
6. Bjarnadóttir TK, Fredriksson R, Schiöth HB. 2007. The adhesion GPCRs: a unique family of G protein-coupled receptors with important roles in both central and peripheral tissues. *Cell. Mol. Life Sci.* 64, 2104-2119. (doi:10.1007/s00018-007-7067-1)
7. Yona S, Lin H-H, Siu WO, Gordon S, Stacey M. 2008. Adhesion-GPCRs: emerging roles for novel receptors. *Trends Biochem. Sci.* 33, 491-500.
(doi:10.1016/j.tibs.2008.07.005)
8. Paavola KJ, Hall RA. 2012. Adhesion G protein-coupled receptors: signaling, pharmacology, and mechanisms of activation. *Mol. Pharmacol.* 82, 777-783.
(doi:10.1124/mol.112.080309)
9. Cardoso JCR, Pinto VC, Vieira FA, Clark MS, Power DM. 2006. Evolution of secretin family GPCR members in the metazoa. *BMC Evol. Biol.* 6, 1-16. (doi:10.1186/1471-2148-6-108)
10. Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 11, 97-108. (doi:10.1038/nrg2689)
11. Acemel RD, Maeso I, Gómez-Skarmeta JL. 2017. Topologically associated domains: a successful scaffold for the evolution of gene regulation in animals. *Wiley Interdiscip. Rev. Dev. Biol.* 6, 1-19. (doi:10.1002/wdev.265)
12. Hill CA, et al. 2002. G protein-coupled receptors in *Anopheles gambiae*. *Science* 298, 176-178. (doi:10.1126/science.1076196)
13. Robertson HM. 1998. Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement, and intron loss. *Genome Res.* 8, 449-463.
(doi:10.1101/gr.8.5.449)
14. Thomas JH, Robertson HM. 2008. The *Caenorhabditis* chemoreceptor gene families. *BMC Biol.* 6, 1-17. (doi:10.1186/1741-7007-6-42)
15. Niimura Y, Nei M. 2007. Extensive gains and losses of olfactory receptor genes in mammalian evolution. *PLoS ONE* 2, e708. (doi:10.1371/journal.pone.0000708)

16. Fredriksson R, Lagerström MC, Schiöth HB. 2005. Expansion of the superfamily of G-protein-coupled receptors in chordates. *Ann. N. Y. Acad. Sci.* 1040, 89-94.
(doi:10.1196/annals.1327.011)
17. Lee PN, Callaerts P, de Couet HG, Martindale MQ. 2003. Cephalopod Hox genes and the origin of morphological novelties. *Nature* 424, 1061-1065.
(doi:10.1038/nature01809.1.)
18. Hochner B. 2008. Octopuses. *Curr. Biol.* 18, R897-R898.
(doi:10.1016/j.cub.2008.07.057)
19. Young JZ. 1971. *Anatomy of the nervous system of octopus vulgaris*. Oxford, UK: Oxford University Press.
20. Hanlon RT, Messenger JB. 2018. *Cephalopod behaviour*, 2nd edn. Cambridge, UK: Cambridge University Press.
21. Cloney RA, Florey E. 1968. Ultrastructure of cephalopod chromatophore organs. *Zeitschrift für Zellforsch. und Mikroskopische Anat.* 89, 250-280.
(doi:10.1007/BF00347297)
22. Nödl M-T, Kerbl A, Walzl MG, Müller GB, de Couet HG. 2016. The cephalopod arm crown: appendage formation and differentiation in the Hawaiian bobtail squid *Euprymna scolopes*. *Front. Zool.* 13, 1-16. (doi:10.1186/s12983-016-0175-8)
23. Nixon M, Young JZ. 2003. *The brains and lives of cephalopods*. Oxford, UK: Oxford University Press.
24. Kingston ACN, Kuzirian AM, Hanlon RT, Cronin TW. 2015. Visual phototransduction components in cephalopod chromatophores suggest dermal photoreception. *J. Exp. Biol.* 218, 1596-1602. (doi:10.1242/jeb.117945)
25. Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, Brenner S, Ragsdale CW, Rokhsar DS. 2015. The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* 524, 220-224.
(doi:10.1038/nature14668)
26. The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45, D158-D169. (doi:10.1093/nar/gkw1099)
27. Belcaid M, Casaburi G, Mcanulty SJ, Schmidbaur H, Suria AM. 2019. Symbiotic organs shaped by distinct modes of genome evolution in cephalopods. *Proc. Natl Acad. Sci. USA* 1-6. (doi:10.1073/pnas.1817322116)
28. Kim B-M, et al. 2018. The genome of common long-arm octopus *Octopus minor*. *Gigascience* 7, 1-7. (doi:10.1093/gigascience/giy119)
29. Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772-780.
(doi:10.1093/molbev/mst010)

30. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972-1973. (doi:10.1093/bioinformatics/btp348)
31. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587. (doi:10.1038/nmeth.4285)
32. Quang Le S, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* 25, 1307-1320. (doi:10.1093/molbev/msn067)
33. Whelan S, Goldman N. 2001. A General empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* 18, 691-699. (doi:10.1093/oxfordjournals.molbev.a003851)
34. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5, e9490. (doi:10.1371/journal.pone.0009490)
35. Shimodaira H, Hasegawa M. 1999. Letter to the editor: multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16, 1114-1116. (doi:10.1093/oxfordjournals.molbev.a026201)
36. Planet PJ. 2006. Tree disagreement: measuring and testing incongruence in phylogenies. *J. Biomed. Inform.* 39, 86-102. (doi:10.1016/j.jbi.2005.08.008)
37. Ritschard EA, Fitak RR, Simakov O, Johnsen S. 2019. Data from: Genomic signatures of G-protein-coupled receptor expansions reveal functional transitions in the evolution of cephalopod signal transduction. Dryad Digital Repository. (<https://doi.org/10.5061/dryad.d3qh5c8>)
38. Yanai I, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21, 650-659. (doi:10.1093/bioinformatics/bti042)
39. Kryuchkova-Mostacci N, Robinson-Rechavi M. 2016. Tissue-specificity of gene expression diverges slowly between orthologs, and rapidly between paralogs. *PLoS Comput. Biol.* 12, e1005274. (doi:10.1371/journal.pcbi.1005274)
40. Kruskal WH, Wallis WA. 1952. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* 47, 583-621. (doi:10.1080/01621459.1952.10483441)
41. Dunn OJ. 1964. Multiple comparisons using rank sums. *Technometrics* 6, 241-252. (doi:10.1080/00401706.1964.10490181)
42. Holm S. 1979. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 65-70.
43. Finn RD, et al. 2017. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* 45, D190-D199. (doi:10.1093/nar/gkw1107)

44. Jones P, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1-5. (doi:10.1093/bioinformatics/btu031)
45. Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586-1591. (doi:10.1093/molbev/msm088)
46. Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289-290. (doi:10.1093/bioinformatics/btg412)
47. Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34, 609-612. (doi:10.1093/nar/gkl315)
48. Kersey PJ, et al. 2017. Ensembl genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.* 46, D802-D808. (doi:10.1093/nar/gkx1011)
49. Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* 22, 1107-1118. (doi:10.1093/molbev/msi097)
50. Yang Z, Bielawski JR. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15, 496-503. (doi:10.1016/S0169-5347(00)01994-7)
51. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28-36. (doi:10.1111/2041-210X.12628)
52. R Core Team. 2017. R: a Language and Environment for Statistical Computing.
53. Warnes GR, et al. 2016. gplots: various R Programming Tools for Plotting Data. R package version 3.0.1.
54. Jékely G. 2013. Global view of the evolution and diversity of metazoan neuropeptide signaling. *Proc. Natl Acad. Sci. USA* 110, 8702-8707. (doi:10.1073/pnas.1221833110)
55. Hille B. 1992. G protein-coupled mechanisms and nervous signaling. *Neuron* 9, 187-195. (doi:10.1016/0896-6273(92)90158-A)
56. Katz PS, Lillvis JL. 2014. Reconciling the deep homology of neuromodulation with the evolution of behavior. *Curr. Opin. Neurobiol.* 29, 39-47. (doi:10.1016/j.conb.2014.05.002)
57. Wang ZY, Ragsdale CW. 2017. Cadherin genes and evolutionary novelties in the octopus. *Semin. Cell Dev. Biol.* 69, 151-157. (doi:10.1016/j.semcdb.2017.06.007)
58. Schwartz J, Dubos M-P, Pasquier J, Zatylny-Gaudin C, Favrel P. 2018. Emergence of a cholecystikinin/sulfakinin signalling system in Lophotrochozoa. *Sci. Rep.* 8, 1-13. (doi:10.1038/s41598-018-34700-4)

59. Packard A. 1972. Cephalopods and fish: the limits of convergence. *Biol. Rev.* 47, 241-307. (doi:10.1111/j.1469-185X.1972.tb00975.x)
60. Patel MV, Hallal DA, Jones JW, Bronner DN, Zein R, Caravas J, Husain Z, Friedrich M, VanBerkum MFA. 2012. Dramatic expansion and developmental expression diversification of the Methuselah gene family during recent *Drosophila* evolution. *J. Exp. Zool. Part B Mol. Dev. Evol.* 318, 368-387. (doi:10.1002/jez.b.22453)
61. de Mendoza A, Jones JW, Friedrich M. 2016. Methuselah/Methuselah-like G protein-coupled receptors constitute an ancient metazoan gene family. *Sci. Rep.* 6, 4-8. (doi:10.1038/srep21801)
62. Araújo AR, et al. 2013. The *Drosophila melanogaster* methuselah gene: a novel gene with ancient functions. *PLoS ONE* 8, 1-12. (doi:10.1371/journal.pone.0063747)
63. Lynch M. 2007. Genomic expansion by gene duplication. In *The origins of genome architecture*, pp. 193-236. Sunderland, MA: Sinauer Associates, Inc.

Chapter 2: Co-evolutionary signatures of gene family expansions and their contribution to cephalopod nervous system evolution

This chapter has been submitted to the peer-reviewed journal *iScience*.

Citation:

E. A. Ritschard, C. Holzinger, C. B. Albertin, P. Imperadore, G. Ponte, G. Fiorito, O. Simakov, Co-evolutionary signatures of gene family expansions and their contribution to cephalopod nervous system evolution. *iScience*. (submitted)

Co-evolutionary signatures of gene family expansions and their contribution to cephalopod nervous system evolution

Christina Holzinger^{1*}, Elena A. Ritschard^{1,2*}, Pamela Imperadore², Giovanna Ponte², Caroline B. Albertin³, Graziano Fiorito², Oleg Simakov¹

¹ Department of Neuroscience and Developmental Biology, University of Vienna, Vienna, Austria

² Department of Biology and Evolution of Marine Organisms, Stazione Zoologica Anton Dohrn, Napoli, Italy

³ Marine Biological Laboratory, Woods Hole, MA, USA

**these authors contributed equally to this work*

Corresponding Authors:

Oleg Simakov oleg.simakov@univie.ac.at

Elena A. Ritschard elena.ritschard@univie.ac.at

Abstract

Zinc finger proteins comprise a large family of transcription factors involved in diverse processes, from suppression of transposable elements to regulation of neuronal development. In some clades, including vertebrates and coleoid cephalopods, members of the C2H2 class of zinc fingers were shown to be expanded and likely relevant for neuron differentiation. Using genomic data from three octopus and three squid species we show that C2H2 evolutionary history can be subdivided into duplication steps: (i) the coleoid expansions, (ii) octopus- and (iii) squid-specific expansions. While the overall expression patterns are consistent among those expansions and are associated with nervous tissues, the coleoid-specific duplicates showed the highest expression levels compared to all other C2H2 genes. We also found that C2H2 duplicates segregated into different co-expression modules, some associated with other duplicated gene families. We propose a scenario of functional divergence of C2H2s duplicates related to regulation within the cephalopod nervous system.

Introduction

Cephalopods are a class of active predatory molluscs, exhibiting a distinct suite of morphological and behavioural characteristics that clearly sets them apart from other taxa in this Phylum. Among the distinguishing novelties we find i.e., the development of flexible arms with suckers, which provide tactile and chemosensory information from the environment [1] and the ability to camouflage, enabled through contraction and expansion of chromatophore cells [2]. These novel features may have contributed to the evolution of cephalopods' agile predatory skills [1,3]. These behaviours are driven by a large and elaborate nervous system, the largest among invertebrates, which has evolved to acquire great complexity among invertebrates, even paralleling vertebrates [4]. Because of their elaborated behavioural repertoire and morphological novelties, cephalopods are ideal candidate organisms for investigating the interconnectedness of evolutionary mechanisms in forming innovations and their underlying genomic contribution [5], especially with a focus on nervous system evolution.

One of the major mechanisms of genome evolution is gene and genome duplication, which, through creation of redundancy, can enable organisms to adapt to surrounding influences and develop complex regulatory gene networks [5]. The redundancy in newly duplicated genes allows for divergence, without the risk of losing the original gene function and the fate of the copies mostly depends on selection pressure. Two of the hypotheses why redundant genes are kept in the genome are neo- and sub-functionalization [5,6]. Neofunctionalization postulates that one copy retains the original function, while the other one accumulates substitutions and gains a completely new function. In contrast, sub-functionalization posits that redundant duplicated genes evolve to maintain together the original function by dividing the work. Previous genome studies [7,8] have found that cephalopods show several gene family expansions through duplication, which are thought to play a crucial role in the development and maintenance of the nervous system. Some of these expansions were previously believed to be only present in vertebrates, including the Cys2His2- type (C2H2) zinc finger transcription factors, protocadherins and G-protein coupled receptors (GPCRs). While protocadherins and GPCRs expansions have been already thoroughly studied *in silico* [9,10], the evolutionary history and underlying mechanisms of the cephalopod C2H2 repertoire and its functional relation to other gene family expansions remains elusive.

C2H2 zinc finger proteins are the largest family of DNA and chromatin binding transcription factors in many metazoans [11], including humans, and are usually composed of groups of tandem zinc finger domains [12]. Combinations of these domains and their interactions allow

for a large number of potential targets, impacting a wide range of cellular processes [13]. In humans, C2H2 transcription factors are prominently involved in early embryonic development and cell fate determination, especially during neurogenesis [14]. Their outstanding diversification in metazoan genomes, with particular examples of gene duplications in the vertebrates and primate branches, hints to an adaptive role of these transcription factors driving evolutionary innovations [13].

In this study, the evolution of the C2H2 gene repertoire in coleoid cephalopods was assessed by constructing a phylogenetic tree of the gene family using a broad sampling of 20 species, including seven cephalopods. This approach allowed us to differentiate between squid-, octopus- and cephalopod-specific (taxon specific) expansions. Additionally, tissue expression was compared between C2H2 expanded and non-expanded genes in four cephalopod species with transcriptomic data available to assess the functional changes of these duplicated genes. To further shed light into the putative regulatory landscape of C2H2s, the genes were assigned to co-expression modules. Lastly, C2H2 co-evolution with other expanded gene families in cephalopods was explored by using gene ontology (GO) term enrichment analysis of co-expression modules. With this combined approach, it was found that expanded C2H2 genes maintained their expression in the brain in cephalopods, similar to the more conserved non-expanded genes. However, expanded C2H2s contributed to different expression modules - here investigated for the first time - suggesting an underlying shift of gene regulation during the expansion of this gene family. These results contrast with the other hallmark cephalopod expansion of G-protein coupled receptors (GPCRs), where the expression of expanded genes shifted to tissues outside of the brain, which could hint at co-evolutionary connections between expanded gene families during development of the cephalopod nervous system. This work contributes to a deeper insight into C2H2 expansion in cephalopods and builds a foundation for future forays into genomic innovation and its relationship to morphological novelties.

Material and methods

(a) Data acquisition and tree construction

A dataset of protein sequences for the C2H2 zinc finger gene repertoire of 20 animal species was compiled from various sources (Table S1). The species were chosen to cover a wide evolutionary range containing protostomes and deuterostomes, thus enabling a broad cross-phyla comparison. Seven of the 20 species belonged to the coleoid cephalopods; namely *Architeuthis dux* (giant squid), *Callistoctopus minor* (long arm octopus), *Doryteuthis pealeii*

(longfin inshore squid), *Euprymna scolopes* (Hawaiian bobtail squid), *Idiosepius paradoxus* (northern pygmy squid), *Octopus bimaculoides* (California two-spot octopus) and *O. vulgaris* (common octopus). The remaining 13 species were *Anopheles gambiae* (African malaria mosquito), *Branchiostoma floridae* (Florida lancelet), *Caenorhabditis elegans* (roundworm), *Capitella teleta* (polychaete worm), *Crassostrea gigas* (Pacific oyster), *Drosophila melanogaster* (fruit fly), *Helobdella robusta* (Californian leech), *Homo sapiens* (human), *Lottia gigantea* (giant owl limpet), *Mizuhopecten yessoensis* (Japanese scallop), *Saccoglossus kowalevskii* (acorn worm), *Schistosoma mansoni* (blood fluke) and *Strongylocentrotus purpuratus* (purple sea urchin).

The sequences for the non-cephalopod species were gathered from the UniProt database using the Pfam C2H2 zinc finger domain ID (PF00096). If possible, only sequences from a UniProt proteome were used (Table S1). Cephalopod whole-genome protein sequences were taken from published data [7,8,15,16,17,18] (Table S1) and scanned against InterPro signatures using InterProScan [19]. Only sequences that matched the signature PF00096 were selected. The FASTA files were filtered by isoforms to only contain the longest protein sequence per gene. Sequences were then aligned using MAFFT (v7.427) and subsequently trimmed using trimAl (v1.4.rev15) with a gap threshold of 0.25, a residue overlap of 0.25 and a sequence overlap of 90%. To find the right model of molecular evolution for the Maximum-likelihood tree construction, the built-in IQTREE (v1.6.2) method ModelFinder was used and resulted in the Whelan and Goldman model (WAG) to be the most fitting amino acid replacement model for the data set. The tree was then constructed with FastTree (v2.1.10) and re-rooted manually in FigTree (v1.4.4). To know the confidence degree of the resulting topology, branch support values were computed by default with FastTree using the Shimodaira-Hasegawa (S-H) test, which compares multiple topologies based on a non-parametric bootstrap [20,21]. The significance threshold of S-H support values here used was 0.7. The largest monophyletic cluster containing only *B. floridae* sequences (314 in total) was used as an outgroup, because this animal represents the most evolutionarily distant species to the cephalopods from the taxa sampled for our phylogenetic analysis. For tree visualization the R package ggtree [22] was used.

(b) Identification of cephalopod expansions

A list of expansion clusters was created, defined as monophyletic groups containing only sequences from the seven cephalopod species of interest. Such clades represent putative gene expansions through duplication, because in a scenario of no duplications it should contain only one gene per species (*i.e.*, only the orthologues). Depending on the species assemblage of the sequences within the cluster, these were classified as cephalopod-, squid-

or octopus-specific expansions. Due to the lack of data availability of other cephalopod lineages, no further classifications were applied.

Following previous findings on the size of expansions in cephalopods (*i.e.*, GPCRs [9] and protocadherins [10]), monophyletic groups with at least 20 sequences solely belonging to either squid or octopus species were counted as Decapodiformes-specific expansions (or squid-specific) or Octopodiformes-specific expansions (or octopus-specific), respectively. However, if a cluster mostly contained sequences of one lineage (either Decapodiformes or Octopodiformes) and less than four sequences from the other lineage, it was still assigned as lineage-specific (*i.e.*, squid-/octopus-specific). Lastly, clusters with at least 30 sequences belonging to both lineage groups (*i.e.*, squid and octopus) were counted as cephalopod-specific expansions.

Following these classification parameters, all the C2H2 sequences analysed were attributed to being either “non-expanded” (*i.e.*, if not belonging to any expansion cluster) or “expanded” (*i.e.*, if belonging to an expansion cluster, either cephalopod-, squid- or octopus-specific).

(c) Statistical analyses of tissue expression

Tissue expression data can help to gain insights into the putative function of genes, by narrowing down the possible cell types where their product, *i.e.*, the protein, is needed. Thus, the expression of the C2H2 zinc fingers was investigated in cephalopod species with a published genome for which RNA-Seq data of different tissues (Table S2) was publicly available (*i.e.*, *C. minor* [17], *E. scolopes* [8], *O. bimaculoides* [7] and *O. vulgaris* [16]). Normalized expression data was used (transcripts per million or TPM [23] in case of *C. minor* and *O. vulgaris*, copies per million or CPM [24] in *O. bimaculoides* and trimmed mean of M values or TMM [24] for *E. scolopes*) for downstream analyses and only comparisons within species were performed. All species, excepting *O. vulgaris*, had one sample per tissue. Because the *O. vulgaris* dataset had three replicates per tissue, the dataset was merged first previous to any of the following analyses by taking the median values of the replicates.

Expression values between non-expanded C2H2s, cephalopod-specific expanded C2H2s and lineage-specific expanded C2H2s (*i.e.*, octopus- or squid-specific, depending on the species at hand) were firstly compared. To test for significant differences between the expression of these three different C2H2 groups in each tissue, a Levene’s test [26] and the non-parametric Fligner-Killeen test [27] were performed to assess normal distribution and variance homogeneity of the data. Because only seven out of 45 tissue datasets exhibited equal variances, the non-parametric test Kruskal-Wallis [28] was used. For datasets with a

significant p-value in the Kruskal-Wallis test, a Dunn's test [29] using the Bonferroni p-value adjustment for multiple comparison was done to check which of the three C2H2 groups showed the significant differences. Additionally, expression values for each gene were normalized by dividing its TPM/CPM/TMM count in a tissue with its maximal value. The resulting relative expression values, ranging from 0 to 1, were then used to visualize the expression trends of each C2H2 group (*i.e.*, non-expanded, cephalopod-specific and octopus-/squid-specific) using boxplots, sorting these from lowest to highest group mean.

The statistical analyses and boxplots were done in R (v.3.6.1) using the packages ggplot2 [30], dunn.test [31] and car [32].

(d) Identification of gene co-expression modules

To further explore the correlation of expression patterns between the different C2H2s expansion groups and other gene families, a weighted gene co-expression network analysis was performed using WGCNA [33,34] in R. First, the normalized expression tables of *E. scolopes*, *O. bimaculoides*, and *C. minor* were log-transformed to reduce skewness of the data towards large values. The same log-transformation was performed on the merged *O. vulgaris* dataset. The clustering analysis was performed block-wise with the *blockwiseModules* function with a minimal module size of 20 genes and using a signed co-expression measure (*i.e.*, network with correlation values ranging from -1, meaning negative correlation, to 1, meaning positive correlation). As suggested by the developers [33,34] for a signed network with less than 20 samples, the soft threshold power used was 18. Module names were assigned by WGCNA automatically using colour names (*e.g.*, MEpink, MEpurple) and the genes found within each resulting co-expression module can be found in Supplementary Data. To summarize the clustering results, module-trait relationships were drawn by correlating the eigengenes, understood as the average expression profiles of the resulting modules, with each adult tissue. The correlation values were finally displayed in heatmaps.

The distribution of C2H2 genes in the resulting co-expression modules was explored with bar charts using the R package ggplot2 [30]. Only modules with a positive correlation to the brain or other nervous tissues were further examined. We defined brain-related modules as those with a significant correlation coefficient higher than 0.7 to one brain-related tissue (*i.e.*, subesophageal-, supraesophageal mass, optic lobes or the brain as a whole tissue). Because in some species expression data was gathered from the different brain regions separately, some modules were correlated to more than one of those separate tissues (*i.e.*, supraesophageal-, subesophageal mass and optic lobe). This resulted in lower correlation coefficients to brain-related tissues despite a putative generalized brain function of the genes.

Thus, modules were also identified as brain-related if they showed a correlation coefficient greater than 0.3 exclusively with two or more brain-related tissues. Furthermore, brain-related modules were split into three groups: (i) modules solely expressed in brain tissues (whole brain, optic lobes, supraesophageal- , subesophageal mass) were referred to as brain modules (in *O. vulgaris* we used the term central nervous tissues, as it included the gastric ganglion and stellate ganglia); (ii) modules solely connected to peripheral nervous tissues or tissues that are known to have a higher number of peripheral nerve cells (*i.e.*, axial nerve cord, eyes, retina, skin, suckers and arm tip) were referred to as peripheral nervous system modules; and (iii) modules expressed in one or more of the aforementioned brain tissues as well as in one or more of the peripheral nervous tissues were referred to as overlapping nervous system modules. Lastly, the distribution of co-expression modules in the phylogenetic tree was explored, focusing on the expansion clades. To do so, a subset of the original tree containing only the expanded cephalopod C2H2 genes was extracted and visualized with ggtree, using the modules as annotation columns.

(e) GO term analysis

A gene ontology (GO) enrichment analysis was performed to identify potential regulatory relationships of C2H2 zinc finger proteins with other expanded gene families in cephalopods. For each of the four species with expression data, the three modules with the highest percentage of expanded C2H2 genes were selected and checked for GO term enrichment. Briefly, this analysis calculated the statistical significance of how often a GO term from the biological processes or “BF” ontology category occurred in that group of genes (*i.e.*, co-expression module) in comparison to all other genes in the genome. The GO term entries of all genes were firstly isolated from the InterProScan results (see section (a) Data acquisition and tree construction). The GO term enrichment analysis was then performed using the R package topGO [35]. Only GO terms with a p-value smaller than 0.05 were kept for further inspection and comparison.

Results

(a) Phylogenetic analysis reveals C2H2 expansions at different time-points during cephalopod evolution

The final filtered and trimmed dataset (available from the Dryad Digital Repository, doi:10.5061/dryad.jh9w0vtfb) used for tree construction consisted of 8888 C2H2 zinc finger protein sequences, 5603 of which belonged to cephalopod species (Figure 1, Figure S1). The majority of C2H2 protein sequences corresponded to *Doryteuthis pealeii* (2442 sequences),

followed by *Octopus bimaculoides* (1453 sequences) and *Euprymna scolopes* (809 sequences). In concordance with previous studies [7,8,15], coleoid cephalopods showed an increased number of C2H2 zinc fingers compared to other protostome species as well as deuterostome species, some of which also have exhibited zinc finger expansions (e.g., human [36] and *Branchiostoma floridae*).

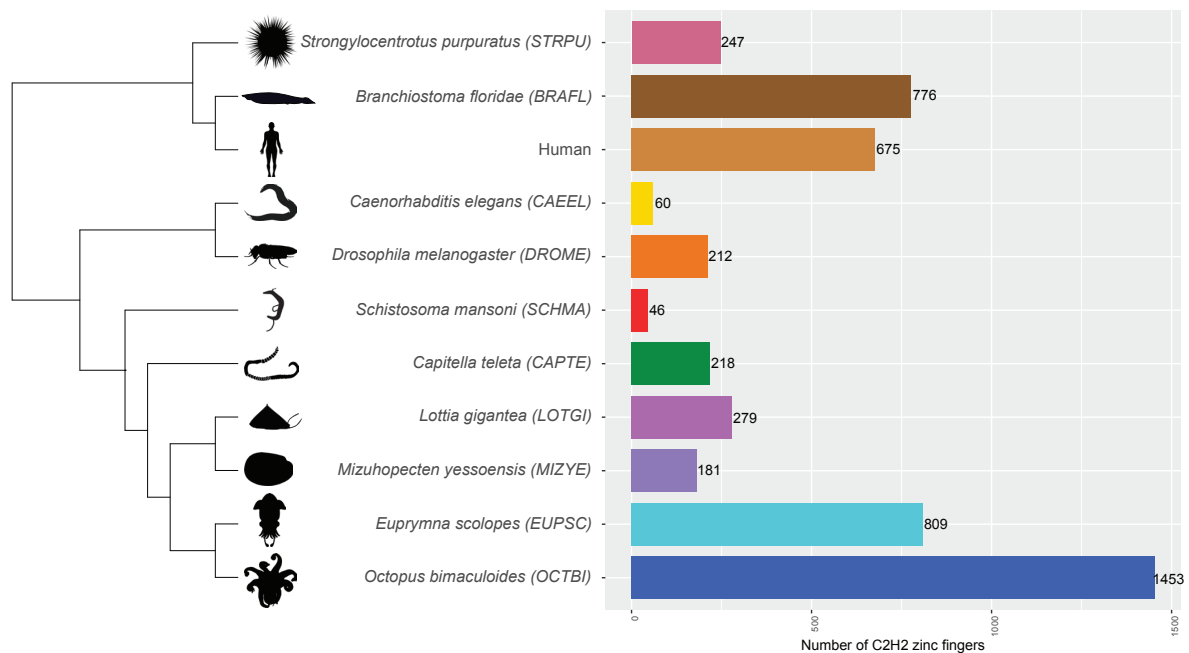
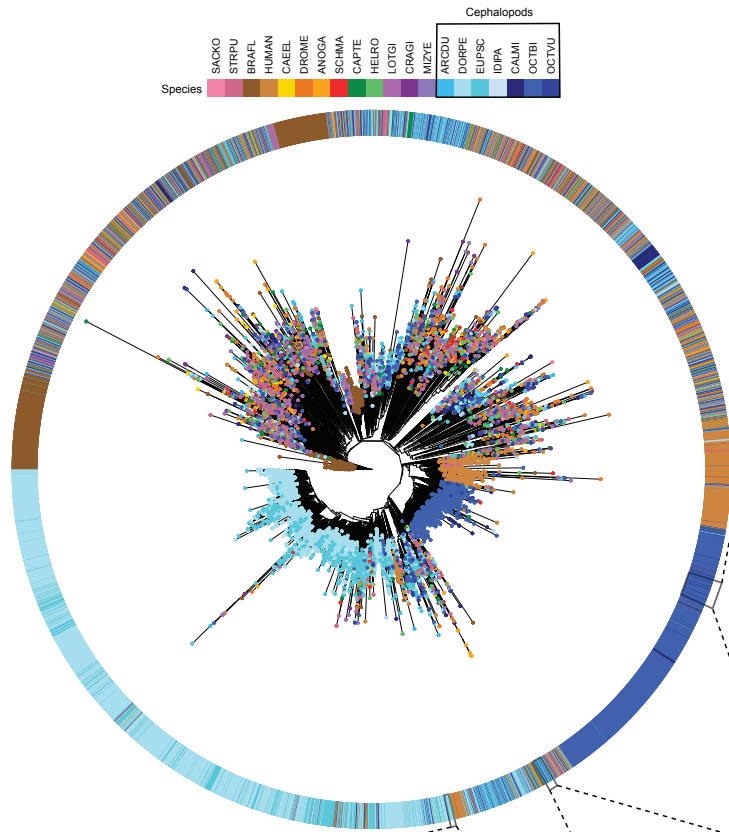


Figure 1. Number of C2H2 protein sequences per species. An assortment of species representing big phyla across the bilaterian clade was chosen from the 20 species used for tree construction to highlight C2H2 expansion in cephalopods. Species were roughly sorted according to their genetic distance, with the three uppermost species (Human, BRAFL, STRPU) being Deuterostomia and therefore acting as sister clade to the other Protostome species, including the coleoid cephalopods. The complete species sampling and their corresponding number of C2H2 protein sequences can be found in Figure S1.

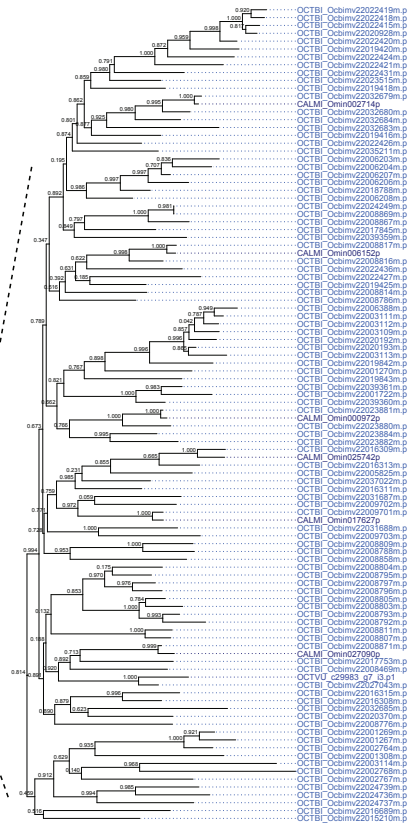
To investigate the C2H2 zinc finger gene expansion in cephalopods more in detail, both as class and in its two main lineages Decapodiformes and Octopodiformes, a phylogenetic tree was constructed. The C2H2 gene family underwent major expansions in cephalopods, evidenced by several big monophyletic groups in the resulting tree (here also referred to as expansion clusters) containing both squid and/or octopus' sequences (Figure 2). In total, 19 expansion clusters met the paralog counting requirements (see Methods, Figure S2, Table S3). The five biggest, in terms of number of paralogues, hold over two thirds of the expanded sequences and were lineage-specific (e.g., octopus-specific cluster 4 with 108 sequences, Figure 2a). We characterized 11 squid-specific expansion clusters, containing mostly *D. pealeii* and *E. scolopes* sequences, with the biggest cluster containing 984 genes (Figure S2a, Supplementary Data). 10 out of those 11 clusters showed a significant degree of confidence

as their S-H support values were above the significant threshold (>0.7). The exception was squid cluster 8 with 36 sequences (Figure S2a, Supplementary Data). Four octopus-specific expansion clusters were identified containing between 104 and 495 sequences (Figure S2b), containing mostly *O. bimaculoides* genes. The biggest octopus cluster (495 sequences) included solely *O. bimaculoides* sequences but did not reach the threshold of significance (<0.7), meaning that there is little signal in the dataset to support this expansion. Furthermore, four smaller cephalopod-specific clusters were found, (Figure S2c) containing 163, 132, 76 and 39 sequences. Three of the four cephalopod clusters had S-H support values greater than 0.7, with cephalopod cluster 1 (163 sequences) being the exception in that group (Figure S2c).

