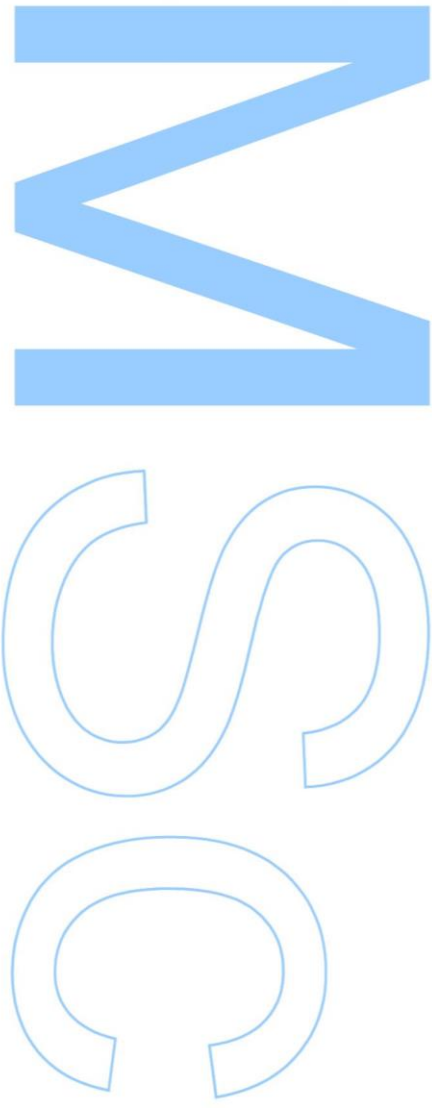
Evolutionary genomics Study of genes Involved in animal Adaptation

Filipe Correia Gonçalves e Silva,
Mestrado de Bioquímica
Departamento de Química e Bioquímica
2014/2015

Orientador
Prof. Doutor Agostinho Antunes,
Investigador/ Professor
Centro Interdisciplinar de Investigação Marinha e Ambiental
Faculdade de Ciências da Universidade do Porto



Todas as correções determinadas pelo júri, e só essas, foram efetuadas.
O Presidente do Júri,

Porto, _____/_____/_____

A proposal submitted to
Faculty of Sciences of the University of Porto
for fulfillment of degree of master of
Biochemistry

Acknowledgements

The Author would like to thank several people who made the completion of this project possible.

First, I want to thank my supervisor, Professor Agostinho Antunes, for the incredible opportunity of working in the vanguard of cephalopod sequencing, his unending patience as he pushed me through my inexperience and his availability to always facilitate my work as much as possible.

Second, I'd like to thank my colleagues who wasted their own work time to help me as they could. Thank you very much Daniela Almeida, Emanuel Maldonado, Tibisay Escalona and Alex Anoop. Thank you all for your help and for just being great people.

Lastly, I'd like to thank my family, for all of the squid food jokes and for their mockery of my chosen organism to study. I love you all and hope that you will continue to watch over me as you've done so far. I know I can be an obnoxious and complicated person at times, but I know you can see past it and put me straight.

Thank you to everyone who helped me make this a reality.

Sumário

A evolução adaptativa é influenciada pela genética populacional, seleção natural e restrições bioquímicas. Variação criada por mutação, uma grande força evolutiva, é traduzida para fenótipos através de vias metabólicas e dinâmicas de redes moleculares. A retenção da variação genética e os efeitos da seleção dependem da dinâmica populacional. A identificação de mutações precisas é necessária para a compreensão dos mecanismos moleculares e evolutivos subjacentes que conduzem a mudanças fenotípicas adaptativas. Neste estudo foram abordados genes envolvidos em fenómenos de adaptação relacionados com a produção de veneno e com processos de desenvolvimento. Os genes Hox definem o eixo anterior-posterior e a identidade dos segmentos nos animais durante as fases iniciais do desenvolvimento, sendo essenciais para corpos bem estruturados. Estes genes alvo de interesse para determinar a origem e causa subjacente de novidades morfológicas devido à possibilidade de mutações resultarem em diferentes. Por outro lado, os venenos fascinaram a humanidade desde tempos antigos e são hoje um pilar em ensaios funcionais de moléculas e desenvolvimento de drogas. São produzidos em glândulas especializadas que libertam um cocktail de componentes biologicamente ativos, importantes em atividades como predação e proteção. Os cefalópodes são o grupo mais diferenciado dos moluscos, representado por invertebrados altamente evoluídos, com formas e tamanhos variados e têm o sistema nervoso e aparelhos sensoriais muito desenvolvidos. Muitos são venenosos e fatais para humanos. Os cefalópodes são um grupo pouco estudado, uma vez que muitas espécies vivem no alto mar e em grandes profundidades, tornando a sua investigação dispendiosa e complicada.

Com o propósito de caracterizar as dinâmicas evolutivas que ocorrem neste grupo de moluscos, foram realizadas análises genómicas comparativas dos genes Hox e dos recém-descobertos genes de veneno, carboxipeptidase e metaloprotease com domínio GON, num genoma de cefalópode e nove transcritomas de cefalópodes. Os resultados revelaram atomatização dos genes Hox, alguns duplicados e outros perdidos, indicando um grau elevado de perturbação da estrutura genómica. As relações filogenéticas das espécies e os potenciais efeitos de mutações nos genes Hox foram analisados. Foram identificadas espécies venenosas putativas e seleção positiva a ocorrer em sítios específicos nos genes de veneno, que se encontram principalmente sob seleção negativa, assegurando a manutenção do esqueleto proteico e a sua atividade.

Palavras-chave

Adaptação ambiental – Desenvolvimento de organismos - Toxinas

Abstract

Adaptive evolution is shaped by population genetics, natural selection and biochemical constraints. Variation created by mutation, a major evolutionary force, is translated into phenotypes through metabolic pathways and dynamics of molecular networks. Retention of genetic variation and the efficiency of selection depend on the population dynamics. The identification of precise mutations is required for a complete understanding of the underlying molecular and evolutionary mechanisms driving adaptive phenotypic change. Genes related to venom production and development processes involved in adaptation were approached in this study. Hox genes define the anterior-posterior axis and segment identity of animals during early development, being essential for properly structured bodies. The possibility that mutations in these genes can lead to different body plans makes the Hox genes a target of interest in determining the origin and underlying cause of morphological novelties. Venoms, on the other hand, have fascinated mankind since early times and even today are a pillar in molecular functions assays and drug development. They are produced in specialized glands that release a cocktail of biologically active components with major relevance in important biological needs, such as predation and protection. Cephalopods are the most distinguished group of molluscs, represented by highly evolved invertebrates of varied shapes and sizes, and a very well developed nervous systems and sensorial apparatus. Many of these species are also venomous and have proven fatal to humans. Cephalopods are also an understudied group, since many species inhabit open ocean waters and the deep-sea, making research difficult and expensive.

Here, we performed comparative genomics analyses of Hox genes and the recently discovered venom genes, carboxypeptidase and metalloprotease GON-domain, in a cephalopod genome and nine transcriptomes of other cephalopod species to characterize the evolutionary dynamics occurring in this unique mollusc clade. We determined that the Hox genes were atomized, some genes duplicated and others lost, indicating a high degree of genomic structure disturbance. We assessed the species phylogenetic relationships and potential effects of the mutational changes in Hox genes. We identified putative venomous species and demonstrated the occurrence of diversifying selection in specific sites in the venomous genes, mostly under purifying selection, ensuring the maintenance of the proteins' scaffold and its activity.

Keywords

Environmental adaptation – Organisms development - Toxins

Contents

Figure Contents

Tables Contents

List of abbreviations

Dfd – Deformed

DNA – Deoxyribonucleic acid

ML – Maximum likelihood

FEL – Fixed effects likelihood

FUBAR – Fast Unconstrained Bayesian AppRoximation

MEME – Mixed Effects Model of Evolution

REL – Random effects likelihood

Scr – Sex combs reduced

SLAC – Single likelihood ancestor counting

Pb – proboscipedia

ADAMTSs – a disintegrin and metalloproteinase domain with thrombospondin type-1 modules

CAP – cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins

Chapter 1 - Introduction

Evolution comprehends a series of processes that promoted the natural changes leading to species' emergence, environment adaptation and eventually extinction, producing over time the diversity of species and organisms we see today. Evolution occurs when DNA suffers change, particularly at gene level in populations, and the frequency of such changes in each gene and type (alleles) are usually influenced by mutations, genetic drift, migration and selection (Barton et al. 2007). Genes encode proteins, which in turn provide the organisms' characteristics. Their expression, which has specific regulation mechanisms, can also be changed. Thus, genetic modification can lead to different body plans and/ or behaviors. Organisms' development and functioning are regulated by these DNA segments, genetically inherited, which determine environmental adaptation and the likelihood of an organism's survival and reproduction (Alberts B. 2002).

An adaptive feature is a characteristic of an organism favoring its survival success in a given location and in a particular fashion. The adaptation can be a physical detail, such as size, shape and bodily mechanisms (such as temperature regulation in homoeothermic animals), or even behavioral, securing safety habits that prolong the organisms' life, namely by avoiding predators. These adaptations resulted from evolution over time and allowed animals of varied shapes and sizes to exist and populate different habitats (Barton et al. 2007).

Within animals, the target of this study centers on Mollusca, animals commonly known as molluscs, which represent an invertebrate group of over 100 000 soft-bodied living species, most aquatic and often possessing a shell of some form, including the well-known scallop, the clam, the oyster, the snail, the slug, the squid, the cuttlefish and the octopus among others (Nielsen 2012). Although there is no definite consensus on their phylogenetic relations, due to erroneous classifications and inconsistent results making their taxonomy a controversial subject, they possess an extensive fossil record and the number of species is estimated to be much larger than currently described (Nielsen 2012). Despite being a group with historical relevance to humans as a source of food, jewelry, tools, and even pets, they are considerably understudied. Within the Mollusca, we mostly focus on the study of cephalopods.

Cephalopoda is an interesting group as it contains highly evolved invertebrates, most commonly known for the features of their skins, with the ability to quickly alter the color and texture for camouflage, mimicry and communication purposes (Hall and Hanlon

2002, Mather and Mather 2004, Hanlon et al. 2009). It is one of the most important clades of Mollusca and generally described as dominant predators, being an essential component of marine systems. The complex behavioral and learning capabilities of cephalopods are associated to a highly developed nervous system that appears to be correlated with their lifestyle. They are also the invertebrates with the largest nervous system known to date (Fiorito et al. 2014).

The flexibility of their behavior is supported by cellular and synaptic plasticity in the central and peripheral nervous system and in neuromuscular junctions. Inhabiting every type of marine habitat has also lead to the evolution of remarkable visual systems. Coleoid eyes are structurally similar to those of vertebrates, possessing a camera-type eye, a spherical lens, a cornea and a retractable pupil, although they are usually monochromatic. The hypothesis that there may be areas of specialization in the retina, linked to pupil shape, resulting in higher levels of acuity and/or sensitivity, has been suggested as adaptive mechanisms to their environment (Talbot and Marshall 2011).

Cephalopods have extremely varied body plans and characteristics. Nautiloids have external shells of calcium carbonate. The shell-less argonaut produces a shell-like structure to brood eggs. The cuttlefish have an internal skeleton. The squids have a reduced internal skeleton of chitin. Finally, octopuses have almost no skeletal elements, with the only hard part in their body being the chitinous beak. Currently, all of the living cephalopods are separated into either Nautiloidea (the basal, more primitive *Nautilus* genus) or Coleoidea, with Ammonoidea being a third and extinct group within the class (Figure 1) (Ruppert 2004). Many cephalopods are also venomous, with toxins such as cephalotoxin, identified in at least four species, being fatal to humans. Venoms are key evolutionary innovations behind the explosive diversification of many animal clades (Ruder et al. 2013).

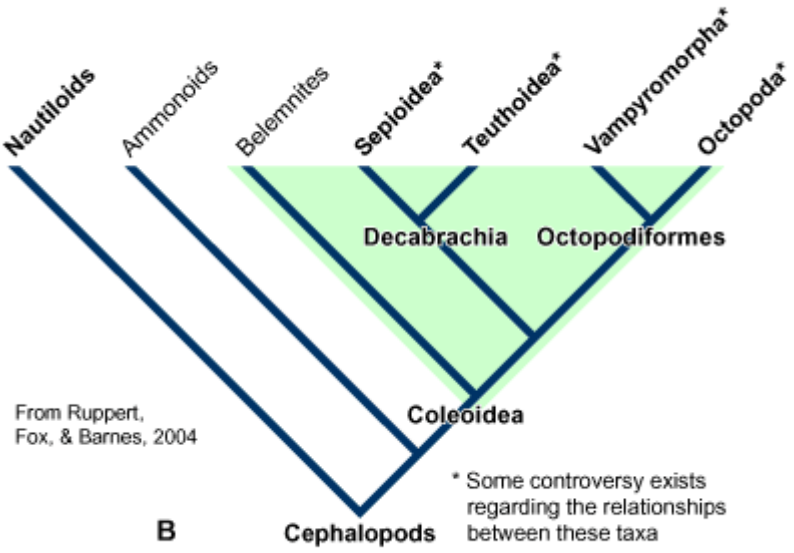


Fig.1 – Schematic representation of the current cephalopod phylogeny. Extant species are represented in bold. Adapted from (Ruppert 2004).

The study of venom has been a matter of great interest among scientists since early times. This was due in part to the inherent danger associated with venomous animals, which make them a subject of public fascination throughout human history (Casewell et al. 2013). Moreover, the frequent bites of these animals contributed to the importance of venom research worldwide (Sunagar et al. 2012). Venom research allowed to understand the mechanism of action of several proteins, venoms included, in binding and interaction in structural and functional biology (Quintero-Hernandez et al. 2013). It also revealed new insights into the evolutionary dynamic of venomous proteins in biology (Sunagar et al. 2013). However, venom research has been strongly biased into very specific taxonomic groups, such as cone snails, scorpions, spiders and snakes, which received much more attention than others (Ruder et al. 2013). Thus, many animal groups with known or suspected venom systems remain poorly understood, making the origin and evolution of venom controversial on a wide number of animals (Fry et al. 2009). Venoms are biochemical cocktails of biologically active components, such as amino acids, peptides, neurotransmitters and others (Fry et al. 2009), produced in a specialized gland. Different animals use them in different ways, with protection and predation being the most common forms (Massey et al. 2012, Casewell et al. 2013). Venom evolution in cephalopods is a poorly explored area. The group has been subject of very little research with most studies performed with octopods, such as the giant Pacific octopus and the blue-ringed octopus (Ruder et al. 2013).

