THE ORIGIN OF THE HOX AND PARAHOX LOCI AND ANIMAL HOMEOBOX EVOLUTION

Olivia Mendivil Ramos

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews

2014

The origin of the Hox and ParaHox loci and animal homeobox evolution

Olivia Mendivil Ramos
(omr3@st-andrews.ac.uk)

This thesis is submitted in partial fullfilment for the degree of
PhD
at the
University of St Andrews

25th April 2013

1. Candidate's declarations

I, Olivia Mendivil Ramos, hereby certify that this thesis, which is approximately 55000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in April, 2009 and as a candidate for the degree of PhD in Biology in April, 2013; the higher study for which this is a record was carried out in the University of St Andrews between 2009 and 2013.

Date 25th April 2013 Signature of candidate

Olivia Mendivil Ramos

2. Supervisors declarations

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in Biology in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date 25th April 2013 Signature of supervisor

Dr. David E. K. Ferrier

Date 25th April 2013 Signature of supervisor

Dr. Daniel Barker

3. Permission for electronic publication

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the electronic publication of this thesis:

Embargo on both part of printed copy and electronic copy for the same fixed period of two years on the following ground: the publication of Chapters 5, 6 and 7 would preclude future publication

Date 25th April 2013

Signature of candidate

Olivia Mendivil Ramos

Signature of supervisors

Dr. David E. K. Ferrier

Dr. Daniel Barker

Abstract

The homeobox superfamily is one of the most significant gene families in the evolution of developmental processes in animals. Within this superfamily the ANTP class has expanded exclusively in animals and, therefore, the reconstruction of its origin and diversification into the different 'modern' families have become prominent questions in the 'evo-devo' field. The current burgeoning availability of animal genome sequences is improving the resolution of these questions, putting them in a genome evolution context, as well as providing the field with a large, detailed and diverse catalogue of animal homeobox complements. Here I have contributed with a new hypothesis on the origin and evolution of the Hox and ParaHox loci and the new term, ghost loci, referring to homologous genome regions that have lost their homeobox genes. This hypothesis proposes that the last common ancestor of all animals had a much more complex genome (i.e. differentiated Hox, ParaHox and NK loci) that underwent a simplification in the early animal lineages of sponges and placozoans. In collaboration with the Adamska group I resolved the orthology of the first ever ParaHox genes reported in calcareous sponges. This finding serves as an independent confirmation of the ghost loci hypothesis and further resolves the events of secondary simplification within the sponge lineage. Finally, I have catalogued the homeobox complement of the newly sequenced arthropod, the myriapod *Strigamia maritima*, and examined the linkage and clustering of these genes. This has furthered our understanding of the evolution of the ANTP class. The diversity of the homeobox complement and the retention in this myriapod and the retention of some homeobox genes not previously described within arthropods, in combination with the interesting phylogenetic position that this lineage occupies relative to other arthropods, makes this complement an important point of reference for comparison within the arthropods and in a broader perspective in the ecdyzosoans. These findings have provided significant further insights into the origin and evolution of the homeobox superfamily, with important implications for animal evolution and the evolution of development.

Acknowledgements

I do not think that there is enough space to describe how grateful I am for the superb supervision that Dr. David Ferrier has done for the last years. Thanks a lot for your immense patience in teaching me how to be a scientist, for the stimulating science conversations and exciting debates raised during all this time. I would also like to thank Dr. Daniel Barker for his bioinformatic and statistical angle that have raised the quality of my thesis. Also for all the help and support at any time during these past years.

I would like to thank the School of Biology of University of St Andrews and the BBSRC DTG for funding me and this project.

To Dr. Maja Adamska and Sofia Fortunato for inviting me to visit SARS and let me help them in solving wonderful mysteries of sponges.

To the eDGE Group, the FMR Group, Dr. Alfredo Ojanguren, Friday Lab Chat Group (in particular Prof. Mike Ritchie and Dr. David Shuker) for pushing me to be a much better scientist.

To Dr. Maria Dornelas and Prof. Anne Magurran to give me chance to work with them in their database project in the last stages of my PhD.

To Dr. Ana Guarner for her infinite support since the end of our undergraduates and the chance of talking back home. To Dr. Tamara Polajnar and Dr. Ronan Cummings for engaging me with computers, teaching me to code and taking care of me when I went to a conference in Galway. Also, to Nina for taking care of me when I went to a conference in Paris. To Mora, Saray, Sandra, Ana, Anita, Sophie and Andy for looking after me when I needed the most.

I would like to thank my husband Wim for the loving support throughout my PhD and for the insightful interdisciplinary scientific/engineer conversations and his constructive criticism towards some aspects of my thesis.

Last but not least, I am specially grateful to dad, mom, Sara and Kiko for everything that they have given to me. I certainly would have not come this far without their vision, direction, education at all levels and/or investment. Thank you!

“Stay hungry, stay foolish”

S.J.

Table of Contents

List of figures

List of tables

Chapter 1

Introduction

Section 1.4 (Genome dynamics) is adapted from Mendivil Ramos, O. & Ferrier, D. E. K. 2012. Mechanisms of Gene Duplication and Translocation and Progress towards Understanding Their Relative Contributions to Animal Genome Evolution. International Journal of Evolutionary Biology, 2012, 10.

1.0 A brief foreword about “evo-devo” and the role of homeoboxes

The modern synthesis framework of evolutionary theory fails to explain the origins and diversity of animal body plans in mechanistic terms (Laubichler and Maienschein, 2007, Reid, 2007, Muller, 2007). In the early 80s, the field of evolutionary developmental biology (“evo-devo”) rose to prominence with the promise to address this failing of the modern synthesis. Facilitating the emergence of this field were the advances performed on molecular techniques for gene cloning and visualization of gene activity in embryonic tissues of different taxa, making possible the comparison of developmental processes at the molecular level. The “Evo-devo” research is characterized by a dialectical approach, one that looks how developmental systems have evolved and another that examines the consequences of these historically established systems for organismal evolution. These approaches pursue the overall question of how the evolutionary developmental interactions relate to environmental conditions. This question explores the development-evolution interface in multiple angles using a plethora of interdisciplinary methods, which eventually will capture the consequences for evolutionary theory (Muller, 2007).

At the same time as the emergence of this field the discovery of the homeobox genes and their functionality (i.e. axial patterning in the embryo in bilaterian animals) allowed the rapid comparison of animal development and its evolution across animal phyla (McGinnis et al., 1984). Then, the formulation of the Zootype hypothesis in which it was proposed that the axial expression of the homeobox genes in an animal embryo is a defining character of animals (Slack et al., 1993), made these genes one of the paradigms to understand the core questions of “evo-devo”: how the homoebox genes originate and evolve, and finally what is their implication in modifying developmental processes.

All the results that are going to be presented in this thesis contribute to our understanding of some of these core questions in the context of the homeobox gene superfamily, focusing on gene origins, diversification and loss.

1.1 Thesis outline

This thesis explores two main aspects of the homeobox gene superfamily: first, how some of the families within this superfamily originated and second, the diversity of homeobox genes and their clustering as presented in a recently sequenced animal genome, *Strigamia maritima.* I use whole animal genome sequences as a means to compare, analyse and improve our current understanding of the origin and evolution of the homeobox gene superfamily within the evolutionary context of the animal genomes that possess them.

Studies of animal comparative genomics have benefited from the ever-growing repertoire of publicly available animal genome sequences. The impressive rate at which new animal genome sequences are being released has impacted the field of evolutionary biology on two counts: first, the reassessment of animal phylogeny and the resolution of some specific key nodes of the tree, and second, the comparison of animal genomes across the animal tree in order to understand the influence of genome architecture on gene regulation as a means to explain the phenotypic diversity of animals (Cañestro et al., 2007).