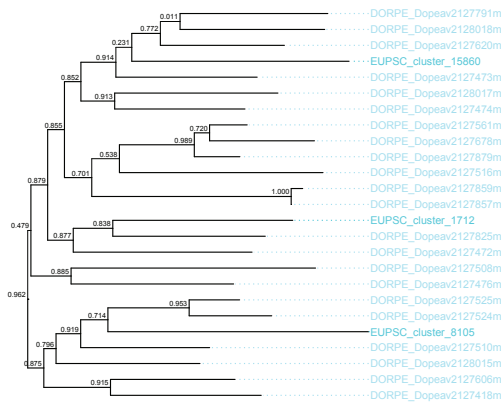
a) C2H2 phylogenetic tree



b) Octo-cluster 4 with 104 sequences



c) Squid-cluster 10 with 25 sequences



d) Cephalopod-cluster 4 with 39 sequences

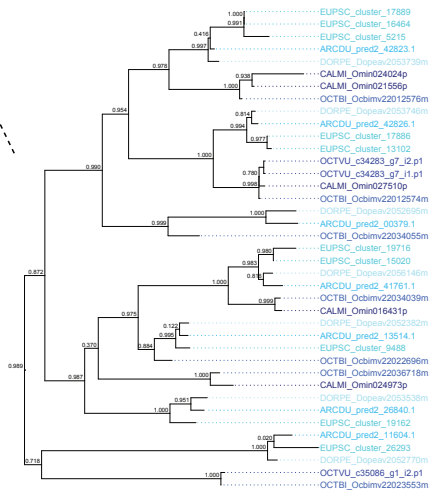


Figure 2. Phylogenetic tree of the C2H2 zinc finger gene family constructed with 20 species. a) Complete tree containing 8888 C2H2 zinc finger protein sequences. Legend uses the following abbreviations: ANOGA: *Anopheles gambiae*, ARCDU: *Architeuthis dux*, BRAFL: *Branchiostoma floridae*, CAEEL: *Caenorhabditis elegans*, CALMI: *Callistoctopus minor*, CAPTE: *Capitella teleta*, CRAGI: *Crassostrea gigas*, DORPE: *Doryteuthis pealeii*, DROME: *Drosophila melanogaster*, EUPSC: *Euprymna scolopes*, HELRO: *Helobdella robusta*, HUMAN, IDIPA: *Idiosepius paradoxus*, LOTGI: *Lottia gigantea*, MIZYE: *Mizuhopecten yessoensis*, OCTBI: *Octopus bimaculoides*, OCTVU: *Octopus vulgaris*, SACKO: *Saccoglossus kowalevskii*, SCHMA: *Schistosoma mansoni* and STRPU: *Strongylocentrotus purpuratus*. Examples of expansion clusters are shown as zoom-ins in panels b, c and d. Node

support values are written above the branches in the trees. All expansion clusters can be additionally found in the Dryad Repository (doi:10.5061/dryad.jh9w0vtfb). b) Octopus-specific expansion cluster containing 104 sequences. c) Cephalopod-specific expansion cluster containing 39 sequences. d) Squid-specific expansion cluster containing 25 sequences.

(b) C2H2 duplicated genes in cephalopods maintained the ancestral brain expression domain but show increased expression intensity

To gain insights about the function of C2H2 zinc fingers in cephalopods and explore the differences between conserved and duplicated genes, expression values of C2H2 genes were analysed in each species individually (*i.e.*, *E. scolopes*, *O. bimaculoides*, *O. vulgaris* and *C. minor*). In *E. scolopes* (Figure S3), *O. bimaculoides* (Figure S4) and *O. vulgaris* (Figure S5), the relative expression mean of all C2H2 genes was found highest in the brain, followed by neural tissues distributed in the periphery (*e.g.*, retina/eyes, skin, gastric ganglion, axial nerve cord). However, in the two octopus' species, which were the only ones with expression data available for each of the three central brain masses, the subesophageal one showed a lower relative expression mean compared to peripheral nervous tissues. Moreover, in *C. minor* the whole central brain had the second highest relative expression mean, following the ovary (Figure S6).

To further investigate putative functional transitions between the different expansion stages, the relative expression of expanded and non-expanded C2H2s genes was analysed separately in each species. In *E. scolopes*, the highest expression of non-expanded and cephalopod-specific C2H2s was found in the brain and the accessory nidamental gland (Figure 3d-e). For the squid-specific expanded genes, the brain was still the tissue with the highest relative expression but followed by the eyes (Figure 3f). In *O. bimaculoides*, the three groups (*i.e.*, non-expanded, cephalopod and octopus expansions) showed similar trends, where the optic lobe, the supraesophageal mass and the axial nerve cord had the highest mean (Figure 3a-c). The relative expression of non-expanded and cephalopod-specific C2H2s in *O. vulgaris* was highest in the optic lobe, the supraesophageal mass and the arm tip (Figure S5). By comparison, the tissues with highest mean of relative expression values in the octopus-specific expansion group were the arm tip and the gastric ganglion, followed by the optic lobe and supra-esophageal mass. In *C. minor*, the tissues with the highest expression in the non-expanded group were the ovary, intestine and testis (Figure S6). The brain was in this group of genes the fourth tissue with highest mean of relative expression. Lastly, both cephalopod and octopus expanded genes showed the highest relative expression values in the ovary and in the brain.

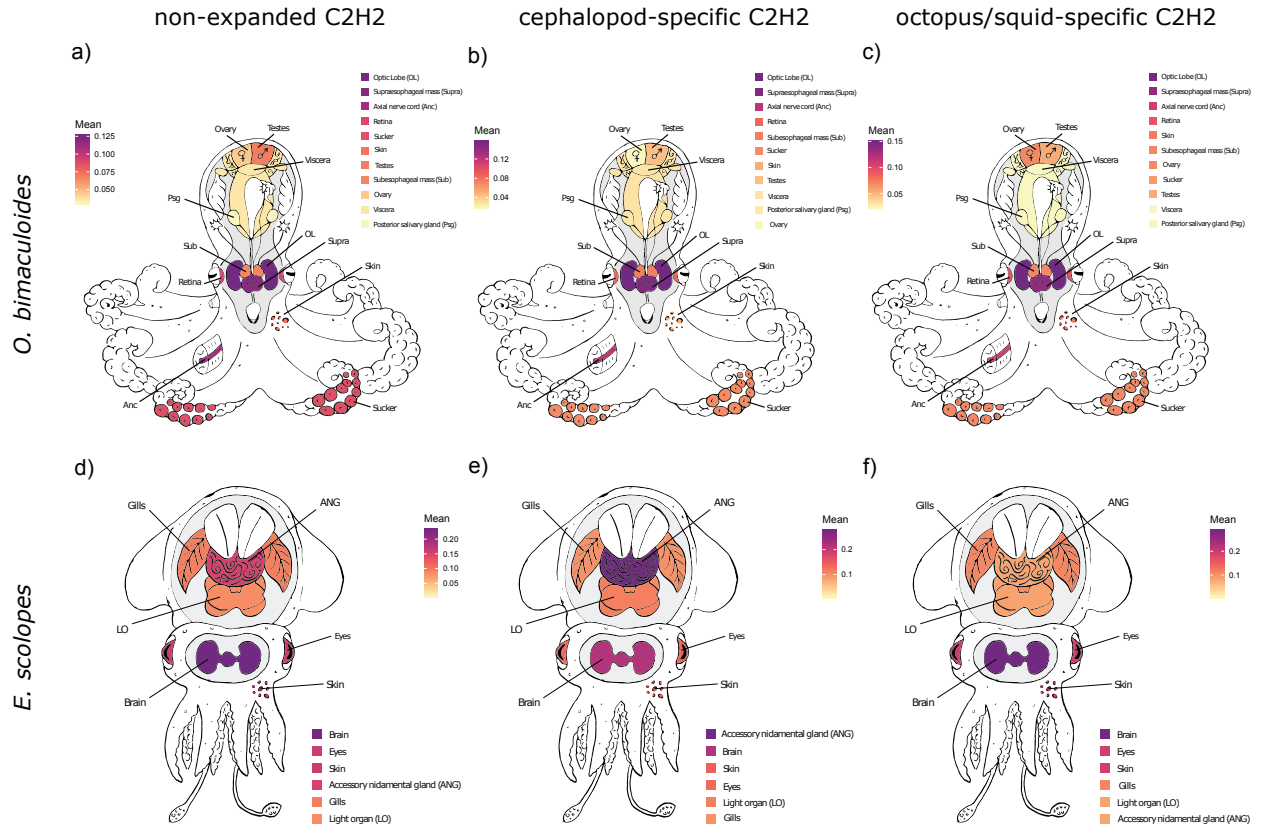


Figure 3. Relative C2H2 tissue expression in *E. scolopes* and *O. bimaculoides*. a-c) Expression of a) non-expanded, b) cephalopod-specific expanded and c) octopus-specific expanded C2H2s in *O. bimaculoides*. Expression of d) non-expanded, e) cephalopod-specific expanded and f) squid-specific expanded C2H2s in *E. scolopes*. Tissues in legend are sorted from highest gene expression (top) to lowest gene expression (bottom). Colours are taken from corresponding boxplots in Supplementary Figures S3-4.

Absolute expression values were also compared statistically between expansion groups but in each tissue (Figure S7-10), to investigate if differences could be found in gene expression intensity. In *E. scolopes* (Figure S7), the cephalopod-expanded C2H2s had the highest absolute expression mean in all tissues. However, significant differences (Table S4) were only found between cephalopod- and squid-specific C2H2s in the gills, skin and light organ (the latter also significantly different between cephalopod- and non-expanded C2H2s), and between all three groups in the accessory nidamental gland. The absolute expression in *O. bimaculoides* (Figure S8) was found significantly higher in the cephalopod-specific expanded C2H2 genes in all but the ovary tissues. These differences were found to be significant between cephalopod-specific and non-expanded C2H2s as well as between cephalopod- and octopus-specific expanded C2H2s in six of the 11 tissues (*i.e.*, optic lobe, supraesophageal mass, axial nerve cord, retina, skin and subesophageal mass). In four tissues (*i.e.*, suckers, testes, viscera and posterior salivary gland), significant differences between all three groups were found. In the ovaries, the expression of octopus-specific expansions was found to be

significantly higher than the other two groups. In the absolute expression of *O. vulgaris* (Figure S9) there were only significant differences in two of 9 tissues (*i.e.*, the optic lobe and the supraesophageal mass), where again the cephalopod-specific expansion was higher than the non-expanded and octopus-specific groups. Despite the lack of statistical significance, in all tissues the cephalopod-specific C2H2s followed the trend in the aforementioned species, with the highest absolute expression mean of all three groups. Lastly, *C. minor* (Figure S10) cephalopod-specific expanded C2H2 genes were also found to be expressed significantly higher in all tissues in comparison to all other expansion groups (only in the liver the significant difference was between the cephalopod-specific and the non-expanded genes).

Altogether, no considerable differences in tissue expression patterns were found between non-expanded, cephalopod-specific and squid/octopus-specific expanded C2H2s in all four cephalopod species, as the highest relative expression mean was generally found in central brain tissues (Figure S3-6, Figure 3). However, in all four species analysed, the expression of C2H2 genes found in cephalopod-specific expansions was consistently higher than the more conserved and octopus/squid expanded genes across all tissues.

(c) Co-expression analysis highlights the close association of expanded C2H2 genes and specific nervous system-related modules

Gene co-expression modules were constructed using a weighted gene co-expression network analysis with WGCNA (Figure 4a) to further investigate putative functional correlation between different C2H2 genes (*e.g.*, non-expanded and expanded) and between C2H2s and other expanded gene families (*e.g.*, GPCRs). The analyses performed for the four species with adult tissue expression data available resulted in 21 modules for *E. scolopes*, 29 for *O. bimaculoides*, 45 for *C. minor* and 75 for *O. vulgaris* (available from the Dryad Digital Repository, doi:10.5061/dryad.jh9w0vtfb).

The distribution of all C2H2s in the resulting co-expression modules was not found to be specific to modules uniquely correlated to nervous system tissues (Figure 4b, Fig. S11-S16). This was particularly true for *E. scolopes*, where only 1 module (*i.e.*, MEgrey, a module with expression in all tissues but the ANG, Figure S13) did not contain any C2H2 genes (Figure 4b). Additionally, the modules showed a general low percentage of C2H2s within them (*i.e.*, number of C2H2s per total number of genes in the module), independent of their expansion state. More specifically, in all four species this percentage was found to be below 10% (Figure 4b, Fig. S11-12). The only exception was the brain-related module MEsalmon in *O. bimaculoides* with 35% of C2H2s (Figure 4b, Figure S14). This outlier case could well be related to the higher number of C2H2 genes found in this species in comparison to the other

three ones (i.e., 1453 in *O. bimaculoides* versus 809, 228 and 280 in *E. scolopes*, *O. vulgaris* and *C. minor*, respectively, Figure S1).

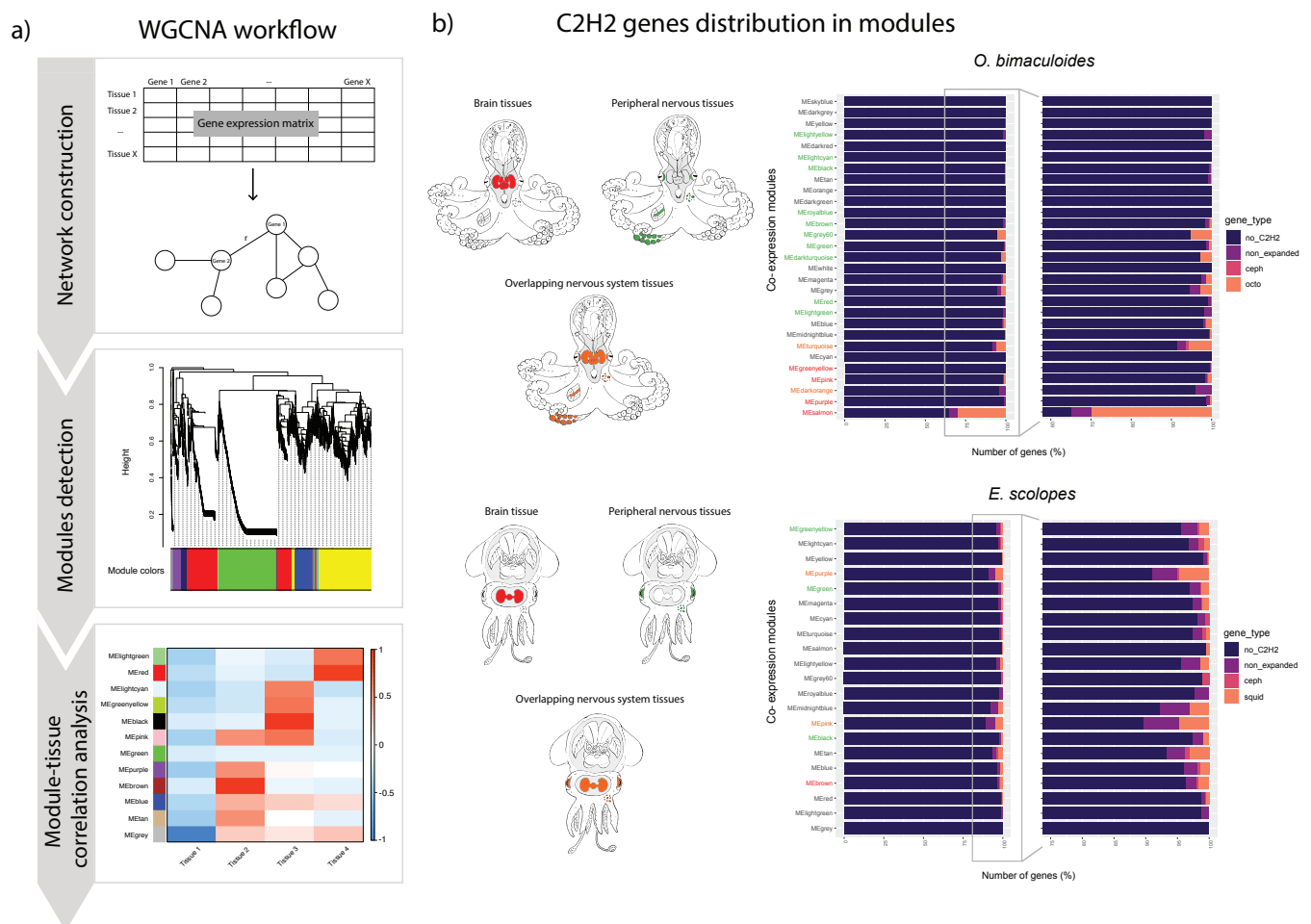


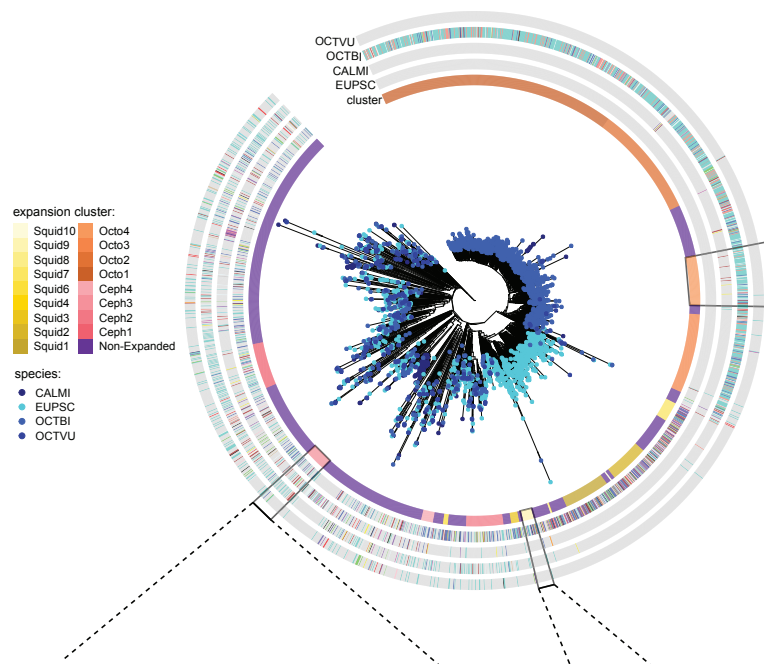
Figure 4. Co-expression analysis with WGCNA and distribution of C2H2 genes in the resulting modules in *O. bimaculoides* and *E. scolopes*. a) A weighted gene correlation network for each species was constructed using a gene expression matrix. The edge weights linking the different nodes (genes) are calculated based on the correlation coefficient (r) between gene expression patterns. Modules are then detected using hierarchical clustering and named after colors (e.g., MEgreen, MEdred). Lastly, correlation between tissues and module eigengenes (i.e., the first principal component of each module) was calculated. b) Distribution of C2H2 genes in the resulting modules in *O. bimaculoides* and *E. scolopes*, shown as percentage of gene numbers (from 0% to 100% in left barplot and zoom-in to a shorter range in right barplot, for greater detail). Module names are color-coded based on their correlation to nervous tissues: red for brain tissues, green for peripheral nervous tissues and orange for overlapping central and peripheral nervous tissues. Modules coloured black were not specifically correlated to nervous tissues.

Despite the lack of a clear association of C2H2s with nervous system-specific modules, some of these modules showed the highest percentages of expanded C2H2 genes. In *O. bimaculoides*, we found MEsalmon (brain-related) and METurquoise (related to both central and peripheral nervous tissues), as these showed the two highest percentages of C2H2s,

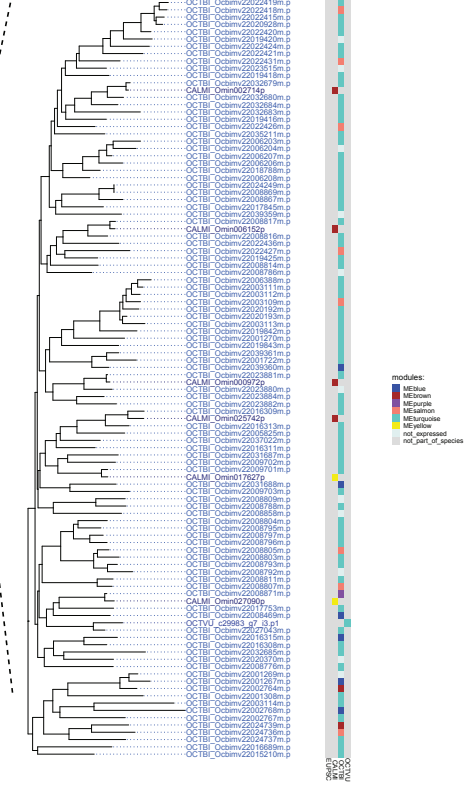
particularly genes from octopus-specific expansions (Figure 4b, Figure S14). In *E. scolopes*, MEpink and MEpurple (both related to central and peripheral nervous tissues) showed the two highest count of C2H2 genes and specifically, squid-specific expanded genes (Figure 4b, Figure S13). Moreover, in *O. vulgaris* we identify MEgreen, MEmagenta and MEturquoise, as all brain-related modules (Figure S15). MEmagenta showed the highest percentage of octopus-specific expanded genes, and cephalopod expanded C2H2s were highest in MEgreen and MEturquoise (Figure S11). Lastly, nervous system-specific modules in *C. minor* containing most cephalopod-expanded C2H2s were the brain-related modules MEbrown, MEturquoise and MEyellow (Figure S12, Figure S16). MEbrown and MEyellow contained additionally octopus-specific expanded genes.

To further investigate the evolutionary relationship of the modules and explore if genes of the same expanded group belonged to the same co-expression module, a phylogenetic tree with only the cephalopod C2H2 genes and their corresponding module was visualized (Figure 5). However, no particular co-expression module corresponded to an expansion group, rather, paralogues belonged to different modules. Altogether, modules with the highest number of expanded C2H2 genes were found to be specifically related to nervous system tissues, but these expression modules are not defined by evolutionary relationships.

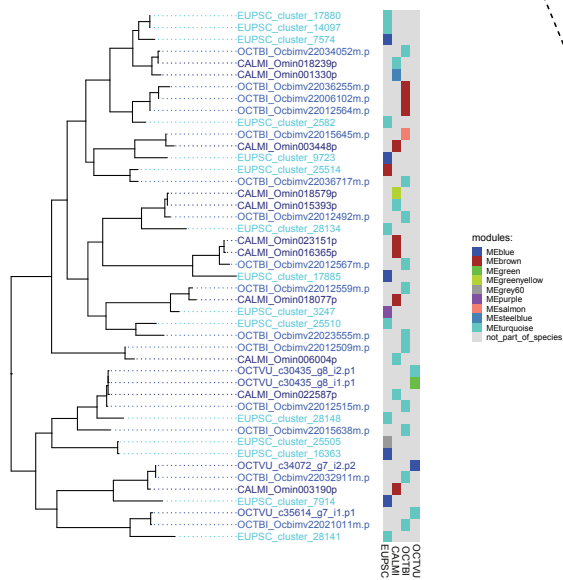
a) Cephalopod C2H2 tree with expression data



b) Octo-cluster 4 with 104 sequences



c) Cephalopod-cluster 3 with 45 sequences



d) Squid-cluster 9 with 28 sequences

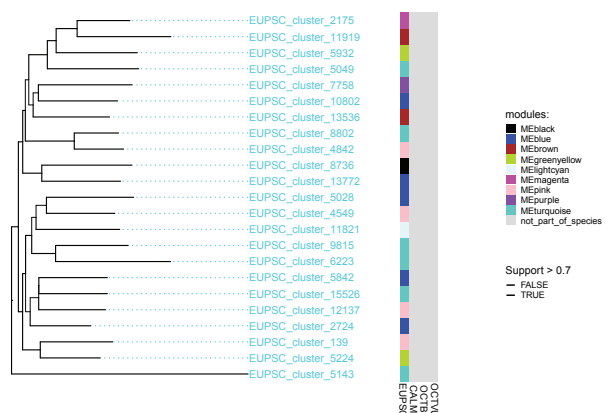


Figure 5. Trimmed tree containing only C2H2 genes of *E. scolopes*, *C. minor*, *O. bimaculoides* and *O. vulgaris* including information on the corresponding co-expression modules. a) Tree with cephalopod C2H2 sequences with expression data. The tree was trimmed to only contain sequences from cephalopod species with expression data, namely *E. scolopes*, *C. minor*, *O. bimaculoides* and *O. vulgaris*. Therefore, zoom-ins on the same clusters as in figure 2 may contain less sequences, if it also holds sequences from a cephalopod species without expression data. Annotation columns give additional information on the co-expression module and the expansion state (cephalopod-specific expansions marked in pink tones; squid-specific expansions marked in yellow tones; octopus-specific expansions marked in orange tones and non-expanded/conserved C2H2s marked in purple). b) Zoom-in on an octopus-specific expansion cluster. c) Zoom-in on a cephalopod-specific expansion cluster. d) Zoom-in on a squid-specific expansion cluster.

(d) Enriched C2H2 modules show putative functional associations with other neuronal gene families

To investigate putative regulatory associations of C2H2 genes in relation to other expanded gene families like the GPCRs, a GO-term analysis was performed in the three modules with highest number of C2H2 genes in all four species (Table S5). These results showed that the most abundant biological process GO-terms were: “GO:0007186 – G protein-coupled receptor signaling pathway” (6 out of 12 modules, Table S5), “GO:0006511 – ubiquitin-dependent protein catabolic process” (4 out of 12 modules), “GO:0006886 – intracellular protein transport” (4 out of 12 modules) and “GO:0006357 – regulation of transcription by RNA polymerase II” (4 out of 12 modules). An additional analysis of the molecular function ontology was done as a control step to identify modules with enriched C2H2 GO-terms. In 7 out of 12 modules significant enrichment of the main C2H2 GO-term “GO:0003676 – nucleic acid binding” was found. Those modules were most of the ones highlighted above in the previous results section, namely MEsalmon and MEturquoise in *O. bimaculoides*, MEpink and MEpurple in *E. scolopes*, MEbrown and MEturquoise in *C. minor*, and MEturquoise in *O. vulgaris*, all of which are either brain or overlapping nervous system modules. Out of the 12 modules with the highest percentages of expanded C2H2s, 2 were found to have enrichments of both C2H2 and GPCR GO terms (*i.e.*, *E. scolopes* MEpurple and *O. vulgaris* MEturquoise), comprising 25 cephalopod-specific and 145 lineage-specific expanded C2H2 genes (Table S5).

Discussion

While previous studies were able to show general association in C2H2 expression with neural tissues in cephalopods [7,8] and strong co-localization of the genes in just a few chromosomes [15], the evolutionary timing and further dissection of their expression patterns remained elusive due to the lack of expression data and genomic sequences of other coleoids. With the inclusion of various squid and octopus' species genomes, we reconstructed expansions of this gene family at different evolutionary periods, distinguishing ancient proteins shared with other metazoans from genes that duplicated during the coleoid radiation. Altogether, our analyses shed light on the evolutionary history and expression dynamics of one of the largest coleoid cephalopod gene families - the zinc finger type C2H2 transcription factors.

(a) Expanded C2H2s in cephalopods maintained their nervous system-related expression

In total, 16 expansion events were found with our phylogenetic approach. The majority of these expansions (*i.e.*, 13 out of 16) originated after the Decapodiformes/Octopodiformes split, while only three emerged in cephalopods before the separation of these major coleoid lineages. Studying the expression of this gene family, we confirmed previous findings [7,8] that C2H2s generally are expressed in nervous tissue. However, the contribution of this gene family differed between the various central brain lobes. In those species with data in hand, namely *O. bimaculoides* and *O. vulgaris*, the contribution of C2H2s was highest in the optic lobe and supraesophageal mass, while the subesophageal mass showed a lower expression than peripheral nervous tissues.

Moreover, no clear shift in tissue expression pattern was found between expansion groups (Figure 3, Fig. S1-4). This result contrasts with the findings on the GPCR gene family expansion in cephalopods, where expanded genes were expressed in other tissues outside of the brain, where conserved genes are consistently expressed [9]. For instance, some of the GPCR duplicates were expressed in the axial nerve cord, the eyes, the skin and the gills. Despite the lack of tissue expression shift, we found a consistent difference between expanded and non-expanded genes in expression levels, as cephalopod-specific C2H2s expression was significantly higher than in non-expanded and octopus-/squid-specific genes in 31 of 45 tissues. Such trends will be interesting to investigate further in the context of neuronal networks where C2H2s are particularly enriched. It is possible that any functional significance of C2H2 expansion may be related to the expansion of the central and peripheral cephalopod nervous system and the associated increase in number of neurons. Further investigation of the expanded genes by *e.g.*, examining the annotation of non-cephalopod sequences of their sister groups in the phylogenetic tree and investigating their expression in other species would also help to better understand the commonalities behind C2H2 zinc finger expansions and contribution to the evolution of nervous systems in metazoans.

(b) Paralogues are present in different gene expression modules suggesting a change in regulation

While overall C2H2 expression remains in the domains of the neuronal tissues and mainly the central brain, further analysis using WGCNA revealed that expanded genes have contributed to different co-expression modules. Even within the same expansion clade, C2H2 genes did not share the same module identity (*e.g.*, Figure 5a). Our prediction is thus that, despite the observation that gene duplicates can occur in tandem [15] and may share regulatory elements, C2H2 duplications in cephalopods have resulted in a switch in their regulation such that

duplicates are often associated with different expression modules. This may be a reflection of selective pressure to diversify regulation that is potentially associated with different neuronal cell types following gene duplication. WGCNA has been previously used to investigate gene regulatory networks in different biological contexts [37], but further analyses would be needed to test our prediction. For instance, because the expression data here analyzed corresponds to whole tissues containing different cell types from independent experimental sources, leveraging single cell transcriptome data from cephalopod brains may provide fruitful insights into the nature of the regulation of these C2H2 module identities.

(c) Contrast to GPCRs - complementary evolutionary paths between gene family expansions in cephalopods?

The paralogous genes of the two largest gene family expansions in cephalopods - C2H2 and GPCRs - have contributed to different tissues. On one hand, expanded GPCRs shifted to expression domains outside the brain possibly through neofunctionalization [9] (Figure 6a). On the other hand, this study found that C2H2s maintained a central nervous-tissue expression. However, this did not imply the preservation of co-expression module identity between paralogous genes (Figure 6b) but rather a divergence into different modules (Figure 6c). These two gene families share the timing of some expansions, as in both GPCRs and C2H2s cephalopod- and octopus-specific expansions were characterized. This combination of shared expansion timing and differences in expression patterns of expanded genes in both families, may suggest a functional link between sensory field expansion in the peripheral nervous system and neuronal diversity. Both C2H2 signatures of brain modules and their evolutionarily coupled GPCR expansions may thus provide cues for when certain circuits have emerged.