Cephalopods are an understudied group in part because many species live at great depths and in the open ocean, making research difficult and expensive (Hoving et al. 2014). The giant squid (Figure 2) (*Architeuthis dux*) was first photographed in the wild only recently (2005) despite years of attempts and resources pooled (Kubodera and Mori 2005). It is the whole genome sequencing of this particular species that this study emphasizes, in comparison with the transcriptomes of several other cephalopods, including *Dosidicus gigas* - the Humboldt squid*, Doryteuthis pealeii* - the longfin inshore squid*, Lolliguncula brevis* - the Atlantic brief squid*, Sepioteuthis lessoniana* - the bigfin reef squid*, Sepia esculenta* - the golden cuttlefish*, Onychoteuthis banksii* - the common clubhook squid*, Bathypolypus arcticus* – a deep sea octopus, unidentified *Grimpoteuthis sp.*, a cirroctopod (dumb octopus) and *Pareledone albimaculata*, an Antarctic octopus, to further the knowledge on the group and complement previous studies on evolutionary development surrounding the Hox genes.

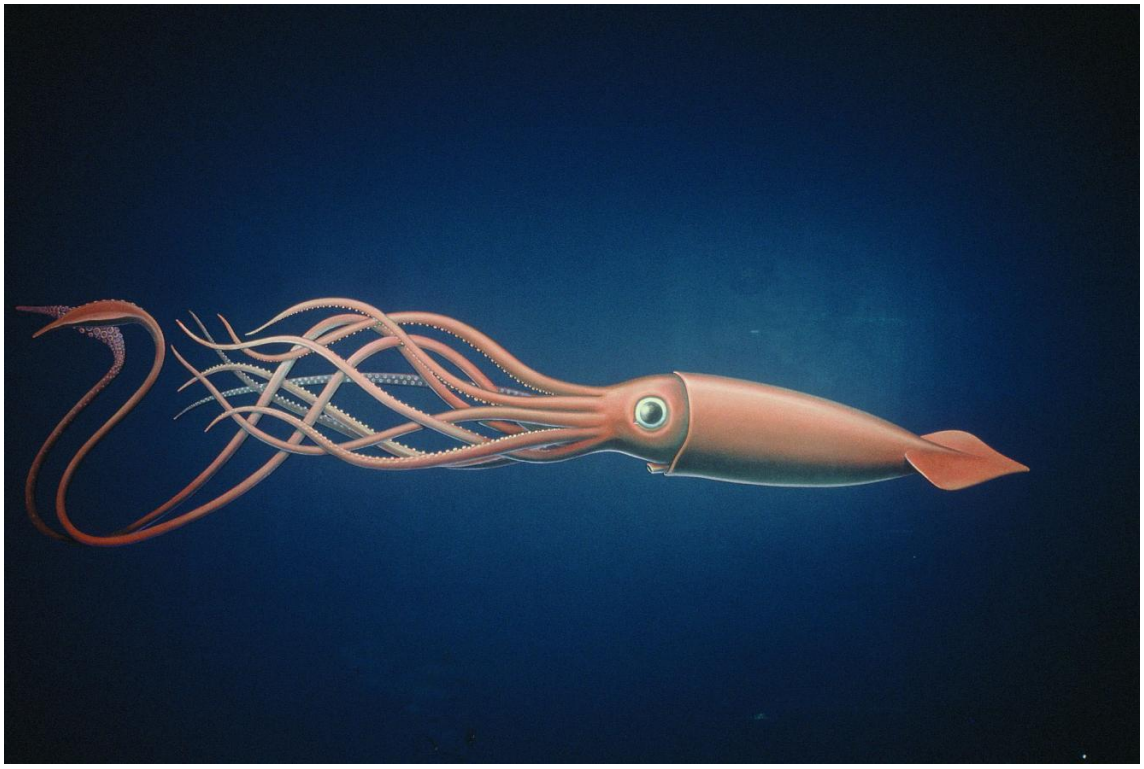


Fig.2 – Illustration of the giant squid, *Architeuthis dux* (retrieved from the Smithsonian Institution and credited to Glenn Rankin). This species is capable of reaching over 10m in length and is observed in depths from 300m up to 1000m. It shares most anatomical feature with other squids and has one of the largest eyes of living creatures. It also boasts a highly developed nervous system and a complex brain.

In August 2015, the *Octopus bimaculoides* genome was published, with the main goal of investigating the molecular basis of cephalopod brain and body innovation (Albertin et al. 2015). They found no evidence for hypothesized whole-genome duplications in the octopus lineage, although they detected massive expansions of certain gene families, previously thought to be a unique case to vertebrates. The study suggests that substantial expansion of a handful of gene families, along with extensive remodeling of the genome and repetitive content, played a critical role in the evolution of cephalopod morphological innovations, including their large and complex nervous systems.

A recent study on Coleoid venom expanded the limited information in this group and identified two novel venom components not yet documented. The molecular evolution of such venom genes showed a small proportion of sites under diversifying selection, most of which were confined to the molecular surface, thus highlighting the importance of variation of the protein's surface (Ruder et al. 2013).

Here, we analyzed the comparative genomics and evolution of the Hox genes and two novel venom genes, carboxypeptidase and metalloprotease GON-domain, in a cephalopod genome and nine other cephalopod transcriptomes in order to characterize the unique properties of these genes in cephalopods. We assessed the presence of genes in the studied species to: 1) identify the genes in groups not previously studied; 2) use the newly-acquired data to further resolve the cephalopod phylogeny; 3) determine how the diverse body structures and plans are connected to the differences found among the species' Hox genes; 4) analyze carboxypeptidase and metalloprotease GON-domain genes to understand their gene recruitment into to the coleoid venom arsenal, and 5) characterize the selective pressure acting on these proteins and its importance.

Chapter 2 - Methods

Hox genes

Genome and Transcriptomes acquisition

The giant squid (*Architeuthis dux*) genome and the transcriptomes of *Dosidicus gigas, Doryteuthis pealeii, Lolliguncula brevis, Sepioteuthis lessoniana, Sepia esculenta, Onychoteuthis banksii, Bathypolypus arcticus, Grimpoteuthis sp.*, and *Pareledone albimaculata* were obtained from an international consortium collaboration to study the giant squid. Illumina sequencing was used and a Trinity (Haas et al. 2013) assembly pipeline was chosen to assemble the sequences into larger scaffolds. EvidentialGene (Gilbert 2013) was used to predict coding regions and associated translations.

Gene identification and extraction

All known molluscan sequences of Hox genes were extracted from the National Center for Biotechnology Information (*www.ncbi.nlm.nih.gov/*) – NCBI – and the BLAST (Boratyn et al. 2013) algorithm from the same center was used to run these sequences, both in DNA and protein, against the genome and the transcriptomes in study to identify candidate scaffolds containing Hox genes. Additional sequences were recovered from Ensembl (Cunningham et al. 2015) Genomes database. Coding region predictions were used when applicable. All scaffolds and coding regions were manually curated to verify the identification. While three organisms were used for an initial confirmation (*Euprymna scolopes* – when not possible, *Nautilus* genus –, *Lottia gigantea* and *Crassostrea gigas*), an alignment containing all known sequences for the specific gene was used for the final decision.

Scaffold organization

Hox genes are commonly arranged in clusters and organized according to their expression. As this has proven to not be the case for all of the molluscans and in particular for cephalopods, a synteny analysis was performed. The scaffolds where the genes were found were searched for other genes, using the BLAST algorithm. When genes were split across several scaffolds, all scaffolds were analyzed, particularly favoring the surrounding region to the coding region.

Sequence alignment and phylogenetic analysis

All of the Hox genes were individually aligned by their full sequence in order to proceed with species phylogenetic analysis. All alignments were performed using seaview (Gouy et al. 2010). Due to incomplete gene sequences and limited sample size, the alignment was restricted to the 180 nucleotides that make up the homeobox. Concatenated alignments were constructed from these smaller sequences as well, due to absence of some genes in different species and to improve phylogenetic signal. Sequence saturation has been shown to be capable of interfering with correct phylogenetic assessment and was determined using DAMBE (Xia 2013). Substantial saturation was found, although not compromising the analyses. In particular, the third codon position was highly saturated. Multiple alignments were constructed, removing species with a high number of dubious sequences, and excluding the third codon position or masking it. Alignments for gene relationships (and definite confirmation of gene identification) were based on the homeobox, fully aligning all genes from all species. Genes from distant species were removed.

Three different methods were used in phylogeny determination: Bayesian inference using the software MrBayes (Ronquist et al. 2012), Maximum Likelihood using iqtree (Nguyen et al. 2014), and the Neighbor-Joining method in seaview. Convergence in Bayesian trees was determined by observing molecular parameters sampling, likelihood score stabilization, random tree sampling equality and external software AWTY (Nylander et al. 2008). Maximum Likelihood and Neighbor-Joining trees were run with 1000 bootstraps. The most adequate evolutionary model for the data in analysis was determined in jmodeltest (Posada 2008) (for nucleotide sequences) and prottest (Abascal et al. 2005) (for peptide sequences).

Each approach is based on different models and assumptions, each one appropriate for certain datasets. As it is difficult to determine the most efficient, all approaches were considered. Bayesian inference is particularly useful as it is a robust method, based on adding a priori probabilities on parameters and trees, following Bayes' theorem, without relying on asymptotic approximation, capable of estimating any functions of parameters directly, while still obeying the likelihood principle. The results are easily interpretable and can be implemented through Markov chain Monte Carlo, sampling and tracking most parametric models. There is, however, no indication of how to select priors and the effect these will have – important as posterior probabilities may be heavily influenced by them. Convergence is also difficult to determine, and most methods search for divergence signs, rather than convergence. It is also usually computationally intensive (Barton et al. 2007).

Maximum Likelihood was another used method, as it provides a consistent approach to parameter estimation. The methods usually have adequate optimality properties and are minimum variance unbiased estimators as sample pool increases. The approximate normal distributions and variances can be used to obtain confidence bounds and hypothesis tests for the parameters. The methods can, however, be heavily biased for small samples. The likelihood equation must be determined for each distribution and estimation, and can be sensitive to starting values. This method can also be quite heavy in terms of computation (Barton et al. 2007).

Lastly, Neighbor-Joining was used. This is a faster method, model based, with clear assumptions, capable of processing high amounts of data, but heavily dependent on the input matrix and with severe difficulty in using sequence information from one distance calculation to another.

Homeodomain patterns determination

A previous study showed that cephalopods had an unexpected variation in the homeodomain motif. We have extended such conclusions using the most recent information available and our own data. Homeodomain alignment of the different species was used to verify specific amino acids modifications between the different mollusc species. Attention was given to amino acid changes that involved modification of the amino acidic properties and its potential implications in Hox proteins structure and function.

Venom genes

Gene identification and extraction

Recently published novel coleoid venom components carboxypeptidase and metalloprotease GON-domain sequences (Ruder et al. 2013) were bait in using the BLAST algorithm against the giant squid genome and the nine cephalopod transcriptomes. Recently published *Octopus bimaculoides* (Albertin et al. 2015) genome was also searched. Identification was positive in cases where homology with the bait sequences was high and conserved regions were kept.

Sequence alignment and phylogenetic analysis

All retrieved sequences were aligned according to the alignment used in the previous study (Ruder et al. 2013). Partial sequences were excluded from further analysis. Sequences with high similarity and lacking conserved regions were noted and removed. These proteins may be different venomous variations or modified enzymes adapted to organisms' environment.

All trees were run using amino acid sequences, given the alignment length permitted it and saturation influence in phylogenetic resolution. Both Maximum Likelihood and Neighbor-Joining trees were created, using 1000 bootstraps, best determined model by prottest and all other parameters defaulted. A non-venomous homolog of the gene was used as an outgroup. ML trees were run using iqtree. Neighbor-Joining was performed using seaview.

Selection analyses

The HyPhy (Pond et al. 2005) package was chosen to run selective pressure analysis and to estimate the ratio between non-synonymous to synonymous nucleotide substitutions rates (dN/dS). In particular, site by site analyses were ran using: 1) SLAC (Single likelihood ancestor counting), a heavily modified and improved derivative of the Suzuki–Gojobori counting approach; 2) FEL (Fixed effects likelihood), a method that estimates directly non-synonymous and synonymous substitution rates at each site; 3) REL (Random effects likelihood), a method which models variation of non-synonymous and synonymous rates across all sites into classes, with the selection pressure at an individual site inferred using an empirical Bayes approach (Kosakovsky Pond and Frost 2005) and 4) FUBAR (Fast Unconstrained Bayesian AppRoximation), a method that can detect sites that are evolving under the influence of pervasive diversifying and purifying selection faster than the other methods and allows us to visualize Bayesian inference for each site (Dutertre et al. 2014). Given that previous methods may fail in identifying positive selection when the lineages in a phylogenetic tree mostly follow the regime of negative selection, hiding positive selection signal, MEME (Mixed Effects Model of Evolution) was also used. MEME can detect episodic diversifying selection, because it allows omega to vary from site to site and from branch to branch at a site (Ruder et al. 2013).

For a dN/dS ratio higher than one, positive selection is identified. Purifying selection if the ratio is lower than 1 and neutral selection should the ratio equal 1.

Due to the effects of recombination masking true positive detection of sites undergoing positive selection, Single Break Point and Genetic Algorithm Recombination Detection algorithms available in the HyPhy package were used. These can detect and correctly compartmentalize any recombinant sequences, allowing for more accurate determination of positive selection sites.

Structural analysis

In order to understand the selection pressures influencing the evolution of venom components, we used Phyre 2 webserver (Kelley et al. 2015) to acquire homology models and putative protein structure. Consurf webserver (Celniker et al. 2013) was used for mapping the evolutionary selection pressures on the three-dimensional homology models. All models visualization was performed with Visual Molecular Dynamics (Humphrey 1996).