Accurate knowledge of the evolutionary relationships amongst animals provides a fundamental framework for understanding the directionality of evolutionary change of a particular animal trait. The phylogenetic tree of metazoans has been extensively refined in the last 25 years (Telford, 2008, 2013). Since the formulation of the new animal phylogeny, based on molecular data, and the inclusion of animal whole genome sequences (i.e. phylogenomics as defined in (Philippe et al., 2005a, Delsuc et al., 2005)) a fairly detailed overview of animal relationships has been developed. However, this type of analysis has some limitations that could easily lead to contradictory results or poorly resolved phylogenetic topologies. These pitfalls have been identified and tackled by recent advances in methods that detect systematic errors, improvements in data quality, wider taxonomic sampling and the identification of new markers of evolutionary history (Philippe and Telford, 2006, Philippe et al., 2011). I describe the recent progress in this research area and the working hypothesis I will be using for the rest of this study in Section 1.2 and summarized in Figure 1.3.

The architecture of the animal genome and its evolution can be a determinant factor in the generation of phenotypic variation. It is noteworthy that the majority of comparative animal genomic studies have focused on descriptive studies of gene homologue content and their linkage, rather than the cis-regulatory landscape of these genomes. The discovery of homologous regions amongst genome sequences via synteny analyses (i.e. preserved gene/orthologue linkage across species) has played a major role in the reconstruction of animal archetypes and landmarks within the radiation of the animal kingdom (Putnam et al., 2007, Putnam et al., 2008). Moreover, the mapping of conserved synteny across animal genomes serves as the basis on which genome rearrangements are estimated (Lv et al., 2011, Irimia et al., 2012, Simakov et al., 2013).

During the first part of this study, in which I will address the origin of Hox and ParaHox loci, I use genomes from the basal lineages of the animal tree: the startlet sea anemone *Nematostella vectensis* (Putnam et al., 2007), the only placozoan representative *Trichoplax adhaerens* (Srivastava et al., 2008) and the desmosponge *Amphimedon queenslandica* (Srivastava et al., 2010), all of them constituting proxies with which to reconstruct ancestral genome states of metazoans. These are compared with amphioxus (*Branchiostoma floridae*), an animal that represents one of the key nodes for understanding the pre-duplication state of vertebrates (Putnam et al., 2008). I will also use the human genome (*Homo sapiens* (Consortium, 2001, Lander, 2011)), as it has the best quality and physical map yet available. In the second part of this thesis I use newly sequenced sponge genomes, *Sycon ciliatum* and *Leucosolenia sp.*, kindly provided by the Adamska group (SARS centre, Norway), to test my hypothesis about Hox and ParaHox origins. In the third part of this study I will use the recently sequenced coastal centipede, *Strigamia maritima*, the only myriapod sequenced so far and a means to investigate a key node within the arthropod phylogeny, and catalogue its homeobox complement. This genome will shed light on arthropod evolutionary history and provide further insights into the evolution of the developmental mechanisms of arthropods.

The homeobox gene superfamily is one of the fundamental types of transcription factor that are needed to direct the correct development of animal embryos. Molecular developmental biology experiments on classical model organisms proved that this superfamily is largely responsible for directing the development of diverse morphologies, and that it is widely represented across the animal kingdom. Within this superfamily there is a family of genes, the renowned Hox genes, that when mutated lead to homeotic phenotypes, i.e. the transformation of one part of the body into another (Lewis, 1978, Akam, 1989). At the sequence level all members of this superfamily possess a highly conserved region, the homeobox region, which encodes a DNA-binding motif, usually of 60 amino-acids, called the homeodomain ((Johnson and Herskowitz, 1985, McGinnis et al., 1984), see Fig. 1.1). This region has been used to classify the different members of this superfamily into classes and, sometimes with the aid of other protein motifs/conserved domains outside the homeodomain, members of the same class into different families (Bürglin, 1994, 2005). The phylogenetic relationships of these sequences imply that this superfamily underwent a drastic expansion in the animals that in large part was specific to the animals and independent from the other expansions of homeobox genes in other eukaryotes. The process of classification into classes and families is not a straightforward one, as in some cases only a limited phylogenetic signal can be obtained from the 60 amino-acid motif, and the ancient nature of the duplications that gave rise to the different classes has eroded the phylogenetic signal from the homeodomain through evolutionary time.

In addition to their abundance, diversity and importance in the evolution of developmental mechanisms in animals, some of the homeobox gene families have the intriguing feature of being arranged in clusters within animal genomes. The clustering arrangement, in combination with the phylogenetic signal, is often integral to hypotheses about the diversification of this superfamily (Pollard and Holland, 2000, Hui et al., 2012). One of the most renowned examples of this clustering is the Hox cluster (Lewis, 1978), but this is not the only one within this superfamily (Kim and Nirenberg, 1989, Brooke et al., 1998, Mazza et al., 2010).

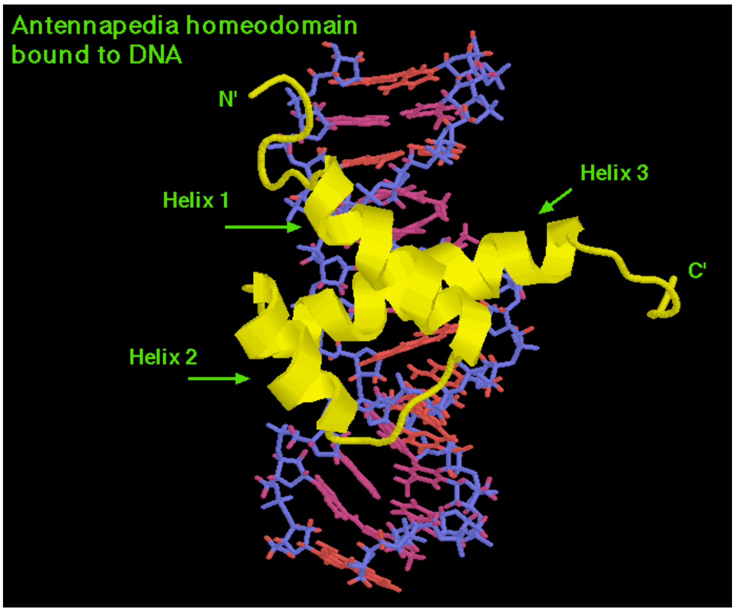


Figure 1.1.- General 3D structure of the ANTP homeodomain. *The homeodomain has the helix-turn-helix motif (in yellow) to bind to the DNA. Adapted from http://www.biosci.ki.se.*

Having briefly introduced the main topics of this thesis I will now examine current views on animal phylogeny, animal genome dynamics and different aspects of the biology and classification of the homeobox gene superfamily, in more detail.

1.2 Animal phylogeny

The constant publication of new genome sequences constitutes a platform for refined comparative studies, including the refinement of animal phylogeny which is fundamental for the formulation of evolutionary statements. In this thesis the underlying phylogeny of the animal kingdom (the metazoans) is crucial for the rest of the work presented. For more than a century, the elucidation of the relationships among phyla of metazoans has been a very dynamic field of research with constant controversy, and clearly represents a challenge that is now benefiting from the large influx of new molecular data (Philippe and Telford, 2006, Telford, 2008, Dunn et al., 2008, Hejnol et al., 2009, Pick et al., 2010).

The traditional, pre-molecular phylogenies tended to be based upon three major concepts ((Adoutte et al., 2000), summarised in (Halanych, 2004)), which are summarised in Libbie Hyman's diagrams (p38, Fig. 5 (1940)):

(i) Evolution proceeds from a simple form to a complex form.