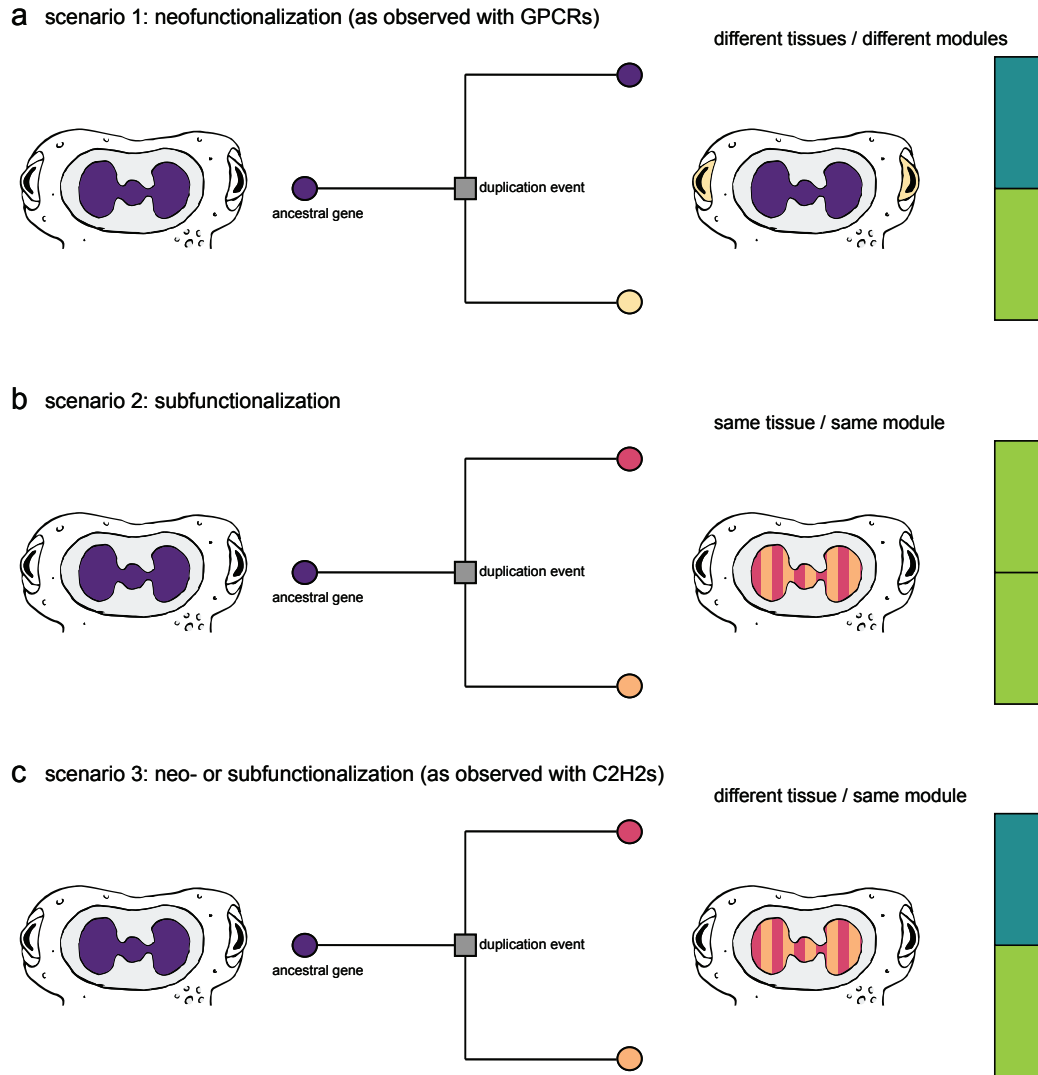


Figure 6. Potential evolutionary scenarios of duplicated genes. a) Scenario 1 represents the classical neofunctionalization hypothesis, where one of the duplicates keeps the ancestral function, while the other gains a new function and is expressed in a different tissue, therefore being allocated to different expression modules (represented here as the green bars on the right side of the figure). This was observed e.g. in cephalopod GPCRs. b) Scenario 2 represents the classical subfunctionalization hypothesis, where the duplicates divide the ancestral function while maintaining expression in the same tissue and are allocated as well in the same co-expression module. c) Scenario 3 represents the results as observed in this study with cephalopod C2H2s, where the duplicates remain mostly expressed in the same tissue as the conserved genes but are allocated to different co-expression modules.

Closing remarks

Our study is the first comprehensive phylogenetic and expression analysis of C2H2 zinc fingers in cephalopods. We characterized expansions that occurred in cephalopods at Decapodiformes (*i.e.*, squid) and Octopodiformes (*i.e.*, octopus) phylogenetic level. We found

that the expanded C2H2 genes maintained the ancestral expression domain: the nervous system and mainly the brain. This result contrasts with previous findings in other expanded gene families, like the GPCRs, where paralogues shifted their expression patterns from the central brain to the peripheral nervous system and non-nervous tissues (Figure 6a). Moreover, further analyses showed us that expanded genes originating from the same expansion events do not belong necessarily to the same co-expression modules (Figure 6c) – here shown for the first time -, suggesting that the functional divergence of paralogues might be related to changes in regulatory processes within the same tissue rather than to the gain of an entirely new function elsewhere. The diversification of recently duplicated C2H2s among distinct expression modules proposes a testable hypothesis of functional diversification, either through neo- or subfunctionalization (Figure 6c). These results set the stage for further dissection of this large gene family and additional data, in particular from single-cell and spatial transcriptomics experiments, will help test these hypotheses and the evolutionary and functional relationship to other gene family expansions.

Data accessibility

Phylogenetic analysis files (sequences, alignments and resulting tree), zoom-ins to the main tree, as well as the WGCNA results were deposited in the Dryad Digital Repository under doi:10.5061/dryad.jh9w0vtfb and can be downloaded using the following link: <https://datadryad.org/stash/share/qR2uGGiAXmCRn6h50aiyyPLcplkjl3a32dtaKw7Yf3I>

Acknowledgements

C.H, E.A.R. and O.S. were supported by the Austrian Science Fund (FWF) grant P30686-B29. Computation was done using the Life Sciences Cluster at the University of Vienna. E.A.R was financially supported by the Stazione Zoologica Anton Dohrn.

Author's contributions

C.H., E.A.R. and O.S. designed the study. C.H. and E.A.R. conducted the analyses and wrote the manuscript, with input from all co-authors.

References

1. Nödl, M. T., Kerbl, A., Walzl, M. G., Müller, G. B., De Couet, H. G. (2016). The cephalopod arm crown: Appendage formation and differentiation in the Hawaiian bobtail squid *Euprymna scolopes*. *Front. Zool.* 13, 44.
2. Cloney, R. A. and Florey, E. (1968). Ultrastructure of cephalopod chromatophore organs. *Zeitschrift für Zellforsch. und Mikroskopische Anat.* 89, 250–280.
3. Packard, A. (1972). Cephalopods and fish: the limits of convergence. *Biological Reviews (Cambridge)* 47, 241–307.
4. Shigeno, S., Andrews, P.L.R., Ponte, G., and Fiorito, G. (2018). Cephalopod Brains: An Overview of Current Knowledge to Facilitate Comparison With Vertebrates. *Frontiers in Physiology* 9(952). 10.3389/fphys.2018.00952.
5. Ritschard, E. A., Whitelaw, B., Albertin, C. B., I. Cooke, R., Strugnelli, J. M. and Simakov, O. (2019). Coupled Genomic Evolutionary Histories as Signatures of Organismal Innovations in Cephalopods: Co-evolutionary Signatures Across Levels of Genome Organization May Shed Light on Functional Linkage and Origin of Cephalopod Novelties. *BioEssays*. 41. 10.1002/bies.201900073.
6. Innan, H. and Kondrashov, F. (2010). The evolution of gene duplications: Classifying and distinguishing between models. *Nat. Rev. Genet.* 11, pp. 97–108.
7. Albertin, C. B., Simakov, O., Mitros, T., Wang, Z. Y., Pungor, J. R., Edsinger-Gonzales, Brenner, E., S., Ragsdale, C. W. and Rokhsar, D. S. (2015). The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*. 524, 220–224. 10.1038/nature14668
8. Belcald, M., et al. (2019). Symbiotic organs shaped by distinct modes of genome evolution in cephalopods. *Proc. Natl. Acad. Sci. U. S. A.* 116, 3030–3035. 10.1073/pnas.1817322116
9. Ritschard, E. A., Fitak, R. R., Simakov, O. and Johnsen, S. (2019). Genomic signatures of G-protein-coupled receptor expansions reveal functional transitions in the evolution of cephalopod signal transduction. *Proc. R. Soc. B Biol. Sci.* 286. 10.1098/rspb.2018.2929.
10. Styfhals, R., Seuntjens, E., Simakov, O., Sanges, R. and Fiorito, G. (2019). *In silico* Identification and Expression of Protocadherin Gene Family in *Octopus vulgaris*. *Front. Physiol.* 9:1905. 10.3389/fphys.2018.01905
11. Najafabadi, H.S., Garton, M., Weirauch, M.T., Mnaimneh, S., Yang, A., Kim, P.M. and Hughes T.R. (2012). Non-base-contacting residues enable kaleidoscopic evolution of metazoan C2H2 zinc finger DNA binding. *Genome Biol* 18, 167.

- 10.1186/s13059-017-1287-y
12. Razin, S. V., Borunova, V. V., Maksimenko, O. G. and Kantidze, O. L. (2012). Cys2his2 zinc finger protein family: Classification, functions, and major members. *Biochem.* 77, pp. 217–226.
 13. Albà, M.M. (2017). Zinc-finger domains in metazoans: evolution gone wild. *Genome Biol* 18, 168. 10.1186/s13059-017-1307-y
 14. Al-Naama, N., Mackeh, R. and Kino T. (2020). C2H2-Type Zinc Finger Proteins in Brain Development, Neurodevelopmental, and Other Neuropsychiatric Disorders: Systematic Literature-Based Analysis. *Front. Neurol.* 11, p. 32.
 15. Albertin, C.B., et al. (2022). Genome and transcriptome mechanisms driving cephalopod evolution. *Nat Commun* 13, 2427. 10.1038/s41467-022-29748-w
 16. Zarrella, I., et al. (2019). The survey and reference assisted assembly of the *Octopus vulgaris* genome. *Sci. Data.* 6, 1–8. 10.1038/s41597-019-0017-6
 17. Kim, B., et al. (2018). The genome of common long-arm octopus *Octopus minor*. *Gigascience.* 7, 1–7. 10.1093/gigascience/giy119.
 18. Da Fonseca, R. R., et al. (2020). A draft genome sequence of the elusive giant squid, *Architeuthis dux*. *Gigascience.* 9, 1–12. 10.1093/gigascience/giz152.
 19. Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. and Lopez, R. (2005). InterProScan: protein domains identifier. *Nucleic Acids Research.* 33:W116–120. 10.1093/nar/gki442
 20. Shimodaira, H. and Hasegawa, M. (1999). Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference, *Molecular Biology and Evolution*, Volume 16, Issue 8. 10.1093/oxfordjournals.molbev.a026201
 21. Planet, P. J. (2006). Tree disagreement: measuring and testing incongruence in phylogenies. *J Biomed Inform.* 39(1):86-102. 10.1016/j.jbi.2005.08.008.
 22. Yu, G., Smith, D. K., Zhu, H., Guan, Y. and Lam, T. T. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28–36. 10.1111/2041-210X.12628
 23. Wagner, G. P., Kin, K. and Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* 131, 281–285. 10.1007/s12064-012-0162-3
 24. Bushel, P. R., Ferguson, S. S., Ramaiahgari, S. C., Paules, R. S. and Auerbach, S. S. (2020). Comparison of Normalization Methods for Analysis of TempO-Seq Targeted RNA Sequencing Data. *Front. Genet.*, 11. 10.3389/fgene.2020.00594
 25. Robinson, M.D. and Oshlack A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 11, R25. 10.1186/gb-2010-11-3-r25

26. Levene, H. (1960). In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, I. Olkin, Ed. Stanford University Press, pp. 278–292.
27. Conover, W. J., Johnson, M. E., Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*. 23, 351–361.
28. Kruskal, W. H. and Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* 47, 583– 621.
29. Dunn, O. J., (1964). Multiple Comparisons Using Rank Sums. *Technometrics*. 6, 241–252.
30. ggplot2 package | R Documentation, (available at <https://www.rdocumentation.org/packages/ggplot2/versions/3.3.3>).
31. dunn.test package | R Documentation, (available at <https://www.rdocumentation.org/packages/dunn.test/versions/1.3.5/topics/dunn.test>).
32. car package | R Documentation, (available at <https://www.rdocumentation.org/packages/car/versions/3.0-10>).
33. Zhang, B. and Horvath S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4. 10.2202/1544-6115.1128.
34. Langfelder, P. and Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*. 9, 559. 10.1186/1471-2105-9-559
35. topGO package | R Documentation, (available at <https://www.rdocumentation.org/packages/topGO/versions/2.24.0>).
36. Liu, H., Chang, L. H., Sun, Y., Lu, X. and Stubbs, L. (2014). Deep vertebrate roots for mammalian zinc finger transcription factor subfamilies. *Genome Biol. Evol.* 6, 510–525. 10.1093/gbe/evu030
37. Singh, A. J., Ramsey, S. A., Filtz, T. M. and Kioussi C. (2018). Differential gene regulatory networks in development and disease. *Cell. Mol. Life Sci.* 75, 1013–1025. 10.1007/s00018-017-2679-6

Chapter 3: Combined contribution of orphan and expanded genes at the developmental and organismal level in cephalopods

Combined contribution of orphan and expanded genes at the developmental and organismal level in cephalopods

Introduction

The previous chapters outlined genomic changes at the coding sequence level in cephalopods via gene duplication, focusing on the GPCRs and C2H2 zinc finger gene family expansions. Although gene duplication and regulatory evolution have been dominant in the understanding of genome diversification and lineage-specific adaptations, orphan genes have started to be considered as another important source for organismal novelty [1]. Orphan genes, also termed taxonomically restricted genes (TRGs), lack homologues in lineages outside of a taxon of interest and have been found to represent up to 20% of every genome [2]. Two main mechanisms for their emergence have been hypothesized: (i) gene duplication and rapid divergence beyond a threshold of sequence similarity and (ii) *de novo* gene birth from ancestral non-coding regions. Evidence in different organisms have shown that the function of the orphan genes is tightly linked to lineage-specific adaptations related to developmental processes and interactions with the environment [1]. For instance, the antifreeze glycoprotein gene AFGP in the Arctic codfishes, which enables their survival in freezing sea temperatures, has been shown to have evolved through *de novo* gene birth [3]. Other examples are orphan genes encoding for antigens involved in host-parasite interactions in parasitic protozoans [4] or vertebrate-specific genes involved in structuring the axon-covering myelin sheaths [5]. Moreover, novel gene emergence throughout evolutionary time has been shown to predate big lineage radiations (e.g., the Mesozoic-Cenozoic radiation and the rosid radiation in flowering plants [1]), hinting not only at a lineage-specific relevance but at their contribution to macroevolutionary events in the history of life on earth.

In cephalopods, orphan genes have been found to be transcriptionally active in e.g., the accessory nidamental gland (ANG) in *Euprymna scolopes* [6], a symbiotic female reproductive organ unique to some squid species [7], and the suckers in *Octopus bimaculoides* [8], multifunctional structures (e.g., tactile and chemosensory) distributed in the ventral part of the cephalopod arms [9]. These first results already give us an idea of the contribution of TRGs to particular cephalopod novelties, but a wider exploration of these genomic features at the developmental and whole organismal level has not been explored yet. Moreover, exploring putative functional linkages of orphan genes with other genomic novelties could help us to

understand how genome evolution has shaped altogether the body plan of these highly divergent mollusks [10]. To which extent have the orphan genes and gene family expansions contributed in combined efforts to the overall evolution of organismal novelties in cephalopods? This chapter approaches this question by characterizing orphan gene complement in different cephalopod species through a wide comparative genomics approach, investigating their transcriptional and regulatory properties through gene co-expression modules and contrasting these results with the previously characterized gene family expansions, *i.e.* GPCRs and C2H2 zinc fingers.

Material and methods

(a) Orthology inference

Proteomes of 30 species covering the major metazoan lineages (Figure 1) were downloaded from various sources (Table 1). 7 of these were cephalopod species, *Callistoctopus minor* (Korean long-arm octopus), *Octopus bimaculoides* (Californian two-spot octopus), *Octopus vulgaris* (Mediterranean common octopus), *Euprymna scolopes* (Hawaiian bobtail squid), *Architeuthis dux* (giant squid), *Doryteuthis pealeii* (longfin inshore squid) and *Idiosepius paradoxus* (northern pigmy squid, species only used for OrthoFinder but not further analyzed due to the lack of a genome and the incompleteness in the dataset). Isoforms were erased to keep only the longest transcript per gene using the corresponding GFF files, if available for the species (Table 1). All-against-all blasts were performed between the proteomes using blastp v2.10.0 and the results used subsequently to infer orthogroups (OGs) and gene orthology with OrthoFinder v2.3.8 [11]. The species tree and gene trees for each OG were constructed in OrthoFinder with FastTree, after performing multiple sequence alignments with MAFFT and trimming with TrimAl using a gap threshold of 0.1.

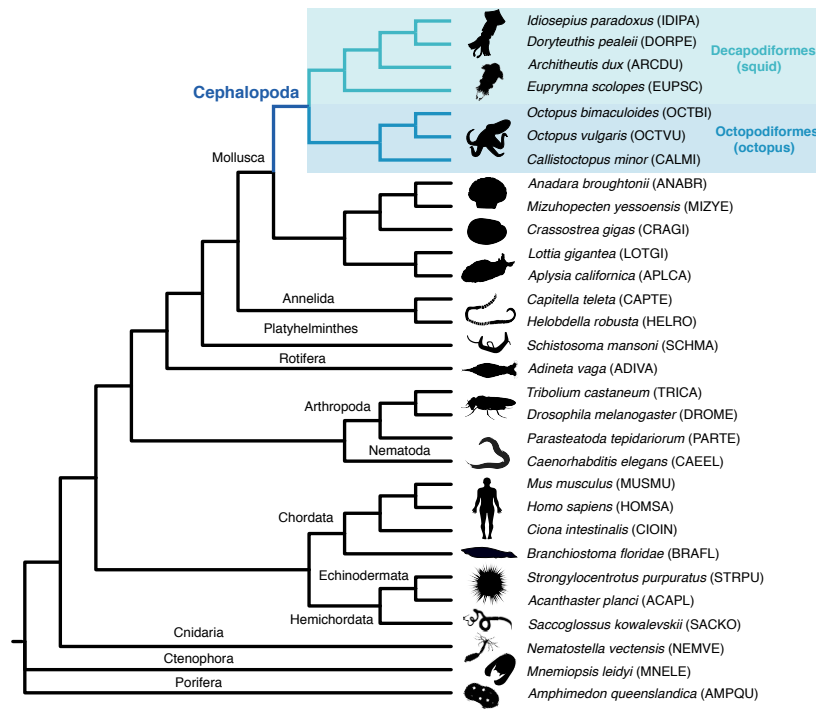


Figure 1. Phylogenetic tree of species sampled for the orthology inference. Transcriptomes of the 30 metazoan species here shown were used as input to OrthoFinder.

(b) Characterization of orphan genes

In order to find genes with no orthologues outside of Cephalopoda (*i.e.*, orphan, also called taxonomically restricted genes), the OGs resulting from the OrthoFinder analysis were firstly characterized based on their species composition. Only OGs with at least two sequences were taken into consideration for the following steps, meaning that unassigned genes were not considered in further analyses. If an OG had no sequences of a non-cephalopod species but (I) at least one sequence of a squid species and one of an octopus' species, it was defined as a *cephalopod-specific* OG, (II) at least one sequence of two different squid species and no octopus' sequence, it was defined as *squid-specific*, (III) at least one sequence of two different octopus species and no squid's sequence, it was defined as *octopus-specific* and (IV) only sequences of one cephalopod species, it was defined as *species-specific*. All other OGs were described as *conserved*.

To double-check the absence of orthology with non-cephalopod species of the resulting putative taxonomically restricted genes defined above, the sequences were additionally blasted against the non-redundant ("nr") database using DIAMOND [12]. OGs containing sequences that only had hits with cephalopod entries (e-value < 0.001) or no hits at all were confirmed to be taxonomically restricted and thus defined as *orphan*. The remaining OGs that did not meet this requirement were characterized as *non-orphan*.

Because cephalopod-specific OGs which had DIAMOND hits non-cephalopod sequences (*i.e.*, non-orphan) could be a product of over-splitting by OrthoFinder, the blast scores of sequences within these OGs were compared with the blast scores of those same sequences with sequences outside the OG. If all within-scores were higher (*i.e.*, better) than the outside-scores, the OG would still be named as *non-orphan* and considered as a cephalopod-restricted family with possible motif similarity to genes outside the lineage. All other OGs that did not fulfill this requirement were characterized as *conserved* in the final annotation.

Lastly, to gain insights into the putative function of the characterized orphan genes at the tissue level of adult animals, gene expression datasets of *E. scolopes*, *C. minor*, *O. bimaculoides* and *O. vulgaris* adult tissues (previously analyzed in Chapter 2) were visualized using the heatmap function as implemented in R.

(c) Assessment of co-expression module correspondence

To explore the correlation of expression patterns between orphan genes and other expanded gene families, a weighted gene co-expression network analysis was performed on the four adult tissue datasets using WGCNA as described in Chapter 2 (*Material and methods, (d) Identification of gene co-expression modules*). Orthology correspondence of the resulting co-expression modules between the four cephalopod species was then assessed by performing pairwise comparisons (6 in total). The number of shared orthologues between modules was counted using the list of orthologues resulting from OrthoFinder. The lineage characterization (*i.e.*, cephalopod-, squid- or octopus-specific orphan genes) was not taken into account for the counts but just if these were orphan genes or not. The percentage of shared orthologues per module was then calculated in each species, using as a denominator the number of genes in that module that have orthologues in both species (*i.e.*, the genes that belong to a module but had no orthologue in the other species were not counted). Because in each species the percentage of shared orthologues in module comparison could result in differing numbers (*e.g.*, in species 1 the compared module was smaller than the module in species 2), only the highest percentage between those two modules was taken as a result and used to visualize in a heatmap. For the latter, heatmap was used.

(d) Collection of *E. scolopes* embryos

E. scolopes cultured eggs were obtained from the Marine Biological Laboratory (MBL) Cephalopod Initiative in Woods Hole, MA, USA. Animals were collected from the same clutch at 8 different developmental stages, following Lee *et al.*, 2009 [13] staging series. Namely, stages 14 and 16, representing gastrulating embryos, 18 and 20, representing early organogenesis, 22 and 24, representing late organogenesis, and 27 and 28, representing pre-

hatchlings. Per stage 3 replicates were sampled, each containing 5 embryos. The animals were collected in Eppendorf tubes after removing the jelly layers, egg cases and the yolk. Seawater was removed as much as possible from the tubes, 50 µl of TRIzol® Reagent (Life Technologies, Carlsbad, California, United States) was added and samples were flash frozen in liquid nitrogen and stored in -80°C until RNA-Seq library preparation.

(e) RNA extraction and sequencing

To extract RNA from the developmental stages of *E. scolopes*, frozen samples were first grinded with a sterile (*i.e.*, RNase-free) pestil. 950 µl of cold TRIzol® Reagent was then added (for stages 14, 16, 18 and 20 only 450 µl), mixed and let sit for 5 minutes at room temperature. 200 µl of chloroform isoamyl alcohol was added to the samples and shaken for 15 seconds (for stages 14, 16, 18 and 20 only 100 µl). After 10 minutes of rest at room temperature, the samples were centrifuged for 15 minutes at 12'500 revolutions per minute (rpm) and 4°C. The supernatant was then transferred to a new Eppendorf tube, 400 µl of isopropanol was added (for stages 14, 16, 18 and 20 only 200 µl), shaken for 15 seconds and let sit for 10 minutes at room temperature. After centrifuging the samples for 10 minutes (12'500 rpm and 4°C), a pellet could be seen at the bottom of the tubes and the supernatant was discarded. 700 µl of cold 70% ethanol was added to the tube (for stages 14, 16, 18 and 20 only 350 µl), vortexed for 15 seconds and centrifuged for 5 minutes (12'500 rpm and 4°C). The supernatant was discarded, the tube was let open upside down for around 10 minutes at room temperature until the pellet was dry. Finally, the pellet was resuspended in 20 µl of double-distilled water, ddH₂O (for stages 14, 16 and 18 only 10 µl). Successful RNA extractions were evidenced in an agarose gel and subsequently stored at -80°C until sequencing.

Illumina mRNA library was prepared using the NEBNext® Poly(A) mRNA Magnetic Isolation Module (New England BioLabs® Inc.) and sequenced by 75 bp paired end mode with NextSeq 500 Mid Output using 300 cycles at the Vienna BioCenter Core Facilities (VBCF) NGS. Quality control of the reads was performed with FastQC and results were aggregated with MultiQC to compile one single HTML report for all sequenced samples. The reads were mapped to the *E. scolopes* genome [6] with STAR [14] using the ENCODE default settings. The count matrix was then generated with featureCounts as implemented in the subread package [15] using a non-redundant GTF file (*i.e.*, with overlapping exons merged into one) and the transcript ID as the attribute type. Raw counts were then normalized to Transcripts per Million (TPMs). Gene expression of orphan genes, GPCRs and C2H2s in the *E. scolopes* developmental stages was visualized using the pheatmap function as implemented in R.

Results and discussion

- (a) Cephalopod-, squid- and octopus- specific orphan genes contribute to a wide arrange of tissues

The gene orthology inference with OrthoFinder allowed us to characterize the putative evolutionary origin of the genes in the cephalopod species based on the OG to which these belonged to (Table 1). The final annotation of OGs after checking for DIAMOND hits with non-cephalopod sequences resulted in 1013 non-orphan OGs, from which 790 were renamed to conserved after assessing possible over-splitting of gene families. Thus, our mixed approach reduced the number of putative taxonomically restricted genes and ultimately resulted in the characterization of around 500 orphan genes in the three squid species analyzed and almost the double (around 1000) in the octopus' species (Table 1, Figure 2). In the three octopus' species analyzed, the number of octopus-specific orphan genes was higher than the cephalopod-specific ones. The contrary was found for the squid species (Figure 2b). These results could suggest a higher contribution of orphan genes to the evolution of octopus' species in comparison to squids. However, the differences in time divergence of the lineages represented by our taxon sampling could also be affecting the differing gene numbers. The three squid species here used share their latest common ancestor at the Decapodiformes split, which happened around 100-120 million years ago (mya), while the last common ancestor of *O. bimaculoides*, *O. vulgaris* and *C. minor* dates around 80-90 mya at the Octopodoidea node [16]. The more recent split of the Octopodoidea lineage could influence on the higher number of octopus-specific orphan genes, as these could represent younger genes that haven't been lost in comparison to an older lineage like the Decapodiformes. Performing further phylostratigraphic analyses could help to better understand the rates of orphan gene emergence across cephalopod evolution and their differential contribution to specific lineages within [17].

Table 1. Summary of the cephalopod-restricted genes characterization in each cephalopod genome analyzed. The number of genes by OG type is previous to performing DIAMOND searches and assessing over-splitting. Species abbreviations are as follows: ARCDU: *A. dux*, DORPE: *D. pealeii*, EUPSC: *E. scolopes*, OCTVU: *O. vulgaris*, OCTBI: *O. bimaculoides*, CALMI: *C. minor*.

SPECIES	# GENES BY OG TYPE				# GENES BY FINAL ANNOTATION					TOTAL OF GENES
	non-ceph	ceph	squid/octopus	species	conserved	non-orphan	ceph orphan	octo/squid orphan	species orphan	
ARCDU	29994	1156	608	124	31172	143	321	231	15	31882
DORPE	18893	723	422	6	19396	114	310	218	6	20044
EUPSC	19712	949	643	85	20791	108	230	246	14	21389
OCTVU	15301	643	898	85	15689	66	289	800	83	16927
OCTBI	20733	921	1241	41	21455	124	367	966	24	22936
CALMI	25290	1178	1203	80	26581	129	356	643	42	27751

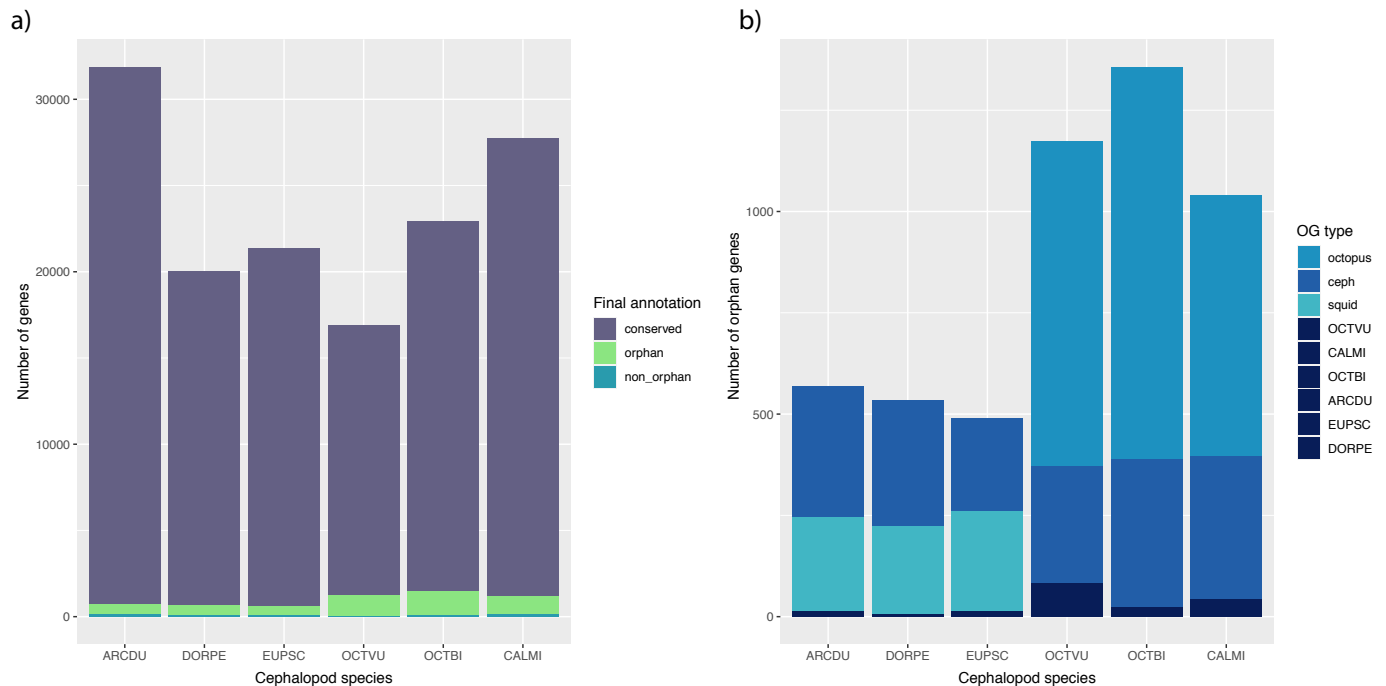


Figure 2. Number of conserved, non-orphan and orphan genes in each species. a) Total number of genes color-coded by their final annotation. b) Number of characterized orphan genes color-coded by the OG type they belong to. Species abbreviations are as follows: ARCDU: *A. dux*, CALMI: *C. minor*, DORPE: *D. pealeii*, EUPSC: *E. scolopes*, OCTBI: *O. bimaculoides*, OCTVU: *O. vulgaris*.

To gain insights into the putative function of the characterized orphan genes, their expression in adult tissues of four species with data available (*i.e.*, *E. scolopes*, *C. minor*, *O. bimaculoides* and *O. vulgaris*) was explored (Figure 3 and 4). In all species, orphan genes could be found to be expressed in a wide variety of tissues and did not show co-expression patterns associated to gene type, as genes showing high expression in the same tissue were annotated

as cephalopod-, squid/octopus- and species-specific (Figure 3 and 4). This suggests that the emergence of orphan genes at a particular evolutionary time (e.g., at the Coleoidea node or in more recent lineage splits) did not contribute to the evolution of a singular tissue in cephalopods.

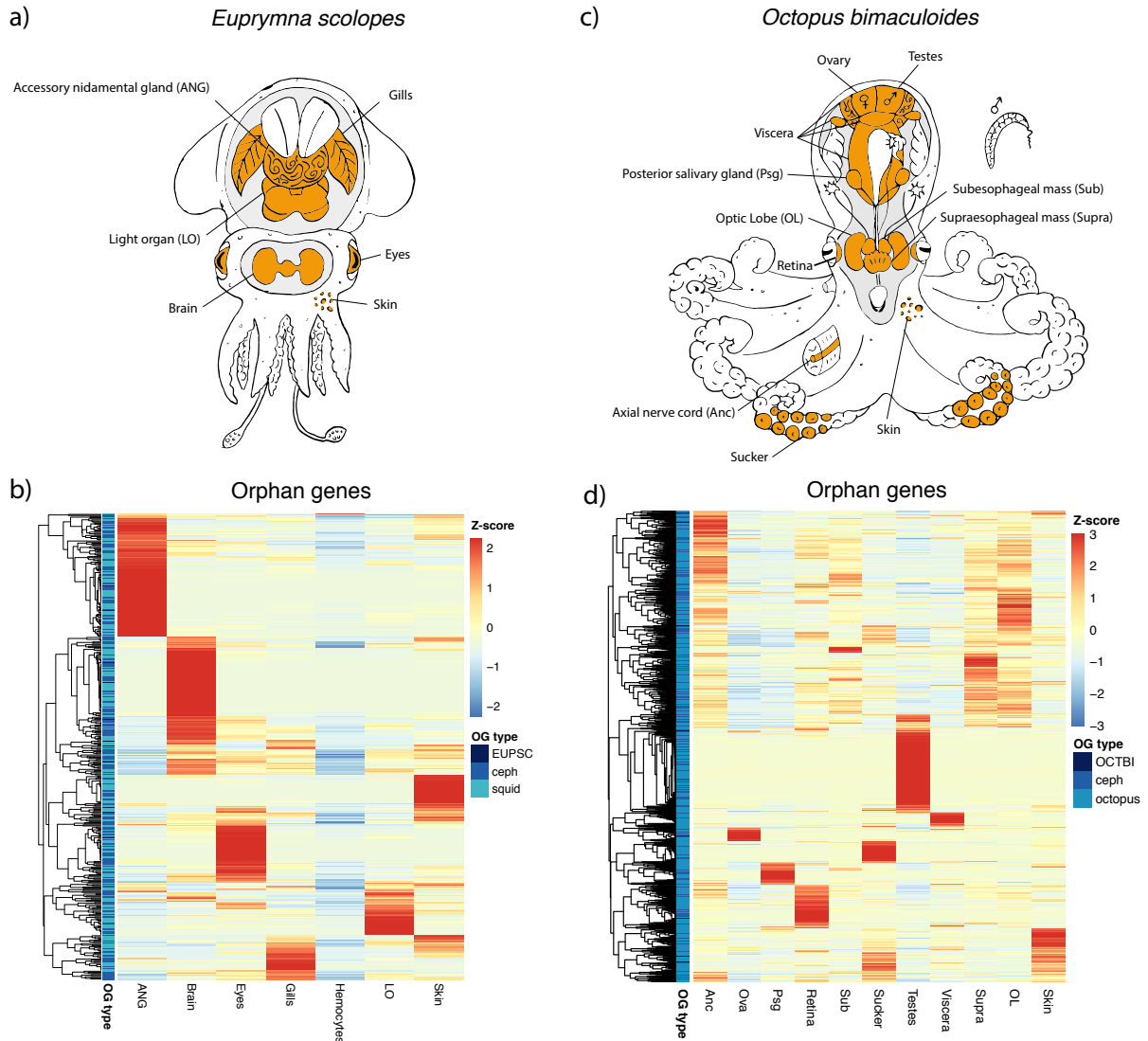


Figure 3. Expression of orphan genes in *E. scolopes* (a-b) and *O. bimaculoides* (c-d). Tissues in the heatmap columns are represented in the animal sketches, colored in orange (excepting hemocytes, the blood cells, of *E. scolopes*). The annotation column in the heatmaps show the final characterization of the orphan genes, defined by which type of OG these belong to (i.e., cephalopod-, octopus-/squid- or species-specific). Species abbreviations are as follows: EUPSC: *E. scolopes*, OCTBI: *O. bimaculoides*.