Chapter 3 – Hox genes results and discussion

Development organization of the diverse animal body plans is controlled by a conserved cluster of homeotic genes, which includes among others the Hox genes. As it has been suggested, the recruitment of regulatory genes or regulation networks, rather than novel genes, is a potential explanation for the divergence of body plans (Callaerts et al. 2002). Although not all patterns are fully adjustable to this model, such divergence may be related with the Hox genes organization and mutation patterns.

Hox genes are a group of related genes that define the anterior-posterior axis and segment identity of metazoan organisms during early embryonic development, determining the number and placement of segment structures. Hox genes are characterized by a DNA sequence referred to as homeobox, comprising 180 nucleotides that code a protein domain known as homeodomain. The homeodomain motif is highly conserved across vast evolutionary distances, as it is found mostly unchanged throughout Bilaterians (Suman Pratihar 2010). The fact that a fly has no changes to its protein functions when its Hox protein is replaced by that of a chicken (Lutz et al. 1996), despite the last common ancestor between both species being over 670 million years ago, shows that a Hox protein in chicken and their homologous protein in fly are so similar that they may be replaced by each other. Although the protein sequence is highly conserved, the same is not true for the DNA sequence, potentially due to codon degeneracy. This level of conservation is likely related to the protein function, controlling eyes and limbs formation in the adequate body area, as even single point mutations in these genes could strongly affect the organism fitness. Hox genes encode transcription factors, molecules that can bind to DNA, regulating the gene expression (Suman Pratihar 2010).

The homeodomain 60 amino acid sequence represents a helix-turn-helix protein structure (Figure 3) and act as a switch for gene transcription by binding to enhancers of a gene, activating or repressing it. The same protein can act as a repressor and an activator. In *Drosophila melanogaster*, Antennapedia activates genes involved in the structures of the second thoracic segment and represses genes involved in eye and antenna formation. The anterior-posterior axis specification and segment identity is functionally conserved across Bilaterians and during vertebrate evolution this gene group has been recruited to perform other functions, involving duplication events, which are relevant further along the organisms' development (Suman Pratihar 2010).

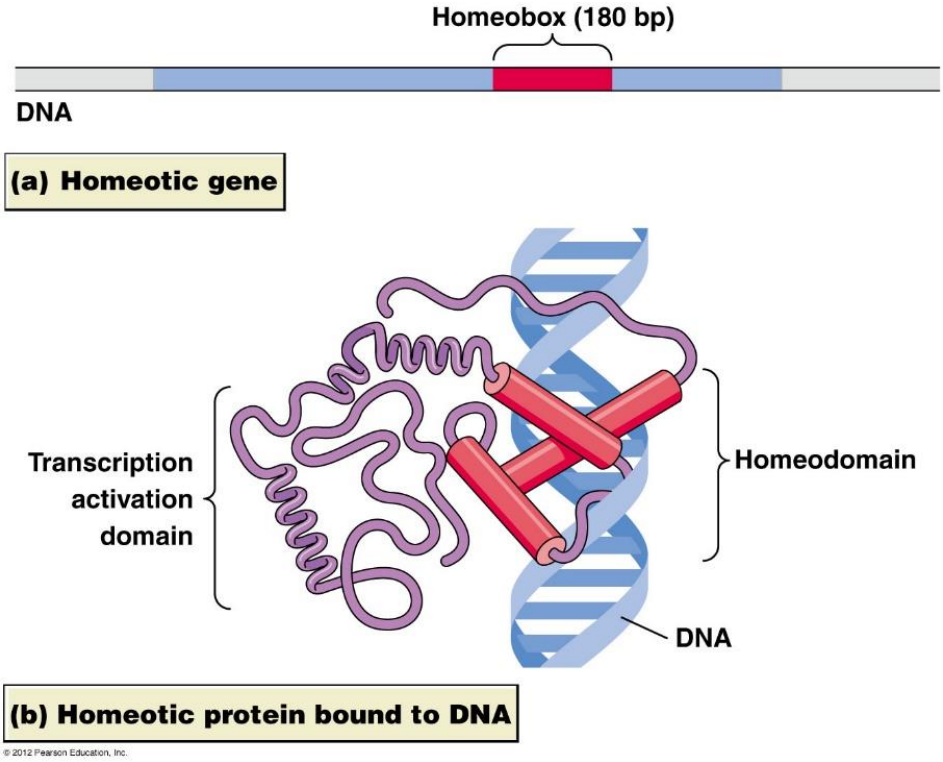


Fig.3 – Representation of a Hox protein, with homeobox and homeodomain stressed out in red (retrieved from Pearson Education, Inc.).

The Hox genes exhibit an unusual property: they are usually clustered together in the genome. Within the cluster, there is usually another impressive demonstration of order in that they are expressed in the same sequence as they are ordered in the DNA (colinearity) (Fröbius et al. 2008). There is also the fact that different combinations of Hox proteins located in the same tissue will lead to different structure identity (Mallo et al. 2010). This makes Hox genes a key target to understand the vast number of forms found in cephalopods.

A number of Hox gene expression studies have been published for various Bilaterians clades, but data on Lophotrochozoans outside the Annelida is scarce. To date, 11 Hox genes have been identified in most Molluscan classes, but expression studies are restricted to two groups, these being Gastropoda (snails and slugs) and Cephalopoda (squids and octopuses). The expression data are limited to the gastropods *Haliotis asinina*, *Haliotis rufescens* and *Gibbula varia*, and to developing embryos of the cephalopods *Euprymna scolopes* and *Sepia officinalis* (Fritsch et al. 2015). The recent *Octopus bimaculoides* genome data also brought some insight on the expression of these genes. The data showed that the Hox genes are involved in the formation of specific morphological features such as the shell, the ganglionic nervous system, the

tentacles and the funnel. Most important, the Hox genes were found to be expressed in a non-co-linear pattern and are not fully organized into clusters as in other studied animals (Albertin et al. 2015). They were also found in sensory organs (e.g. in the apical organ and the statocyst in gastropod larvae, as well as in the light organ of squid embryo). These studies suggest that Hox genes may have been co-opted into the formation of novel Molluscan morphological features (Fritsch et al. 2015).

We identified the Hox genes in the giant squid's genome and nine other cephalopod species with their transcriptome. We also analyzed genomic organization of the genes and compared the cluster structure to other Lophotrochozoans and suggested potential consequences of any differences found. The obtained genes were used in phylogenetic analyses to better understand the cephalopod evolutionary history. The Hox genes DNA binding site motif was analyzed to assess conservation within Mollusca and homeodomain variability. Variations found were verified for the possibility that the multitude of body plans between the different species in Mollusca were associated.

Recovered genes and scaffold organization

Molluscan Hox gene sequences were recovered from NCBI and Ensembl databases and were used to perform BLAST searches against the assembled giant squid genome and nine other cephalopod transcriptomes. Genes were identified according to their alignment to known sequences. The giant squid sequences for Hox 1, ANTP and Post 2 were not considered for further analyses. These genes were identified by *Euprymna scolopes* sequences but no homeodomain was found. The high similarity to the known sequences suggest a correct identification. The reason explaining why no homeodomain was found may be related to assembly difficulties, sequencing deficiencies or pseudogenization.

Anterior group orthologs were recovered for *Architeuthis dux*, *Dosidicus gigas* and *Doryteuthis pealeii*. *Architeuthis dux* sequence was identified by homology to a different region of the gene found in *Euprymna scolopes*, which did not include the homeodomain. Similarity was significant and identification should be accurate, as the bait sequence was recovered from an expression study. No Hox 2 orthologs were found in any of the datasets available. The recently published octopus genome also did not present this gene. Hox 2 has yet to be identified in Coleoids.

Paralog Group 3 orthologs were found in *Architeuthis dux*, *Dosidicus gigas*, *Onychoteuthis banksii* and *Doryteuthis pealeii*, thus complementing previously identified

Hox 3 in *Euprymna scolopes* and *Sepia officinalis*. Curiously, no Hox 3 was identified in *Sepia esculenta*. However, transcriptomes may not be able to recover all existing genes due to expression dependency. Although no Hox 3 gene was detected in *Octopus bimaculoides*, we successfully identified the gene in a cirroctopod dumb octopus (*Grimpoteuthis* sp.).

Central group orthologs for Hox 4 were only found in the *Architeuthis dux*; for Hox 5, *Architeuthis dux*, *Dosidicus gigas*, *Sepia esculenta* and *Doryteuthis pealeii*; for Lox 2, orthologs were determined for *Sepia esculenta*, *Sepioteuthis lessoniana* and *Doryteuthis pealeii*; for Lox 4, *Architeuthis dux*, *Sepia esculenta* and *Doryteuthis pealeii*; for ANTP, *Architeuthis dux*, not identified by homeodomain, *Dosidicus gigas* and *Sepia esculenta*; for Lox 5, *Architeuthis dux*, *Dosidicus gigas*, *Lolliguncula brevis*, *Sepia esculenta*, *Sepioteuthis lessoniana*, *Grimpoteuthis* sp. and *Pareledone albimaculata*. Notably, likely due to the chosen tissue samples and current growth stage, all of the central group genes were found for *Sepia esculenta*, with the exception of Hox 4. Given that the dataset refers to transcripts, gene loss cannot be unambiguously confirmed. Further research into Hox gene regulation and protein interaction on cephalopods should be able to provide evidence to support if the gene may be present, but is not expressed. It is conceivable that a certain combination of other Hox genes could at a certain development stage repress Hox 4. However, there is currently not enough support to fully understand whether the results are due to the transcriptome nature or molecular interaction and cellular regulation.

Posterior group orthologs were not determined for *Lolliguncula brevis*, *Onychoteuthis banksii*, *Sepia esculenta* and *Grimpoteuthis* sp. Post 1 was found for *Sepioteuthis lessoniana* and *Bathypolypus arcticus*, and Post 2 for *Doryteuthis pealeii*. All other species had both orthologs identified. *Architeuthis dux* Post 2 gene had no homeodomain found.

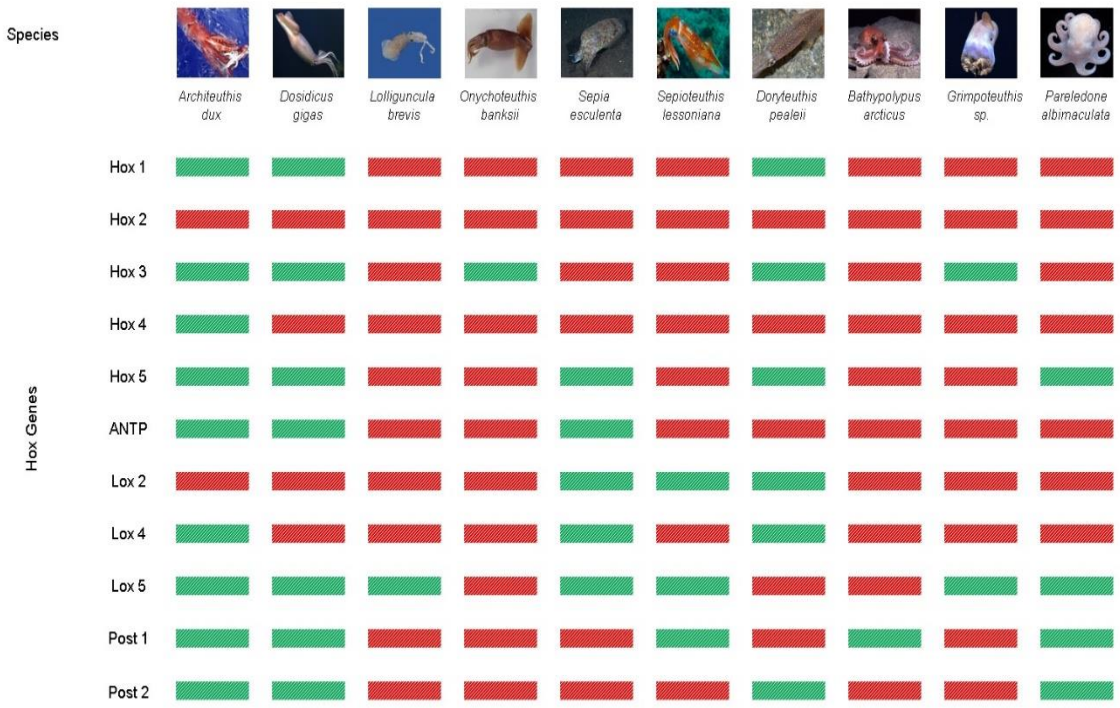


Fig.4 – Representation of the Hox gene distribution in the studied cephalopod species. Green indicates presence of the gene. Red indicates lack of evidence for the presence of the gene.

The gene distribution (Figure 4) reveals that the Hox 2 gene (homolog of proboscipedia) is absent in all of the sequenced cephalopods analyzed in this study. This is in accordance with previous works, whereas no Hox 2 gene or a putative homolog was found in the cephalopods *Eurpymna scolopes*, *Nautilus macromphalus* and *Sepia officinalis*. With the exception of *Nautilus pompilius*, no other cephalopod has the Hox 2 gene reported. Current phylogeny considers nautiloids to be the most basal group of Cephalopoda (Almeida et al. 2015).

Hox 2 (*proboscipedia*, pb) is a unique case among the Hox genes in that it does not confer segment identity but modifies segment structures. It has been mostly associated with head development and oral components (mouthparts in insects *Drosophila melanogaster*, *Tribolium castaneum* and *Oncopeltus fasciatus* (Rusch and Kaufman 2000); pharyngeal arches in *Danio rerio* (Hunter and Prince 2002) and hindbrain formation in the early embryo of *Mus musculus* (Monica Lynn Boyle)) and is believed to have interaction with *Deformed* (Dfd/ Hox4) and *Sex combs reduced* (Scr/ Hox5) (Rusch and Kaufman 2000). In insects, both Scr and pb must be expressed to confer proboscis identity (DeCamillis et al. 2001). The lack of either protein or both proteins will result in different structures. They were expressed in the nervous system of *Haliotis asinina* and *Gibbula varia* and thus involved in neurogenesis processes, in

particular in the pedal ganglia (foot region) (Hinman et al. 2003, Samadi and Steiner 2010).