(ii) A set of conserved embryological features (e.g. cleavage pattern, blastopore fate and mode of coelom formation).

(iii) Overall body architecture (e.g. segmentation and type of coelom).

In particular, animal phylogeny tended to be based upon the form of the body cavity (the coelom), with acoelomates being the basal bilaterian lineages, followed by pseudocoelomates, with the coelomate phyla being seen as the most highly evolved (Fig. 1.2A). In the Hyman-like mindset the bilaterians are split into protostomes and deuterostomes, based on the blastopore fate, and this has traditionally been applied to coelomate animals ((Adoutte et al., 2000, Halanych, 2004); Fig 1.2A). According to this definition, if the ultimate fate of the blastopore (i.e. opening of the archenteron during the embryonic stages) is the mouth and anus then an animal belongs to the clade Protostomia and if the ultimate fate of the blastopore is the anus alone then an animal belongs to the clade Deuterostomia, with the mouth developing from a secondary invagination. Much of this classification system has now been modified due to the impact of molecular data.

As molecular techniques became more sophisticated, the first comprehensive molecular phylogenetic analyses of the major animal groups were based upon the small subunit ribosomal RNA gene (SSU rRNA) ((Field et al., 1988); Fig 1.2B). With increased taxon sampling and improved sequence evolution models our understanding of the interrelations among metazoans have undergone several changes. One of the major contributions to such changes was by Halanych et al. (1995), who recognised the super-phylum Lophotrochozoa within the protostomes. The Lophotrochozoa are an amalgamation of phyla with either of two characteristics: a ciliated feeding structure called the lophophore (distinguished by the presence of a lumen derived from the middle coelomic cavity), or trochophore type larva. The Lophochotrozoa thus unites the annelids, molluscs, platyhelminthes and the lophophorate phyla (brachiopods, phoronids and bryozoans/ectoprocts) along with several other phyla. Another of the major contributions was the work of Aguinaldo et al. (1997), in which the super-phylum Ecdyzosoa was defined. The Ecdyzosoa are animals which share a characteristic moulting of the cuticle (ecdysis), and thus unites arthropods with pseudocoelomate nematodes and priapulids. Further support for the Ecdyzosoa/ Lophotrochozoa split came from analyses of Hox genes (de Rosa et al., 1999), horseradish peroxidase (HRP) antibody staining (Haase et al., 2001), large subunit ribosomal RNA (LSU) (Mallatt and Winchell, 2002), myosin heavy chain (Ruiz-Trillo et al., 2002) and sodium/potassium ATPase (Anderson et al., 2004).

This contradicted the traditional pre-molecular phylogenies, in which annelids and arthropods were put into the Articulata clade due to their segmentation. Halanych et al. (1995) and Aguinaldo et al. (1997) placed the annelids within the Lophotrochozoa and arthropods in the Ecdysozoa. Moreover, the acoelomate and pseudocoelomate phyla of Platyhelminthes, Nemertea and Nematoda are now placed in amongst the coelomate groups (Aguinaldo and Lake, 1998), and chaetognaths and lophophorates, which were classically allied with deuterostomes, are now placed amongst the protostomes.

As sequencing technologies improved and the first animal genomes were released, new studies based on a restricted number of genomic-scale datasets and limited taxa contradicted the new animal phylogeny (Rogozin et al., 2007, Zheng et al., 2007). This data supported the monophyletic clade of coelomates and proposed a return to traditional topologies (the Coelomata hypothesis, (Hyman, 1940)). However, this outcome was due to a systematic phylogenetic error, long branch attraction (LBA, i.e. a phylogenetic artifact which reflects similarity due to convergent or parallel changes that have been accumulated on particularly divergent or fast evolving lineages producing an artifactual phylogenetic grouping of taxa due to an inherent bias in the estimation procedure (Philippe et al., 2011)). In this case this was caused by the inclusion of the fast evolving nematode *Caenorhabditis elegans* (Copley et al., 2004, Philippe et al., 2005b, Irimia et al., 2007). The work of Philippe et al. (Philippe et al., 2005b) led to some improvement, involving a comprehensive study of metazoan relationships comprising 146 genes from 35 species. They address the problem of LBA by removing the rapidly evolving taxa from the analysis, and using a better model of sequence evolution, the CAT model (Lartillot and Philippe, 2004). The improved analyses of Philippe et al. (Philippe et al., 2005b) confirm the placement of nematodes in the Ecdysozoa and platyhelminthes in the Lophotrochozoa.

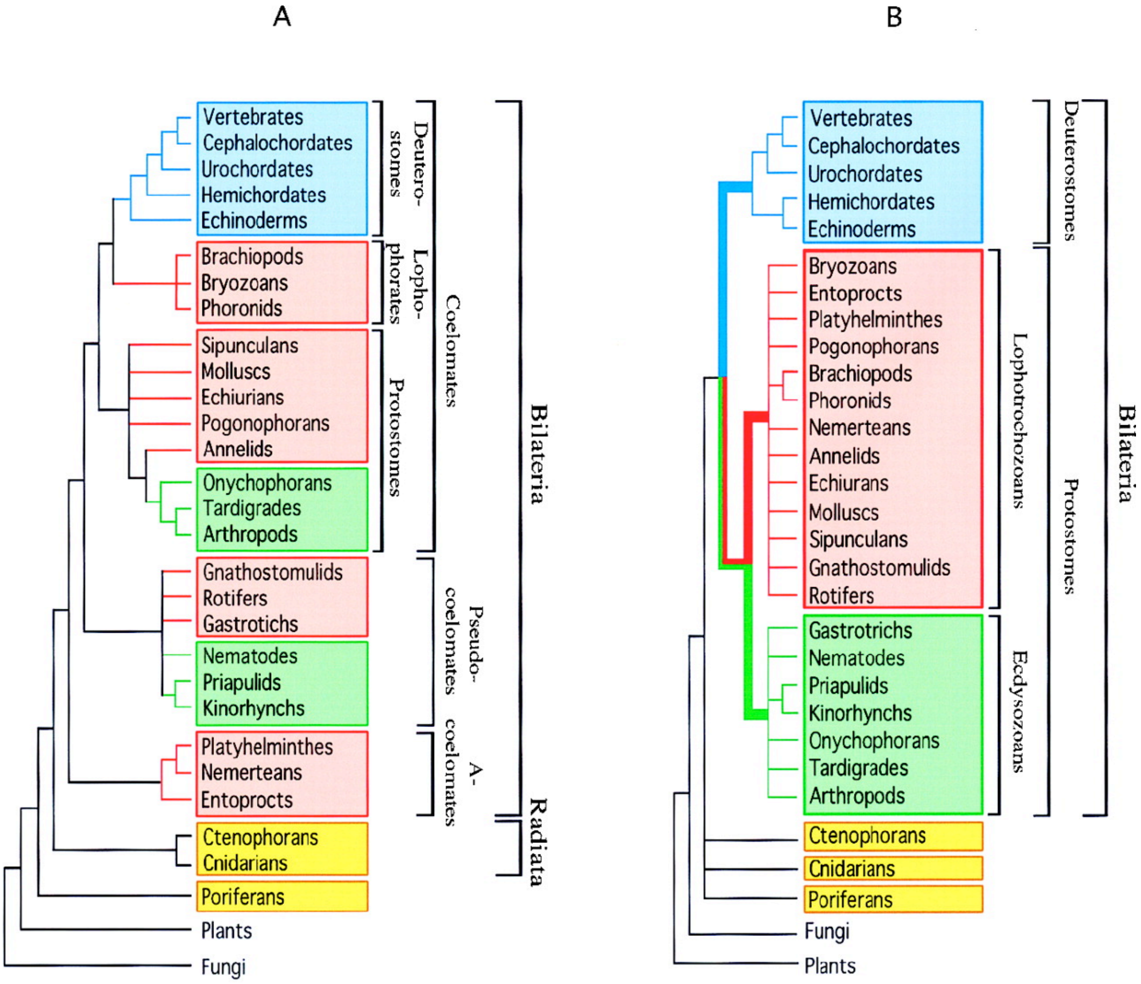


Figure 1.2.- Comparison of the metazoan phylogenies. *(A) Traditional metazoan phylogeny based on embryology and morphology (adapted from Hyman* ***(1940)****). (B) New animal phylogeny based on the molecular sequences of rRNA (phylogenies expanded from Field et al.* ***(1988)*** *to what has been reported by Halanych et al.* ***(1995)*** *and Halanych et al.* ***(1997)****). Diagram adapted from Adoutte et al.* ***(2000)****.*