Despite the lack of a singular contribution of orphan genes to a few tissues, we could find different groups of genes with a tissue-specific expression pattern. For instance, in *E. scolopes*, the ANG and the brain showed big clusters of orphan genes expression (Figure 3a-b), consistently with Belcaid *et al.* 2019 [6]. The retina in *O. bimaculoides* (Figure 3c-d) and

eyes in *E. scolopes* and *C. minor* (Figure 4c-d) also showed a big cluster of orphan gene expression. Further investigation of these results could help to understand the role of genomic novelties in the convergent evolution of visual systems of cephalopods and vertebrates. Moreover, in *O. vulgaris*, a group of orphan genes could be found with high and specific expression in the arm tip (Figure 4a-b). This is a tissue with a high neuron-to-muscle ratio, meaning that a group of orphan genes may be contributing to the peripheral nervous tissue associated with the octopus' arms [18]. Lastly, in both *O. bimaculoides* (Figure 3c-d) and *C. minor* (Figure 4c-d), a high number of orphan genes showed a specific high expression in the testes. This has been shown to be a trend in newly evolving genes and can be associated to the hypertranscriptional state of cells in this tissue, which increases the chances of mutations, and sexual selection mechanisms positively selecting these new genes [1].

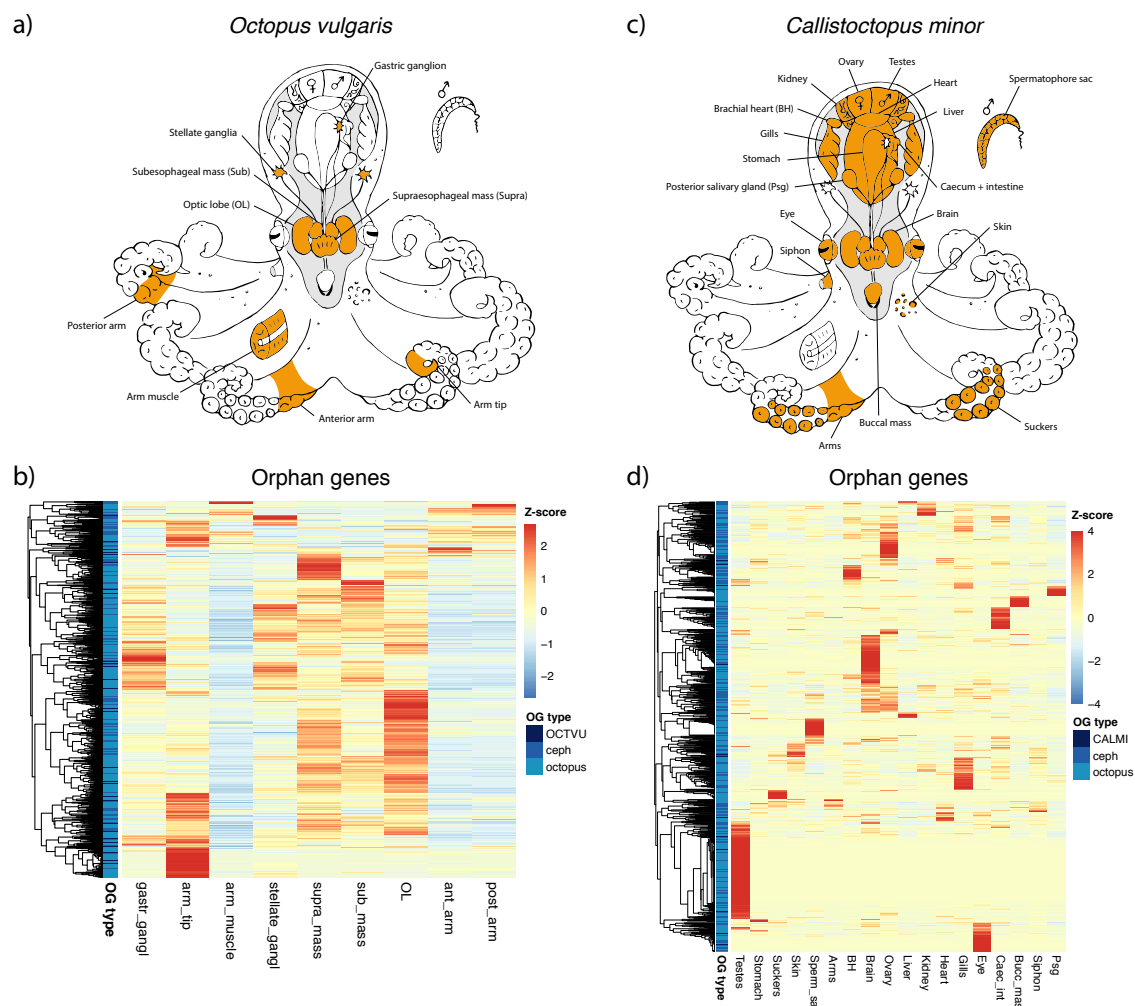


Figure 4. Expression of orphan genes in *O. vulgaris* (a-b) and *C. minor* (c-d). Tissues in the heatmap columns are represented in the animal sketches, colored in orange. The annotation column in the heatmaps show the final characterization of the orphan genes, defined by which type of OG these belong to (*i.e.*, cephalopod-, octopus- or species-specific). Species abbreviations are as follows: CALMI: *C. minor*, OCTVU: *O. vulgaris*.

(b) The evolution of gene novelties in nervous tissues is consistent between octopus' species

We sought to gain further knowledge on the evolution of expanded and orphan genes in cephalopods by understanding whether their tissues' contribution is consistent between species. To do so, WGCNA co-expression modules of *O. bimaculoides*, *O. vulgaris*, *C. minor* and *E. scolopes* were contrasted using the orthology results from OrthoFinder by counting orthologues shared between modules of the different species.

Correspondence of orthologues, not taking into account the type of genes these were, was found between tissue-specific modules but with a percentage usually lesser than 50% (Figure 5). These highest correspondences were found between species that shared similar tissue sampling like *C. minor*, *O. bimaculoides* and *E. scolopes* (Figure 5a, d, e). Interestingly, the separate brain masses did not show a particular high percentage of shared orthology between species, whereas the brain tissue as a whole did show a higher one. This could be related to the differences in mass composition of the species that are hypothesized to be a consequence of developmental- and/or ecological adaptations [19].

From the counts of shared orthologues between modules, the annotation of these genes was further checked to see how much co-expression overlap could be found between gene novelties (*i.e.*, expanded GPCRs, C2H2s and orphan genes). Comparing the resulting modules in the 3 octopus' species, highest numbers (*i.e.*, between 30 and 60 genes) of orthologous orphan genes and C2H2s were found to be overlapping in central nervous system-related modules (*i.e.*, expressed in all sub- and supraesophageal masses, optic lobe in *O. vulgaris* and *O. bimaculoides*, and brain in *C. minor*). Around 20 orthologous orphan genes were additionally found to be enriched in modules specific to the eye/retina, suckers/arm tip and the posterior salivary gland each. Another specific case of high orphan genes orthology found was between testes-specific modules of *C. minor* and *O. bimaculoides*, with up to 50 shared orthologues. GPCRs did not overlap in enrichment of co-expression modules with the other gene novelties and between species the tissue expression was mostly not found to be similar.

When comparing octopus' species with *E. scolopes*, the tissue correspondence tends to disappear. Orthologous genes annotated as expanded C2H2s and orphan genes are shared in modules with differing tissue expression; while being expressed in nervous tissues in octopus' species, the orthologous corresponding modules were expressed in a variety of tissues, mainly in the ANG. Orthologous genes of expanded GPCRs were also found to be

shared on modules with very divergent tissue expression, when comparing between octopus' species and with *E. scolopes*. While this could be an indication of differential specific selective pressures in the cephalopod lineages (Decapodiformes and Octopodoidea), this result should be taken precautionous because it can be also an artifact from the tissue sampling differences between the species.

buccal mass, Syst_heart: systemic heart, Branch_heart: branchial heart, Cec_int: Caecum and intestine, Sp_sac: spermatophore sac, Psg: posterior salivary gland, Stell_gangl: stellate ganglia, Gastr_gangl: gastric ganglion, Post_arm: posterior arm, Ant_arm: anterior arm, Anc: axial nerve cord, Supra and Supra_mass: supraesophageal mass, Sub and Sub_mass: subesophageal mass, OL: optic lobe, LO: light organ and ANG: accessory nidamental gland. The viscera tissue in *O. bimaculoides* consists of the heart, kidney, and hepatopancreas.

Altogether, we found consistent patterns of novel gene evolution across octopus' species in relation to their nervous systems, with (i) C2H2s and orphan genes specifically contributing to the brain and (ii) other orphan genes enriched in tissues with high number of peripheral nerve cells, like the eyes and suckers. Further dissection of such genes at the cellular level with single cell and spatial transcriptomics, as well as explorations of gene location within the highly rearranged genomic architecture of cephalopods, would help to unravel putative functional links of gene novelties within and among nervous tissues, contributing to a better understanding of the genomic contribution to neuronal networks in cephalopods.

(c) Gene novelties in the early development of *E. scolopes*: differential contributions of expanded GPCRs, C2H2s and orphan genes

Gene regulation offers a relevant insight into the understanding of gene novelties evolution, as it represents functional linkages between genomic elements through their interactions [10]. One starting approach to regulation is by exploring expression dynamics throughout development. Therefore, RNAseq developmental time-course of the model species *E. scolopes* was generated and the expression of the gene novelties characterized in Chapters 1 (*i.e.*, GPCRs), 2 (*i.e.*, C2H2s) and here (*i.e.*, orphan genes) was investigated.

Conserved GPCRs were found to be mostly expressed at the latest stage, stage 28 (Figure 6), when the animal has reached the features of the adult form and is ready to hatch [13]. Despite not having grown to the number of nerve cells that an adult brain would have, by that timepoint the animal would have developed the structures that make up a functional adult brain, as shown in other cephalopods (*i.e.*, *Octopus vulgaris* [20]). A small group of GPCRs, though, showed also high and very specific expression at the earliest stage sampled (*i.e.*, st. 14), a 40% gastrulation stage. The expanded GPCRs showed a similar trend, where most peaked in expression at the pre-hatchling stages (st. 27-28) and just two in stage 14. Some genes were also expressed in stage 24, when eyes are morphologically distinct, and pigmentation of the retina increases. Most of these genes showing expression during the early development of *E. scolopes* are from the expansion group 1, a group of cephalopod expanded neuropeptide receptors that showed high expression in the eyes and the ANG of the Hawaiian bobtail squid [21]. Further dissection of these results with a more in-detail exploration of each

genes annotation, could provide insights into the development of the eyes and the ANG in *E. scolopes*, as well as the role of novel GPCRs in gastrulation.

Conserved and expanded C2H2 genes showed a broader gene expression in comparison to GPCRs, as these genes also get activated in middle stages and showed less stage-specific expression profiles (Figure 7). One stage that stands up in high expression of squid-specific C2H2s is stage 20, an early organogenetic stage in which the optic vesicles are almost complete and arm buds are distinguishable. Deryckere *et al.* 2021 [22] showed that in a comparable early organogenetic stage of *O. vulgaris* neuronal differentiation is already present, a process in which C2H2s have been found to be relevant [23]. Moreover, enriched expression of zinc fingers has been found in precursor cells of the *O. vulgaris* developing brain, highlighting the importance of these gene novelties during embryogenesis of cephalopod nervous systems [24]. Moreover, as it was in the case of the expression of C2H2s in adult stages (Chapter 2), genes belonging to the same expansion cluster did not share the same expression pattern, following the hypothesis of regulatory divergence of C2H2 paralogues.

Lastly, around a third of orphan genes from all three categories (*i.e.*, cephalopod-, squid- and species-specific) showed a specific contribution to the latest stage 28 and another third seems to be active from stage 27 on, while the minority of orphan genes were found to be expressed in early to middle stages (Figure 8). Such pattern is consistent with findings that highlight adult (in this case, an embryonic stage resembling already the adult) transcriptomes are younger than those of early embryonic stages and that middle stages represent a conserved embryonic form among species either because of strong constraints or the lack of need for adaptations to environmental conditions [25].

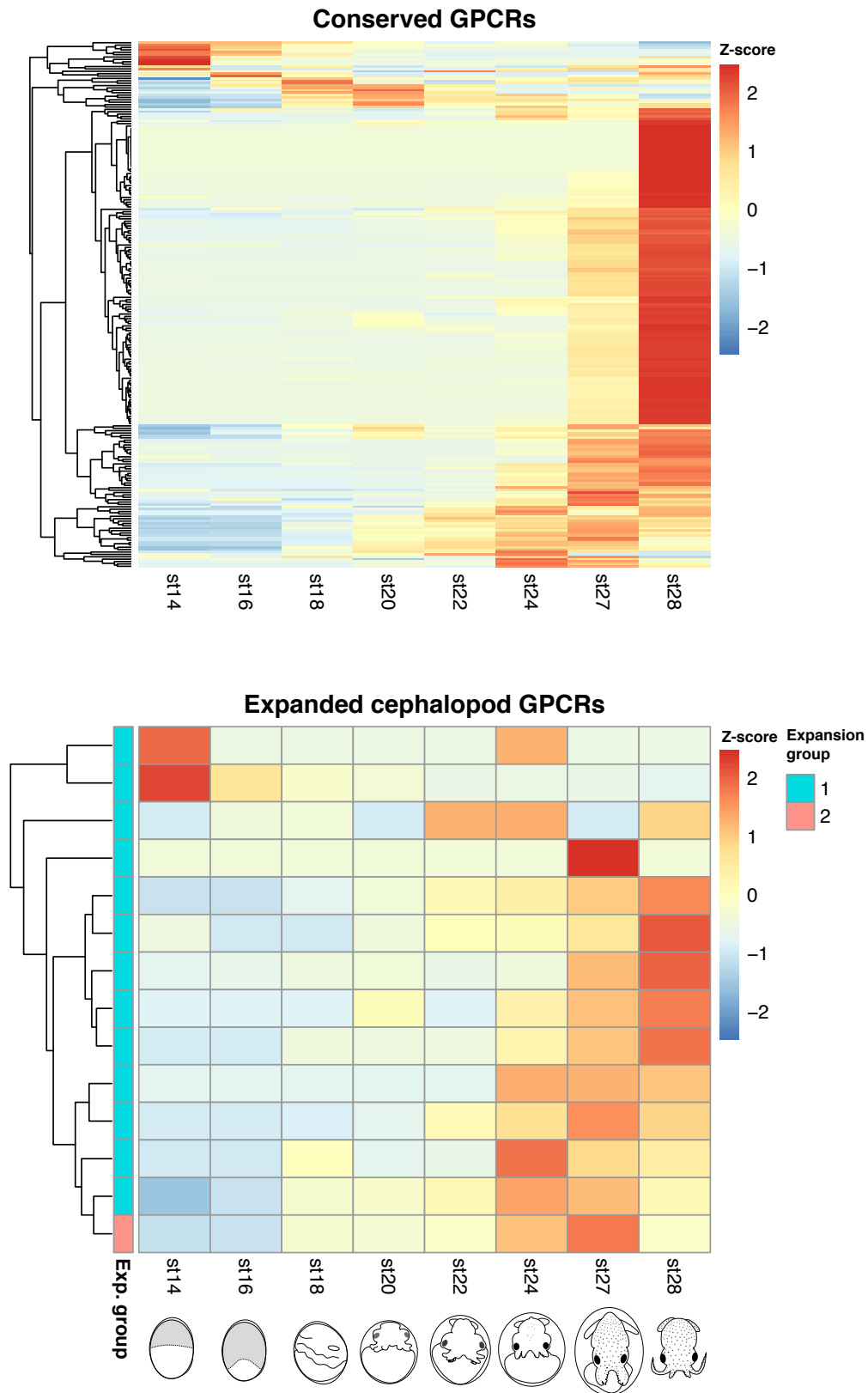


Figure 6. Gene expression of conserved and expanded GPCRs in the early development of *E. scolopes*. The annotation column in the heatmaps show the expansion group as characterized in Chapter 1 [20] to which the genes belong to.

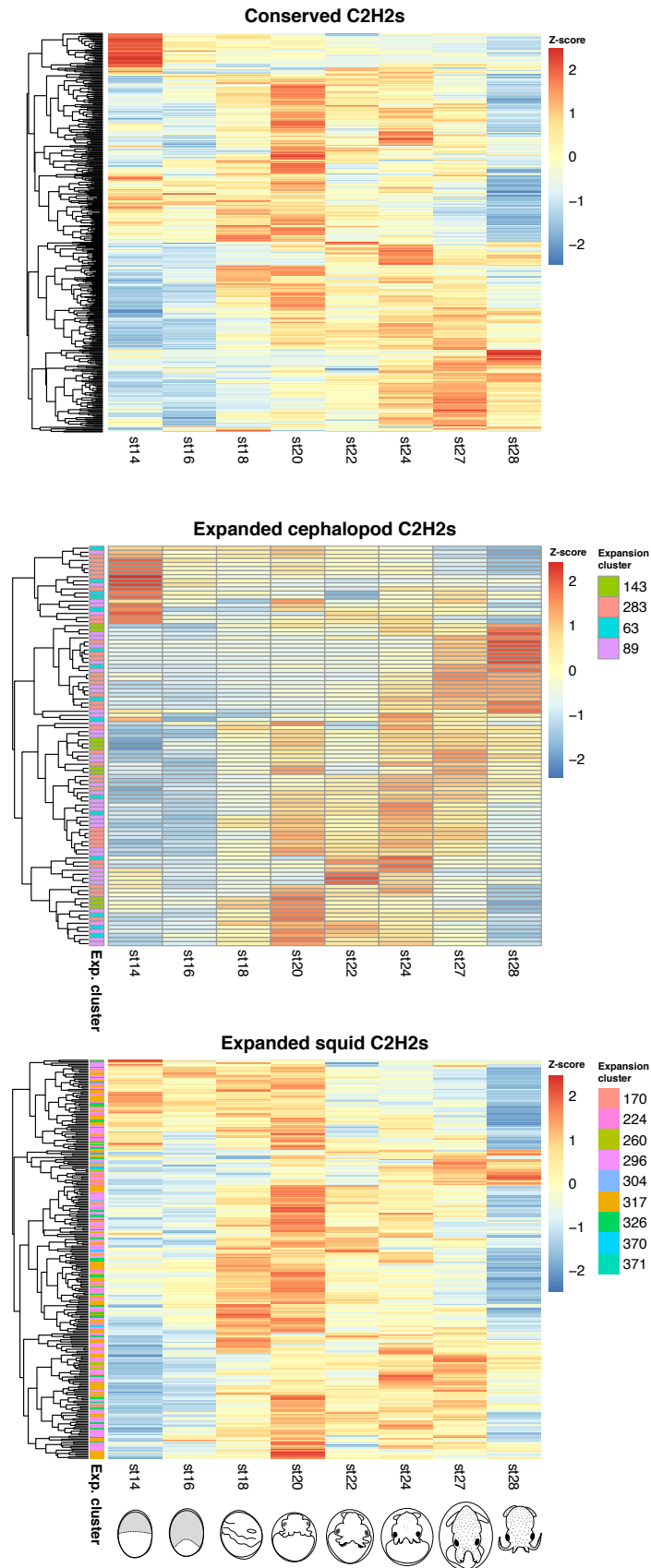


Figure 7. Gene expression of conserved and expanded C2H2s in the early development of *E. scolopes*. The annotation column in the heatmaps show the expansion cluster as characterized in Chapter 2 to which the genes belong to.

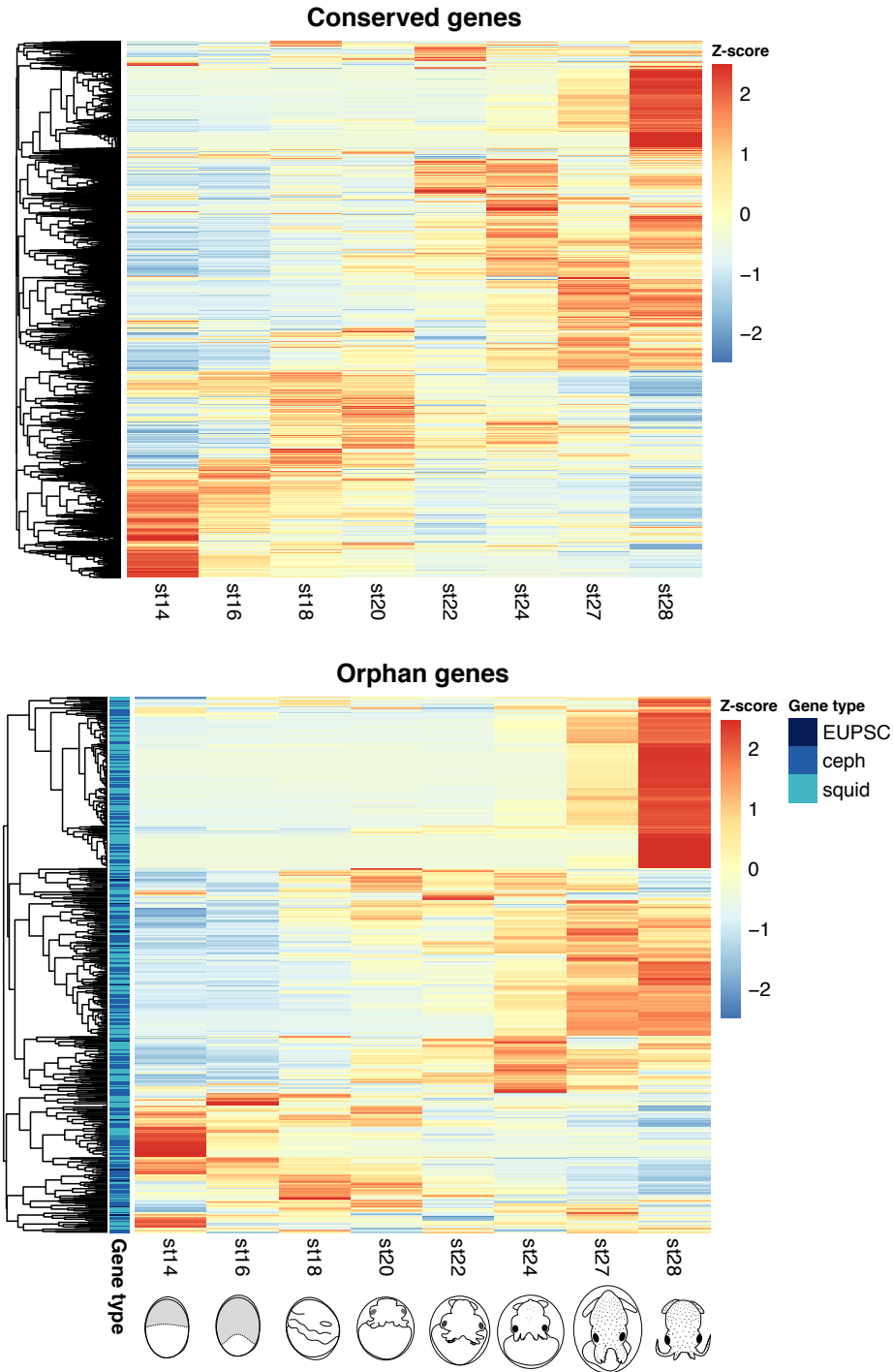


Figure 8. Gene expression of conserved and orphan genes in the early development of *E. scolopes*. The annotation column in the heatmaps show the final annotation of the orphan gene: cephalopod-, squid- or species-specific (*i.e.*, EUPSC: *E. scolopes*).

Altogether, GPCRs and orphan genes showed a higher contribution to the adult-resembling stage, in contrast with C2H2 genes which seem to be more relevant throughout early development of *E. scolopes*. Together with the previously found co-expression of gene novelties in adult tissues, this additional wide overlook at the developmental level offers the

opportunity to further investigate their regulatory interactions, which ultimately would help to understand their combined contribution to the evolution of the cephalopod nervous system and other unique morphological features.

Conclusion

The first examinations of taxonomically restricted genes in cephalopod genome papers was a preamble to the understanding of their contribution to these animals' morphological novelties. However, no in-depth cross-species analysis was previously performed in cephalopods. Here we characterized for the first time TRGs in different cephalopod species, dissected their expression in adult tissues and throughout early development, and investigated how consistent orthologous orphan genes contributed to morphological novelties in these animals. Our findings show that, while the expression of orphan genes is not limited to a few tissues, the only genes maintaining a consistent contribution among octopus' species were expressed in the posterior salivary gland, testes, the brain and tissues related to the peripheral nervous system, like the eyes and the suckers. Additionally, the brain-related co-expression modules in *O. bimaculoides*, *O. vulgaris* and *C. minor* with enriched orthologous orphan genes also showed enrichment of expanded C2H2 zinc fingers. Moreover, the examination of orphan gene expression in the *E. scolopes* developmental time-course showed early to middle-stage expression of expanded C2H2s, possibly related to neuronal differentiation, and a later expression of cephalopod-specific GPCRs and orphan genes in fully developed pre-hatchlings. All in all, these results provide us an insight into putative functional and regulatory linkages between gene novelties, which would be worth investigating further with single cell and/or spatial transcriptomic approaches, as well as with chromatin accessibility (e.g., ATAC-Seq) and conformation (e.g., HiC) assays to explore their interactions within the genomic regulatory landscape.

References

1. Tautz, D. and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nat Rev Genet* 12, 692–702. 10.1038/nrg3053
2. Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. and Bosch, T. C. G. (2009). More than just orphans: are taxonomically-restricted genes important in evolution? *Trends in Genetics*, 25, 9, 404–413. 10.1016/j.tig.2009.07.006.

3. Zhuang, X., Yang, C., Murphy, K.R. and Cheng, C.-H.C. (2019). Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. *Proc. Natl. Acad. Sci.* 116, 4400–4405. 10.1073/pnas.1817138116
4. Kuo, CH. and Kissinger, J.C. (2008). Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*. *BMC Evol Biol* 8, 108. 10.1186/1471-2148-8-108
5. Gould, R., Oakley, T., Goldstone, J., Dugas, J., Brady, S., and Gow, A. (2008). Myelin sheaths are formed with proteins that originated in vertebrate lineages. *Neuron Glia Biology*, 4(2), 137-152. 10.1017/S1740925X09990238
6. Belcaid M., et al. (2019). Symbiotic Organs Shaped by Distinct Modes of Genome Evolution in Cephalopods. *Proc. Natl. Acad. Sci. U. S. A.* 116, 3030–3035. 10.1073/pnas.1817322116
7. Kerwin A. H., McAnulty S.J. and Nyholm S.V. (2021). Development of the Accessory Nidamental Gland and Associated Bacterial Community in the Hawaiian Bobtail Squid, *Euprymna scolopes*. *Biol Bull.* 240(3):205-218. 10.1086/713965.
8. Albertin, C., et al. (2015). The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* 524, 220–224. doi.org/10.1038/nature14668
9. Fouke K. E. and Rhodes H. J. (2020). Electrophysiological and Motor Responses to Chemosensory Stimuli in Isolated Cephalopod Arms. *The Biological Bulletin.* 238: 1-11. 10.1086/707837
10. Ritschard, E. A., Whitelaw, B., Albertin, C. B., Cooke, I. R., Strugnell, J. M., and Simakov, O. (2019). Coupled Genomic Evolutionary Histories as Signatures of Organismal Innovations in Cephalopods: Co-evolutionary Signatures across Levels of Genome Organization May Shed Light on Functional Linkage and Origin of Cephalopod Novelties. *Bioessays* 41 (12), 1900073. 10.1002/bies.201900073
11. Emms, D.M. and Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 20, 238. 10.1186/s13059-019-1832-y
12. Buchfink, B., Xie, C. and Huson, D. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12, 59–60. 10.1038/nmeth.3176
13. Lee, P. N., Callaerts, P. and De Couet, H. G. (2009). The embryonic development of the Hawaiian bobtail squid (*Euprymna scolopes*). *Cold Spring Harb Protoc.* 2009(11):pdb.ip77. 10.1101/pdb.ip77. PMID: 20150049.
14. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 29(1):15-21. 10.1093/bioinformatics/bts635.

15. Liao, Y., Smyth, G. K. and Shi, W. (2014). featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923-30. 10.1093/bioinformatics/btt656
16. López-Córdova, D. A., Avaria-Llautureo, J., Ulloa, P. M., Braid, H. E., Revell, L. J., Fuchs, D., and Ibáñez, C. M. (2021). Mesozoic origin of coleoid cephalopods and their abrupt shifts of diversification patterns. *Molecular Phylogenetics and Evolution*. 10.1016/j.ympev.2021.107331
17. Domazet-Loso, T., Brajković, J. and Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* 23(11):533-9. 10.1016/j.tig.2007.08.014
18. Margheri, L., Ponte, G., Mazzolai, B., Laschi, C., and Fiorito, G. (2011). Non-invasive study of *Octopus vulgaris* arm morphology using ultrasound. *The Journal of Experimental Biology* 214, 3727-3731. 10.1242/jeb.057323
19. Ponte, G., et al. (2020). Cerebrotypes in cephalopods: Brain diversity and its correlation with species habits, life history, and physiological adaptations. *Front Neuroanat* 14: 105. 10.3389/fnana.2020.565109
20. Deryckere, A., et al. (2020). A practical staging atlas to study embryonic development of *Octopus vulgaris* under controlled laboratory conditions. *BMC Dev Biol* 20, 7. 10.1186/s12861-020-00212-6
21. Ritschard, E. A., Fitak, R. R., Simakov, O. and Johnsen, S. (2019). Genomic signatures of G-protein-coupled receptor expansions reveal functional transitions in the evolution of cephalopod signal transduction. *Proc. R. Soc. B Biol. Sci.* 286. 10.1098/rspb.2018.2929.
22. Deryckere, A., Styfhals, R., Elagoz, A. M., Maes, G. E. and Seuntjens, E. (2021) Identification of neural progenitor cells and their progeny reveals long distance migration in the developing octopus brain. *eLife* 10:e69161. 10.7554/eLife.69161
23. Al-Naama, N., Mackeh, R. and Kino T. (2020). C2H2-Type Zinc Finger Proteins in Brain Development, Neurodevelopmental, and Other Neuropsychiatric Disorders: Systematic Literature-Based Analysis. *Front. Neurol.* 11, p. 32. 10.3389/fneur.2020.00032
24. Styfhals, R., et al. (2022). Cell type diversity in a developing octopus brain. *Nat Commun* 13, 7392. 10.1038/s41467-022-35198-1
25. Tautz, D. (2019). The continued mystery of the phylotypic stage. In *Perspectives on Evolutionary and Developmental Biology: Essays for Alessandro Minelli*; Fusco, G., Ed.; Padova University Press: Padova, Italy, pp. 141–149.

General discussion

With the increasing availability of cephalopod genomes and transcriptomes thanks to advancements in sequencing technologies and bioinformatic tools, we can start unravelling the genomic basis behind cephalopod organismal novelties through comparative studies. Such endeavors not only permit us to have a better biological understanding of this fascinating group of mollusks, but also contributes to the construction of knowledge on how DNA information changes over large evolutionary periods and translates into the phenotypic diversity that dazzles us about life on earth.