Even though only *Octopus bimaculoides* and *Architeuthis dux* have genomic information available and there is a possibility of gene expression regulation leading to transcriptome absence, the addition of nine species with identified Hox genes within the Coleoid subclass strengths the hypothesis of the loss of the Hox 2 gene in parallel with morphological novelties. More importantly, it suggests that Hox genes have a unique pattern in this group. Given that these transcription factors are known to relate with morphological characteristics, any changes within highly important conserved regions (such as DNA-binding motif, homeodomain, which has also been implied in protein-protein interaction) may result in different body structures.

Our results and the genome information made available of *O. bimaculoides* (Albertin et al. 2015) suggest that the Hox 2 gene was likely lost in most cephalopods, which may be related to the morphological disparity found between Nautiloids (with a true external shell) and the more recent cephalopods. Given the known distribution of Hox 2 expression in other molluscs and its function in other invertebrates, it is possible that Hox 2 may be involved in the differences found in the "arms", the muscular head appendages of Nautiloids and Coleoid. Nautiloids tentacles lack suckers and can be retractable to a certain degree, as opposed to the Coleoid, which all have tentacles with suckers. The number of appendages is also different, with Coleoid presenting between eight and 10, while Nautiloids can reach around 90 appendages (Toll and Binger 1991). Moreover, not only the appendages associated to the brachial crown are related to the foot (anterior foot) but the funnel as well (posterior foot) (Toll and Binger 1991). Given its role in insects, Hox 2 can also be responsible for the correct formation of mouthparts. Knowing that the Hox genes can regulate other Hox genes and interact with other proteins, correct segment identity might be the result of a specific combination. Further functional studies are necessary to determine the potential of currently known cephalopod Hox genes to interact with a protein possessing similar properties to known Hox 2 proteins.

Hox 2 is known to interact with other cell regulating signals, as it appears to negatively regulate *dachshund* expression directly and withholds positive instructions from *wingless* and *decapentaplegic* morphogens, influencing cell identity. The exact processes that Hox 2 mediates in cephalopods is currently unknown and is essential to fully understand its impact in Nautiloids and the translated differences in Coleoids. The latter do not seem to have the Hox 2 gene, suggesting that certain mechanisms were

interrupted or replaced. As in insects, the lack of Hox 2 expression could still lead to the formation of particular structures depending on other expressed proteins (Joulia et al. 2005).

It is also possible that expression regulation prevented Hox 2 from being detected in the transcriptomes. The sequencing may also have not been exhaustive enough, explaining why Hox 2 has not been found in the giant squid's genome as well. The possibility of the gene being located in a region that is not easily amplified and sequenced cannot be discarded.

We have also found a Hox 3 gene in an octopus species, the first observation of this gene in the Octopodiforme group and it was detected in a cirroctopod. Little is known regarding Hox 3 function in cephalopods, although it is involved in the pedal ganglion of *Euprymna scolopes*, *Sepia officinalis* and in the abalone *Haliotis asinine* (Hinman et al. 2003, Lee et al. 2003, Fritsch et al. 2015). Hox 3 is also expressed in a small number of cells in the inferior frontal lobe of supraesophageal mass in the cuttlefish. There are also results that suggest the Hox 3 protein as part of molecular signature involved in the functional partitioning of the cephalopod brain rather than positional axial information in the cuttlefish (Focareta et al. 2014, Fritsch et al. 2015). It is not known if Hox 3 has the same or similar function in other cephalopods, although the high conservation and closeness of the species would suggest a comparable role. More functional studies are required to fully understand the Hox 3 gene role and the implications of its modification or loss.

Lox 2, previously undetected in *Euprymna scolopes* and difficult to detect in *Sepia officinalis*, has now been found in four other species (Callaerts et al. 2002, Lee et al. 2003, Focareta et al. 2014). However, this gene has been undetected in the giant squid genome and in many other transcriptomes. It is possible that the gene was not found due to limitations of expression studies. These results show that there are particular patterns in cephalopods and more studies are required to fully understand them. The clade is notorious for having some of the most evolved invertebrate marine organisms (Fiorito et al. 2014) and their unique body plan may correlate with the changes observed at these genes.

It is also noteworthy the possibility of functional duplication. Pseudogenization, assembly artifacts and inaccuracy of the coding region prediction cannot be discarded but our methodology allowed the detection of Lox 5 in different scaffolds. The Hox genes duplication has been suggested to be correlated with some of the vertebrate complexity. Thus, cephalopod complexity, in particular Decapodiforme’s body structure, may in part

be due to duplication of the Hox genes. This duplication is also consistent with a suspected chromosomal duplication during the evolution of the Decapodiforme (Hallinan and Lindberg 2011).

It is important to note the importance of duplication as an evolutionary event. Duplications have long been regarded as a very potent driving force in evolution. Duplications can lead to gene neofunctionalization, where the duplicated gene, not being under heavy selective pressure, accumulates enough mutations promoting new functionalities, often with loss of the previous one (Teshima and Innan 2008, Innan and Kondrashov 2010). This duplication event is one possible form of gene recruitment. Recruitment of genes into new functions is not an unusual phenomenon, and is particularly observed in the evolution of venoms. Many components found in venoms are modified counterparts of proteins that perform other functions in an organism and are required for its proper activity (Fry et al. 2009). Another example would be the Hox genes found in vertebrates. These genes were duplicated and perform different functions in the body development (Suman Pratihar 2010). Duplications can also lead to subfunctionalization, where both (or more) copies suffer alterations and they all become essential for the organism as the original function is now distributed between the two (Innan and Kondrashov 2010). The Hox 3 ortholog found in the insects is a good example. The Hox 3 gene was duplicated and can be found as bicoid and zerknullt, with its original function in specifying segment identity lost and acquiring a role in the morphogenesis and specification of the extraembryonic tissues (SF. 2000). The other possible occurrence is gene loss, where the variation becomes detrimental and eventually disappears from the population (Figure 5). Many duplications have no positive or negative effects, but are rather subjected to random modifications resulting from genetic drift (Innan and Kondrashov 2010).

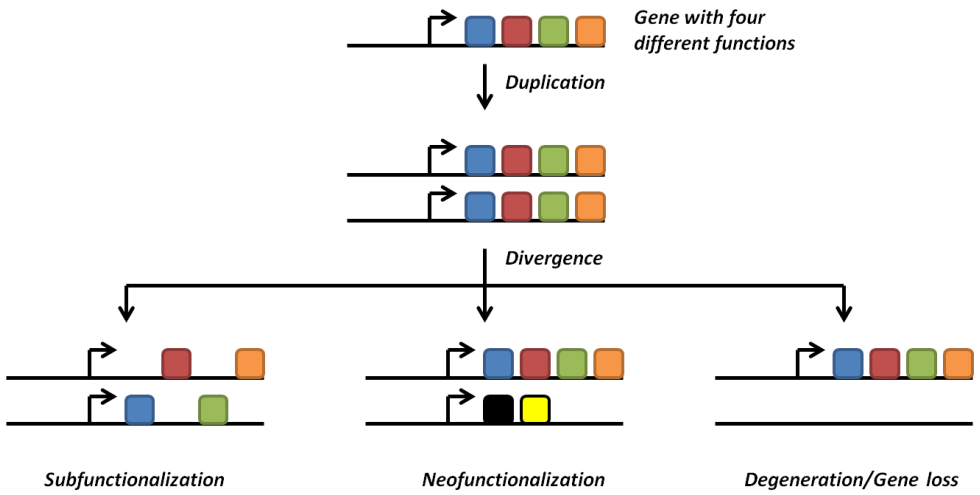


Fig.5 – Simplified diagram with an explanation of the three evolutionary fates of duplicate genes.

Considering the potential effects of duplications, it is important to determine whether or not the Hox genes were duplicated in cephalopods. Their organization in the genome can also provide evidence of gene dynamics and change of genomic information. A synteny analysis provides understanding of the current structure of the Hox gene cluster. Expression analysis would be required in order to assess tissue expression, development dynamics and function. Expression studies in *Euprymna scolopes* demonstrated that colinearity, commonly verified in vertebrate Hox genes, is not always verified (Lee et al. 2003). There is also strong evidence suggesting that molluscs should only have one Hox gene cluster or very small fragmentation (Biscotti et al. 2014). The most recent work in *Octopus bimaculoides* shows possibly that all of the genes have been atomized. The Hox genes were not found in a single cluster and were all separated and individualized (Albertin et al. 2015).

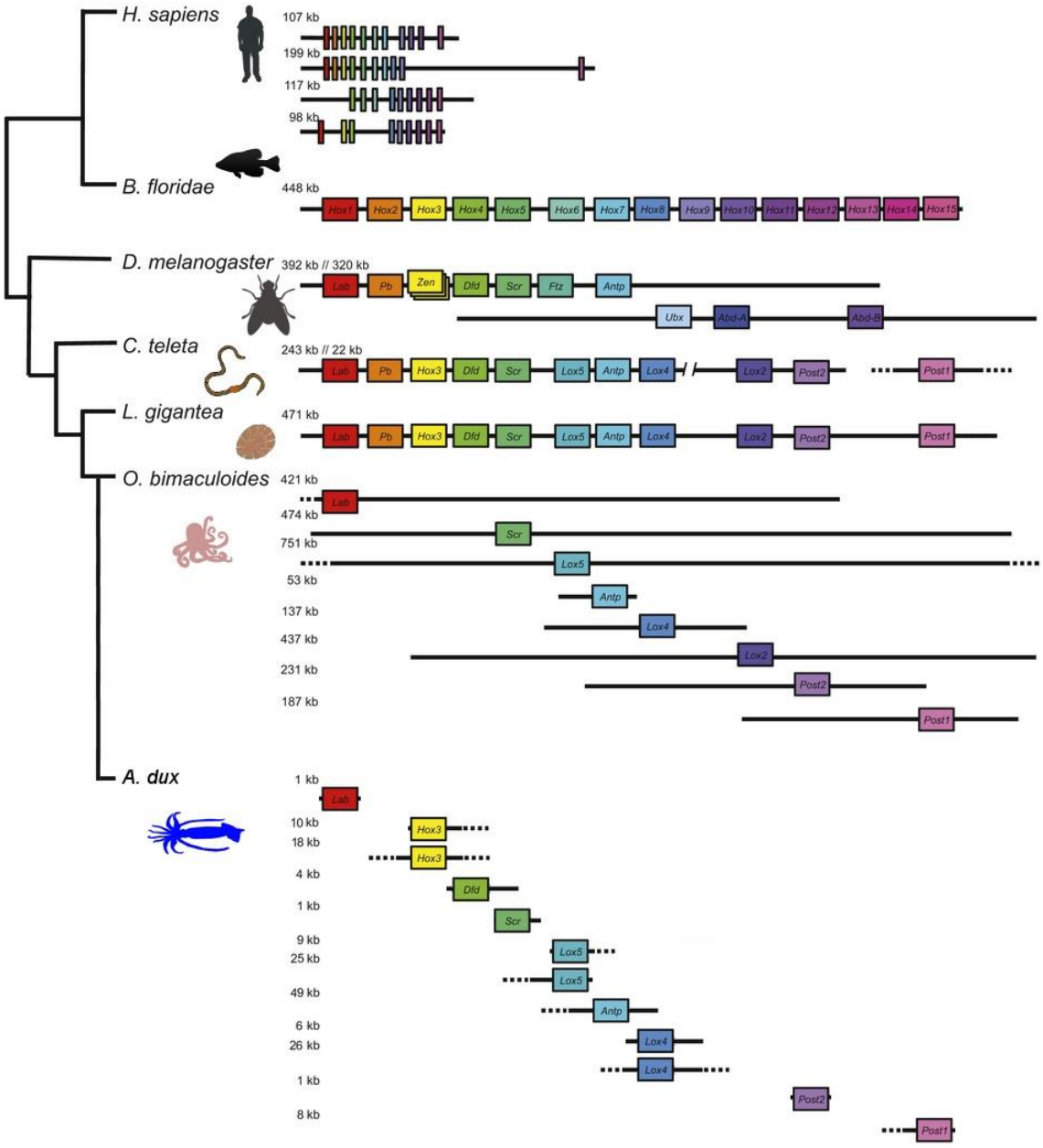


Fig.6 – Schematic representation of the Hox gene organization in varied groups, including the most recently sequenced cephalopods (adapted from (Albertin et al. 2015); icons provided by Daniela Almeida).

Our results place the genes in very small scaffolds when comparing to other organisms (100 to 400 Kbp against 5 to 20 Kbp) but still suggest atomization in this group. Although the interpretation should be done cautiously, the total scaffold size of all the Hox genes suggest they are not all located on the same cluster (Figure 6). The clusters may have been fragmented or rearranged. Duplication should also be considered, increasing the possibility of modifications influencing the Hox genes cluster.

Given previous knowledge on Hox genes and their regulatory mechanisms and functions, it is possible that these genomic shifts have altered Hox gene dynamics,

contributing to the sophisticated morphological organization and complexity that the clade now presents. It may also have interfered with interactions with other proteins and recruitment of those into the regulatory network, significantly changing the development program.

Species and Gene phylogeny

In order to reconstruct the evolutionary history of the Hox genes, a phylogenetic analysis was performed. All of the Hox genes were aligned according to the homeodomain and each alignment was verified to remove sequences that could reduce signal. Given that sequence saturation could preclude an accurate phylogeny, sequences had their third codon removed or masked.

The amino acid based trees for species were unresolved, as the algorithms could not determine a well-supported phylogeny, given the limited number of samples and the small sequence size (180 nucleotides, 60 amino acids when complete – many sequences in databases are incomplete). Most alternative alignments resulted in polytomies. Gene trees, however, were resolved, albeit still with poor support, by Maximum Likelihood methods and Neighbor-Joining. Nucleotide based gene trees were unsuccessfully resolved, with known gene sequences being paired incorrectly in nodes and had poor branch support. The nucleotide based species trees were resolved by Bayesian inference. Maximum Likelihood methods potentially failed due to the small amount of sequences and the effect of some saturation. Neighbor-Joining methods also provided results with poor support.