Despite the great progress from the original molecular phylogeny of Katherine Field et al. (Field et al., 1988), there are still some portions of the animal tree that remain unresolved (Telford, 2008, 2013). The incorporation of large scale genome sequence data makes it possible to identify and address the principal problems affecting phylogenetic analyses (e.g. systematic errors in phylogeny resulting from the usage of homoplastic characters (i.e. molecular characters in which the same nucleotides are independently acquired by distantly related species because the G+C content of their genomes is similar (Telford, 2008)), stochastic errors due to small data samples, and the effects of data partitioning when different models of sequence evolution are used and thus, impacting tree topologies (e.g. CAT model supports the Ecdysozoa/ Lophotrochozoa split, whereas the WAG model supports the Coelomata hypothesis (Lartillot and Philippe, 2008)) (Philippe and Telford, 2006, Philippe et al., 2011). In this vein, using more genes, more taxa and better models Dunn et al. (2008) covered a great diversity of taxa in their analyses. The analyses recovered the Lophotrochozoa and the Ecdysozoa, which together form the monophyletic clade Protostomia, with strong support. However, within this analysis there are still poorly resolved areas within the Deuterostomia clade (Philippe et al., 2007, Lartillot and Philippe, 2008, Dunn et al., 2008, Hejnol et al., 2009). After the study of Dunn et al. (2008), several studies in the same vein have been published with the aim of resolving some of the unclear areas by improving the taxon sampling of the Dunn et al. (2008) dataset and testing alternative modes of phylogenetic reconstruction (Hejnol et al., 2009, Pick et al., 2010, Philippe et al., 2011).

To resolve the ancient splits from the base of the animal tree using molecular sequence data from dipoblastic animals has always been difficult. These difficulties stem from the erosion of phylogenetic signal due to evolutionary time along with long branches caused by multiple substitutions causing non-phylogenetic signal (Philippe et al., 2011). The combination of the long branches along with other sequences that have accumulated few changes (i.e. short internal branches (Philippe et al., 2011)) bias the internal branches to achieve a highly supported phylogenetic signal (Philippe et al., 2011). One of the most surprising results from the phylogenies of Dunn et al. (2008) and Hejnol et al. (2009) is the placement of ctenophores as the most basal lineage. This result is contradicted by the work of Philippe et al. (2009) and Pick et al. (2010), in which the lack of resolution of the basal animal lineages in the studies of Dunn et al. (2008) and Hejnol et al. (2009) is highlighted. Also, Philippe and colleagues question the controversial conclusion of Dellaporta et al. (2006),who placed Placozoa at the base of the animal tree based upon mitochondrial genomes sequences, and the recent study of Schierwater et al. (2009), who recovered a clade of diploblastic animals (Placozoa branching off first) as the sister group to the triploblastic Bilateria. The works of Philipe et al. (2011) and Pick et al. (2010) cautiously examined the controversial conclusions of these phylogenetic studies, and carefully re-analysed the same datasets, avoiding all possible artefacts that could generate non-phylogenetic signal (e.g. in the Schierwater et al. (2009) analysis the supermatrix used to retrieve the phylogenetic analysis was composed of genes with questionable orthology, frameshift errors, point mutations as well as some 'contaminations' of unrelated genes). These analyses of Philippe and colleagues restored a monophyletic Porifera as the basal lineage of the Metazoa, as well as uniting the Ctenophora and Cnidaria into the Coelenterata, which forms a sister clade to the Bilateria. Finally, Philippe and colleagues place Placozoa as the sister to the Eumetazoa (i.e. Eumetazoa entails Coelenterata and Bilateria (Hatschek, 1888)). The working hypothesis I favour and that I will be adopting for the rest of this thesis is the one retrieved by Pick et al. (2010) summarized in Fig. 1.3.

Within the basal animal lineages there remains a debate as to whether the interrelationships of the poriferan classes are monophyletic or paraphyletic (Fig. 1.4). The cladistic analyses based on morphological characters (e.g. biphasic life cycle, filter-feeding, sessile adult form, pinacocytes, choanocytes and aquiferous system (Böger, 1983, Ax, 1996, Reitner and Mehl, 1996)) supported sponge monophyly. As the molecular phylogenetic analyses started to become more prominent, the early 1990s molecular studies of sponges recovered a paraphyletic topology of sponges. In these early studies proposing the paraphyly of sponges there are a few problems (Wörheide et al., 2012):

(i) Absence of significant support values for the hypothesis.

(ii) Hampered by insufficient data (e.g. sparse taxon sampling).

(iii) Methodological shortcomings (e.g. usage of simple sequence evolution models to reconstruct phylogenies).

On the other hand, sponge monophyly is supported by more recent, careful phylogenomic studies (Philippe et al., 2009, Pick et al., 2010, Philippe et al., 2011), which are now congruent with cladistic analyses of morphological characters. However, it is noteworthy that these phylogenomic studies recover a sister relationship of Calcarea and Homoscleromorpha (Dohrmann et al., 2008, Philippe et al., 2009, Pick et al., 2010, Erwin and Thacker, 2007) which is currently difficult to ally with morphological synapomorphies.

Another on-going phylogenetic debate is the one dealing with the relationships among the major Arthropod lineages. Traditionally, as mentioned