Along this thesis, I characterized independent genomic novelties of coleoids (*i.e.*, two gene family expansions and orphan genes) and started drawing connections between them to approach putative functional linkages. I used phylogenetic construction and sequence similarity-based orthology inference methods with a thorough sampling of metazoan genomes to 1) define cephalopod-, octopus- and squid-specific orphan genes and expansions of GPCRs and C2H2s. I further explored gene expression of those characterized novel genes in octopus and squid adult tissues and in early developmental stages of *E. scolopes* to 2) gain insights into the putative function of the genes and describe the evolutionary transitions from conserved to novel sequences. Finally, I investigated correspondence between co-expression networks of different squid and octopus adult tissues datasets using orthology and inspected enrichment of the characterized novel genes within them to 3) infer functional linkages between them. With this approach, I could dissect functional relationships between these three groups of gene novelties and show a shared contribution to the cephalopod nervous system, putatively at the regulatory level. These results provide new insights worth studying in the future as of how, or if, expanded and orphan genes co-evolved promoting neural wiring and the development of lineage-specific adaptive features, as discussed in the next sections.

Contribution of expanded adhesive proteins in neural wiring of the octopus axial nerve cord

Variations in size of G-protein coupled receptors (GPCRs) repertoires by duplication and losses have been described in different species and associated with sensory adaptations to the environment [1-4]. GPCR expansions were firstly reported in cephalopod genomes' publications, but the evolutionary dynamics behind this gene family were not studied in detail until the publication of my first chapter. In this paper, we studied the expression transitions

between GPCR genes shared with other bilaterians and expanded ones in cephalopod adult tissues [5]. We revealed functional transitions in the evolution of cephalopod signal transduction, starting with conserved receptors having brain-related functions. Cephalopod-specific duplications followed, which were identified as neuropeptide receptors and diverged in expression across different tissues, particularly gills and testes. Finally, octopus-specific expansions showed most expression in tissues related to the peripheral nervous system, such as the suckers and the axial nerve cord. Most expansions showed some degree of co-localization and signs of positive selection, suggesting a scenario of tandem duplications and functional adaptation in the process of acquiring novel expression domains.

One of the most recent octopus-specific GPCR expansions (*i.e.*, group 3) was annotated as Adhesion GPCRs. In cephalopods, another group of adhesion proteins has been found to be expanded and also highly expressed in the axial nerve cord (ANC): protocadherins [6]. Adhesive molecules can be involved in neural wiring, forming or avoiding links between particular neurons [7,8]. Therefore, these two expansion events could have complemented each other in the evolution of ANC neural wiring. Using the previous annotation of expanded protocadherins from Styfals *et al.*, 2019 [6], the expression of these two groups of genes could be explored in single cell expression data, reconstructing, as in other chapters here, gene regulatory networks or expression modules. This could help elucidate the nature of this hypothesized complementation, which could be either positive (*e.g.*, if both expanded families appear to be co-expressed in same neural cells) or negative (*e.g.*, if the use of one type of adhesion protein restricts the use of the other one and thus contributes to different cell type identities) [9]. Moreover, comparing duplication events and positive selection signatures of ANC-specific adhesion proteins could give insights into the possible scenario of co-evolution of protocadherin and GPCRs gene families. For instance, if there was co-evolution, we can expect to see the same selection signatures and mirror phylogenies of the expanded clades of both gene families. If the proteins have physical interactions, the latter can be a result of a compensatory changes (*i.e.*, a beneficial mutation in one protein compensating changes in the other protein when forming molecular complexes) [10]. Another scenario would be the mutual involvement in the same cellular process that is under selective pressures. Aside from single cell expression data, spatial transcriptomics (*i.e.*, a molecular profiling method to measure and map in a tissue sample all gene activity, see <https://www.10xgenomics.com/spatial-transcriptomics>) could help elucidate the aforementioned possible co-evolutionary scenarios. Further functional dissection of the genes would be also needed to understand their involvement in neural identity. Altogether, investigating such an event could help us understand how expanded genes have diversified neural wiring and contributed to new functions in the ANC.

Unravelling cephalopod neuronal network connectivity through contrasting C2H2 and GPCR expansion histories

In contrast with the evolutionary dynamics of GPCRs, the expanded C2H2 transcription factors showed no evident transitions in expression domains (Chapter 2). Conserved genes, as well as cephalopod-, octopus- and squid- specific expanded genes, maintained an expression in the central nervous system. Despite the lack of expression divergence, paralogues within a same expansion clade were assigned to differing co-expression modules. With these results we hypothesized that paralogues diverged at the cell-type and/or the regulatory level within the same tissue. Because the analyses performed are based on whole tissues which contain a mixture of different cells, here again, expression data at the single cell level would help to dissect the contribution of expanded *versus* conserved genes to neural cell identity within the central nervous system. It could ultimately allow us to confirm or discard the absence of functional transition in the evolution of C2H2 expansions found in my second chapter and test whether these evolved via neo- or sub-functionalization. Furthermore, the regulatory landscape around the genes could be further addressed by exploring their co-localization and the coding and noncoding sequences in the vicinity, as well as inspecting chromatin availability and physical interactions of the genes with ATAC-Seq [11] and HiC [12] data. Such further investigations could help us to test the regulatory divergence hypothesis here proposed and, overall, gain a better understanding of the evolutionary dynamics of this gene family expansion.

Moreover, co-evolutionary patterns between the C2H2 and GPCR expansions may help us to understand how network connectivity in cephalopods nervous systems evolved and underlie the development of their rich behavioral repertoire [9]. Brain pathway controlling complex behaviors, like vocal learning, has been proposed to have evolved by whole brain circuit duplication [13]. At the molecular level, a hypothesis posits that gene duplication of transcription factors and axon guidance genes underlies the development of such duplicated brain pathways. Given this, studying co-evolution signatures of the expanded C2H2 zinc fingers transcription factors highly contributing to the central brain, and expanded genes putatively involved in axon guidance and neural wiring exclusively in the arms and distal structures (*i.e.*, the expanded GPCRs in ANC, but also suckers and skin) may help address such hypothesis, not only at the brain level but the circuitries connecting neuronal periphery and centers [9]. As proposed above, mirror phylogenies and shared positive selection signatures might hint us to such scenario of gene duplication shaping circuit duplication, but

further studies of expression location at the cellular level in the developing nervous system would be needed.

Orphan genes and C2H2 zinc finger expansions in the evolution of octopus brain regulatory networks

Contrary to gene duplications, orphan genes were not considered as a major evolutionary driver until recently [14-17]. These can arise either through rapid evolution after gene duplication or *de novo* from noncoding sequences and their function has been hypothesized to be relevant for taxon-specific adaptations [18,19]. In cephalopod genomes, distinct signatures of orphan gene contribution in certain organs was already found (*e.g.*, the symbiotic accessory nidamental gland in *E. scolopes* [20], the suckers in *O. bimaculoides* [21]). However, a comparative study using a thorough bioinformatic approach to characterize orphan genes and focusing on their evolutionary contribution across different squid and octopus' species was missing until now. In my third chapter, I showed that taxonomically restricted genes (TRGs) specific to cephalopods, octopus and squid have contributed to a wide range of tissues in all species investigated. Particular tissue contributions were found to be consistent only between octopus' species, where orthologous orphan genes showed same expression domains (*i.e.*, testes, posterior salivary gland, brain, suckers and eyes). Exploring each of these gene groups and co-expression modules more in detail would help elucidate to what extent these are related to octopus-specific adaptations, and how orphan genes that evolved at the coleoid cephalopod node have diverged over time. Nevertheless, further studies would benefit from more consistent tissue sampling between species to do expression comparisons, as the datasets here analyzed were considerably different, possibly biasing the results and/or leading to underestimation of shared tissue contributions.

Further examining the correspondence of co-expression between species, I found that the brain-related modules did not only show a high contribution of orphan genes among octopuses, but also of expanded C2H2 zinc fingers. This was, again, only found in the octopus' and not squid species investigated, highlighting differences of the central nervous system between cephalopod lineages. While the function of those genes remains to be investigated, the correlated expression of these two types of novelty may be explained by shared regulatory networks and hints again to a scenario of co-evolution between gene novelties. The orphan and C2H2 zinc finger genes could contribute together to the evolution of new or, rather, act as an addition to the already existing core gene regulatory networks of the central nervous system [9]. Moreover, gene expression of all characterized gene novelties in this thesis in a

developmental time-course of *E. scolopes* showed early to middle-stage contribution of expanded C2H2s, possibly relating these genes to neuronal differentiation, and a later expression of cephalopod-specific GPCRs and orphan genes in fully developed pre-hatchlings. These results already allow to further group particular orphan genes and expanded genes in co-regulatory units, defined by developmental activation. Finally, to address the co-evolutionary hypotheses and putative regulatory interactions between TRGs and gene duplicates here drawn, aside from expression data at the single cell level and chromatin architecture assays, it becomes vital to explore the actual function of these characterized genes. Only recently, the first successful gene knockout with CRISPR-Cas9 in the squid *Doryteuthis pealeii* has been published [22], opening a new era for functional approaches in cephalopods that can help us to reach a better understanding on the genomic background of organismal novelties.

Concluding remarks

The multi-scale study of genome organization levels offers a powerful approach to understand how genomic features have synergically contributed to functionally linked structures and the overall evolution of organismal novelty. This thesis provides a first glimpse at the relevance of novel genes, characterized here as orphan and duplicate genes, in the evolution of cephalopod innovations. Particular tissues enriched in expression of these novel genes were part of the nervous system, both central and peripheral. Cephalopods have been already for years of particular interest in the evolutionary biology field due to their complex nervous systems and outstanding behavioral repertoires, both features highly divergent to their molluscan relatives. However, molecular mechanisms behind the evolution of this hallmark are just starting to be understood. By further dissecting the evolutionary relationships between orphan genes and the C2H2 zinc fingers, GPCRs and protocadherin gene family expansions, we may near us to understand the concomitant contribution of genomic novelty to neural circuit evolution in the cephalopod nervous system and lineage-specific adaptations that have made these animals in their diversity take advantage of every niche in the immense sea.

References

1. Hill, C.A. *et al.* (2002). G protein-coupled receptors in *Anopheles gambiae*. Science 298, 176-178. 10.1126/science.1076196

2. Robertson, H.M. (1998). Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement, and intron loss. *Genome Res.* 8, 449-463.
10.1101/gr.8.5.449
3. Thomas, J.H. and Robertson, H.M. (2008). The *Caenorhabditis* chemoreceptor gene families. *BMC Biol.* 6, 1-17. 10.1186/1741-7007-6-42
4. Niimura, Y. and Nei, M. (2007). Extensive gains and losses of olfactory receptor genes in mammalian evolution. *PLoS ONE* 2, e708. 10.1371/journal.pone.0000708
5. Ritschard, E. A., Fitak, R. R., Simakov, O. and Johnsen, S. (2019). Genomic signatures of G-protein-coupled receptor expansions reveal functional transitions in the evolution of cephalopod signal transduction. *Proc. R. Soc. B Biol. Sci.* 286. 10.1098/rspb.2018.2929.
6. Styfhals, R., Seuntjens, E., Simakov, O., Sanges, R. and Fiorito, G. (2019). *In silico* Identification and Expression of Protocadherin Gene Family in *Octopus vulgaris*. *Front. Physiol.* 9:1905. 10.3389/fphys.2018.01905
7. Serizawa, S., Miyamichi, K., Takeuchi, H., Yamagishi, Y., Suzuki, M. and Sakano, H. (2006). A neuronal identity code for the odorant receptor-specific and activity-dependent axon sorting. *Cell.* 127(5):1057-69. 10.1016/j.cell.2006.10.031.
8. Hasegawa, S., Hamada, S., Kumode, Y., Esumi, S., Katori, S., Fukuda, E., Uchiyama, Y., Hirabayashi, T., Mombaerts, P. and Yagi, T. (2008). The protocadherin-alpha family is involved in axonal coalescence of olfactory sensory neurons into glomeruli of the olfactory bulb in mouse. *Mol Cell Neurosci.* 38(1):66-79. 10.1016/j.mcn.2008.01.016.
9. Ritschard, E. A., Whitelaw, B., Albertin, C. B., I. Cooke, R., Strugnelli, J. M. and Simakov, O. (2019). Coupled Genomic Evolutionary Histories as Signatures of Organismal Innovations in Cephalopods: Co-evolutionary Signatures Across Levels of Genome Organization May Shed Light on Functional Linkage and Origin of Cephalopod Novelties. *BioEssays.* 41. 10.1002/bies.201900073.
10. Juan, D., Pazos, F. and Valencia A. (2008). Co-evolution and co-adaptation in protein networks. *FEBS Lett.* 582(8):1225-30. 10.1016/j.febslet.2008.02.017.
11. Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Current protocols in molecular biology*, 109(1), 21-29. 10.1002/0471142727.mb2129s109
12. Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*, 58(3), 268-276. 10.1016/j.ymeth.2012.05.001

13. Chakraborty, M., and Jarvis, E. D. (2015). Brain evolution by brain pathway duplication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1684), 20150056. 10.1098/rstb.2015.0056
14. Domazet-Lošo, T., and Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, 468(7325), 815-818. 10.1038/nature09632
15. Khalturin, K., Anton-Erxleben, F., Sassmann, S., Wittlieb, J., Hemmrich, G., and Bosch, T. C. G. (2008). A novel gene family controls species-specific morphological traits in *Hydra*. *PLoS biology*, 6(11), e278. 10.1371/journal.pbio.0060278
16. Knowles, D. G., and McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. *Genome research*, 19(10), 1752-1759. 10.1101/gr.095026.109
17. Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *Genome research*, 20(10), 1313-1326. 10.1101/gr.101386.109
18. Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., and Bosch, T. C. (2009). More than just orphans: are taxonomically-restricted genes important in evolution?. *Trends in Genetics*, 25(9), 404-413. 10.1016/j.tig.2009.07.006
19. McLysaght, A., & Hurst, L. D. (2016). Open questions in the study of de novo genes: what, how and why. *Nature Reviews Genetics*, 17(9), 567-578. 10.1038/nrg.2016.78
20. Belcaid, M., *et al.* (2019). Symbiotic organs shaped by distinct modes of genome evolution in cephalopods. *Proc. Natl. Acad. Sci. U. S. A.* 116, 3030–3035. 10.1073/pnas.1817322116.
21. Albertin, C. B., Simakov, O., Mitros, T., Wang, Z. Y., Pungor, J. R., Edsinger-Gonzales, E., Brenner, S., Ragsdale, C.W. and Rokhsar, D. S. (2015). The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*, 524(7564), 220-224. 10.1038/nature14668
22. Crawford, K., Quiroz, J. F. D., Koenig, K. M., Ahuja, N., Albertin, C. B., and Rosenthal, J. J. (2020). Highly efficient knockout of a squid pigmentation gene. *Current Biology*, 30(17), 3484-3490. 10.1016/j.cub.2020.06.099

Appendix

Supplementary material in Chapter 1

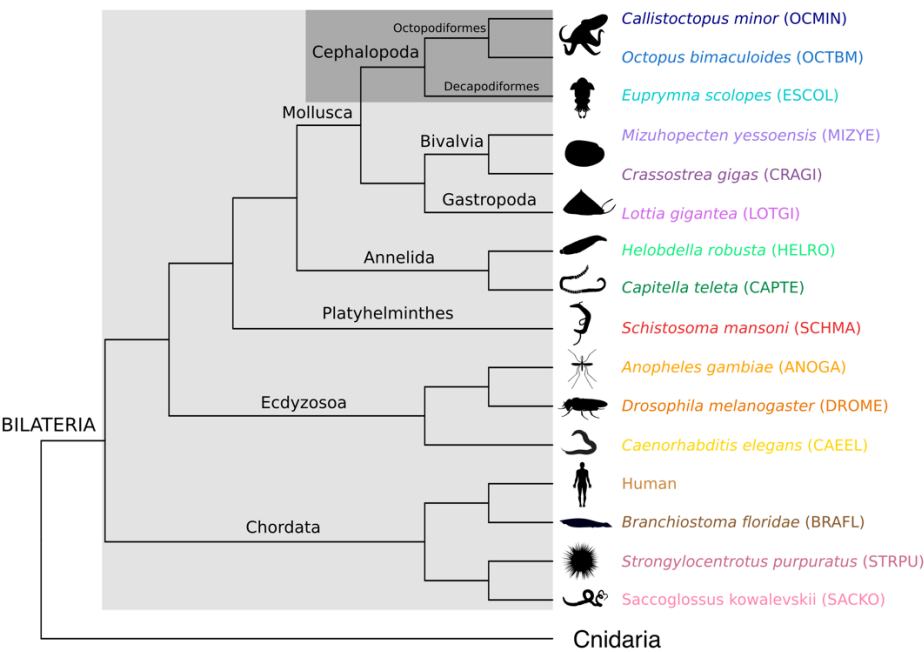


Figure S1. Phylogeny of the taxa sampled. GPCR repertoires from 16 bilaterian species here shown were used in our phylogenetic analyses. See Methods for more details. Octopus and squid illustrations were designed by Hannah Schmidbaur, all others were downloaded from Phylopic (<http://phylopic.org>).

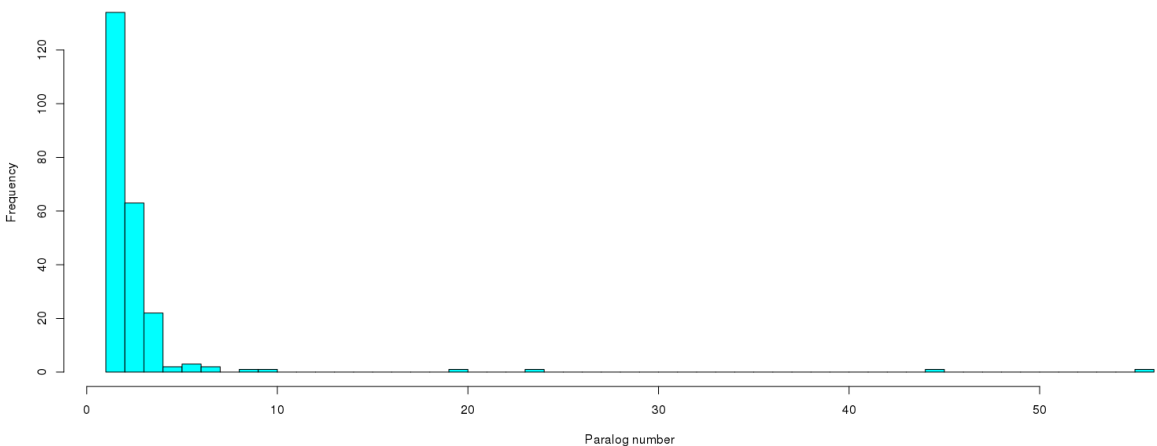


Figure S2. Cut-off for the identification of expanded groups in our phylogenetic analysis. The four bars with over 10 paralog sequences (*i.e.*, sequences belonging to one, two or the three cephalopod species used in the phylogenetic analysis) correspond to groups 1-4.

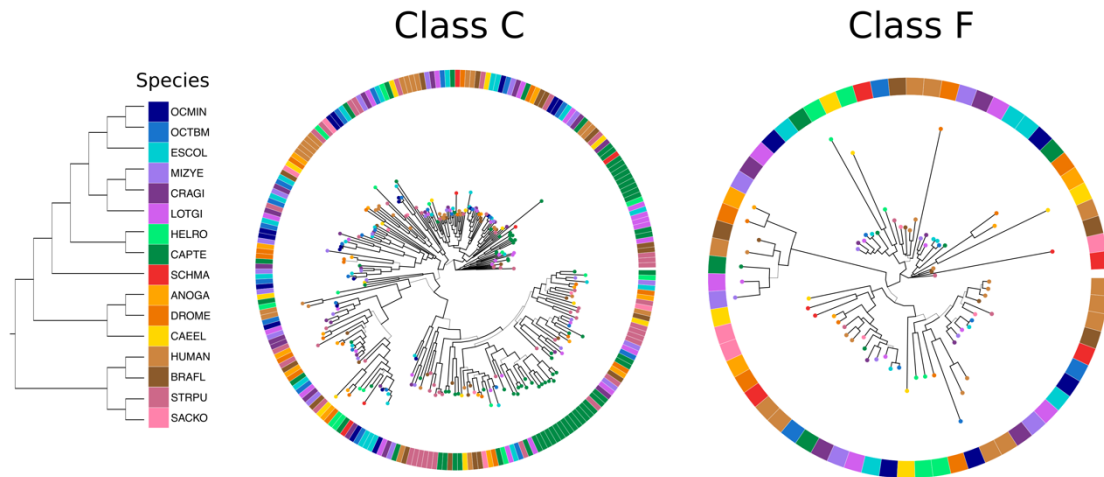
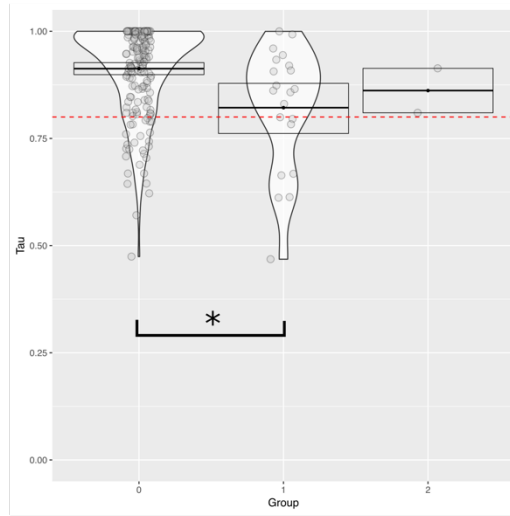
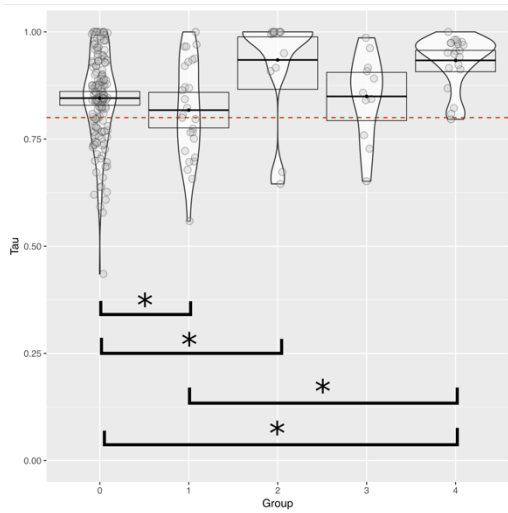


Figure S3. Phylogenetic trees for class C and F. Thicker branches represent significant S-H supports (>0.7) and each color corresponds to a species. ANOGA: *Anopheles gambiae*, BRAFL: *Branchiostoma floridae*, CAEEL: *Caenorhabditis elegans*, CAPTE: *Capitella teleta*, CRAGI: *Crassostrea gigas*, DROME: *Drosophila melanogaster*, ESCOL: *Euprymna scolopes*, HELRO: *Helobdella robusta*, HUMAN, LOTGI: *Lottia gigantea*, MIZYE: *Mizuhopecten yessoensis*, OCMIN: *Callistoctopus minor*, OCTBM: *Octopus bimaculoides*, SACKO: *Saccoglossus kowalevskii*, SCHMA: *Schistosoma mansoni*, STRPU: *Strongylocentrotus purpuratus*. No cephalopod expanded groups were found in these two GPCR classes.

E. scolopes



O. bimaculoides



C. minor

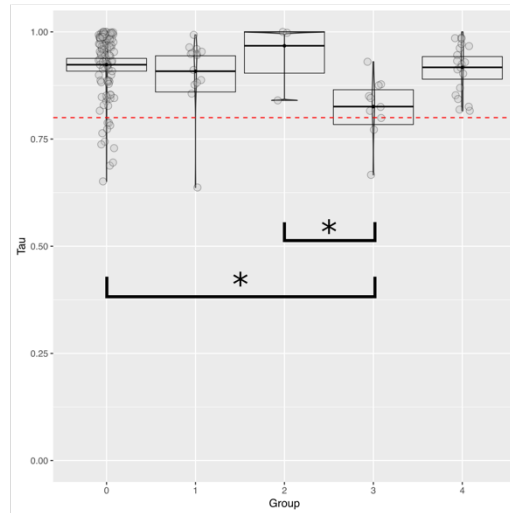


Figure S4. Tau values in non-expanded GPCR_S (0) and expanded groups 1-4 for *E. scolopes*, *O. bimaculoides* and *C. minor*. Boxes represent the 95% confidence interval of the data distribution, black line in the middle the mean and the violin shape the density of data points (*i.e.*, more data points around a Tau value will generate a wider violin shape). The red dashed line represents the tissue-specificity threshold (Tau value 0.8). Bars with asterisks represent significant differences between the groups (Dunn's test: p-value <0.025).

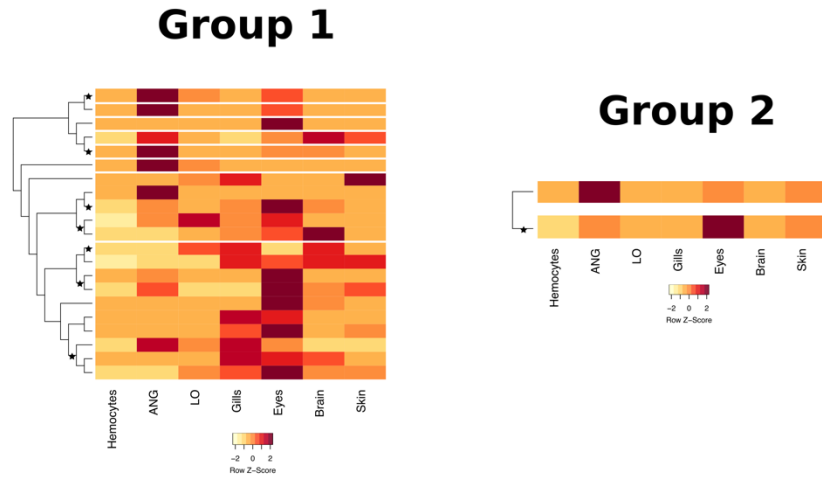


Figure S5. Expression profiles and positive selection results cephalopod-specific expansions (groups 1 and 2) for *E. scolopes*. Genes (rows) in heatmaps are clustered following the trees resulting from the phylogenetic analyses. Spaces between rows indicate the presence of other cephalopod species' sequences as resulted in the phylogenetic analyses. Black stars represent positive selection ($\omega>1$).

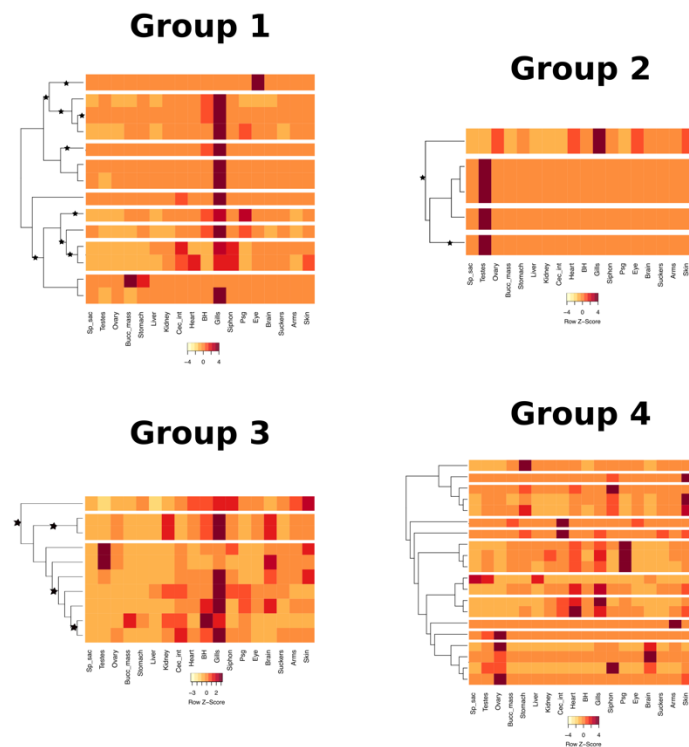
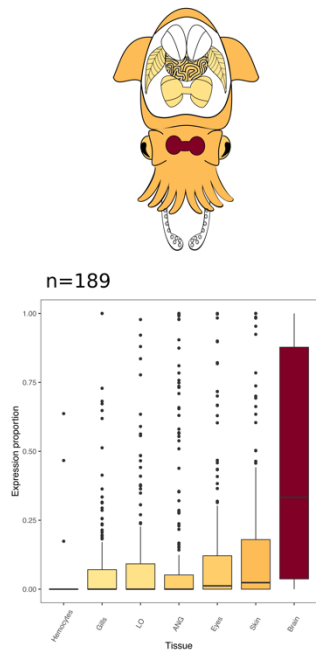


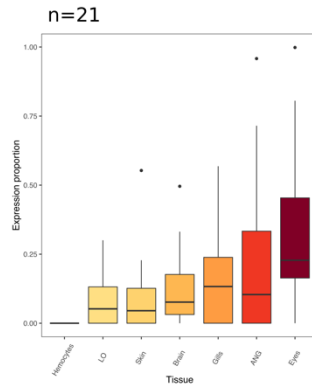
Figure S6. Expression profiles and positive selection results cephalopod-specific expansions (groups 1 and 2) and octopus-specific expansions (groups 3 and 4) for *C. minor*. Genes (rows) in heatmaps are clustered following the trees resulting from the phylogenetic analyses. Spaces between rows indicate the presence of other cephalopod species' sequences as resulted in the phylogenetic analyses. Black stars represent positive selection ($\omega>1$).

Non-expanded GPCRs



Cephalopod-specific expansions

Group 1



Group 2

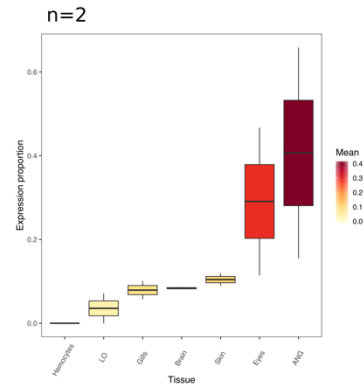
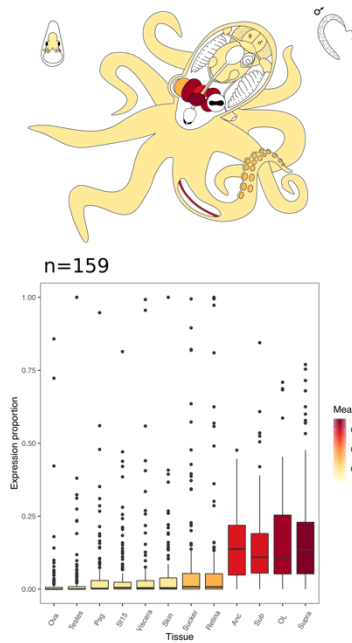


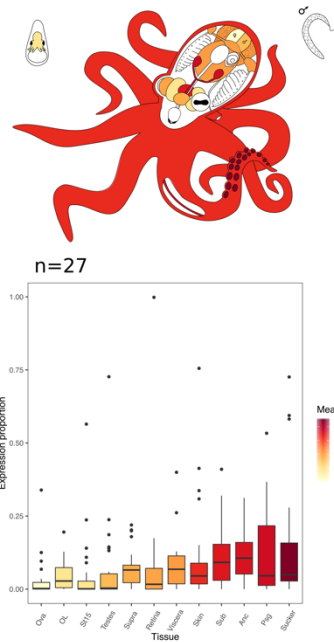
Figure S7. Expression proportions of *E. scolopes* genes in each tissue in non-expanded GPCRs and cephalopod-specific expansions. Boxplots represent data between the first and third quartiles of the distribution and the median. Coloration gradient of boxplots follow the mean values of gene expression proportion in each tissue and was used to colour the corresponding tissues in the octopus' diagrams above. Tissues: hemocytes, light organ (LO), accessory nidamental gland (ANG), skin, gills, brain and eyes. Number of data points in each boxplot: n.