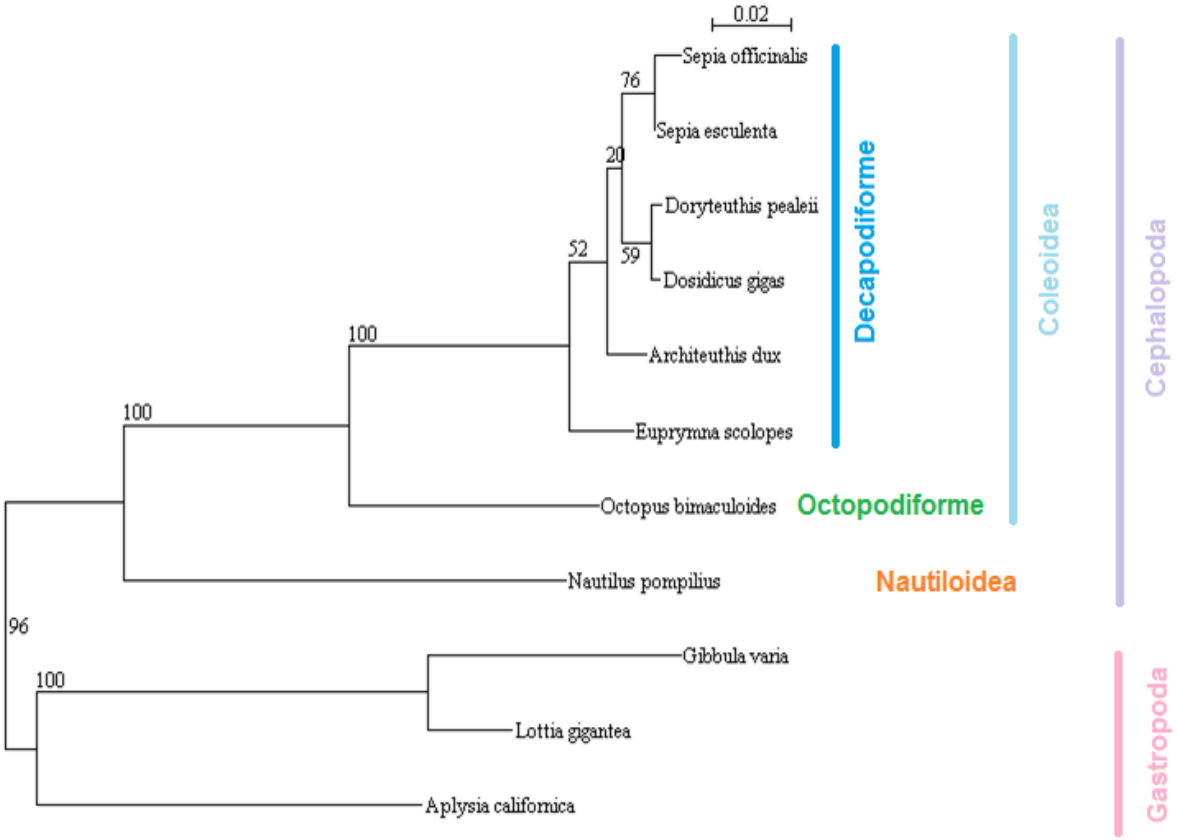


Fig.7 – Phylogeny of the Cephalopods based on the Hox genes (1,980 bp corresponding to 11 concateneted Hox genes). ML trees were constructed based on the best determined model and 1000 bootstraps. Percentual bootstraps are shown. Gastropods were used as outgroup.

Mollusca phylogenetics is currently not fully understood and there is some disagreement regarding its members' classifications. Currently, Mollusca is divided into seven or eight classes, depending on the author and the classification criteria used (Nielsen 2012, Biscotti et al. 2014). Our results show that for species trees, a larger number of genes is required to resolve evolutionary history, particularly in groups with complex relationships, such as molluscs. A mostly untreated alignment, with no codon modification, was capable of accurately separating cephalopods from other molluscs, although the distinction between Bivalvia, Gastropoda, Scaphopoda and Polyplacophora is not clear. Removing most of the incomplete sequences reduce information and groups become undistinguishable. Multiple alignments were made to reach a balance point and simpler trees were acquired. Although cephalopod history is slightly less informative, these trees distinguish most groups accurately and with good support. Interestingly, Maximum Likelihood methods were capable of reaching better supported trees, as well as with the more curated dataset, despite the significant loss of information (Figure 7).

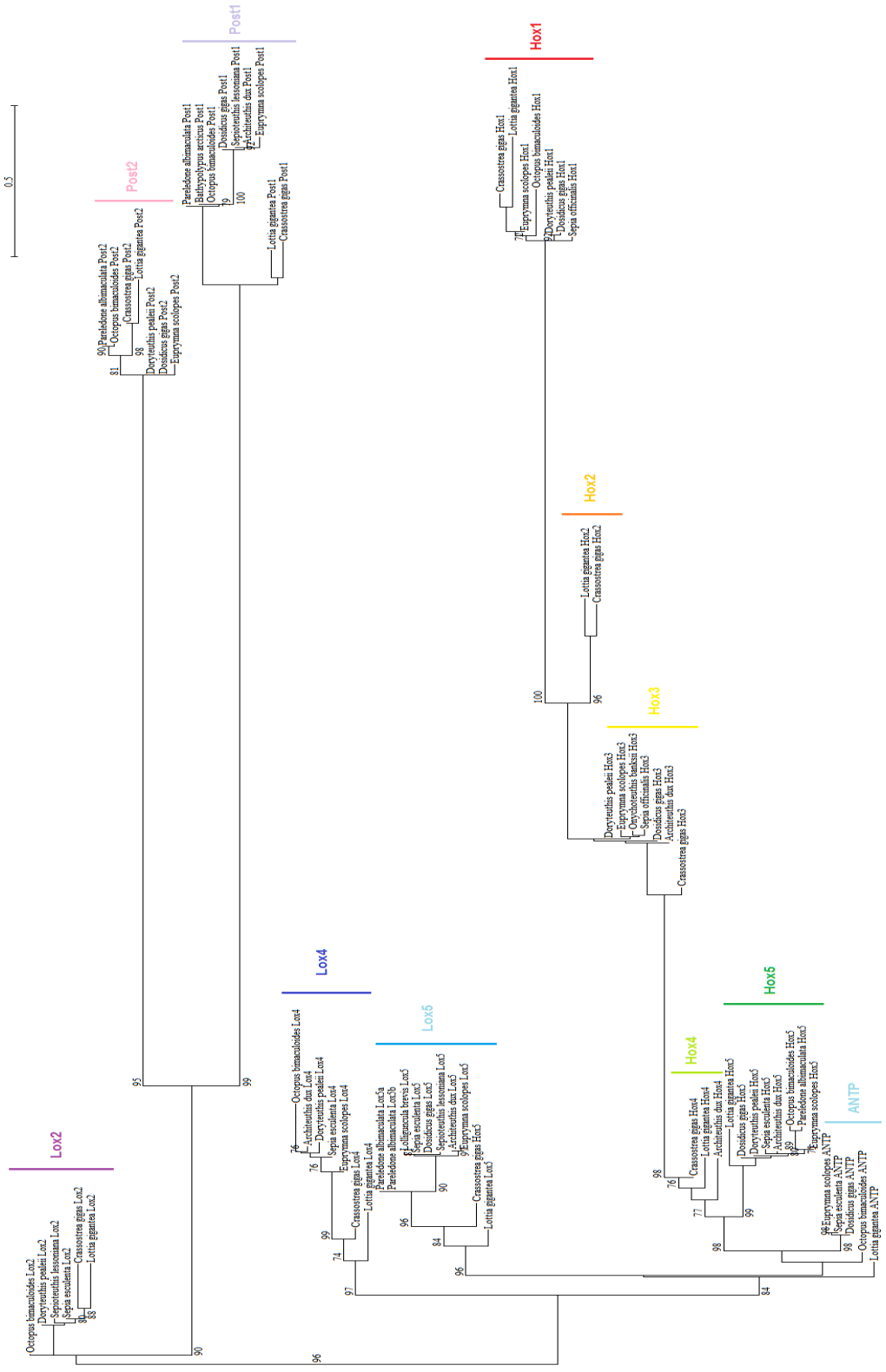


Fig.8 – Phylogeny of the Hox genes. Alignments are 60 amino acids long. ML trees were constructed based on the best determined model and with 1000 bootstraps. Percentual bootstraps are shown.

The gene trees mostly support an accurate identification and categorization of the newly sequenced data when compared to annotated sequences (Figure 8). Highly

divergent groups were excluded. Gene trees separated the posterior group from the others, kept Hox 3 and Hox 2 closely related, as in previous studies, with Hox 1 diverging from them. The central group was observed mostly together.

The species trees show some variation to other phylogenetic studies, but still within acceptable limits due to the nature of the sequences and the reduced number of genes involved.

Homeodomain site-specific variation

The homeodomain of all the obtained Hox genes in cephalopods was compared to those of other molluscs. The inspection of the homeodomain alignment revealed few alterations that could potentially influence the protein structure and interaction. Previously, *Drosophila melanogaster* Hox 3 homolog, bicoid was shown to have slight structural variation in accordance with specific amino acid changes when compared to other homeodomain proteins. In particular, helix flexibility and proximity alterations due to mutations in the amino acids involved in the turns between helices. Functional implications are not known (Figure 9) (Baird-Titus et al. 2006). It is known that the homeodomain is responsible not only for the protein-DNA interaction, but for protein-protein interaction as well. In particular, homeodomain-homeodomain and homeodomain-paired domains interactions have been demonstrated (Ohneda et al. 2000). Therefore, specific amino acid substitutions may imply the gain or loss of a particular interaction and result in morphological modification.

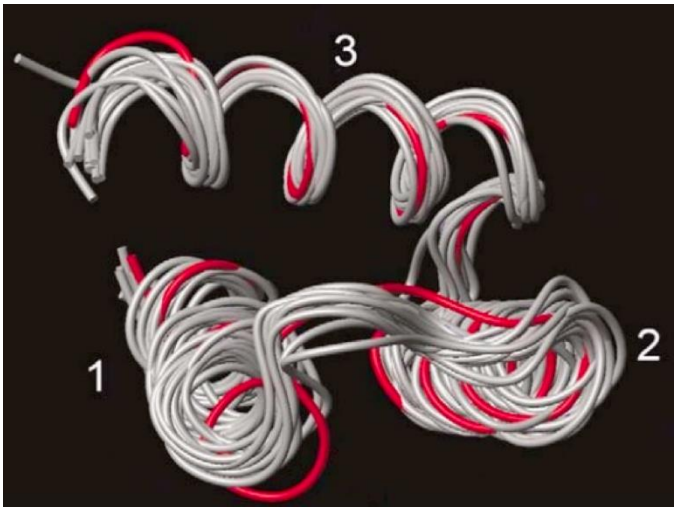


Fig.9 – Modeled homeodomain of Drosophila superimposed over other homeodomains. Amino acid variation leads to structure variation. Adapted from (Baird-Titus et al. 2006).

A previous study aimed to assess this variation and its implications. We reviewed previous data (Pernice et al. 2006) in light of our own results. No significant modifications were observed in Hox 1, Hox 4, ANTP, Lox 2 and Post 2.

In Hox 3, a specific Coleoid signature previously identified (Pernice et al. 2006) is maintained at position 24 of the homeodomain, reinforcing the basal placement of Nautiloids and that Hox may be associated with differences between the two groups. If Hox 2 was effectively lost, then the accumulation of differences in the Hox genes may be able to partly justify the evolutionary road from Nautiloids to Coleoids. This specific differentiation found in Hox 3 (and other genes described below) is in tandem with most recent phylogenetic analyses and our results that place the *Nautilus* genus in a basal position within the cephalopods.

Hox 5 sites 22 and 24, tyrosine and lysine, respectively, are substituted by phenylalanine and arginine. These two sites were previously considered to be specific signals allowing to distinguish molluscs from other phyla (Pernice et al. 2006). However, the phenylalanine modification was found in a significant number of molluscs. Thus, it is no longer possible to use these sites to accurately differentiate Mollusca. Even more, a combination of both sites is also not a viable possibility, as *Crassostrea gigas* was found to possess both modifications.

In the Lox 5 gene, at position 33, aspartic acid is found in Coleoids, whereas Nautiloids and other Mollusca have glutamic acid. Although the amino acids share the same property, being negatively charged, it is a very specific change present in all of the currently studied molluscs. Position 22 presents a phenylalanine in all cephalopods, while in some other molluscs and other invertebrates is replaced by either tyrosine or serine. At position 41, molluscs, with the sole exception of cephalopods and of the gastropod, *Aplysia californica*, have threonine. Both cephalopods and *Aplysia californica* possess serine. The site can therefore be used to separate cephalopods from other groups other than gastropods. These modifications are less likely to induce homeodomain interaction changes due to the amino acids' nature, but the combination of both sites can distinguish cephalopods from other molluscs, which may prove useful in further studies.

The Lox 4 gene was the most variable among groups. Four sites were found with distinct amino acids conserved across groups, with two having amino acid properties changed. Position 39 has a highly conserved cysteine that in Coleoids has been replaced by an asparagine. Position 24 has asparagine in Coleoids and histidine or arginine in the other groups. These two sites may be responsible for substantial differentiation in Hox

gene dynamic and interaction. These four sites also alienate Nautiloids from other cephalopods, further reinforcing their basal position within Cephalopoda. At position 22, a phenylalanine is replaced by a tyrosine in Coleoids. At position 41, a serine is found in Coleoids, while a threonine exists in other invertebrates. Position 45 is also Coleoid specific, having a valine instead of an isoleucine. The Coleoid specific trend reinforces the idea of a unique evolutionary pattern of the Hox genes in cephalopods.

Post 1 presents a highly conserved leucine in Cephalopoda at position 38, while other molluscs have other amino acids with hydrophobic side chains and similar properties. The Cephalopoda class is effectively separated (previously suggested as a signature site capable of distinguishing Mollusca from other phyla) (Pernice et al. 2006).

Chapter 4 – Venom genes results and discussion

Carboxypeptidase

Natural evolution has generated a large variety of forms among protein functional families, and metallocarboxypeptidases have also followed this trend. Metallocarboxypeptidases are an important class of enzymes that catalyze the hydrolysis of peptide bonds at the C-terminus of peptides and proteins. Their action causes strong effects in the biological activity of their peptide and protein substrates (Alonso-del-Rivero et al. 2009, Alonso-del-Rivero et al. 2012).