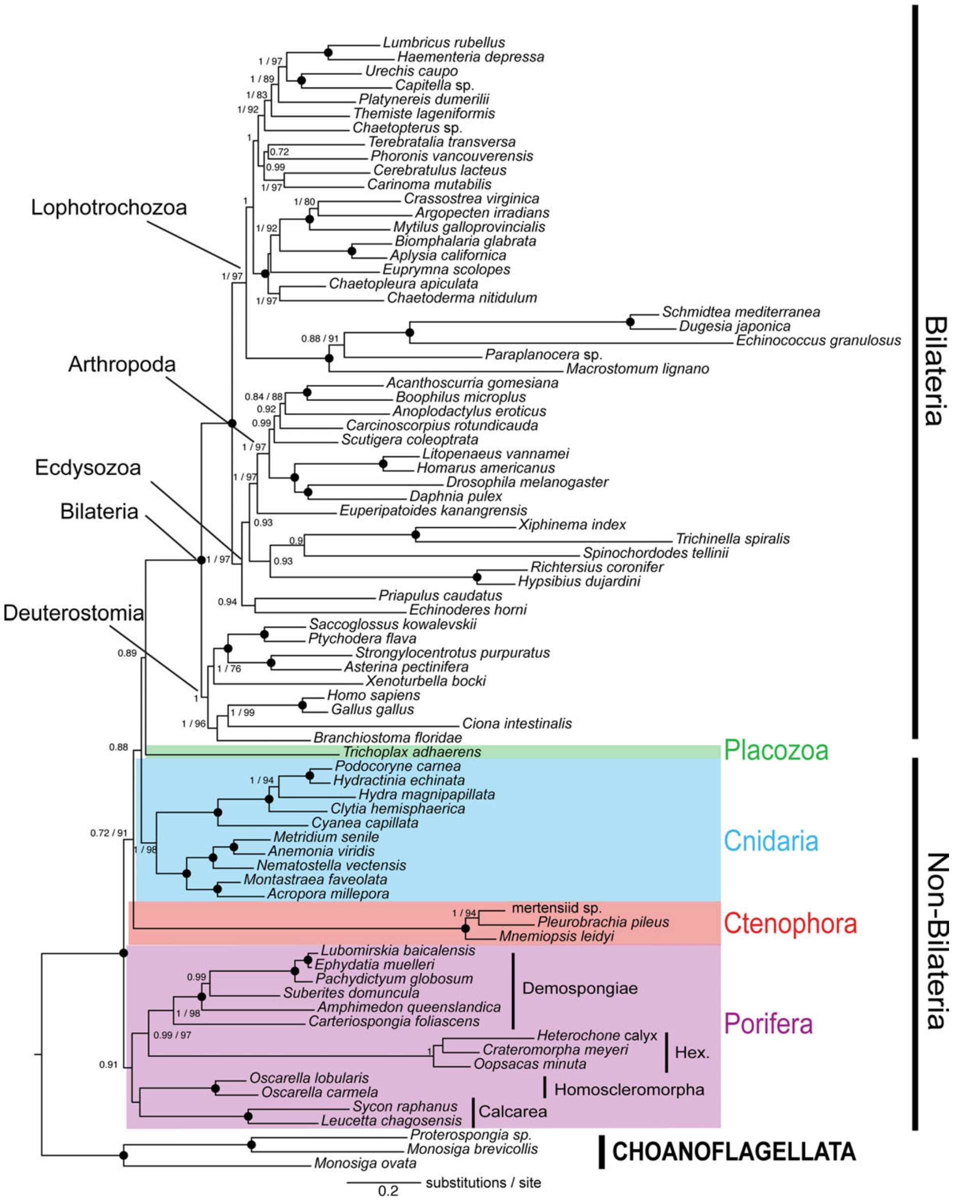


Figure 1.3.- Animal phylogeny adapted from Pick et al. (2010)

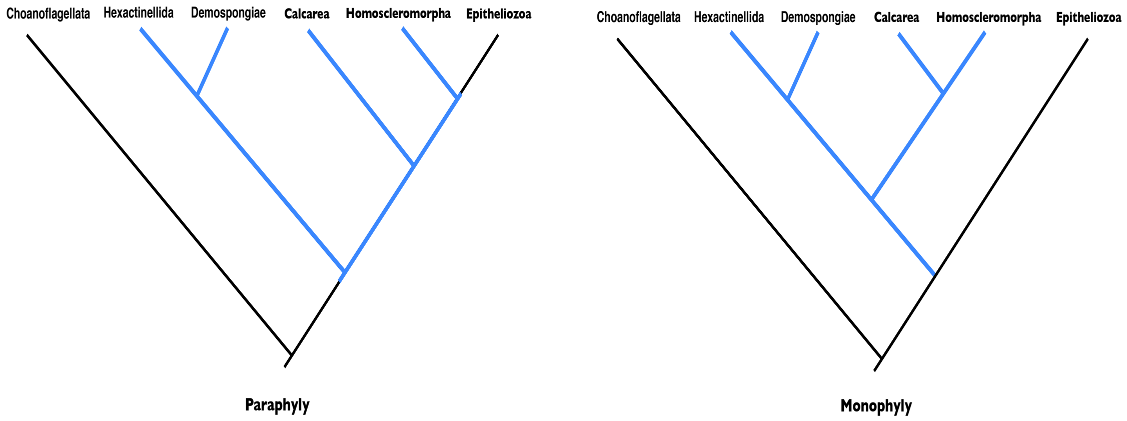


Figure 1.4.- Monophyletic and Paraphyletic sponge relationships adapted from Wörheide et al. (2012).

above, arthropods (as well as onychophorans and tardigrades) were grouped in the Articulata clade. This grouping was based upon the segmented body plan in all the Articulata phyla, but this grouping was broken up (Schmidt-Rhaesa et al., 1998, Scholtz, 2002, Giribet, 2003). The monophyletic group Arthropoda within the Edyzosoa has been supported by a number of characters (e.g. shared presence of sclerotized exoskeleton, legs composed of sclerotized podomeres separated by arthrodial membranes, muscles that attach at intersegmental tendons, and segmentation gene characters amongst others (Giribet and Edgecombe, 2012)). However, the relationships among major arthropod lineages (i.e. Pycnogonida, Euchelicerata (i.e. nonpycnogonid chelicerates (Giribet and Ribera, 2000)), Myriapoda, Crustacea and Hexapoda) have been debated for centuries (Giribet and Edgecombe, 2012). For a long time the monophyly of the clade Atelocerata (i.e. a group that included Hexapoda and Myriapoda) was broadly accepted, but with the addition of new molecular, developmental and anatomical data new topologies have been invoked. For instance, based on the presence of four crystalline cone cells in the compound eye ommatidia in Hexapoda and Crustacea has led to these classes being grouped together into the clade Tetraconata or Pancrustacea (Richter et al., 2009). At the moment there is an unrooted arthropod tree that is congruent with all the data available and agreed on by most authors in this field ((Giribet and Edgecombe, 2012); see Fig.1.5). However, the problem contemplated by most authors is a rooting problem of the five taxa mentioned above (Giribet et al., 2005, Caravas and

Friedrich, 2010). The problem has been narrowed down to three possible topologies: Mandibulata (((Pycnogonida,Euchelicerata),(Myriapoda,(Hexapoda, Crustacea))) Cormogonida ((Pycnogonida,(Euchelicerata,(Myriapoda, (Hexapoda, Crustacea))))) (Zrzavý et al., 1998, Giribet et al., 2001) and Myriochelata, also known as Paradoxopoda ((Hexapoda,Crustacea),(Myriapoda, (Pycnogonida,Euchelicerata))) (Mallatt et al., 2004, Pisani et al., 2004) (see Fig. 1.5) with the Mandibulata hypothesis highly supported by multiple data of different nature (Regier et al., 2010, Rota-Stabelli et al., 2011).

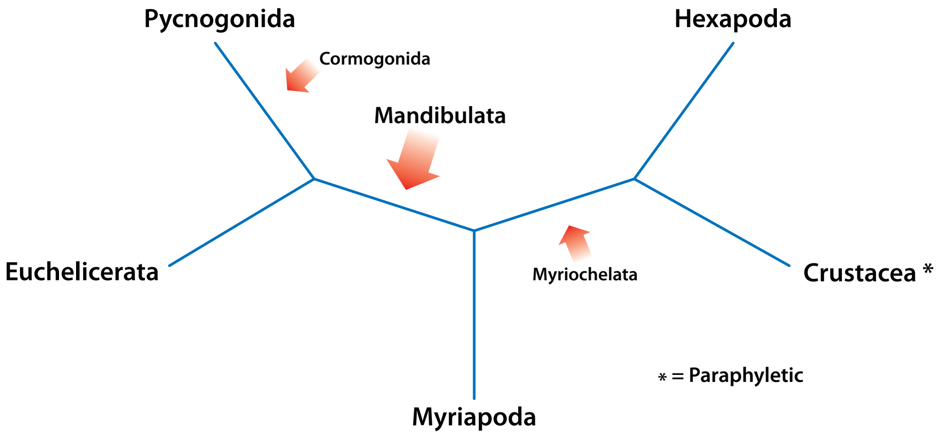


Figure 1.5.- Schema of the unrooted tree of arthropods and its three rooting possibilities indicated by the red arrows. *The different sizes of the red arrows demotes the level of support for each one of the rooting hypotheses. Adapted from Giribet and Edgecombe* ***(2012)***.