Non-expanded GPCRs

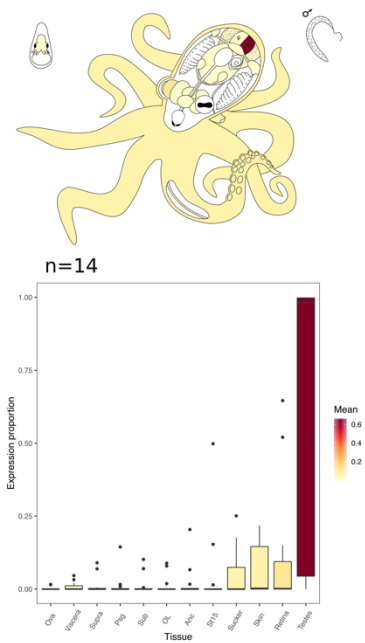


Cephalopod-specific expansions

Group 1

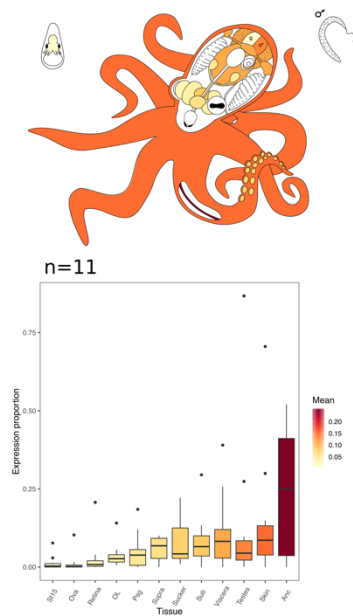


Group 2



Octopus-specific expansions

Group 3



Group 4

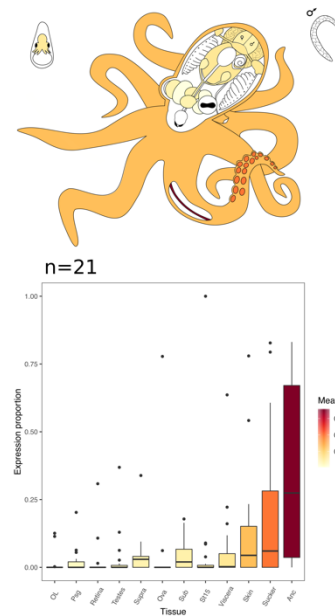
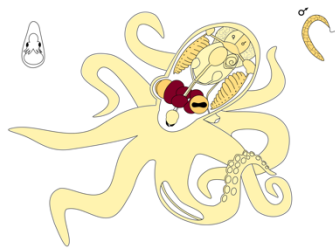
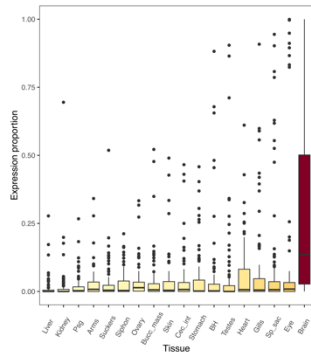


Figure S8. Expression proportions of *O. bimaculoides* genes in each tissue in non-expanded GPCRs, cephalopod-specific expansions and octopus-specific expansions. Boxplots represent data between the first and third quartiles of the distribution and the median. Coloration gradient of boxplots follow the mean values of gene expression proportion in each tissue and was used to colour the corresponding tissues in the octopus' diagrams above. Tissues: sucker, testes, stage 15 (St15) embryo, ova, skin, posterior salivary gland (Psg), viscera (heart, kidney and hepatopancreas), subesophageal brain (Sub), supraesophageal brain (Supra), optic lobe (OL), axial nerve cord (Anc) and retina. Number of data points in each boxplot: n.

Non-expanded GPCRs

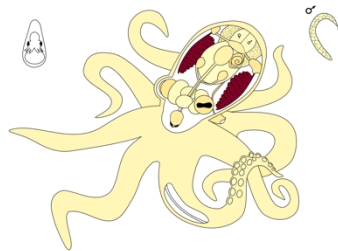


n=106

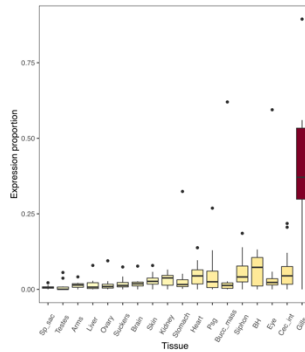


Cephalopod-specific expansions

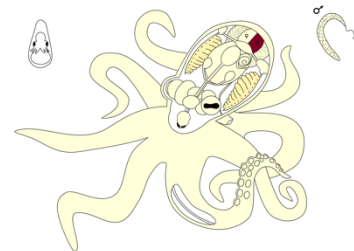
Group 1



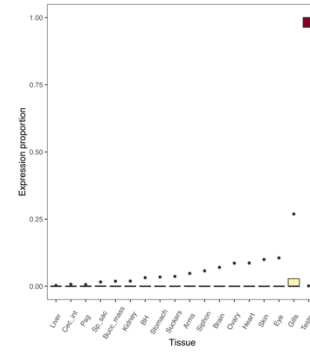
n=14



Group 2



n=5

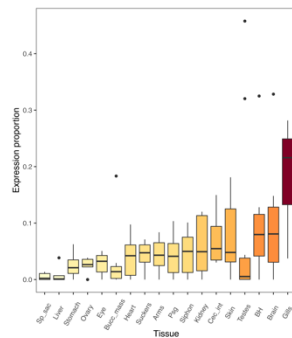


Octopus-specific expansions

Group 3



n=10



Group 4



n=20

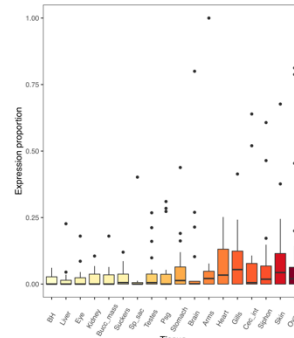


Figure S9. Relative expression of *C. minor* genes in each tissue in non-expanded GPCRs, cephalopod-specific expansions and octopus-specific expansions. Boxes represent data between the first and third quartiles of the distribution and the black middle line the median. Coloration gradient of boxplots follow the mean values of gene expression proportion in each tissue and was used to colour the corresponding tissues in the octopus' diagrams above. Tissues: liver, kidney, stomach, caecum intestine (Cec_int), posterior salivary gland

(Psg), buccal mass (Bucc_mass), branchial heart (BH), systemic heart (heart), suckers, arms, skin, gills, siphon, brain, eye, spermatophore sac (Sp_sac), testes and ovary. Number of data points in each boxplot: n.

Table S1. Summary of sequences used in the phylogenetic analyses. The number of sequences is taken from final datasets after filtering duplicated gene entries (*i.e.*, isoforms).

Species	Proteome entry	Number of sequences				
		Class A	Class B	Class C	Class F	Total
Human	UP000005640	291	50	22	11	374
<i>Branchiostoma floridae</i>	UP000001554	561	81	17	5	664
<i>Strongylocentrotus purpuratus</i>	UP000007110	1019	213	32	5	1269
<i>Saccoglossus kowalevskii</i>	-	19	3	4	4	30
<i>Caenorhabditis elegans</i>	UP000001940	129	5	8	4	146
<i>Anopheles gambiae</i>	UP000007062	82	13	9	6	110
<i>Drosophila melanogaster</i>	UP000000803	73	17	11	5	106
<i>Schistosoma mansoni</i>	UP000008854	76	5	3	7	91
<i>Helobdella robusta</i>	UP000015101	195	29	9	5	238
<i>Capitella teleta</i>	UP000014760	935	39	51	5	1030
<i>Lottia gigantea</i>	UP000030746	276	59	17	5	357
<i>Crassostrea gigas</i>	UP000005408	362	63	14	4	443
<i>Mizuhopecten yessoensis</i>	-	357	46	15	5	423
<i>Euprymna scolopes</i>	-	232	39	19	7	297
<i>Callistoctopus minor</i>	-	190	66	19	2	277
<i>Octopus bimaculoides</i>	UP000053454	233	81	19	6	339
Total number of sequences		5030	809	269	86	6194

Table S4. Summary of the positive selection analysis results. Abbreviations: M0- null model, Ma-alternative model, np- number of parameters, df- degrees of freedom, LRT- Likelihood Ratio Test. Calculations can be found in “PositiveSelection_calculations.xlsx” in Dryad.

		Group 1 and 2	Group 3	Group 4
Branch-model test	M0 likelihood	-31390.1701	-3636.4977	-4193.2279
	M0 np	281	59	119
	Ma likelihood	-30860.40443	-3552.5782	-4150.234
	Ma np	435	96	202
	LRT	1059.531342	167.839006	85.987776
	df	154	37	83
	p-value	0	0	3.89E-01
Sites-model test	M0 likelihood	-31390.17	-3636.4977	-4193.2279
	M0 np	281	59	119
	Ma likelihood	-31356.554	-3566.5901	-4153.9666
	Ma np	284	62	122
	LRT	67.23259	139.815336	78.522504
	df	3	3	3
	p-value	1.67E-14	0	1.11E-16

Tables S2 and S3 can be found as separate files of the Supplementary Material in the published paper. Table captions as follow:

Table S2. Summary of the expression analyses and gene locations. Excel file contains spreadsheets with Tau and proportion of expression for each gene of 1) *Euprymna scolopes*, 2) *Octopus bimaculoides* and 3) *Callistoctopus minor*. Gene locations for *O. bimaculoides* are also shown here.

Table S3. Summary of the annotation results. Excel file contains spreadsheets with 1) InterProScan matches of the sequences of groups 1-4 for *Euprymna scolopes*, *Octopus bimaculoides* and *Callistoctopus minor* and 2) UniProt database descriptions of the nearest sequences to the expanded groups (1-4).

Supplementary material in Chapter 1

Table S1. Species used in the phylogenetic analysis and the corresponding source of their sequences.

**Idiosepius paradoxus* dataset was taken from the NCBI database under the BioSample accession number SAMN00152410.

ABBREVIATION	SPECIES NAME	COMMON NAME	SOURCE	PROTEOME
ANOGA	<i>Anopheles gambiae</i>	African malaria mosquito	UniProt	Yes
ARCDU	<i>Architeuthis dux</i>	Giant squid	[S1]	Yes
BRAFL	<i>Branchiostoma floridae</i>	Florida lancelet	UniProt	Yes
CAEEL	<i>Caenorhabditis elegans</i>	Round worm	UniProt	Yes
CALMI	<i>Callistoctopus minor</i>	Long arm octopus	[S2]	Yes
CAPTE	<i>Capitella teleta</i>	Polychaete worm	UniProt	Yes
CRAGI	<i>Crassostrea gigas</i>	Pacific oyster	UniProt	No
DORPE	<i>Doryteuthis pealeii</i>	Longfin inshore squid	[S3]	Yes
DROME	<i>Drosophila melanogaster</i>	Fruit fly	UniProt	Yes
EUPSC	<i>Euprymna scolopes</i>	Hawaiian bobtail squid	[S4]	Yes
HELRO	<i>Helobdella robusta</i>	Californian leech	UniProt	Yes
HUMAN	<i>Homo sapiens</i>	Human	UniProt	Yes
IDIPA	<i>Idiosepius paradoxus</i>	Northern pygmy squid	NCBI*	Yes
LOTGI	<i>Lottia gigantea</i>	Giant owl limpet	UniProt	Yes
MIZYE	<i>Mizuhopecten yessoensis</i>	Japanese scallop	UniProt	Yes
OCTBI	<i>Octopus bimaculoides</i>	California two-spot octopus	[S5]	Yes
OCTVU	<i>Octopus vulgaris</i>	Common octopus	[S6]	Yes
SACKO	<i>Saccoglossus kowalevskii</i>	Acorn worm	UniProt	No
SCHMA	<i>Schistosoma mansoni</i>	Blood fluke	UniProt	Yes
STRPU	<i>Strongylocentrotus purpuratus</i>	Purple sea urchin	UniProt	Yes

Table S2. Tissues from the different cephalopod species available as expression data and used in this study. “X” means data available. Similar tissues were grouped together and sorted starting with the most relevant ones for this study (i.e., brain, eye, ganglia, skin and arm tissues). Species abbreviations are as follows: CALMI: *C. minor*, EUPSC: *E. scolopes*, OCTBI: *O. bimaculoides*, OCTVU: *O. vulgaris*. Tissue abbreviations are as followed: ANC: axial nerve cord, ANG: accessory nidamental gland, ant_arm: anterior arm, post_arm: posterior arm, BH: branchial heart, OL: optic lobe, sub: subesophageal brain, supra: supraesophageal brain, Bucc_mass: buccal mass, Cec_int: caecum intestine, Gastr_gangl: gastric ganglion, heart: sytemic heart, LO: light organ, Psg: posterior salivary gland, Sp_sac: spermatophore sac, Stellate_gangl: stellate ganglion and viscera: heart, kidney, and hepatopancreas.

TISSUES:	CALMI	EUPSC	OCTBI	OCTVU
BRAIN	X	X		
OL			X	X
SUB			X	X
SUPRA			X	X
EYES	X	X		
RETINA			X	
GASTR_GANGL				X
STELLATE_GANGL				X
SKIN	X	X	X	
ARM	X			
ANT_ARM				X
POST_ARM				X
ARM_TIP				X
ARM_MUSCLE				X
ANC			X	
SUCKER	X		X	
ANG		X		
BH	X			
BUCC MASS	X			
CEC INT	X			
GILLS	X	X		
HEART	X			
HAEMOCYTES		X		
KIDNEY	X			
LIVER	X			
LO		X		
OVARY	X		X	
PSG	X		X	
SIPHON	X			
SP_SAC	X			
STOMACH	X			
TESTIS	X		X	
VISCERA			X	

Table S3. Number of sequences in the resulting expansion clusters. Cluster ID recalls whether the expansion was characterized as cephalopod-, octopus- or squid-specific. Column named “S-H support” stands for the support value of the cluster in the phylogenetic tree using the Shimodaira-Hasegawa test as described in the Methods. Species abbreviations are as follows: ARCDU: *Architeuthis dux*, DORPE: *Doryteuthis pealeii*, EUPSC: *Euprymna scolopes*, IDIPA: *Idiosepius paradoxus*, CALMI: *Callistoctopus minor*, OCTBI: *Octopus bimaculoides*, OCTVU: *Octopus vulgaris*.

CLUSTER ID	S-H SUPPORT	TOTAL NR. OF SEQUENCES	SQUID	OCTOPUS	ARCDU	DORPE	EUPSC	IDIPA	CALMI	OCTBI	OCTVU
CEPH 1	0,236	163	106	57	31	37	38	0	11	35	11
CEPH 2	0,835	132	89	43	26	28	35	0	6	27	10
CEPH 3	0,809	76	46	30	15	16	15	0	11	15	4
CEPH 4	0,989	39	24	15	7	7	10	0	5	7	3
OCTO 1	0,465	495	1	494	1	0	0	0	0	494	0
OCTO 2	0,787	235	1	234	0	1	0	0	9	223	2
OCTO 3	0,851	167	3	164	0	1	2	0	1	157	6
OCTO 4	0,814	104	0	104	0	0	0	0	6	97	1
SQUID 1	0,767	984	984	0	4	882	96	2	0	0	0
SQUID 2	0,778	499	499	0	0	415	82	2	0	0	0
SQUID 3	0,887	161	161	0	0	144	17	0	0	0	0
SQUID 4	0,771	107	107	0	0	106	1	0	0	0	0
SQUID 5	0,845	47	47	0	0	47	0	0	0	0	0
SQUID 6	0,888	46	44	2	0	36	8	0	0	1	1
SQUID 7	0,864	42	42	0	0	1	41	0	0	0	0
SQUID 8	0	36	36	0	0	33	3	0	0	0	0
SQUID 9	0,832	28	28	0	0	5	23	0	0	0	0
SQUID 10	0,962	25	25	0	0	22	3	0	0	0	0
SQUID 11	0,794	20	20	0	0	20	0	0	0	0	0

Table S4. Results of the Kruskal-Wallis and Dunn’s tests with Bonferroni correction. Statistical significance (p-values) associated with comparisons of expression values between expansion groups (non-expanded, cephalopod-expanded and octopus/squid-expanded C2H2s) in each tissue for *E. scolopes* (EUPSC), *C. minor* (CALMI), *O. bimaculoides* (OCTBI) and *O. vulgaris* (OCTVU). Cells with significant p-values (< 0.05) are coloured green. *This table is found in the Dryad link provided in the manuscript.

Table S5. Summary of GO term analysis results. Expression modules used in the GO term enrichment analysis (1st column), their correlated tissues (2nd column), the total number of genes per module (3rd column), the number of lineage-specific (squid- or octopus-specific), cephalopod-specific and non-expanded C2H2 genes (4th, 5th and 6th column). The last three columns show whether the C2H2 GO:0003676 and the GPCR GO:0007186 were enriched (“y” for yes, “n” for no), and how often the GPCR GO term was found in the respective module. Tissue abbreviations are as follows: ANC: axial nerve cord, OL: optic lobe, supra: supraesophageal brain, Gastr_gangl: gastric ganglion, LO: light organ, Sp_sac: spermatophore sac. Species abbreviations associated with module names are as follows: EUPSC for *E. scolopes*, CALMI for *C. minor*, OCTBI for *O. bimaculoides* and OCTVU for *O. vulgaris*.

Modules	Tissue	Size	Squid /Octo	Ceph	Non-exp	C2H2 enriched	GPCR enriched	# GPCR GO term
EUPSCpurple	Skin, Brain	596	28	2	24	y	y	17
EUPSCpink	Eyes, Brain	705	33	0	40	y	n	0
EUPSCtan	LO, Brain	325	10	2	10	n	y	12
CALMIsteelblue	Brain, Sp_sac	138	0	1	1	n	y	7
CALMIbrown	Brain, Ovary	2205	4	9	28	y	n	0
CALMITurquoise	Brain, Ovary	2986	1	13	39	y	n	0
OCTBIsalmon	OL, Supra, Ovary	507	151	1	25	y	n	0
OCTBITurquoise	OL, Supra, Anc	11164	646	64	224	y	n	0
OCTBIgrey60	Sucker, Testis	59	3	0	0	n	n	0
OCTVUmagenta	Gastr_gangl	1288	1	0	1	n	y	6
OCTVUgreen	OL	5013	0	3	15	n	y	19
OCTVUturquoise	OL, Supra	57662	7	23	121	y	y	120

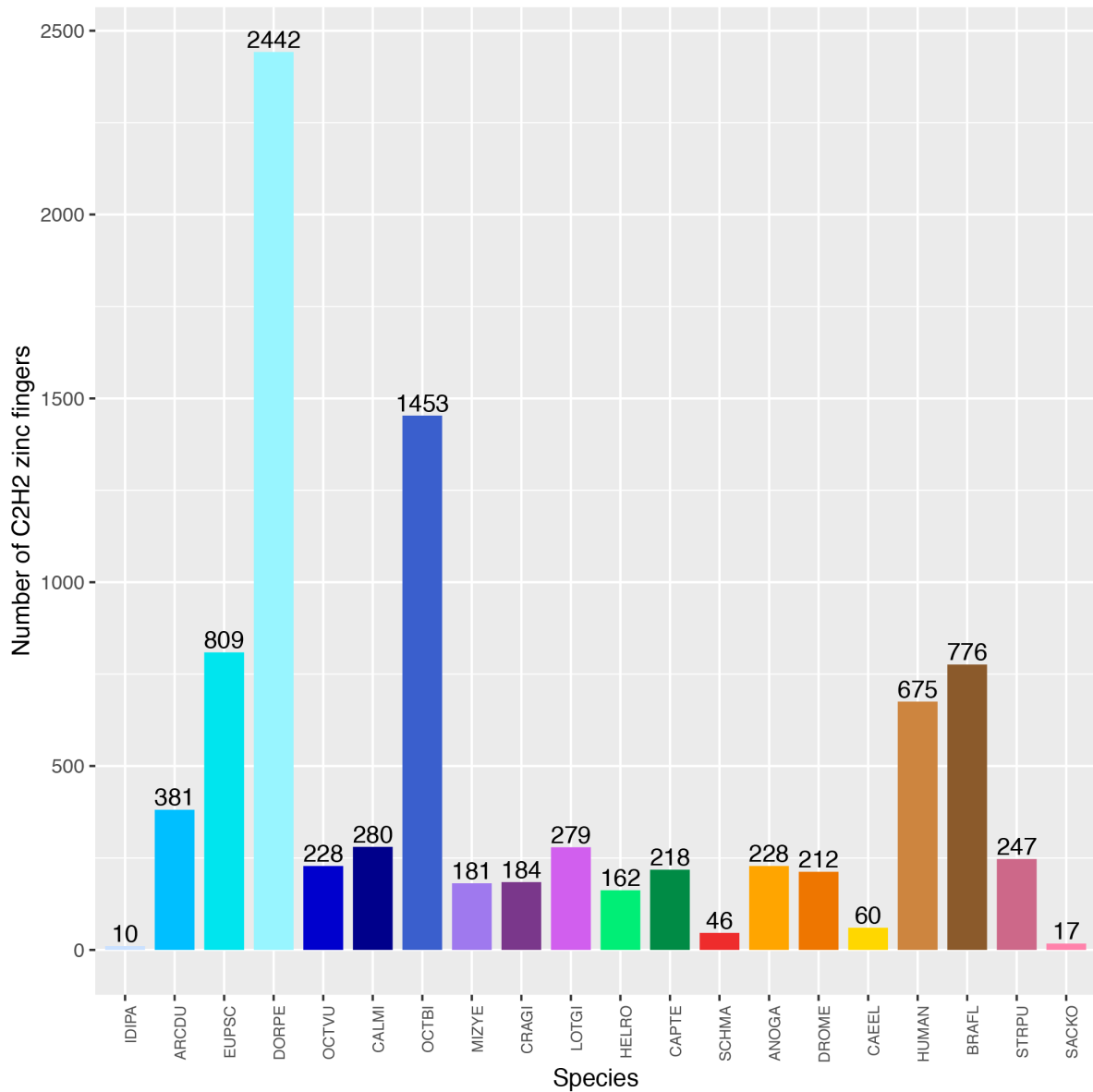
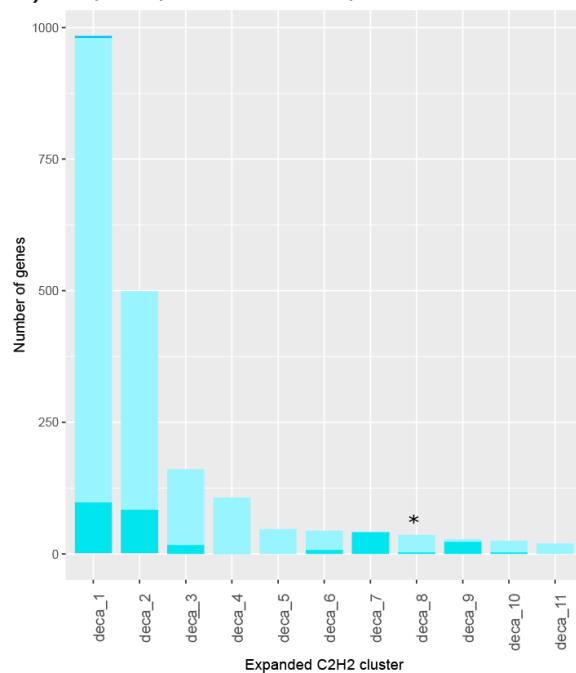
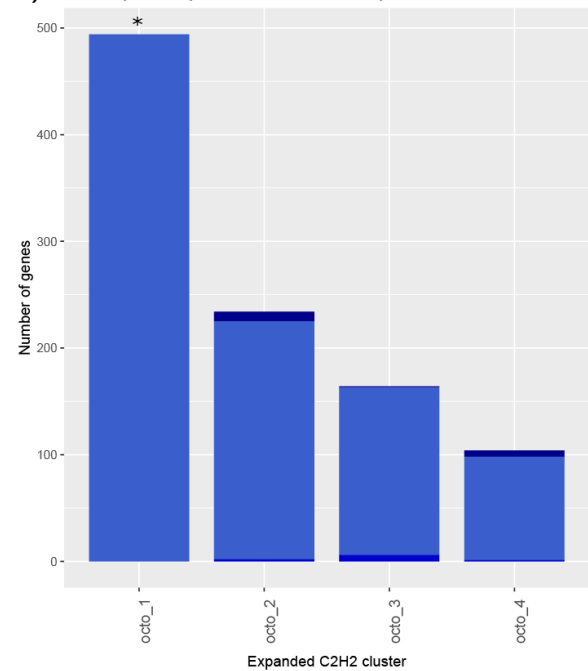


Figure S1. Number of C2H2 protein sequences per species used in tree construction. Species were sorted according to their evolutionary distance, with the four rightmost species (HUMAN, BRAFL, STRPU, SACKO) being deuterostomes and therefore acting as sister clade to the other protostome species, including the coleoid cephalopods (IDIPA, ARCDU, EUPSC, DORPE, OCTVU, CALMI, OCTBI). Abbreviations are as follows: ANOGA: *Anopheles gambiae*, ARCDU: *Architeuthis dux*, BRAFL: *Branchiostoma floridae*, CAEEL: *Caenorhabditis elegans*, CALMI: *Callistoctopus minor*, CAPTE: *Capitella teleta*, CRAGI: *Crassostrea gigas*, DORPE: *Doryteuthis pealeii*, DROME: *Drosophila melanogaster*, EUPSC: *Euprymna scolopes*, HELRO: *Helobdella robusta*, HUMAN, IDIPA: *Idiosepius paradoxus*, LOTGI: *Lottia gigantea*, MIZYE: *Mizuhopecten yessoensis*, OCTBI: *Octopus bimaculoides*, OCTVU: *Octopus vulgaris*, SACKO: *Saccoglossus kowalevskii*, SCHMA: *Schistosoma mansoni* and STRPU: *Strongylocentrotus purpuratus*.

a) Squid-specific C2H2 expansions



b) Octopus-specific C2H2 expansions



c) Cephalopod-specific C2H2 expansions

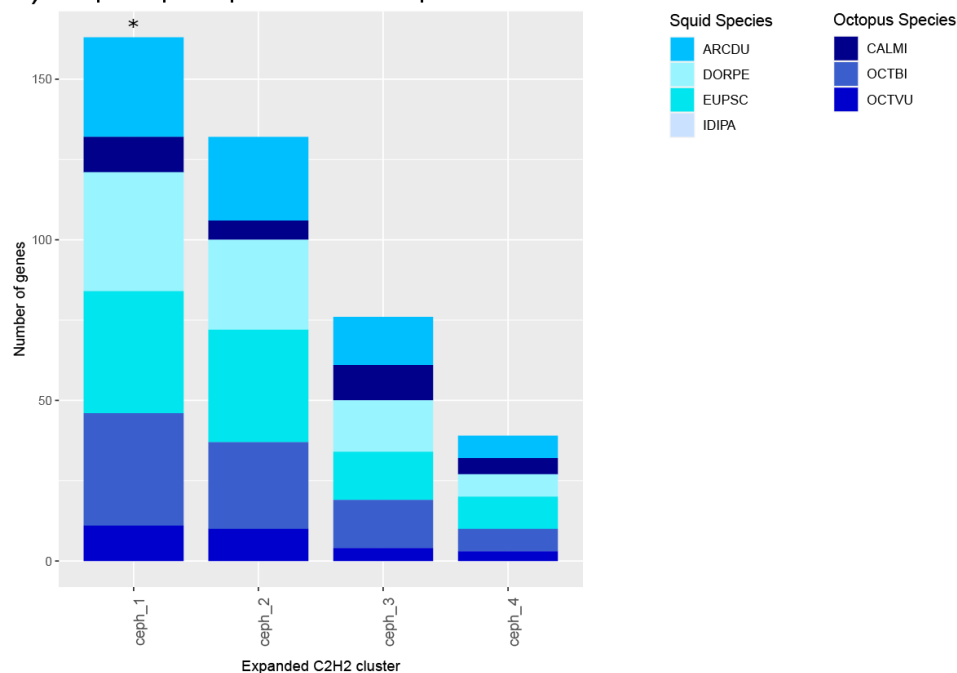
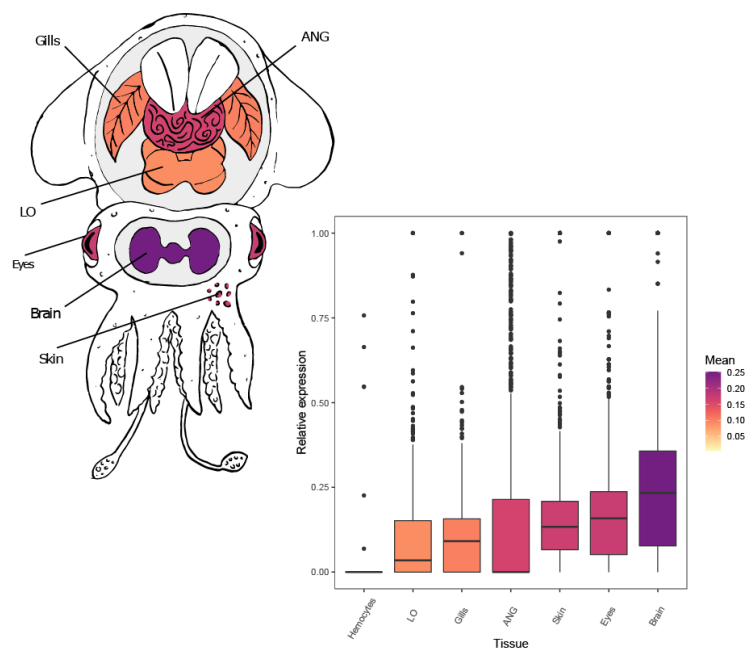
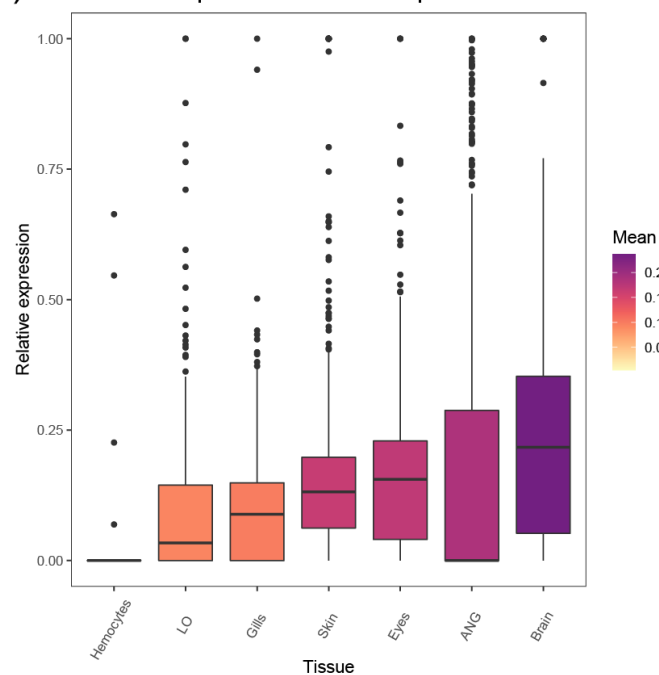


Figure S2. Number of genes per species in squid-, octopus- and cephalopod-specific C2H2 expansions. a) Squid-specific expansion clusters, ranging from 20 to 984 sequences. b) Octopus-specific expansion clusters, ranging from 104 to 495 sequences. c) Cephalopod-specific expansion clusters ranging from 39 to 163 sequences. All expansions had an SH support value greater than 0.7, with the exception of those marked with an asterisk (*). Abbreviations are as follows: ARCDU: *Architeuthis dux*, CALMI: *Callistoctopus minor*, DORPE: *Doryteuthis pealeii*, EUPSC: *Euprymna scolopes*, IDIPA: *Idiosepius paradoxus*, OCTBI: *Octopus bimaculoides*, OCTVU: *Octopus vulgaris*.