The subfamily that includes the digestive enzymes CPA1, CPA2, and CPB1 is referred to as the "A/B" subfamily. All members of the A/B subfamily contain a 90 amino acid-long N-terminal pro-peptide region that functions as a chaperone in the folding of the active carboxypeptidase domain. This domain also functions as an inhibitor of the enzyme, with full activity requiring cleavage of the pro domain in one or more places. Cleavage releases the active domain, approximately 300 residues long (Alonso-del-Rivero et al. 2009, Alonso-del-Rivero et al. 2012).

Besides a role in digestive protein degradation, these enzymes are also key elements of selective proteolysis-regulated physiological processes, such as blood coagulation, inflammation, neuropeptide processing and attack and defensive strategies occurring between plant and insects, among others (Alonso-del-Rivero et al. 2009, Alonso-del-Rivero et al. 2012).

Overall, our knowledge about metallocarboxypeptidases in invertebrates is very limited, in part due to huge diversity of these organisms compared to the much larger number of characterized carboxypeptidases from vertebrates (Alonso-del-Rivero et al. 2009, Alonso-del-Rivero et al. 2012).

Recently, a metallocarboxypeptidase was characterized in tissue of a venom gland extracted from cephalopods (Ruder et al. 2013). This molecule was identified as a novel component in Coleoid venom, namely in *Octopus cyanea* and *Sepioteuthis australis*.

Metalloprotease GON-domain

Metalloproteases are the most diverse of the four main types of proteases, with more than 50 families identified to date. A divalent cation activates the water molecule, being held in place by amino acid ligands, usually three in number. The known metal ligands are histidine, glutamic acid, aspartic acid or lysine and at least one other residue is required for catalysis, which may play an electrophillic role (Apte 2009, Brunet et al. 2015).

The ADAMTSs (a disintegrin and metalloproteinase domain with thrombospondin type-1 modules) are a family of zinc dependent metalloproteinases that are involved in several normal and pathological conditions. These enzymes have a well-defined domain organization including signal sequence, pro-peptide, metalloproteinase, disintegrin-like domain, central TS-1 motif, cysteine-rich region, and a variable number of TS-like repeats at the C-terminal region. The GON domain is a 200-residue region, whose presence is the hallmark of a subfamily of structurally and evolutionarily related ADAMTSs, called GON- ADAMTSs. The GON domain is characterized by the presence of several conserved cysteine residues and is likely to be globular, although this has yet to be confirmed (Apte 2009, Brunet et al. 2015).

Research efforts in invertebrates to fully explore GON-domain metalloproteases are incredibly scarce and little is known about their role other than some functions played in *Caenorhabditis elegans* regarding morphogenesis control. Recently, a type of metalloprotease GON-domain was identified in venom glands' tissue of octopus, squids and cuttlefish (Albertin et al. 2015).

Here, we analyzed the molecular evolution of both carboxypeptidase and metalloprotease GON-domain genes. We characterized the molecules with additional Coleoid species recently sequenced. We interpreted the implications of these genes

existence in cephalopods and conduct further analysis to comprehend their evolutionary origin. We also determined the selective pressures acting on these genes and assess its implications. We also modeled the genes against known protein sequences to understand the relevance of changes in protein structure, its interactions and functions.

Sequences and phylogenetic analysis

Metalloprotease GON-domain was found in *Octopus bimaculoides*, *Onychoteuthis banksii*, *Doryteuthis pealeii* and *Architeuthis dux*. In the transcriptomes of the *Grimpoteuthis sp.* and *Dosidicus gigas* a match was also obtained. Despite high similarity, careful inspection would reveal that it was not the same gene. Preliminary tree reconstructions placed these genes as an outgroup, corroborating the previous interpretation. Considering the high similarity, these genes could be a modified version of the same protein with alternative functions. It could also be the same gene, but modified to retain the same function towards different targets. Two candidate sequences were found in the giant squid, both highly similar to the novel venom. Considering the phylogenetic analysis (Figure 10) did not place those genes in a basal position, or close to the outgroup, it can be assumed that *Architeuthis dux* has, at the very least, two genes highly similar to a putative venomous metalloprotease GON-domain gene. According to our phylogenetic analysis, these two genes are more closely related to each other than to any other sequence in the dataset, suggesting a duplication event. Thus, the possibility of neofunctionalization cannot be excluded.

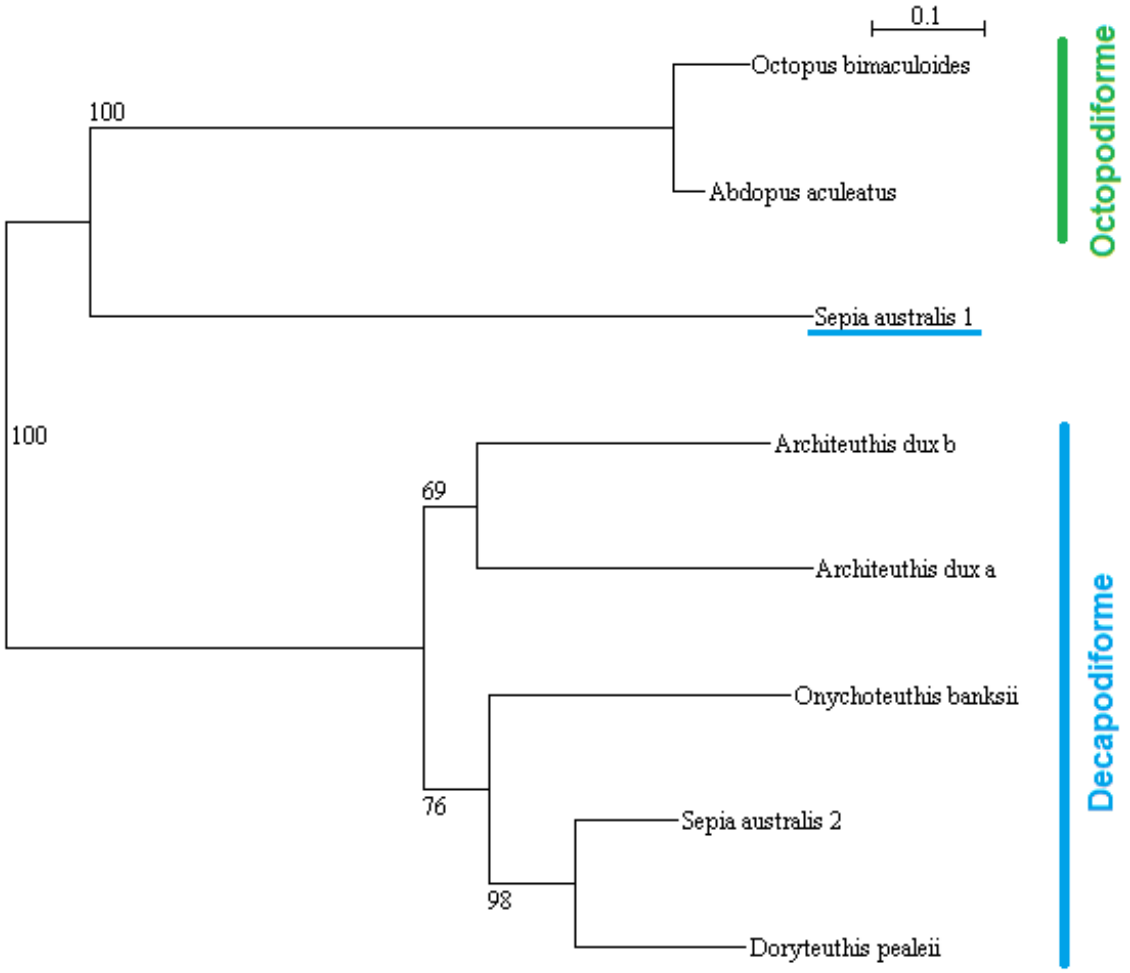


Fig.10 – Coleoid phylogeny based on metalloprotease GON-domain. Alignment is 530 amino acids long. ML trees were constructed based on the best determined model and with 1000 bootstraps. Percentual bootstraps are shown.

In the carboxypeptidase analysis (Figure 11), candidate genes were found in *Onychoteuthis banksii*, *Octopus bimaculoides*, *Doryteuthis pealeii*, *Pareledone albimaculata* and *Stenoteuthis oualanensis* (an additional transcriptome provided for this study). Partial genes were found in *Dosidicus gigas* and *Lolliguncula brevis*. Similar but different genes were found in *Pareledone albimaculata* and *Grimpoteuthis sp.*

The current findings strongly suggest that *Octopus bimaculoides*, *Onychoteuthis banksii* and *Doryteuthis pealeii* possess venomous genes. No venomous cocktail has been characterized and no distinct venom gland analysis has been reported for these species. It is possible that despite the high similarity, these genes do not possess the exact same function. However, many Coleoids that had not been documented as venomous were incorrectly categorized. In order to correctly characterize these organisms, sequencing of venomous genes should be performed in the future, as well as research focused on structure-function relationship of the currently known genes.

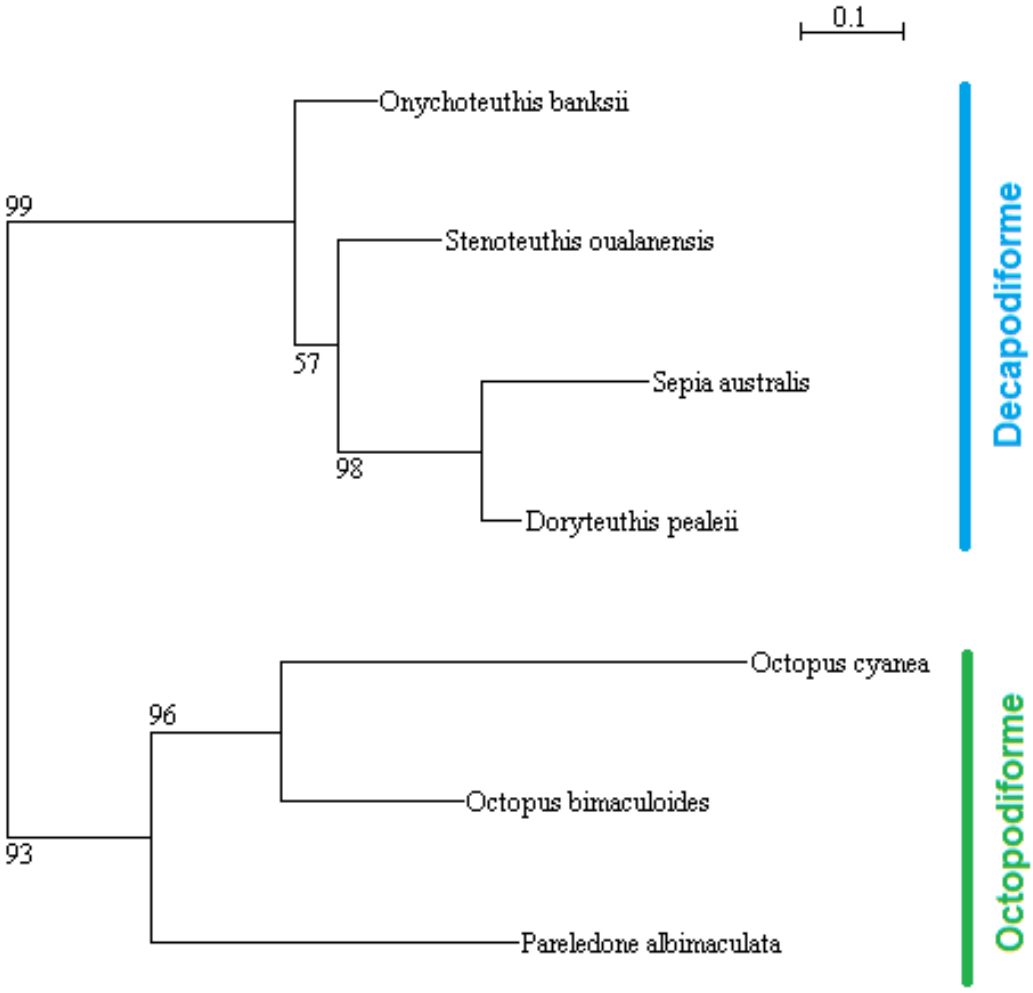


Fig.11 – Coleoid phylogeny based on carboxypeptidase. Alignment is 430 amino acids long. ML trees were constructed based on the best determined model and with 1000 bootstraps. Percentual bootstraps are shown.

Two other cases of interest include *Dosidicus gigas* and *Grimpoteuthis sp*. The cirroctopod does not seem to possess any of these venomous genes, although he has very similar ones. Further information would be needed in the future to properly identify these genes. *Dosidicus gigas* had one different gene detected in metalloprotease GON-domain analysis and a partial match to carboxypeptidase. Partial matches from transcripts suggest that there is expression of genes that code for a similar protein. The partial match does not allow to fully verify if *Dosidicus gigas* would be a new addition to the venomous Coleoids ranks. These two species should be further examined, as they suggest the presence of novel genes derived from modified venom genes.

The previous study on Coleoid venoms acquired a surprisingly large variety and number of putative venom components despite its small sampling. It is suggested that further sampling will provide novel isoforms of known venoms or entirely new classes (Ruder et al. 2013). Given the minimal information available in the area of venoms and

cephalopods, as well the recent results in sequencing efforts, the possibility cannot be understated.

The variant found in *Pareledone albimaculata* also suggests unexplored genes that may be helpful in further understanding cephalopod history, dietary habits and behavior. The existence of several species possessing genes associated with venom and somewhat different genes that may hold similar functions is incredible and promotes additional effort into understanding these highly evolved and complex animals. In this particular species, it could be related to its environment. The *Pareledone* genus, composed of Antarctic octopuses, has been shown to possess structural modifications in its proteins, which may correlate with temperature adaptation (Oellermann et al. 2015) – i.e. the different gene found may result from modifications required to allow function in arctic scenarios. In-depth molecular characterization is required to fully understand the consequences of such changes.

It is difficult to correctly classify *Architeuthis dux*, *Lolliguncula brevis* and *Stenoteuthis oualanensis*, given that only one venom gene was characterized, even if partially. However, the discovery of those sequences makes these species excellent candidates for new studies on toxicogenomics. The possibility of discovering more venom genes or closely related genes enforces the need of additional data to fully resolve the genes', and species', relationships.