1.3 The current sequenced animal genomes

With the summary of the overview of the animal phylogeny in hand, here I will present the availability of genomes at key nodes within it. To date, there is a great diversity in genomes that have been sequenced and many others are anticipated. This is a review (see Fig. 1.6) of the current situation of the genomes released and some of them will be used as a platform for my comparative analysis.

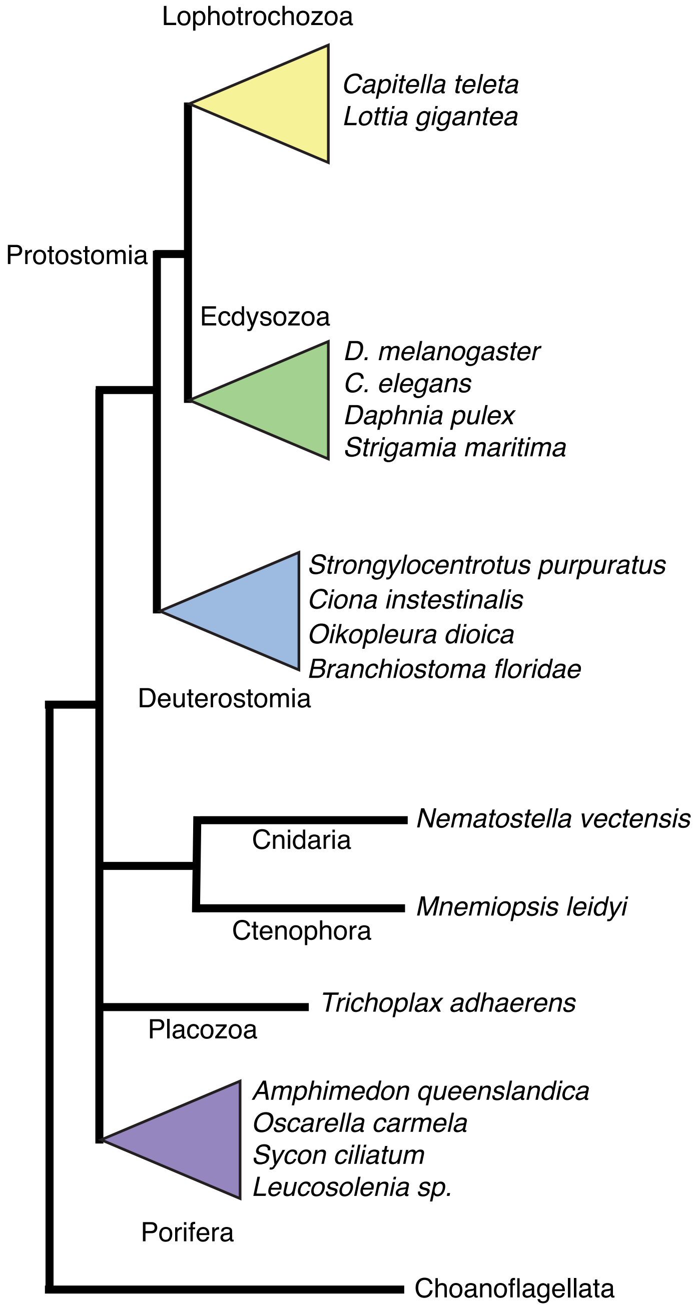


Figure 1.6.- Current sequenced genomes used in this thesis. ***Porifera***: Amphimedon queenslandica *aided our understanding of the ancestral features in animals (Srivastava et al., 2010).* Oscarella carmela *(Feuda et al., 2012) the first homoscleromorph sponge sequenced so far.* Sycon ciliatum *and* Leucosolenia sp. *representatives of the Calcarea group of sponges whose their genome sequences are to be appear eminently.* ***Placozoa****: The recent publication of the genome of the placozoan* Trichoplax adhaerens *(Srivastava et al., 2008) has renewed the interest of many evolution studies aiming to shed light on the primitive structure of genomes, as this genome has not experienced the same degree of intron loss and gene reordering as* C. elegans *and* D. melanogaster. ***Cnidaria****: The completion of the genome of* Nematostella vectensis *revealed a complex genome*

structure including a gene repertoire, exon-intron structure and large-scale gene linkage more similar to vertebrates than to flies and roundworms, suggesting that the eumetazoan ancestor was similarly complex (Putnam et al., 2007). ***Ctenophora***: the much anticipated genome of Mnemiopsis leidyi is still publicly unavailable. ***Lophotrochozoa***: Within this superclade genome sequences have been completed for Capitella teleta and Lottia gigantea by the DOE Joint Genome Institute aiming to provide a better understanding of the major genomic events that took place in this lineage prior to the evolution of the great diversity of body plans(Simakov et al., 2013). ***Ecdysozoa***: Within this superclade lie two of the most powerful genetic model systems D. melanogaster and C. elegans. These two species have been, and will continue to be, premier genetic systems for mechanistic and detailed studies in many fields of biology. It has become clear that they have many unusual traits that make them quite different from other members of this superclade and have biased many of our views on animal evolution. Sequencing projects have been completed of closely related species of these two model organisms, for instance C. briggsae as a closely related species of C. elegans (Stein et al., 2003). Also following this line, the parallel sequencing of the twelve drosophilid genomes (Stark et al., 2007) has provided an opportunity to perform comparative analysis and to analyse chromosomal rearrangements in a genome-scale fashion. The recently sequenced, and about to be released, coastal centipede Strigamia maritima. The ongoing sequencing project of the penis worm Priapulid caudatus should be of evolutionary interest as it occupies a the phylogentic position at the base of the Ecdysozoa. ***Deutorostomia***: The echinoderm Strongylocentrotus purpuratus genome (Consortium et al., 2006) and the ongoing sequencing project of the hemichordate Saccoglossus kowalevskii promises to yield insights about the origin of deuterostomes and chordates and the ancestral state of their genomes as both represent an outgroup for chordates. In addition, the urochordate genomes of Ciona instestinalis (Dehal et al., 2002), Ciona savignyi (Small et al., 2007) and Oikopleura dioica (Seo et al., 2001) can also yield insights about the origin of chordates. The recent sequencing of the cephalochordate Branchiostoma floridae genome (Putnam et al., 2008) possibly provides the most certain platform to understand the evolutionary history of deuterostomes, chordates and vertebrates since its genome contains the basic set of chordate genes involved in development and cell signalling.

1.4 Genome dynamics

Comparative genomics can be used to improve our understanding of the possible mechanisms by which variability and changes in genome architecture are generated. The relative contributions of different mechanisms to the evolution of an animal genome at a macro- and micro-scale remains poorly understood. Duplication of the genetic material is a fundamental route to genetic change, in terms of scale of events as well as its rates of occurrence. The terminology used to describe duplications is varied and sometimes confusing (Mendivil Ramos and Ferrier, 2012). I will not discuss all of this terminology here, however, I provide detailed definitions for each one of the terms in Appendix A (as reviewed in Mendivil Ramos and Ferrier (2012)). Here it is more important to examine the biological processes and evolutionary events involved in duplications, especially whole genome duplication and those duplication events occurring at a sub-chromosomal level, in order to provide a background for interpretation of the evolution of a gene family/superfamily such as the homeoboxes.