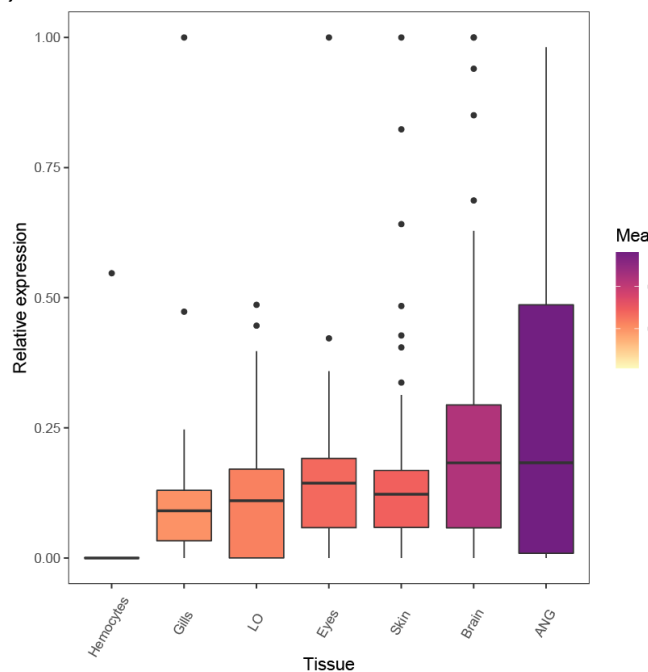
a) Relative expression of all C2H2 genes in *E. scolopes*



b) Relative expression of non-expanded C2H2s



c) Relative expression of cephalopod-specific C2H2s



d) Relative expression of squid-specific C2H2s

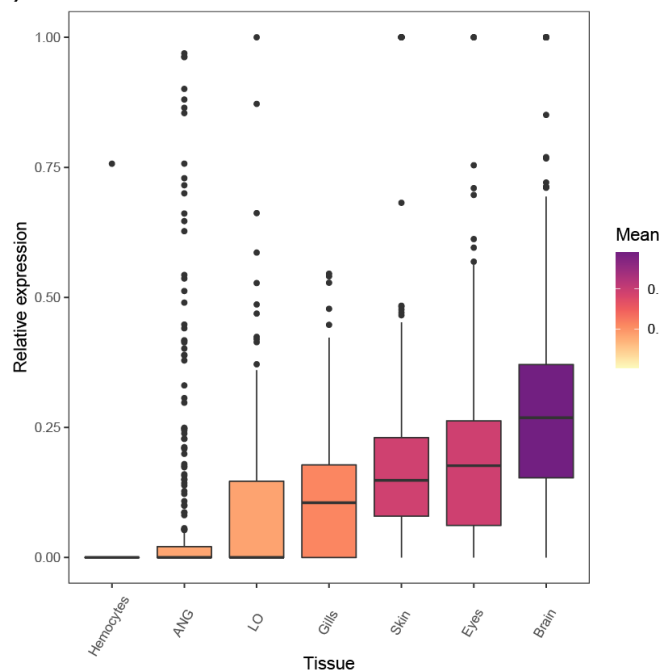
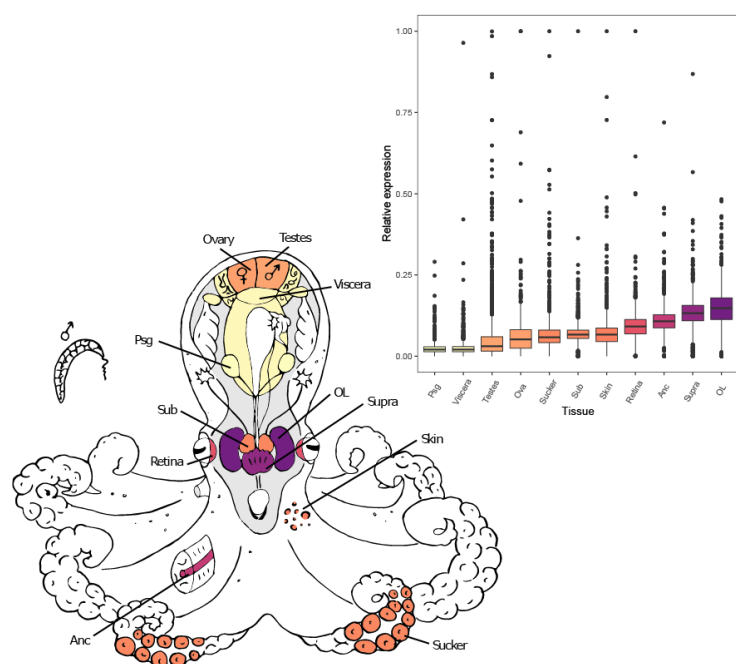
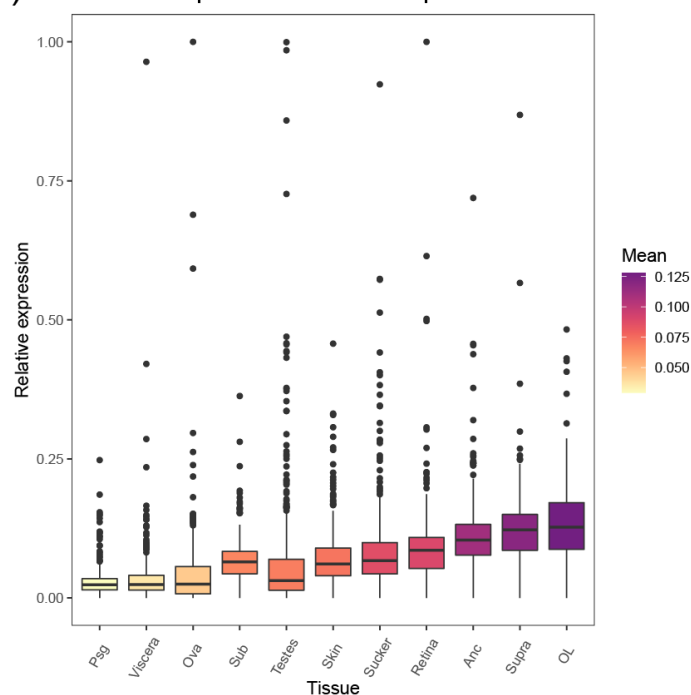


Figure S3. Relative expression of a) all C2H2 genes, b) non-expanded, c) cephalopod- and d) squid-specific expanded C2H2s in *E. scolopes*. Tissues in boxplots are sorted using their means, from lowest (left) to highest (right). The box colours were then used to colour the corresponding tissues in the animal sketches as in (a) and Figure 2d-f. Illustrations by Elena Ritschard.

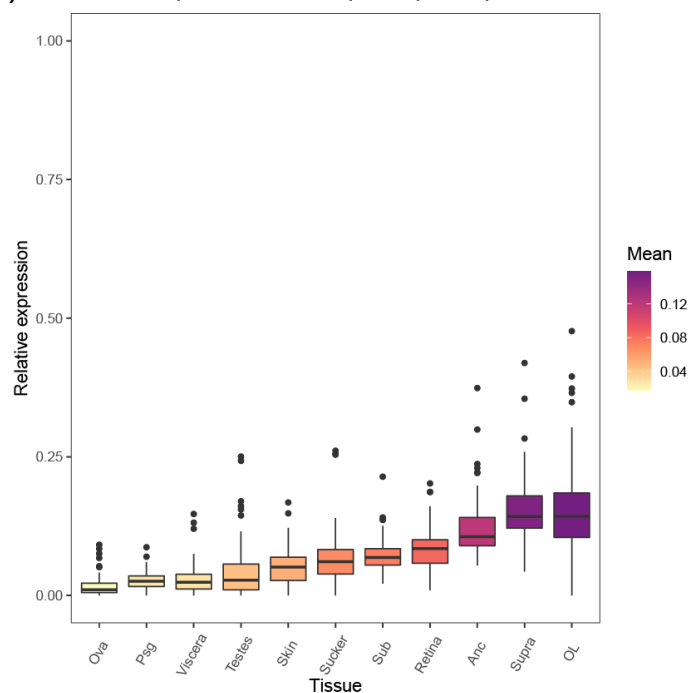
a) Relative expression of all C2H2 genes in *O. bimaculoides*



b) Relative expression of non-expanded C2H2s



c) Relative expression of cephalopod-specific C2H2s



d) Relative expression of octopus-specific C2H2s

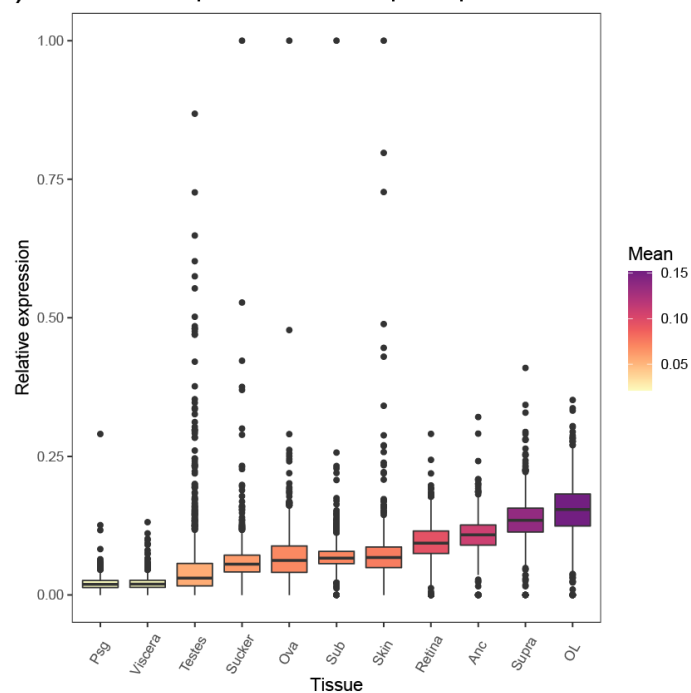
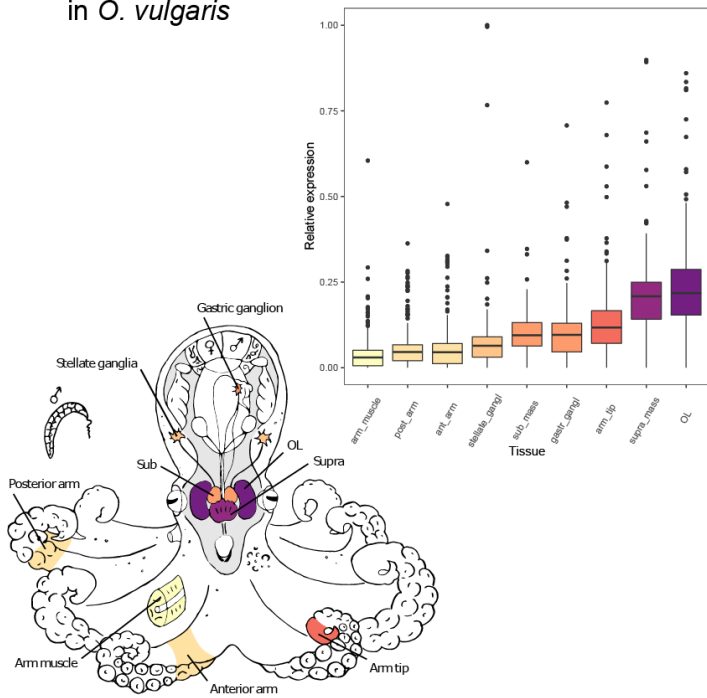
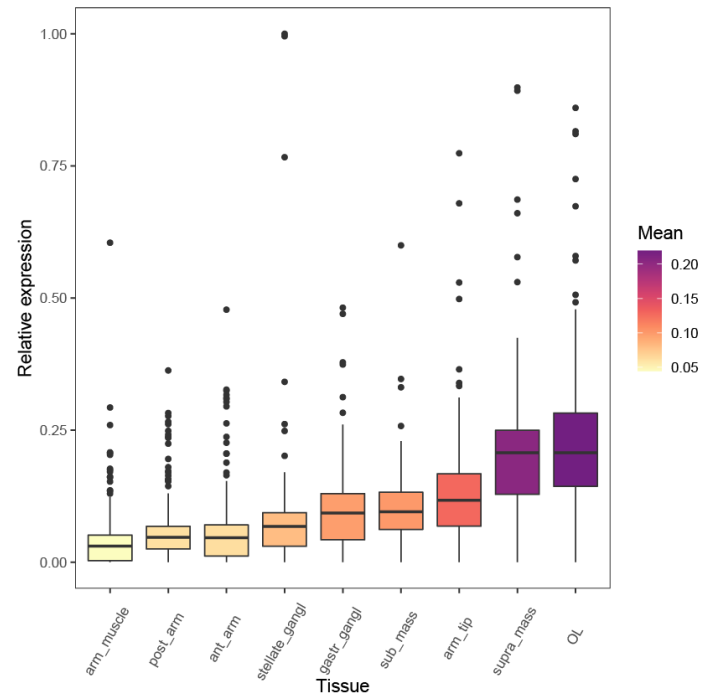


Figure S4. Relative expression of a) all C2H2 genes, b) non-expanded, c) cephalopod- and d) octopus-specific expanded C2H2s in *O. bimaculoides*. Tissues in boxplots are sorted using their means, from lowest (left) to highest (right). The box colours were then used to colour the corresponding tissues in the animal sketches as in (a) and Figure 2a-c. Illustrations by Elena Ritschard.

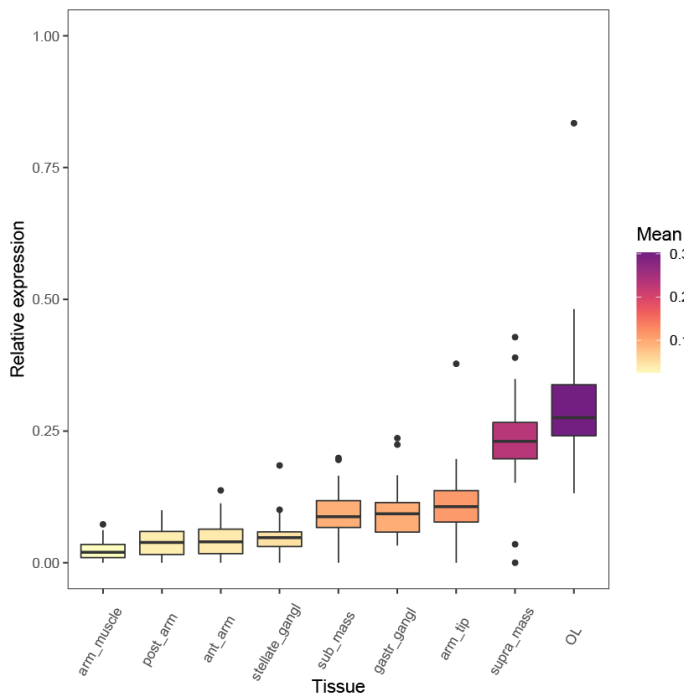
a) Relative expression of all C2H2 genes in *O. vulgaris*



b) Relative expression of non-expanded C2H2s



c) Relative expression of cephalopod-specific C2H2s



d) Relative expression of octopus-specific C2H2s

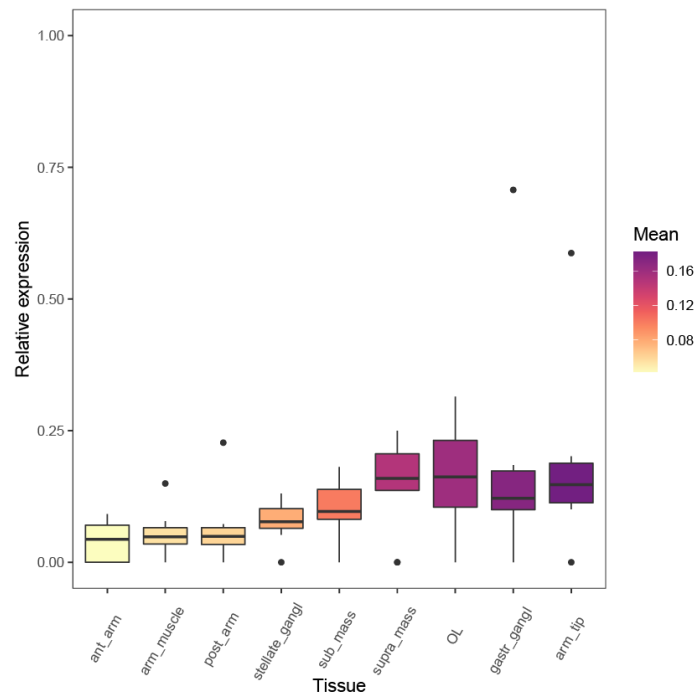
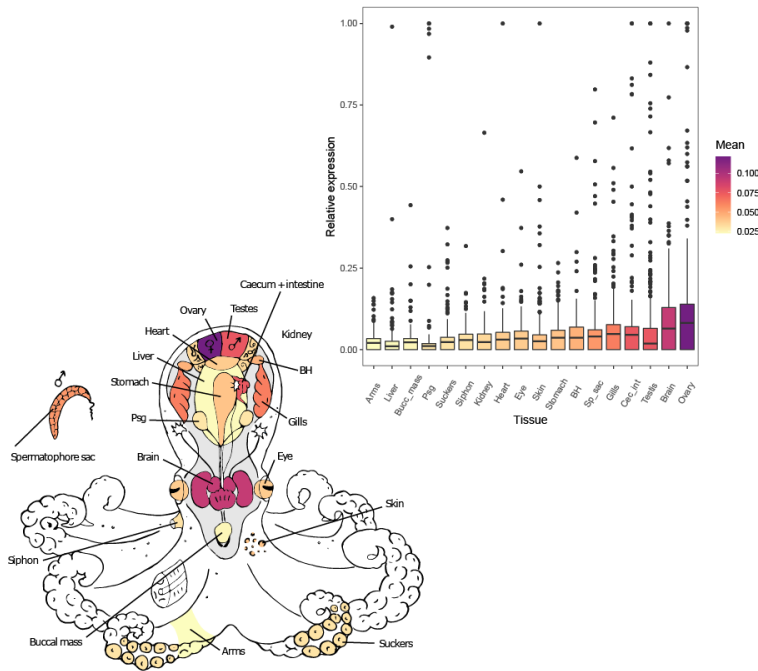
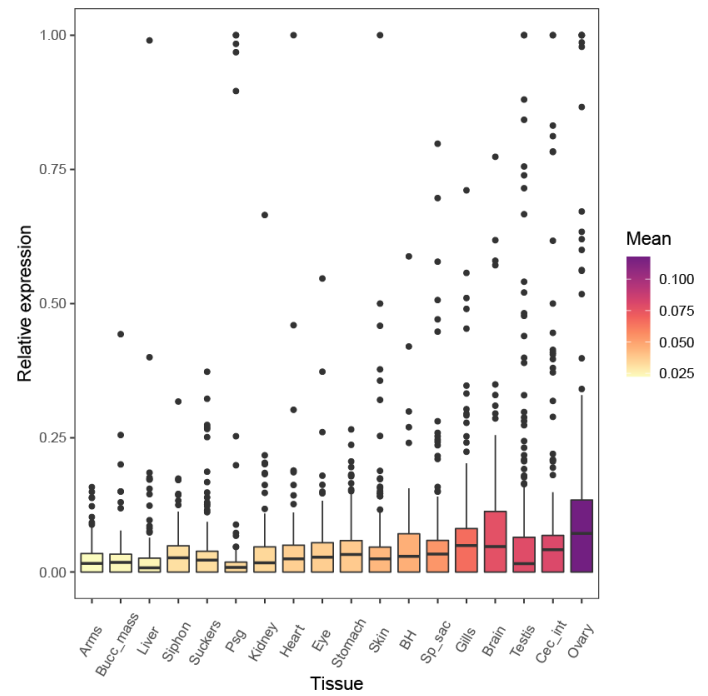


Figure S5. Relative expression of a) all C2H2 genes, b) non-expanded, c) cephalopod- and d) octopus-specific expanded C2H2s in *O. vulgaris*. Tissues in boxplots are sorted using their means, from lowest (left) to highest (right). a) The box colours were then used to colour the corresponding tissues in the animal sketch. Illustrations by Elena Ritschard.

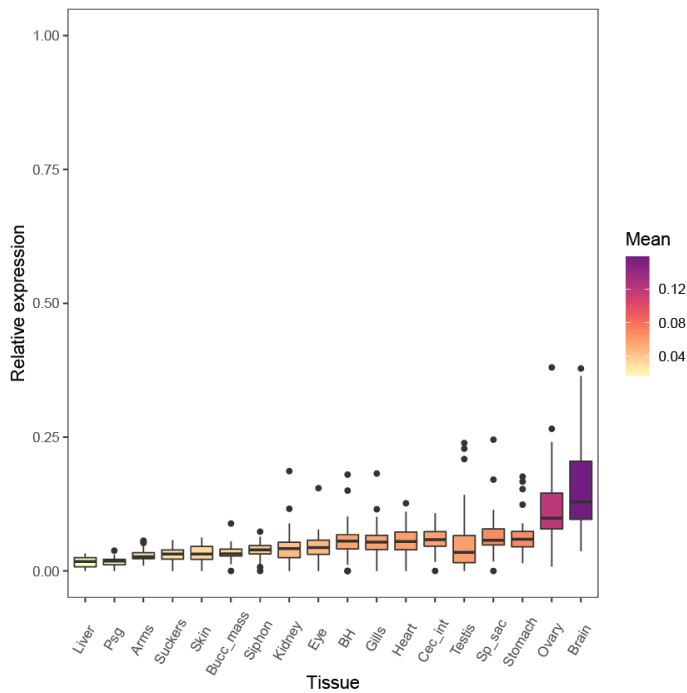
a) Relative expression of all C2H2 genes in *C. minor*



b) Relative expression of non-expanded C2H2s



c) Relative expression of cephalopod-specific C2H2s



d) Relative expression of octopus-specific C2H2s

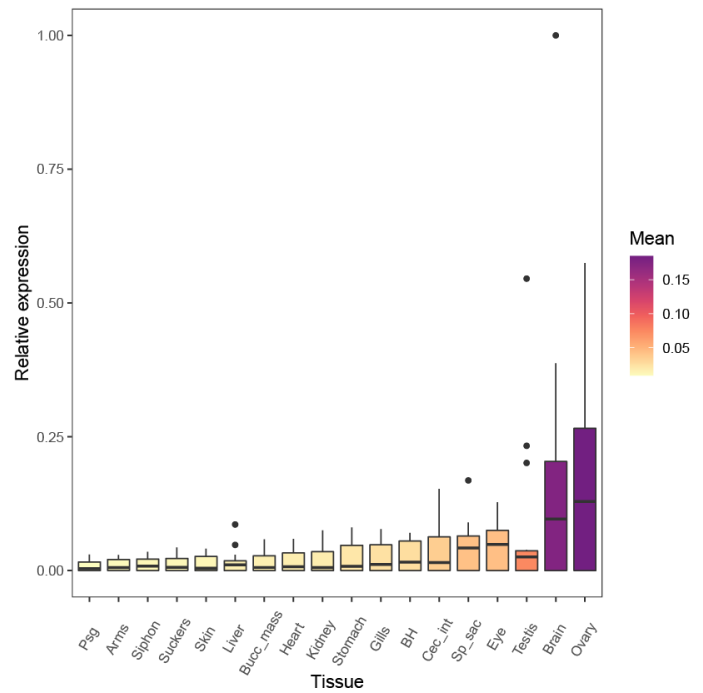


Figure S6. Relative expression of a) all C2H2 genes, b) non-expanded, c) cephalopod- and d) octopus-specific expanded C2H2s in *C. minor*. Tissues in boxplots are sorted using their means, from lowest (left) to highest (right). a) The box colours were then used to colour the corresponding tissues in the animal sketch. Illustrations by Elena Ritschard.

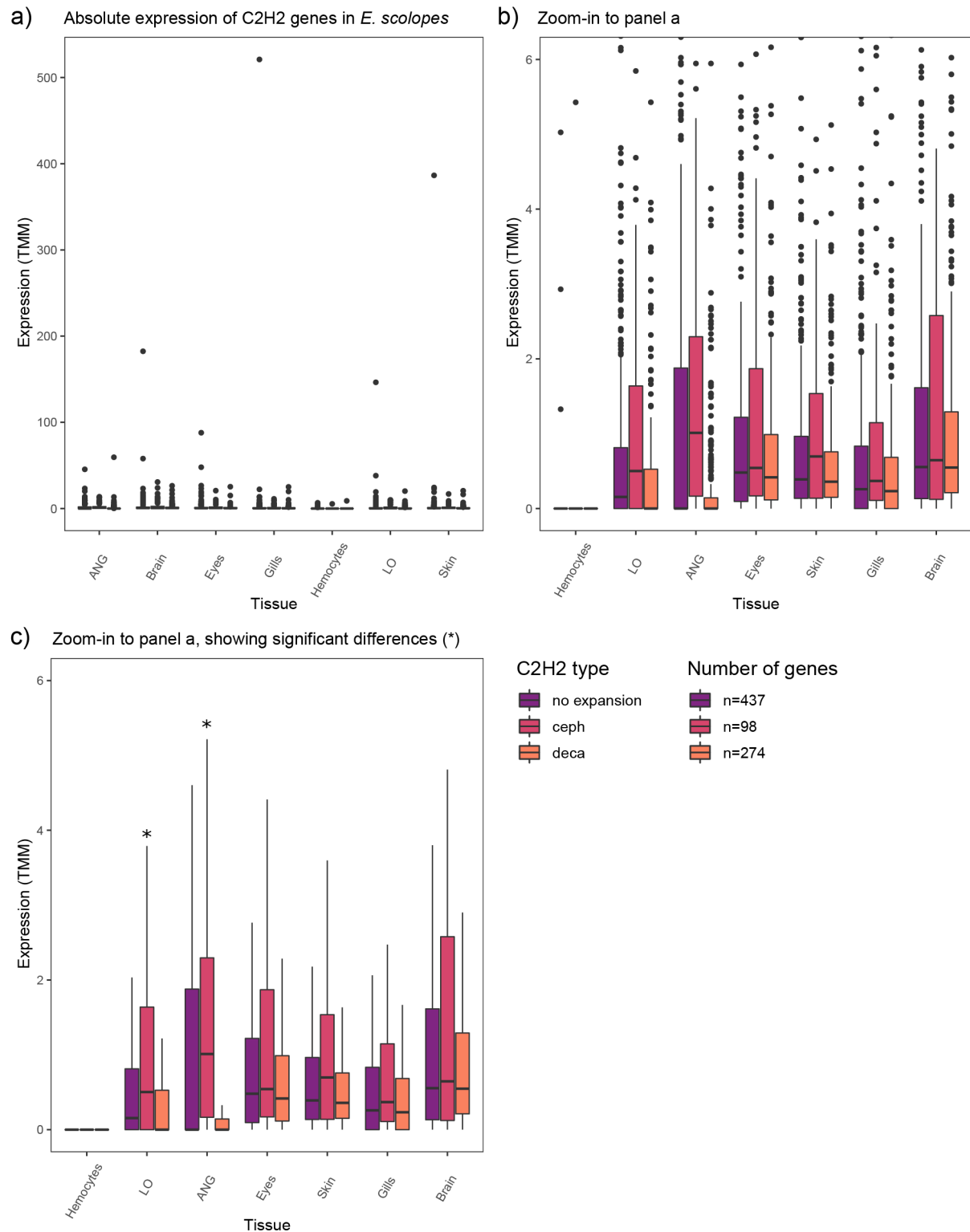


Figure S7. Absolute expression of *E. scolopes* non-expanded, cephalopod- and squid-expanded C2H2 genes in each tissue. Tissues are sorted using the group mean for each tissue, from lowest (left) to highest (right). a) Expression boxplot including all outliers. b) Zoom-in to panel a in the range 0 to 6 TMM expression values, c) Zoom-in to panel a in the range 0 to 6 TMM expression values but without outliers. Tissues with a significantly higher expression of cephalopod-specific expanded C2H2 genes in comparison to both non- and squid-expanded genes were marked with an asterisk (*).

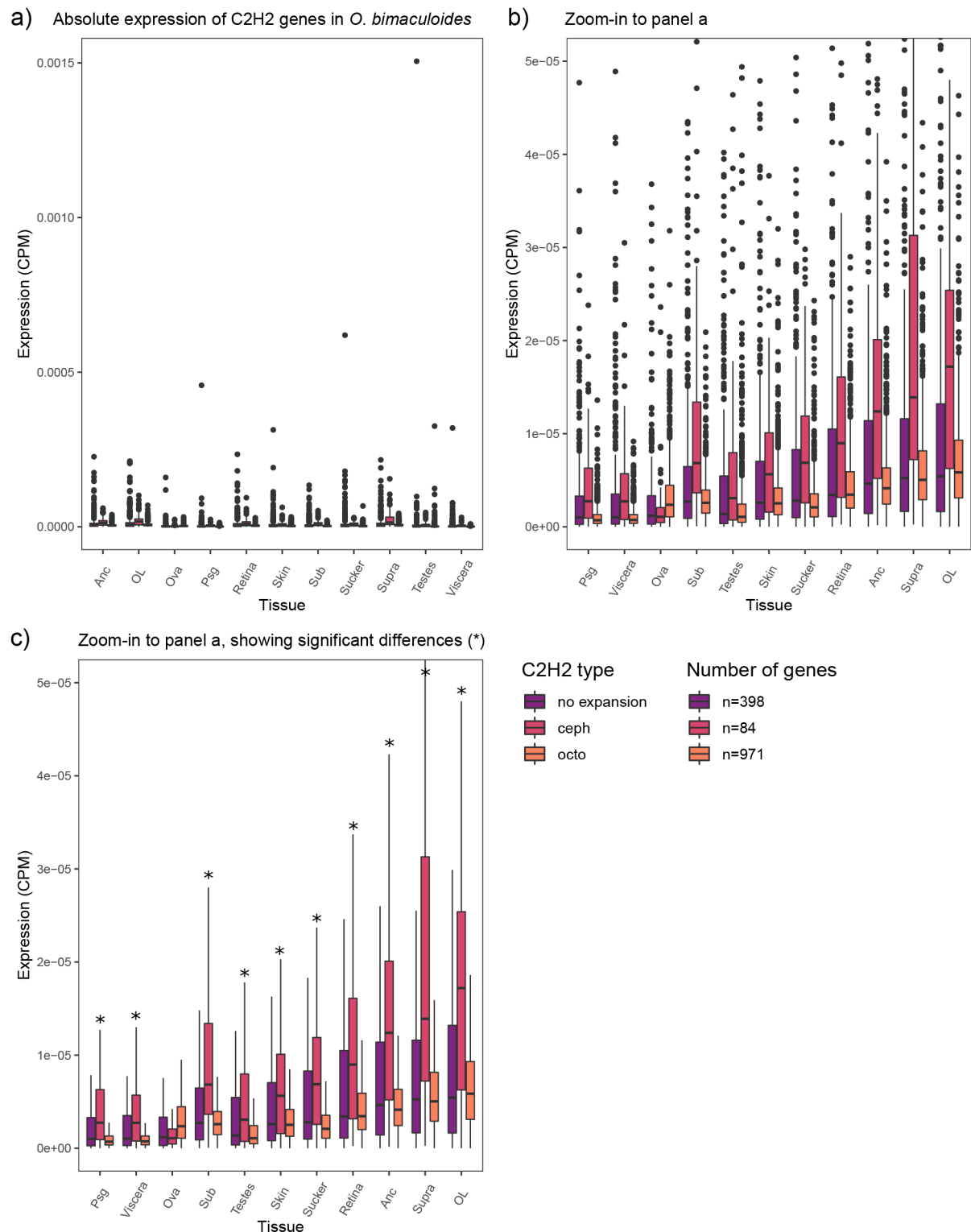


Figure S8. Absolute expression of *O. bimaculoides* non-expanded, cephalopod- and octopus-expanded C2H2 genes in each tissue. Tissues are sorted using the group mean for each tissue, from lowest (left) to highest (right). a) Expression boxplot including all outliers. b) Zoom-in to panel a in the range 0 to 5e-5 CPM expression values, c) Zoom-in to panel a in the range 0 to 5e-5 CPM expression values but without outliers. Tissues with a significantly higher expression of cephalopod-specific expanded C2H2 genes in comparison to both non- and octopus-expanded genes were marked with an asterisk (*).

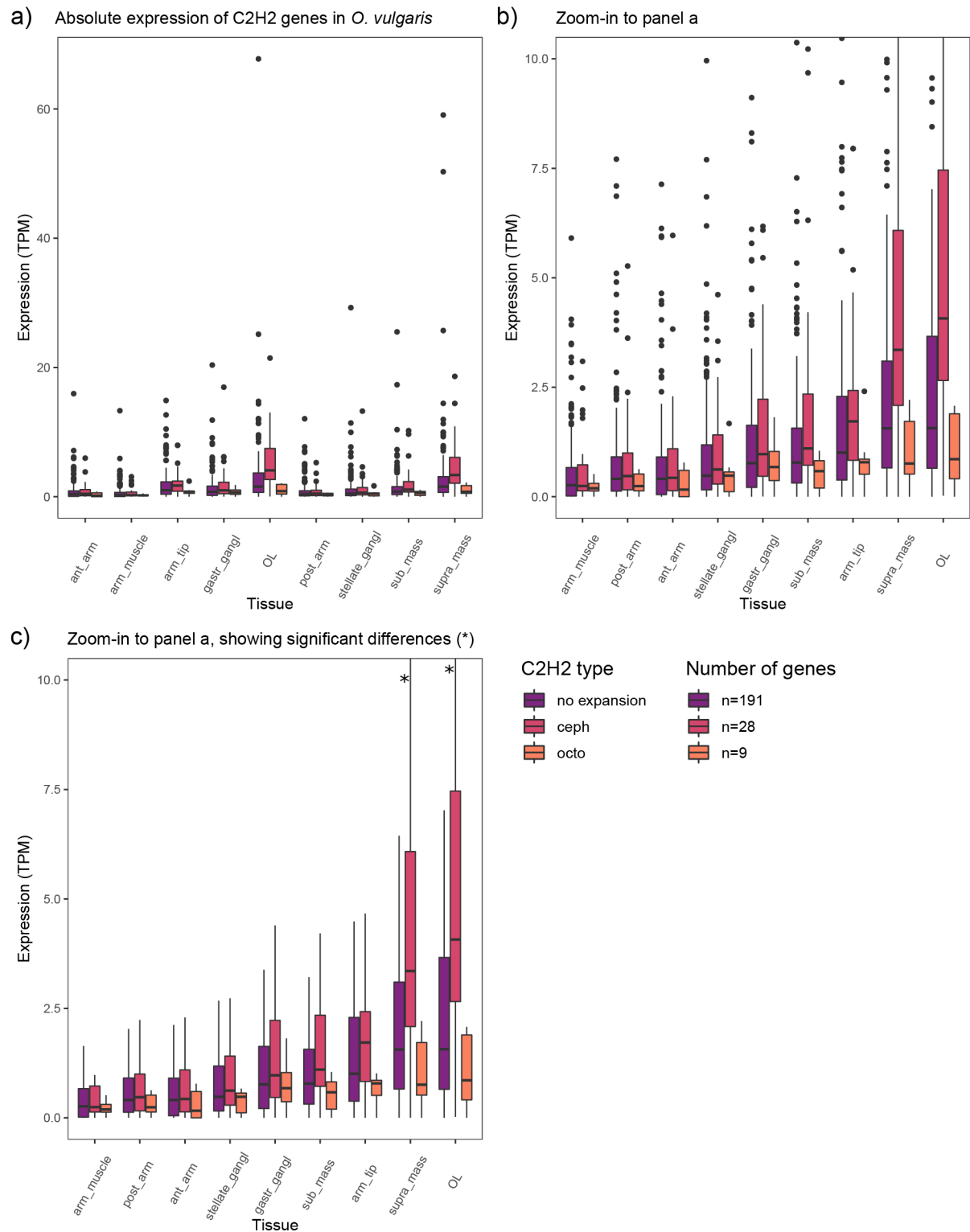


Figure S9. Absolute expression of *O. vulgaris* non-expanded, cephalopod- and octopus-expanded C2H2 genes in each tissue. Tissues are sorted using the group mean for each tissue, from lowest (left) to highest (right). a) Expression boxplot including all outliers. b) Zoom-in to panel a in the range 0 to 10 TPM expression values, c) Zoom-in to panel a in the range 0 to 10 TPM expression values but without outliers. Tissues with a significantly higher expression of cephalopod-specific expanded C2H2 genes in comparison to both non- and octopus-expanded genes were marked with an asterisk (*).