It should be noted that these two genes are not an immediate definition of venomous animals. Venomous genes could have been recruited from other genes, found in the common ancestral of Coleoids. Therefore, it is possible that non-venomous species possess genes similar to venomous species. The genes could also have a different function, with each species having their own environment and unique selective pressures, leading to different adaptations of the same gene. Confirmation with search for further venomous genes is required. In particular, the major Coleoid toxins such as phospholipase A2, serine proteases, pacifastin and cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins – CAP.

The phylogenetic analysis showed two fully distinct groups. All trees correctly distinguished the outgrup from the dataset and the Octopodiformes from Decapodiformes. One metalloprotease GON-domain sequence from *Sepia australis*, seems to be more closely related to Octopodiforme sequences, as it was paired with the group. Phylogenetic analyses suggest that venom genes originated in the common ancestor of Coleoids. The pairing of one of the sequences of the venomous Decapodiforme with the Octopodiforme suggests that these genes have not

accumulated sufficient mutations, suggesting that their common ancestor had the venom gene. Bootstrap support indicates that the interpretation of this phylogeny should be done with care, as some values are under 70. This suggests ambiguity in the topology of some branches.

Selection analysis

Finding which positions are under selective pressure is important as it may provide evidence for toxin functions. Sites under strong positive selections were previously demonstrated to be important for selectivity and potency for some venom proteins (Kaas and Craik 2015). For that reason, four methods (FEL, REL and FUBAR) were used to estimate the ratio between nonsynonymous to synonymous nucleotide substitutions rates (w=dN/dS). This ratio is an indicator of selective pressure at the protein level, as higher than one indicates that the position is under positive selection, equal to 1 means neutral selection and lower than 1 means purifying selection.

Results show that most sites in metalloprotease GON-domain are under purifying selection and heavily conserved (Table 1). No indication of positive selection was found except in FEL and in FUBAR. FUBAR detected only one site coincident with one of the detections by FEL. It is possible that the identifications are false positives or the other methods could not detect diversifying selection. Practical statistical inference is never 100% accurate and the p-value should only be taken as a guideline, as the statistical properties of the test may vary from alignment to alignment.

Table 1 - Number of positive and negative selected sites detected in metalloprotease GON-domain using the HyPhy package. W – dN/dS ratio. SLAC, FEL and MEME results correspond to p-value ≤ 0.05. REL results correspond to 50 Bayes factor.

SLAC		FEL		REL		FUBAR		Integrative		MEME
w>1	w<1	w>1	w<1	w>1	w<1	w>1	w<1	w>1	w<1	
0	45	3	104	0	All	1	147	18	149	18

Previous studies suggested a strong influence of negative selection on the Coleoid toxins (Ruder et al. 2013) and that it may hide positive selection from the site-by-site methods. However, MEME should be capable of detecting positive selection even when hidden under purifying selection in the lineages (Ruder et al. 2013). This method supported three results obtained from FEL and the one common to FUBAR, codon 411.

Carboxypeptidase results show convincing evidence of positive selection on at least three sites, although a total of 19 others were identified (Table 2 and 3). These methods are known for working better with a high number of samples and may produce artifacts with a very low amount of sequences (Samira 2010). Therefore, only sites that were detected by multiple methods and statistically significant were considered.

Table 2 - Number of positive and negative selected sites detected on carboxypeptidase using the HyPhy package. W – dN/dS ratio. SLAC, FEL and MEME results correspond to p-value ≤ 0.05. REL results correspond to 50 Bayes factor.

SLAC		FEL		REL		FUBAR		Integrative		MEME
w>1	w<1	w>1	w<1	w>1	w<1	w>1	w<1	w>1	w<1	
1	33	3	111	9	143	3	183	19	209	14

Table 3 - Sites under diversifying selection in carboxypeptidas according to the integrative analysis. *: not statistically significant.

Codon	SLAC p-value ≤ 0.05	FEL p-value ≤ 0.05	REL Bayes Factor ≥ 50	FUBAR Pr[β>α] ≥ 0.9	MEME p-value ≤ 0.05
80	0.143*	0.026	178.764	0.934	0.022
141	0.243*	0.028	96.567	0.926	0.043
157	0.039	0.020	119.107	0.929	0.005

These results show that despite a strong purifying selection pressure, there are sites evolving dynamically. Given the results from the MEME analysis, the episodic nature of selection may have ensured that the molecular scaffold of the toxins remained extremely conserved over time, while subtle accumulation of advantageous changes occurred in certain regions of the toxin. These changes are, perhaps, key in ensuring that the organisms are capable of predating specific species or defending themselves from others, thus ensuring their survival.

Structural Analysis

In order to assess the effects of selective pressure in proteins, homology models were created using other known protein structures as a base. It was not possible to model the metalloprotease GON-domain, as the different methods used provided 23% and 55% of the sequence modeled with 90% or higher confidence. Over 200 residues were modelled using *ab initio* methods, which are generally unreliable for such analyses. The overall confidence was too low to consider a viable structure (under 70%).

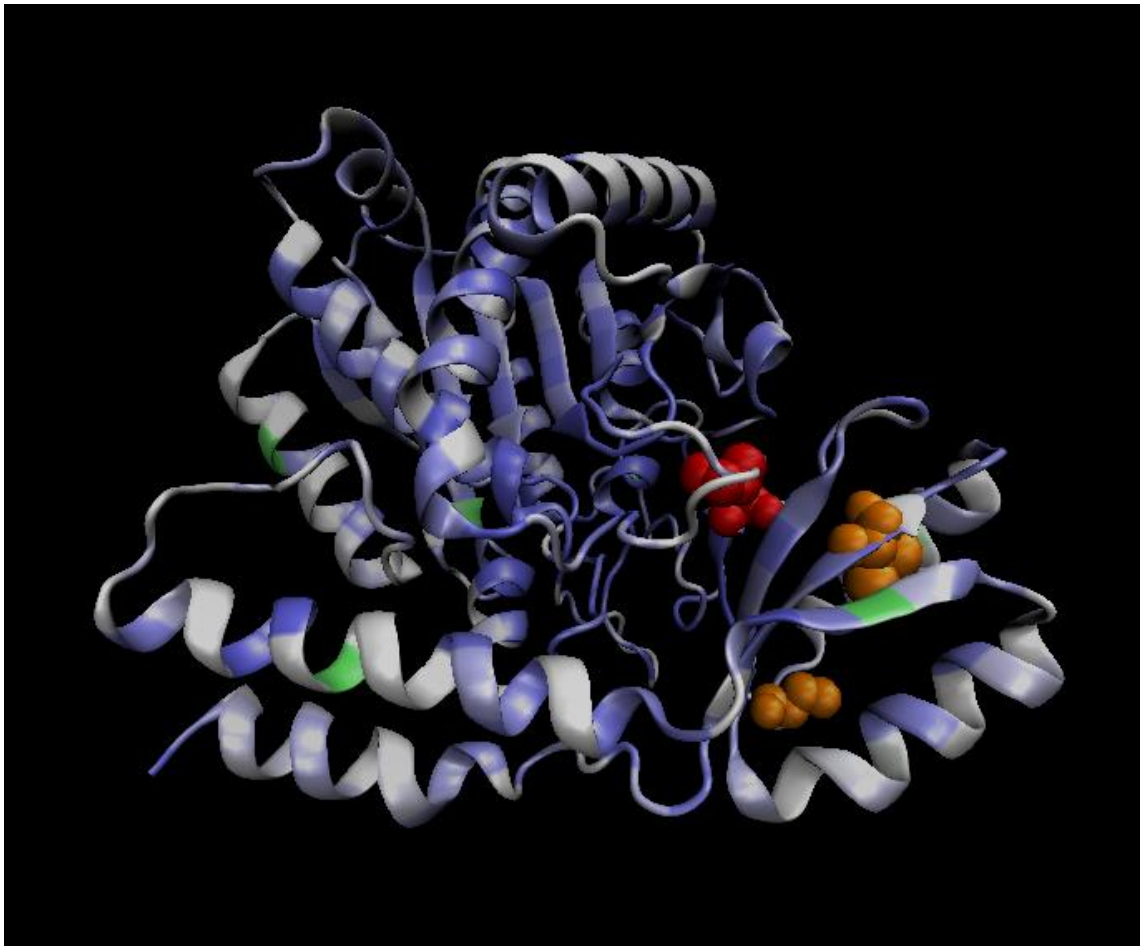


Fig.12 – Carboxypeptidase model based on a human procarboxypeptidase A2. Green regions represent sites where more than 10% of the sequences were available for calculation and are therefore not categorized; white regions represent poorly conserved residues; blue regions represent highly conserved residues; spheres represent sites detected under positive selection by site by site analysis. Red spheres are the residue under diversifying selection close to the putative binding site.

Carboxypeptidase was successfully modelled, with over 95% of the sequence modelled with 90% or higher confidence. The base protein used was a human procarboxypeptidase A2 (Figure 12). We mapped the three sites under diversifying selection in this structure and found them to be all in the same region of the protein. They are also in a very external position, with one being close to the interior of the protein. Codon 157, corresponding to amino acid 53. A prediction using 3DLigandSite suggests that his region is very close to the substrate interaction. It could be that this region is involved in substrate binding and this variability is an adaptation mechanism.

All of the sequences were overlapped and the model was tested for conservation using Consurf. The highest variation exists closer to the surface area of the protein, with the most conserved domains being found on the inside.

Chapter 5 - Conclusion

Hox genes are an essential molecular component of the current animal life form, whose numbers seem to have increased from the most primitive bilaterian to the most complex vertebrate. It is possible that the genomic duplications creating the multiple Hox gene and its groups (anterior, paralog group 3, central and posterior) are directly responsible for specific body plans and its adjustment to each particular animal environment.

Patterns indicating potential loss of Hox genes in Cephalopoda and genome rearrangement can lead to changes in gene dynamics, which can potentially have an impact on the development path. This idea is corroborated by the very recently published in-depth study of *Octopus bimaculoides* (Albertin et al. 2015). Site-specific mutations in the highly conserved homeodomain are also potential candidates for the underlying cause in the emergence of morphological novelties.

We present a fully atomized Hox gene group and the possibility of a duplication event is reinforced by our data, strengthening likely a chromosomal duplication in the evolutionary history of Decapodiforme. The octopus genome also presented an atomized Hox gene cluster but no signs of duplication in the Octopodiforme lineage.

Hox gene cephalopod phylogeny was not fully resolved using the homeodomain and the evolutionary relationship of the genes within the clade is not yet fully understood. The role performed by the genes in these organisms is not completely explored either. Gene trees were successful in supporting gene identification, although the history between the genes could not be confidently determined.

Hox gene expression pattern analysis can be used to identify homologous structures and candidate genes, allowing the determination of gene regulation changes that were involved in Molluscan evolution. Further identification of more Hox genes and their properties in other molluscs is also essential to fully comprehend all of the modifications. Understanding the effects of differentially regulated target genes will also contribute to our understanding of the morphological diversification within the phylum Mollusca and the Cephalopoda class.

Recent studies revealed the existence of novel components in cephalopod venom. We detected these same molecules in several other cephalopods, providing insight into cephalopod history and their venom evolutionary origins. The presence of these genes might also be relevant in further understanding dietary habits, behavior and environment adaptation in cephalopods.

We also demonstrated that there are several other species that deserve attention and should be further studied, such as *Doryteuthis pealeii*, *Octopus bimaculoides* and *Onychoteuthis banksii*, as they may possess yet other unidentified molecules that could play an essential role in toxic activity.

These genes retain similar evolutionary traces to the generally accepted cephalopod evolution. Phylogenetic trees were resolved with most cephalopods being paired with their closer relatives.

Despite the high degree of purifying selection found in venom genes, episodic positive selection was still detected by MEME at several sites and other likelihood methods were also capable of positively identify diversifying selection. This suggests that these molecules could be involved in species adaptation in their environment. The small variations that occur could allow for a greater variety of prey or for more effectiveness against predators.

These variations were mostly found in the surface area of the protein and slight variations close to the putative binding site. Similarly to other venom components, the scaffold of most Coleoid toxins remains extremely conserved, while subtle accumulations occur in certain regions of the toxin, driven by natural selection (Ruder et al. 2013).

References

Abascal, F., R. Zardoya and D. Posada (2005). "ProtTest: selection of best-fit models of protein evolution." Bioinformatics **21**(9): 2104-2105.

Albertin, C. B., O. Simakov, T. Mitros, Z. Y. Wang, J. R. Pungor, E. Edsinger-Gonzales, S. Brenner, C. W. Ragsdale and D. S. Rokhsar (2015). "The octopus genome and the evolution of cephalopod neural and morphological novelties." Nature **524**(7564): 220-224.

Alberts B., J. A., Lewis J (2002). Molecular Biology of the Cell. 4th edition.

Almeida, D., E. Maldonado, V. Vasconcelos and A. Antunes (2015). "Adaptation of the Mitochondrial Genome in Cephalopods: Enhancing Proton Translocation Channels and the Subunit Interactions." PLoS ONE **10**(8): e0135405.

Alonso-del-Rivero, M., S. A. Trejo, M. L. Reytor, M. Rodriguez-de-la-Vega, J. Delfin, J. Diaz, Y. González-González, F. Canals, M. A. Chavez and F. X. Aviles (2012). "Tri-domain Bifunctional Inhibitor of Metallocarboxypeptidases A and Serine Proteases Isolated from Marine Annelid Sabellastarte magnifica." The Journal of Biological Chemistry **287**(19): 15427-15438.

Alonso-del-Rivero, M., S. A. Trejo, M. Rodríguez de la Vega, Y. González, S. Bronsoms, F. Canals, J. Delfín, J. Diaz, F. X. Aviles and M. A. Chávez (2009). "A novel metallocarboxypeptidase-like enzyme from the marine annelid Sabellastarte magnifica– a step into the invertebrate world of proteases." FEBS Journal **276**(17): 4875-4890.

Apte, S. S. (2009). "A Disintegrin-like and Metalloprotease (Reprolysin-type) with Thrombospondin Type 1 Motif (ADAMTS) Superfamily: Functions and Mechanisms." The Journal of Biological Chemistry **284**(46): 31493-31497.

Baird-Titus, J. M., K. Clark-Baldwin, V. Dave, C. A. Caperelli, J. Ma and M. Rance (2006). "The solution structure of the native K50 Bicoid homeodomain bound to the consensus TAATCC DNA-binding site." J Mol Biol **356**(5): 1137-1151.