1.4.1 Whole genome duplication

One of the most striking characteristics of the human genome, which also extends to other members of the subphylum Vertebrata, is the prominent occurrence of paralogons, homologous regions of chromosomes that are related via duplication events rather than speciation events (Furlong and Holland, 2002). These paralogons largely arose after the occurrence of two rounds of whole genome duplication at the origin of the vertebrates. This hypothesis, named the 2R hypothesis by Ohno (Ohno, 1970), stems from the observation of four paralogons for each region of the human genome being considered. In the first instance a whole genome duplication results in extensive genetic redundancy, which in many instances can gradually be removed by loss of genes, such that only around 30% of ohnologues (i.e. paralogues resulting from 2R) now remain intact in the human genome (Makino and McLysaght, 2010). This results in the paralogous genes created after the 2R whole genome duplications now existing as groups of two to four ohnologues (Furlong and Holland, 2004).

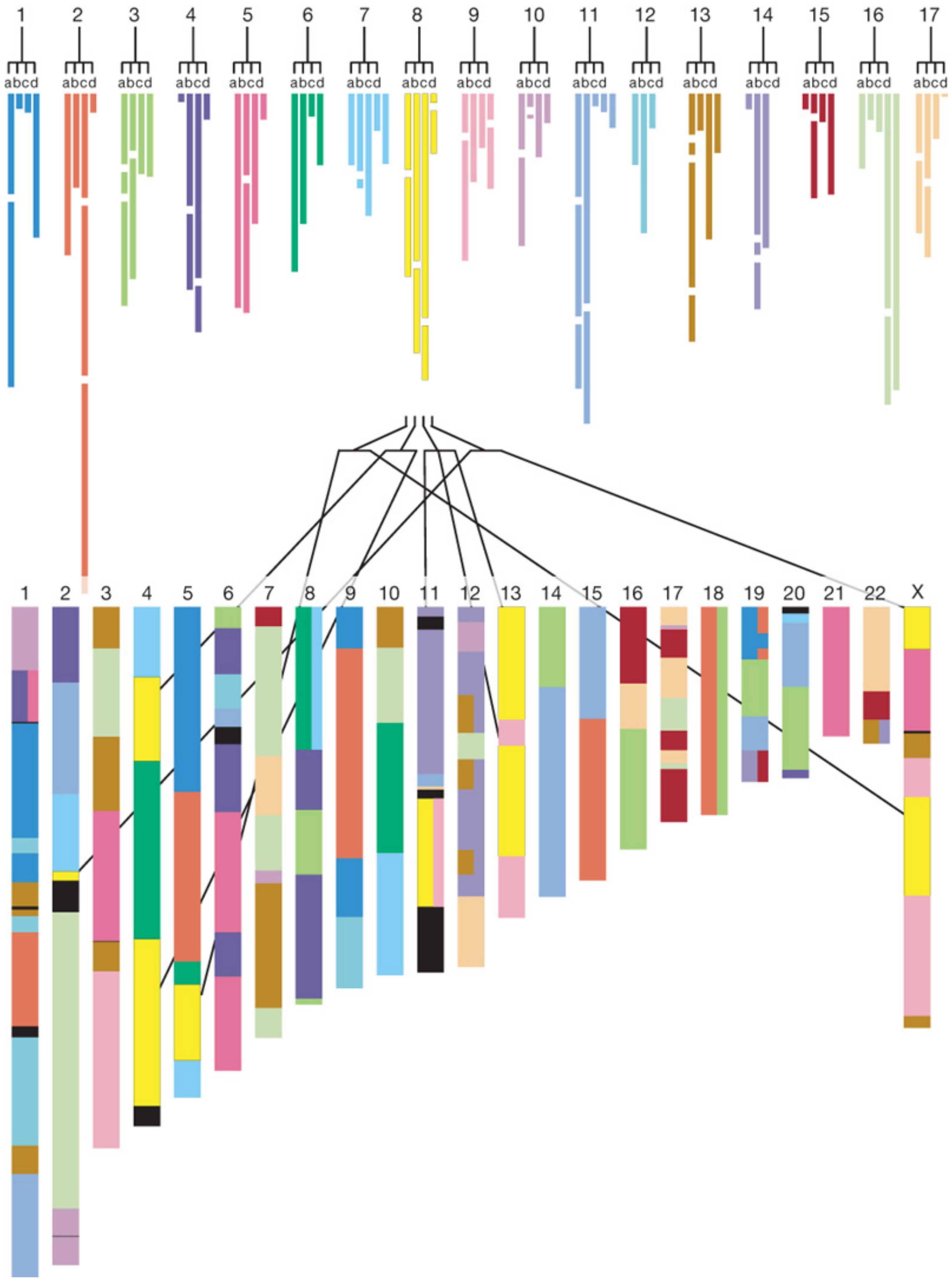


Figure 1.7.- Quadruple conserved synteny in the genome of Homo sapiens relative Branchiostoma floridae as proof of the 2R*. The top part of the figure represents the 17 Ancestral Linkage Groups (AGL) derived from clustering scaffolds according to their synteny with human chromosome segments. The letters a to d represent the four products resulting from the 2R of genome duplication. The bottom part of the figure are the human chromosomes. The colouring of bars and chromosomes segments show identity of the AGL (bottom part of the figure). Adapated from Putnam et al.****(2008)****.*

Upon its formulation, the validity of the 2R hypothesis was often questioned, based on the grounds of phylogenetic inference and strict interpretation of tree topologies of the paralogous families (Hughes, 1999). These strict interpretations of tree topologies incorporated the assumption that the post-duplication paralogues evolve at an equal rate (Hughes, 1999, Abbasi, 2010). This argument has lost strength as evidence increased for unequal or asymmetric evolution of many duplicated genes (Conant and Wagner, 2003). Furthermore, additional evidence in the form of extensive large-scale, genome-wide quadruple conserved synteny in the American amphioxus (*Branchiostoma floridae*) relative to tetrapods destroyed the initial controversy questioning the 2R hypothesis (Putnam et al., 2008) (see Fig. 1.7). However, some authors are trying to revive the debate, arguing for segmental duplications rather than whole genome duplications (Abbasi, 2010). Their interpretations of the molecular phylogenies contain a number of errors (e.g. deductions based on support values at inappropriate nodes, questionable rooting strategies and incomplete datasets), and their model for segmental duplications (SDs) producing quadruple conserved synteny is far less parsimonious than the 2R model.

Furthermore, the process of whole genome duplication or polyploidization is frequent in animal genomes, with increasing numbers of examples being found ((Le Comber and Smith, 2004, Mable, 2004) see Table 1.1). Polyploidizations clearly do occur, have a prominent role in shaping animal genomes, and provide a reasonable explanation for the composition of vertebrate genomes.