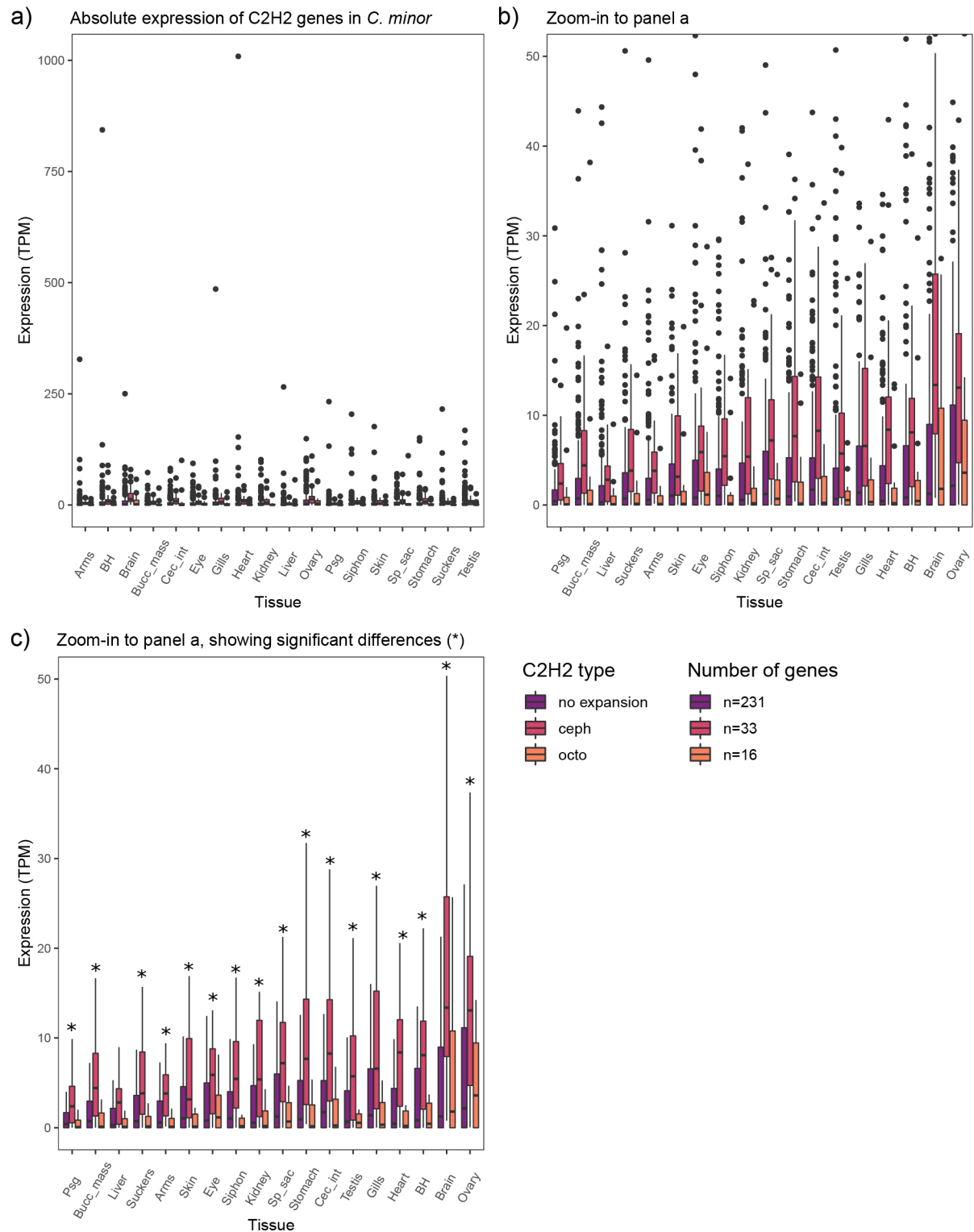


Figure S10. Absolute expression of *C. minor* non-expanded, cephalopod- and octopus-expanded C2H2 genes in each tissue. Tissues are sorted using the group mean for each tissue, from lowest (left) to highest (right). a) Expression boxplot including all outliers. b) Zoom-in to panel a in the range 0 to 50 TPM expression values, c) Zoom-in to panel a in the range 0 to 50 TPM expression values but without outliers. Tissues with a significantly higher expression of cephalopod-specific expanded C2H2 genes in comparison to both non- and octopus-expanded genes were marked with an asterisk (*).

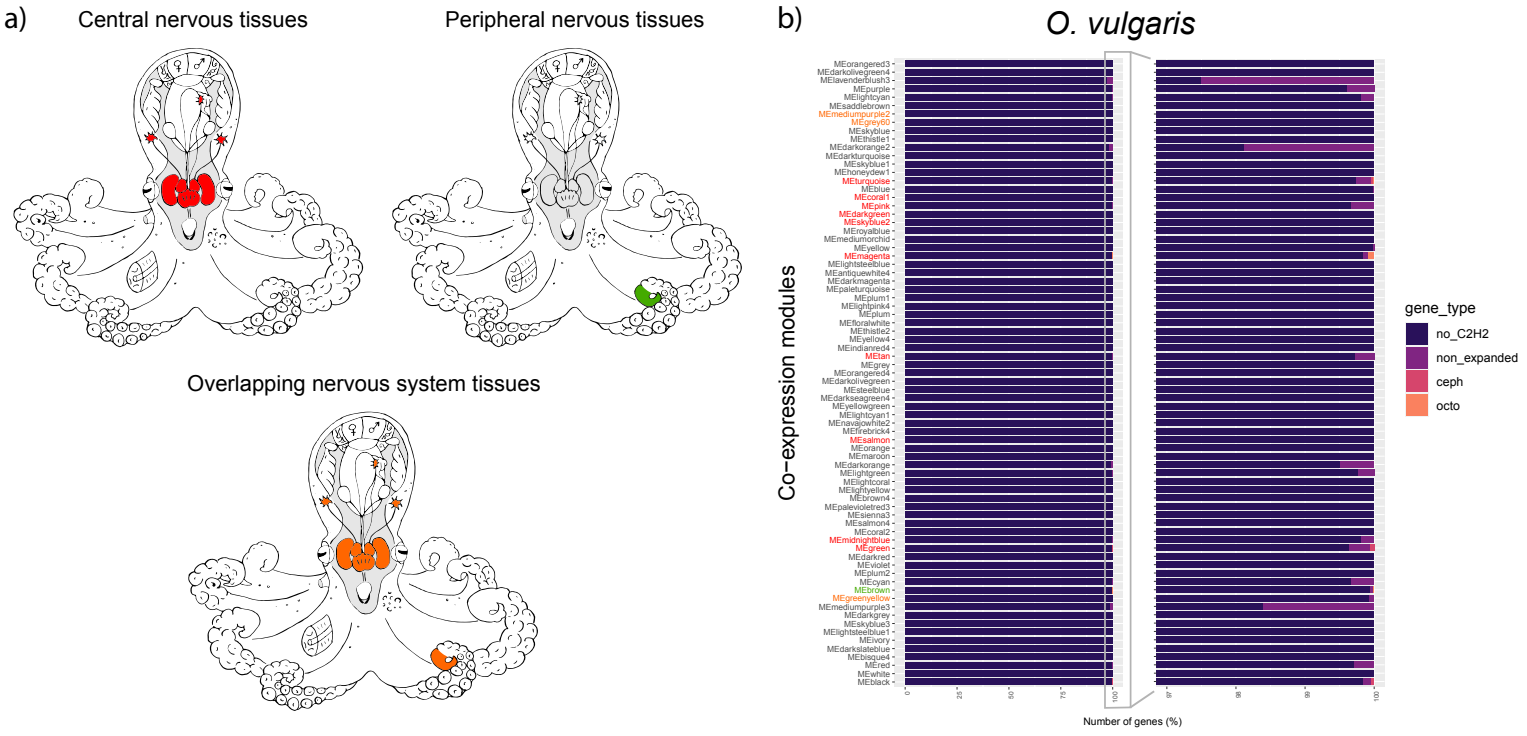


Figure S11. Distribution of C2H2 genes in the resulting WGCNA co-expression modules in *O. vulgaris*. Number of genes are shown as percentages (from 0% to 100% in left barplot and zoom-in to a shorter range in right barplot, for greater detail). Module names are color-coded based on their correlation to nervous tissues: red for central nervous tissues (brain, gastric ganglion and stellate ganglia), green for peripheral nervous tissues (arm tip) and orange for overlapping central and peripheral nervous tissues. Modules coloured black were not specifically correlated to nervous tissues.

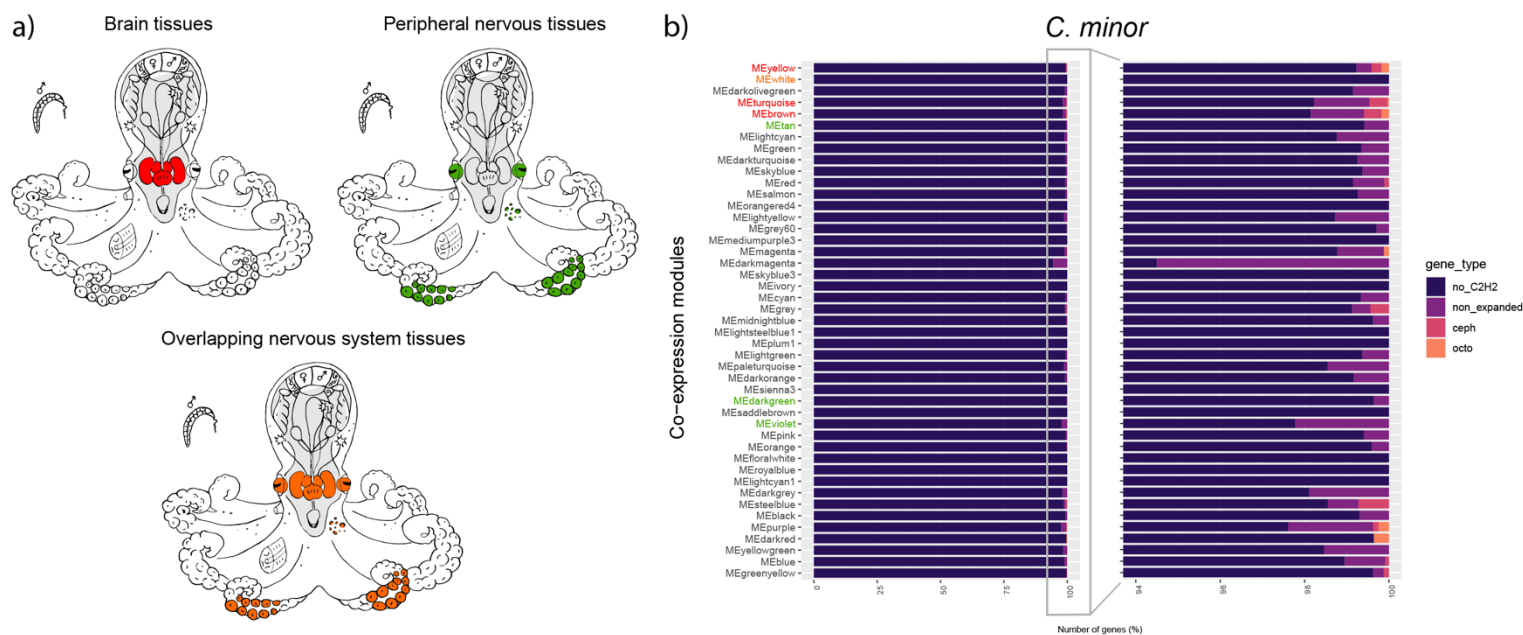


Figure S12. Distribution of C2H2 genes in the resulting WGCNA co-expression modules in *C. minor*. Number of genes are shown as percentages (from 0% to 100% in left barplot and zoom-in to a shorter range in right barplot, for greater detail). Module names are color-coded based on their correlation to nervous tissues: red for brain tissues, green for peripheral nervous tissues and orange for overlapping central and peripheral nervous tissues. Modules coloured black were not specifically correlated to nervous tissues.

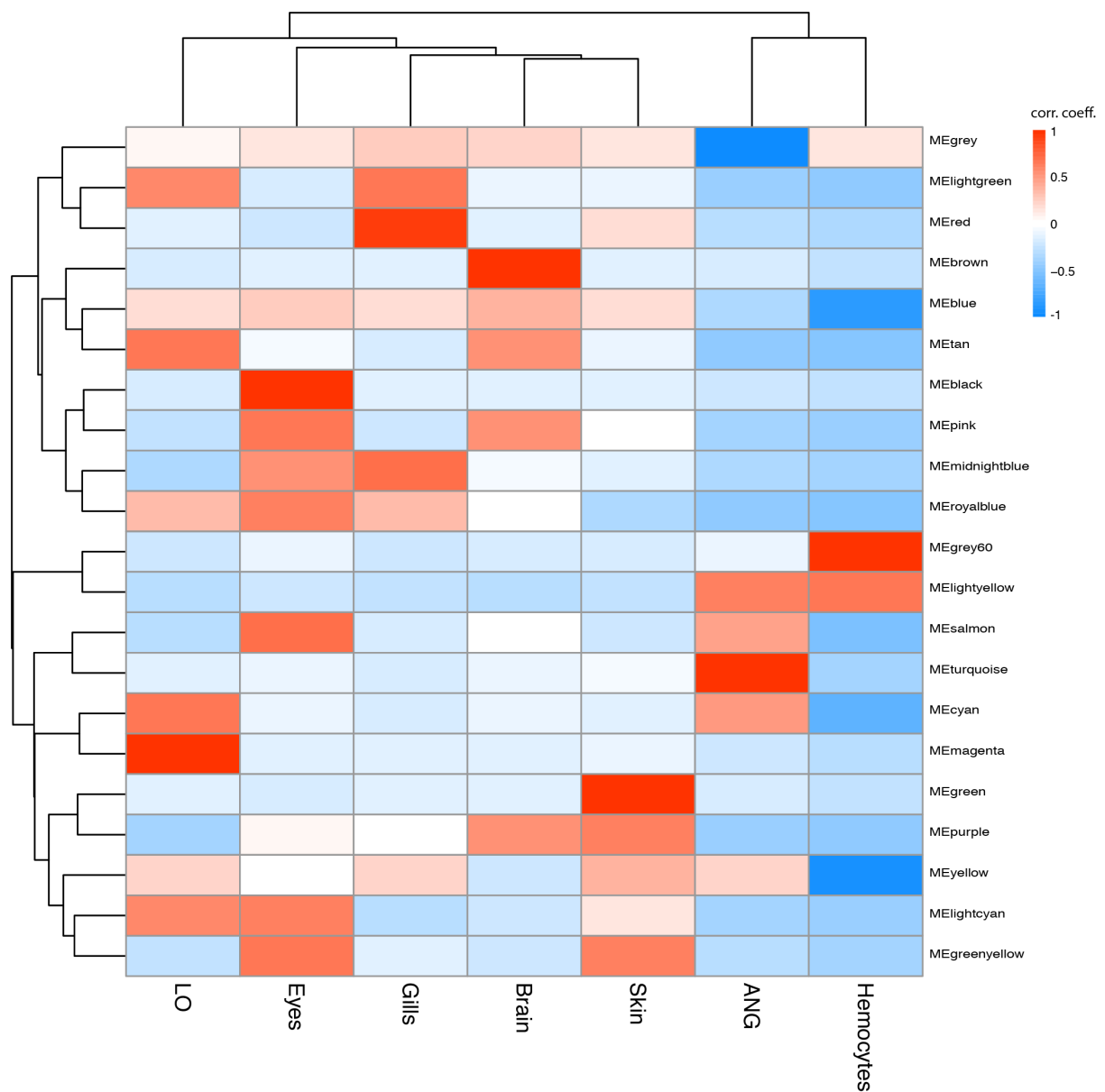


Figure S13. WGCNA module-trait relationship heatmap of *E. scolopes*. The heatmap visualizes the calculated correlation between eigengenes (y-axis), understood as the average expression profiles of the resulting modules, and each adult tissue (x-axis). Colours represent the correlation coefficient values, ranging from -1 (*i.e.*, negative correlation) to 1 (*i.e.*, positive correlation).

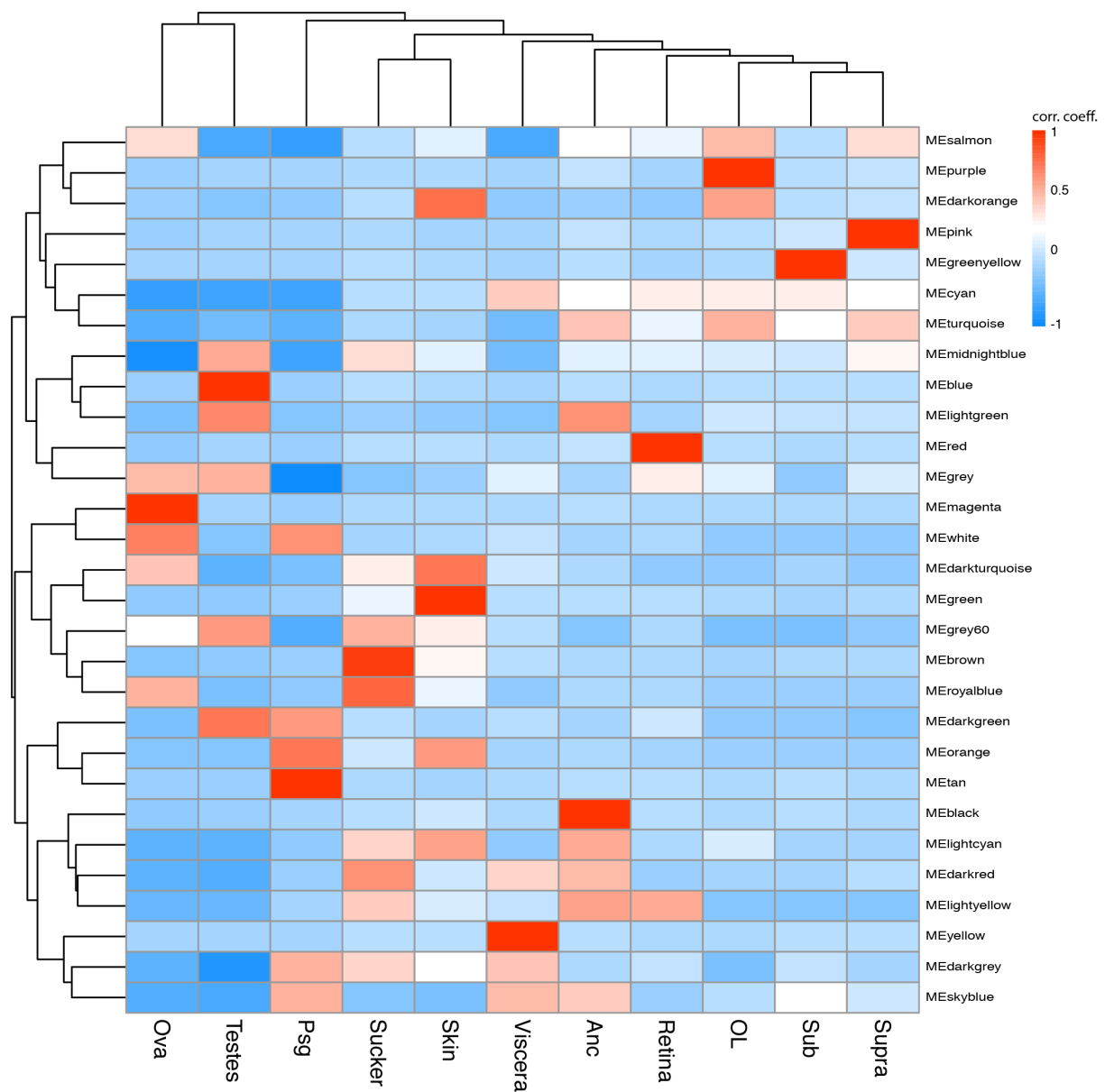


Figure S14. WGCNA module-trait relationship heatmap of *O. bimaculoides*. The heatmap visualizes the calculated correlation between eigengenes (y-axis), understood as the average expression profiles of the resulting modules, and each adult tissue (x-axis). Colours represent the correlation coefficient values, ranging from -1 (*i.e.*, negative correlation) to 1 (*i.e.*, positive correlation).

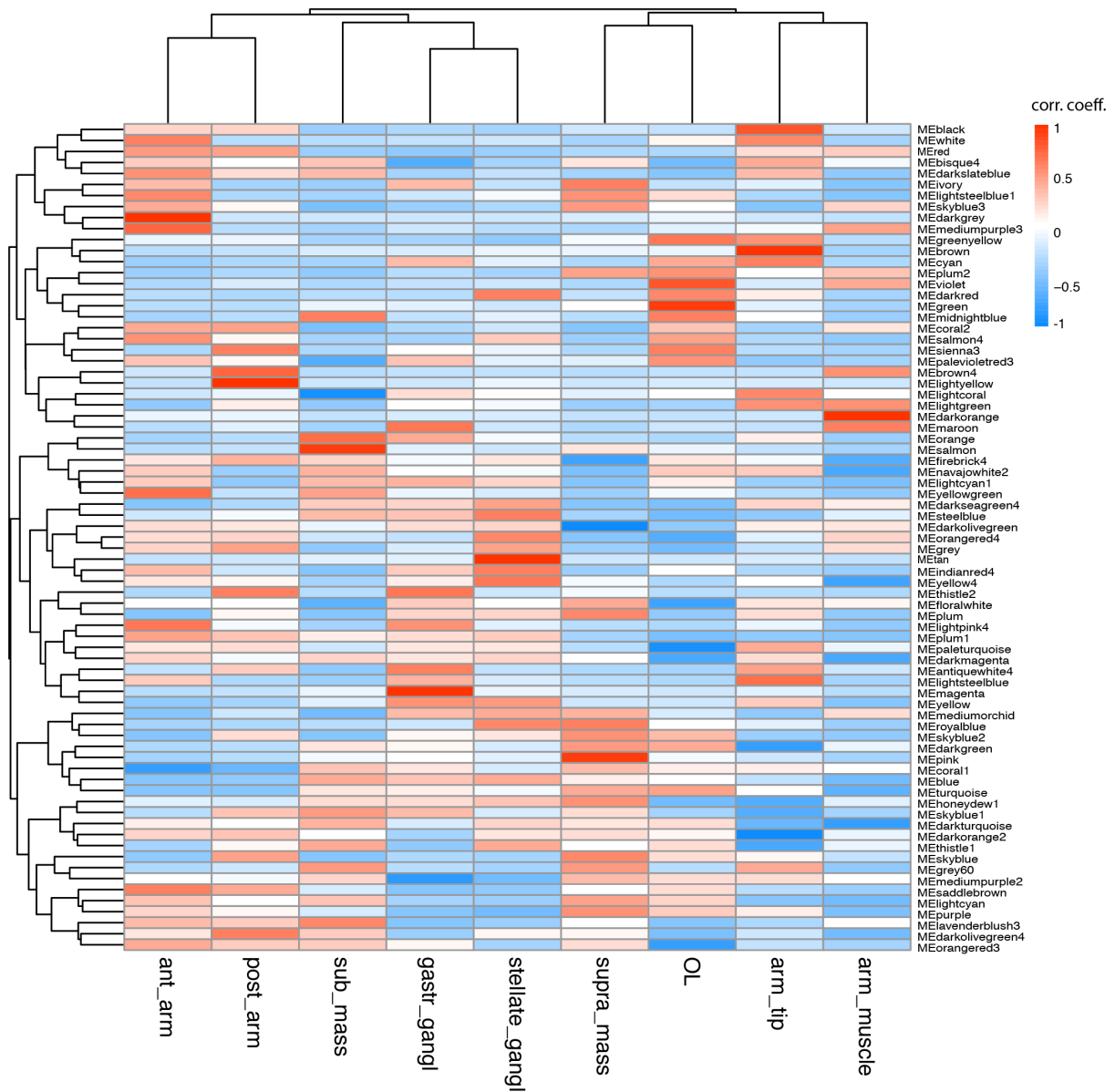


Figure S15. WGCNA module-trait relationship heatmap of *O. vulgaris*. The heatmap visualizes the calculated correlation between eigengenes (y-axis), understood as the average expression profiles of the resulting modules, and each adult tissue (x-axis). Colours represent the correlation coefficient values, ranging from -1 (*i.e.*, negative correlation) to 1 (*i.e.*, positive correlation).

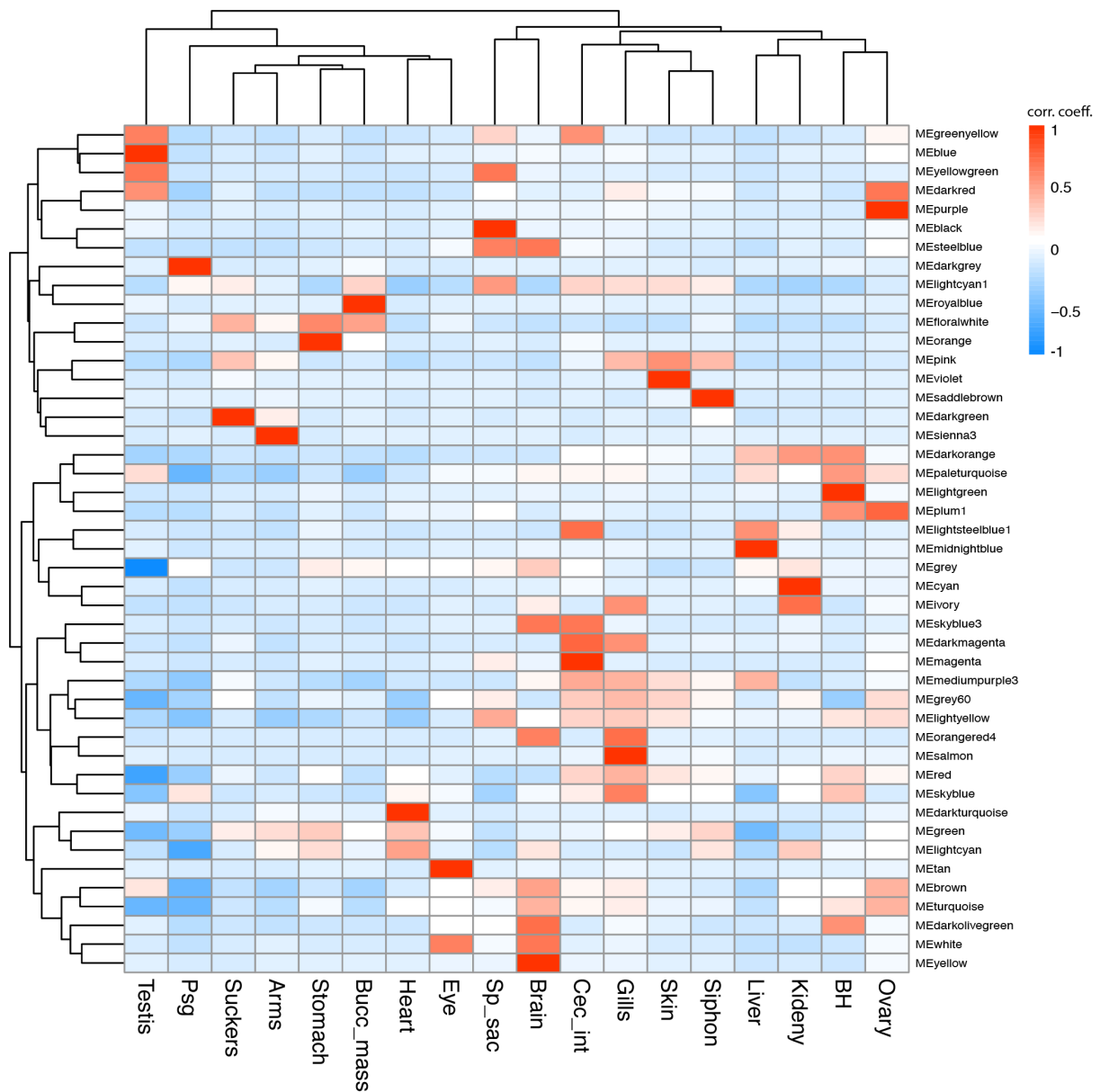


Figure S16. WGCNA module-trait relationship heatmap of *C. minor*. The heatmap visualizes the calculated correlation between eigengenes (y-axis), understood as the average expression profiles of the resulting modules, and each adult tissue (x-axis). Colours represent the correlation coefficient values, ranging from -1 (*i.e.*, negative correlation) to 1 (*i.e.*, positive correlation).

References

- S1. R. R. Da Fonseca, A. Couto, A. M. Machado, B. Brejova, C. B. Albertin, F. Silva, P. Gardner, T. Baril, A. Hayward, A. Campos, Â. M. Ribeiro, I. Barrio-Hernandez, H. J. Hoving, R. Tafur-Jimenez, C. Chu, B. Frazão, B. Petersen, F. Peñaloza, F. Musacchia, G. C. Alexander, H. Osório, I. Winkelmann, O. Simakov, S. Rasmussen, M. Z. Rahman, D. Pisani, J. Vinther, E. Jarvis, G. Zhang, J. M. Strugnell, L. F. C. Castro, O. Fedrigo, M. Patricio, Q. Li, S. Rocha, A. Antunes, Y. Wu, B. Ma, R. Sanges, T. Vinar, B. Blagoev, T. Sicheritz-Ponten, R. Nielsen, M. T. P. Gilbert, A draft genome sequence of the elusive giant squid, *Architeuthis dux*. *Gigascience*. 9, 1–12 (2020).
- S2. B. M. Kim, S. Kang, D. H. Ahn, S. H. Jung, H. Rhee, J. S. Yoo, J. E. Lee, S. Lee, Y. H. Han, K. Bin Ryu, S. J. Cho, H. Park, H. S. An, The genome of common long-arm octopus *Octopus minor*. *Gigascience*. 7, 1–7 (2018).
- S3. C.B. Albertin, S. Medina-Ruiz, T. Mitros, H. Schmidbaur, G. Sanchez, Z. Y. Wang, J. Grimwood, J. J. C. Rosenthal, C. W. Ragsdale, O. Simakov, D. S. Rokhsar. Genome and transcriptome mechanisms driving cephalopod evolution. *Nat Commun* 13, 2427 (2022). doi:10.1038/s41467-022-29748-w
- S4. M. Belcaid, G. Casaburi, S. J. McAnulty, H. Schmidbaur, A. M. Suria, S. Moriano-Gutierrez, M. Sabrina Pankey, T. H. Oakley, N. Kremer, E. J. Koch, A. J. Collins, H. Nguyen, S. Lek, I. Goncharenko-Foster, P. Minx, E. Sodergren, G. Weinstock, D. S. Rokhsar, M. McFall-Ngai, O. Simakov, J. S. Foster, S. V. Nyholm, Symbiotic organs shaped by distinct modes of genome evolution in cephalopods. *Proc. Natl. Acad. Sci. U. S. A.* 116, 3030–3035 (2019).
- S5. C. B. Albertin, O. Simakov, T. Mitros, Z. Y. Wang, J. R. Pungor, E. Edsinger-Gonzales, S. Brenner, C. W. Ragsdale, D. S. Rokhsar, The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*. 524, 220–224 (2015), doi:10.1038/nature14668
- S6. I. Zarrella, K. Herten, G. E. Maes, S. Tai, M. Yang, E. Seuntjens, E. A. Ritschard, M. Zach, R. Styfhals, R. Sanges, O. Simakov, G. Ponte, G. Fiorito, The survey and reference assisted assembly of the *Octopus vulgaris* genome. *Sci. Data*. 6, 1–8 (2019).