Barton, N. H., D. E. G. Briggs, J. A. Eisen, D. B. Goldstein and N. H. Patel (2007). Evolution, Cold Spring Harbor Laboratory Press.

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

Boratyn, G. M., C. Camacho, P. S. Cooper, G. Coulouris, A. Fong, N. Ma, T. L. Madden, W. T. Matten, S. D. McGinnis, Y. Merezhuk, Y. Raytselis, E. W. Sayers, T. Tao, J. Ye and I. Zaretskaya (2013). "BLAST: a more efficient report with usability improvements." Nucleic Acids Research **41**(W1): W29-W33.

Brunet, F. G., F. W. Fraser, M. J. Binder, A. D. Smith, C. Kintakas, C. M. Dancevic, A. C. Ward and D. R. McCulloch (2015). "The evolutionary conservation of the A Disintegrin-like and Metalloproteinase domain with Thrombospondin-1 motif metzincins across vertebrate species and their expression in teleost zebrafish." BMC Evolutionary Biology **15**: 22.

Callaerts, P., P. N. Lee, B. Hartmann, C. Farfan, D. W. Choy, K. Ikeo, K. F. Fischbach, W. J. Gehring and H. G. de Couet (2002). "HOX genes in the sepiolid squid Euprymna scolopes: implications for the evolution of complex body plans." Proc Natl Acad Sci U S A **99**(4): 2088-2093.

Casewell, N. R., W. Wuster, F. J. Vonk, R. A. Harrison and B. G. Fry (2013). "Complex cocktails: the evolutionary novelty of venoms." Trends Ecol Evol **28**(4): 219-229.

Celniker, G., G. Nimrod, H. Ashkenazy, F. Glaser, E. Martz, I. Mayrose, T. Pupko and N. Ben-Tal (2013). "ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function." Israel Journal of Chemistry **53**(3-4): 199-206.

Cunningham, F., M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, A. K. Kähäri, S. Keenan, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, B. Overduin, A. Parker, M. Patricio, E. Perry, M. Pignatelli, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, B. L. Aken, E. Birney, J.

Harrow, R. Kinsella, M. Muffato, M. Ruffier, S. M. J. Searle, G. Spudich, S. J. Trevanion, A. Yates, D. R. Zerbino and P. Flicek (2015). "Ensembl 2015." Nucleic Acids Research **43**(D1): D662-D669.
DeCamillis, M. A., D. L. Lewis, S. J. Brown, R. W. Beeman and R. E. Denell (2001). "Interactions of the Tribolium Sex combs reduced and proboscipedia Orthologs in Embryonic Labial Development." Genetics **159**(4): 1643-1648.
Dutertre, S., A.-H. Jin, I. Vetter, B. Hamilton, K. Sunagar, V. Lavergne, V. Dutertre, B. G. Fry, A. Antunes, D. J. Venter, P. F. Alewood and R. J. Lewis (2014). "Evolution of separate predation- and defence-evoked venoms in carnivorous cone snails." Nat Commun **5**.
Fiorito, G., A. Affuso, D. B. Anderson, J. Basil, L. Bonnaud, G. Botta, A. Cole, L. D'Angelo, P. De Girolamo, N. Dennison, L. Dickel, A. Di Cosmo, C. Di Cristo, C. Gestal, R. Fonseca, F. Grasso, T. Kristiansen, M. Kuba, F. Maffucci, A. Manciocco, F. C. Mark, D. Melillo, D. Osorio, A. Palumbo, K. Perkins, G. Ponte, M. Raspa, N. Shashar, J. Smith, D. Smith, A. Sykes, R. Villanueva, N. Tublitz, L. Zullo and P. Andrews (2014). "Cephalopods in neuroscience: regulations, research and the 3Rs." Invertebrate Neuroscience **14**(1): 13-36.
Focareta, L., S. Sesso and A. G. Cole (2014). "Characterization of Homeobox Genes Reveals Sophisticated Regionalization of the Central Nervous System in the European Cuttlefish <italic>Sepia officinalis</italic>." PLoS ONE **9**(10): e109627.
Fritsch, M., T. Wollesen, A. L. de Oliveira and A. Wanninger (2015). "Unexpected co-linearity of Hox gene expression in an aculiferan mollusk." BMC Evolutionary Biology **15**(1): 151.
Fröbius, A. C., D. Q. Matus and E. C. Seaver (2008). "Genomic Organization and Expression Demonstrate Spatial and Temporal Hox Gene Colinearity in the Lophotrochozoan Capitella sp. I." PLoS ONE **3**(12): e4004.
Fry, B. G., K. Roelants, D. E. Champagne, H. Scheib, J. D. Tyndall, G. F. King, T. J. Nevalainen, J. A. Norman, R. J. Lewis, R. S. Norton, C. Renjifo and R. C. de la Vega (2009). "The toxicogenomic multiverse: convergent recruitment of proteins into animal venoms." Annu Rev Genomics Hum Genet **10**: 483-511.
Fry, B. G., K. Roelants and J. A. Norman (2009). "Tentacles of venom: toxic protein convergence in the Kingdom Animalia." J Mol Evol **68**(4): 311-321.
Gilbert, D. (2013). "Gene-omes built from mRNA seq not genome DNA.".
Gouy, M., S. Guindon and O. Gascuel (2010). "SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building." Molecular Biology and Evolution **27**(2): 221-224.
Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. LeDuc, N. Friedman and A. Regev (2013). "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis." Nat. Protocols **8**(8): 1494-1512.
Hall, K. and R. Hanlon (2002). "Principal features of the mating system of a large spawning aggregation of the giant Australian cuttlefish Sepia apama (Mollusca: Cephalopoda)." Marine Biology **140**(3): 533-545.
Hallinan, N. M. and D. R. Lindberg (2011). "Comparative Analysis of Chromosome Counts Infers Three Paleopolyploidies in the Mollusca." Genome Biology and Evolution **3**: 1150-1163.
Hanlon, R. T., C. C. Chiao, L. M. Mäthger, A. Barbosa, K. C. Buresch and C. Chubb (2009). "Cephalopod dynamic camouflage: bridging the continuum between background matching and disruptive coloration." Philosophical Transactions of the Royal Society B: Biological Sciences **364**(1516): 429-437.
Hinman, V. F., E. K. O'Brien, G. S. Richards and B. M. Degnan (2003). "Expression of anterior Hox genes during larval development of the gastropod Haliotis asinina." Evol Dev **5**(5): 508-521.
Hoving, H. J., J. A. Perez, K. S. Bolstad, H. E. Braid, A. B. Evans, D. Fuchs, H. Judkins, J. T. Kelly, J. E. Marian, R. Nakajima, U. Piatkowski, A. Reid, M. Vecchione and J. C. Xavier (2014). "The study of deep-sea cephalopods." Adv Mar Biol **67**: 235-359.

Humphrey, W., Dalke, A. and Schulten, K., (1996). "VMD - Visual Molecular Dynamics." J. Molec. Graphics **14**: 33-38.
Hunter, M. P. and V. E. Prince (2002). "Zebrafish Hox Paralogue Group 2 Genes Function Redundantly as Selector Genes to Pattern the Second Pharyngeal Arch." Developmental Biology **247**(2): 367-389.
Innan, H. and F. Kondrashov (2010). "The evolution of gene duplications: classifying and distinguishing between models." Nat Rev Genet **11**(2): 97-108.
Joulia, L., H.-M. Bourbon and D. L. Cribbs (2005). "Homeotic proboscipedia function modulates hedgehog-mediated organizer activity to pattern adult Drosophila mouthparts." Developmental Biology **278**(2): 496-510.
Kaas, Q. and D. J. Craik (2015). "Bioinformatics-Aided Venomics." Toxins (Basel) **7**(6): 2159-2187.
Kelley, L. A., S. Mezulis, C. M. Yates, M. N. Wass and M. J. E. Sternberg (2015). "The Phyre2 web portal for protein modeling, prediction and analysis." Nat. Protocols **10**(6): 845-858.
Kosakovsky Pond, S. L. and S. D. W. Frost (2005). "Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection." Molecular Biology and Evolution **22**(5): 1208-1222.
Kubodera, T. and K. Mori (2005). "First-ever observations of a live giant squid in the wild." Proceedings of the Royal Society B: Biological Sciences **272**(1581): 2583-2586.
Lee, P. N., P. Callaerts, H. G. de Couet and M. Q. Martindale (2003). "Cephalopod Hox genes and the origin of morphological novelties." Nature **424**(6952): 1061-1065.
Lutz, B., H. C. Lu, G. Eichele, D. Miller and T. C. Kaufman (1996). "Rescue of Drosophila labial null mutant by the chicken ortholog Hoxb-1 demonstrates that the function of Hox genes is phylogenetically conserved." Genes Dev **10**(2): 176-184.
Mallo, M., D. M. Wellik and J. Deschamps (2010). "Hox Genes and Regional Patterning of the Vertebrate Body Plan." Developmental biology **344**(1): 7-15.
Massey, D. J., J. J. Calvete, E. E. Sanchez, L. Sanz, K. Richards, R. Curtis and K. Boesen (2012). "Venom variability and envenoming severity outcomes of the Crotalus scutulatus scutulatus (Mojave rattlesnake) from Southern Arizona." J Proteomics **75**(9): 2576-2587.
Mather, J. A. and D. L. Mather (2004). "Apparent movement in a visual display: the 'passing cloud' of Octopus cyanea (Mollusca: Cephalopoda)." Journal of Zoology **263**(1): 89-94.
Monica Lynn Boyle, M. A. F., G R Martin "Isolation of the mouse HOX-2.9 gene: Analysis of embryonic expression suggests that postnatal information along the anterior-posterior axis is specified by mesoderm." Development **110**(2): 589-607.
Nguyen, L.-T., H. A. Schmidt, A. von Haeseler and B. Q. Minh (2014). "IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies." Molecular Biology and Evolution.
Nielsen, C. (2012). Animal evolution: interrelationships of the living phyla, Oxford University Press.
Nylander, J. A. A., J. C. Wilgenbusch, D. L. Warren and D. L. Swofford (2008). "AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics." Bioinformatics **24**(4): 581-583.
Oellermann, M., J. Strugnell, B. Lieb and F. Mark (2015). "Positive selection in octopus haemocyanin indicates functional links to temperature adaptation." BMC Evolutionary Biology **15**(1): 133.
Ohneda, K., R. G. Mirmira, J. Wang, J. D. Johnson and M. S. German (2000). "The homeodomain of PDX-1 mediates multiple protein-protein interactions in the formation of a transcriptional activation complex on the insulin promoter." Mol Cell Biol **20**(3): 900-911.
Pernice, M., J. S. Deutsch, A. Andouche, R. Boucher-Rodoni and L. Bonnaud (2006). "Unexpected variation of Hox genes' homeodomains in cephalopods." Mol Phylogenet Evol **40**: 872–879.
Pond, S. L. K., S. D. W. Frost and S. V. Muse (2005). "HyPhy: hypothesis testing using phylogenies." Bioinformatics **21**(5): 676-679.

Posada, D. (2008). "jModelTest: Phylogenetic Model Averaging." Molecular Biology and Evolution **25**(7): 1253-1256.
Quintero-Hernandez, V., J. M. Jimenez-Vargas, G. B. Gurrola, H. H. Valdivia and L. D. Possani (2013). "Scorpion venom components that affect ion-channels function." Toxicon **76**: 328-342.
Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard and J. P. Huelsenbeck (2012). "MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice across a Large Model Space." Systematic Biology.
Ruder, T., K. Sunagar, E. A. Undheim, S. A. Ali, T. C. Wai, D. H. Low, T. N. Jackson, G. F. King, A. Antunes and B. G. Fry (2013). "Molecular phylogeny and evolution of the proteins encoded by coleoid (cuttlefish, octopus, and squid) posterior venom glands." J Mol Evol **76**(4): 192-204.
Ruppert, E. E., R.S. Fox & R.D. Barnes (2004). Invertebrate Zoology, Thomson - Brooks/Cole.
Rusch, D. B. and T. C. Kaufman (2000). "Regulation of proboscipedia in Drosophila by Homeotic Selector Genes." Genetics **156**(1): 183-194.
Samadi, L. and G. Steiner (2010). "Expression of Hox genes during the larval development of the snail, Gibbula varia (L.)—further evidence of non-colinearity in molluscs." Development Genes and Evolution **220**(5-6): 161-172.
Samira, J. P. (2010). "Evolutionary genetics of the AMT domain of the microcystin gene cluster in various cyanobacteria genera.".
SF., G. (2000). The Origins of Anterior-Posterior Polarity, 6th edition.
Suman Pratihar, R. P. N., Jayanta Kumar Kundu (2010). "HOX GENES AND ITS ROLE IN ANIMAL DEVELOPMENT." International Journal of Science and Nature **1**(2): 101-103.
Sunagar, K., B. G. Fry, T. N. Jackson, N. R. Casewell, E. A. Undheim, N. Vidal, S. A. Ali, G. F. King, K. Vasudevan, V. Vasconcelos and A. Antunes (2013). "Molecular evolution of vertebrate neurotrophins: co-option of the highly conserved nerve growth factor gene into the advanced snake venom arsenalf." PLoS One **8**(11): e81827.
Sunagar, K., W. E. Johnson, S. J. O'Brien, V. Vasconcelos and A. Antunes (2012). "Evolution of CRISPs associated with toxicoferan-reptilian venom and mammalian reproduction." Mol Biol Evol **29**(7): 1807-1822.
Talbot, C. M. and J. N. Marshall (2011). "The retinal topography of three species of coleoid cephalopod: significance for perception of polarized light." Philosophical Transactions of the Royal Society B: Biological Sciences **366**(1565): 724-733.
Teshima, K. M. and H. Innan (2008). "Neofunctionalization of Duplicated Genes Under the Pressure of Gene Conversion." Genetics **178**(3): 1385-1398.
Toll, R. and L. Binger (1991). "Arm anomalies: cases of supernumerary development and bilateral agenesis of arm pairs in Octopoda (Mollusca, Cephalopoda)." Zoomorphology **110**(6): 313-316.
Xia, X. (2013). "DAMBE5: A Comprehensive Software Package for Data Analysis in Molecular Biology and Evolution." Molecular Biology and Evolution.