1.4.2 Subchromosomal duplications

Segmental duplications (SDs) are sections of duplicated DNA of smaller size than a whole chromosome. SDs can vary in size (i.e. from few base pairs up to many kilobases) and may or may not contain intact, functional genes. Also, SDs can be found in different arrangements, which provides evidence of how SDs might have arisen. Thus, adjacent SDs arise from tandem duplication, whilst SDs separated or interspersed along a chromosome can have arisen from a non-tandem, intrachromosomal duplication, and finally SDs found on distinct chromosomes can result from an interchromosomal duplication. The appropriate detection of these categories depends largely on the quality of the genome sequence assembly. For instance, the SDs in the human genome are estimated to be approximately 5-6% (for SDs >= 1kb, with >= 90% sequence identity, and filtered for transposable elements and other high-copy repeats (Bailey and Eichler, 2006)). In comparison, other mammals have lower levels of SDs than human. Although newly revised SD levels in the mouse genome sequence have been reported to be almost 5% and thus comparable to humans (Bailey et al., 2003). When the rates and distributions of SDs in mammals (rodents and dogs) are compared it is the category of tandem duplications that is predominant (She et al., 2008). A prominent example is the cow genome sequence where tandem duplications comprise 75-90% of the SDs (Liu et al., 2009). However, this situation is not reflected in humans, in which SDs are much more frequently interspersed (Bailey and Eichler, 2006). These high probabilities of interspersed SDs are probably the result of an expansion of Alu transposable elements within primates (Bailey et al., 2003, Bailey and Eichler, 2006). Outside the mammals, the fruit fly *D. melanogaster* has 86% of its SDs in the intrachromosomal category and, moreover, these are situated close together, less than 14kb apart and so are presumably mostly tandem duplicates (Fiston-Lavier et al., 2007).

Another aspect that could be inferred from the categories of the SDs (tandem, interspersed intrachromosomal and interchromosomal) may well be the different mechanisms of DNA-based duplication. One of these mechanisms is non-homologous end-joining (NHEJ) which is more likely to account for adjacent duplications (Ranz et al., 2001, Szamlek et al., 2006, Meisel, 2009b) with the repair of DNA breaks being more likely to occur between ends in close proximity. Other mechanisms include the alternative of non-allellic homologous recombination (NAHR), which is likely mediated via repetitive sequences dispersed around the genome and hence is the mechanism that produces interspersed duplications. This mechanism has also been given the name duplication-dependent strand annealing and is described in the work of Fiston-Lanvier et al. (Fiston-Lavier et al., 2007).

Species/Taxon (Common name)	References
Xenopus laevis (African clawed frog)	Morin et al. (2006)
Tympanoctomys barrerae (Red viscacha rat)	Gallardo et al. (1999)
Daphnia pulex (Water flea)	Vergilino et al. (2009)
Schimidtea polychroa (Planarian Flatworm)	D'Souza et al. (2004)
Acipenser brevirostrum (Shortnose Sturgeon)	Fontana et al. (2008)
Scaphirynchus platorhynchus (Shovelnose sturgeon)	Schultz (1980)
Polyodon spathula (American paddlefish)	Schultz (1980)
Menidia sp. (Atlantic silverside)	Echelle and Mosier (1981)
Barbatula barbatula (Stone Loach)	Collares-Pereira et al. (1995)
Catostomidae (Suckers)	Schultz (1980)
Botia spp.(Pakistani Loach)	Yu et al. (1987), Rishi et al. (1998)
Cobitis spp.(Loach)	Schultz (1980), Vriejenhoek et al. (1989), Janko et al. (2007)
Misgurnus anguillicaudatus (Dojo Loach)	Arai et al. (1993)
Misgurnus fossilis (European weather Loach)	Raicu and Taisescu (1972)
Barbodes spp.(Tinfoil)	Chenuil et al. (1999)
Barbus spp. (Barb)	Suzuki and Taki (1981)
Acrossocheilus sumatranus (Large-scale Barb)	Suzuki and Taki (1981)
Aulopyge hugelii (Dalmatian Barbelgudgeon)	Mazik et al. (1989)
Cyprinus carpio (Carp)	Wang et al. (2012)
Carassius auratus (Goldfish)	Schultz (1980), Yu et al. (1987), Shimizu et al. (1993)
Schizothorax spp. (Snowtrouts)	Mazik et al. (1989)
Synocyclocheilus spp. (Barbels)	Yu et al. (1987), Rishi et al. (1998)
Tor spp. (Mahseer)	Gui et al. (1985)
Zacco platypus (Freshwater Minnow)	Yu et al. (1987), Mazik et al. (1989)
Poecilia spp. (Guppy)	Schultz (1980), Vriejenhoek et al. (1989)
Poeciliopsis spp. (Desert Minnows)	Schultz (1980)
Protopterus dolloi (Slender Lungfish)	Vervoort (1980)
Lepisosteus oculatus (Spotted Gar)	Schultz (1980)
Stizostedion vitreum (Walleye)	Ewing et al (1991)
Salmonidae (Salmons)	Allendorf and Thorgaard (1984)
Clarias batrachus (Walking Catfish)	Pandey and Lakra (1997)
Heteropneustes fossilis (Indian Catfish)	Pandian and Koteeswaran (1999)

Table 1.1.- Examples of species undergoing WGD/polyploidy. Adapted from Le Comber and Smith (2004), Mable (2004) and Mendivil Ramos and Ferrier (2012).

In addition to different mechanisms likely giving rise to different duplicate locations, it is notable that the sizes of the SDs differs between the different categories. Also, it is striking that the size of SDs varies in different species. Lanvier (2007) noted that in *D. melanogaster* the mean size of intrachromosomal events is larger than the mean size of interchromosomal events (3.1 kb versus 2.1 kb, respectively). This contrasts with the average size of SD events in the human being approximately 18.6kb and 14.8kb for the intrachromosomal and interchromosomal SDs respectively (Zhang et al., 2005). A further reference point is provided by *C. elegans*, in which the average size of SDs is only 1.4kb (Katju and Lynch, 2003). This difference in the size of intrachromosomal SDs versus interchromosomal SDs may be linked to their different modes of origin. In addition, the different sizes of SDs between different species most likely reflects differences in the structure and organisation of the different genomes, unless there are also different duplication mechanisms operating in distinct species. A factor that is potentially responsible for the size of the SDs is the density and distribution of repetitive sequences, as they are implicated in duplication processes and also vary across the different species. Another factor that might play an important role in the sizes of duplications is the selective pressures that operate in genes when duplicated within SDs. If a gene is duplicated and then expressed it could often disrupt genetic networks and pathways (e.g. dosage imbalance (Qian and Zhang, 2008)), such that there should be a selective pressure against duplications that encompass genes and their regulatory elements, thus reducing the average size of segmental duplicates of animal taxa with smaller and more compact genes (Mendivil Ramos and Ferrier, 2012).

Alongside consideration of the duplication mechanisms within the context of determining the organisation of duplicated genes, one must also consider processes by which segments of DNA or genes can be translocated around the genome. It must be noted that these mechanisms are not neccesarily leading to the generation of duplicated genes or segments themselves, but are leading to the observed distribution of genes or segments. One of these mechanisms is retrotransposition. Although this is one of the common duplication mechanisms it does not necessarily lead to generation of functional duplicated genes unless the retrotransposed gene co-opts regulatory elements in its new environment. Retrotransposition is crucial in distributing duplicated single genes, eeespecially in an inter-chromosomal fashion (Pan and Zhang, 2007, Bhutkar, 2007, Babushok and Kazazian, 2007, Lorenzen et al., 2008). Also, inversions are very common and help to scatter duplicated genes along a particular chromosome arm (Carvalho et al., 2011). Inversions between arms involving the centromere or chromosome fusions and fissions are also known to play a prominent role in karyotype evolution, and reciprocal translocations between chromosome arms are very common (Olivier-Bonet et al., 2002). Rates of reciprocal translocations in humans are surprisingly high, with estimates of around one in 500 newborns carrying such large-scale rearrangements (Ogilvie and Scriven, 2002). This is not necessarily specific to humans, as reciprocal translocations have been estimated to occur at a rate of 1.4 per 1000 in cattle (Chang et al., 2012). These high rates of translocation are thought to be mediated via NAHR using duplicated or repetitive segments located in different chromosomes, which are collectively called interchromosomal low-copy repeats (LCRs) (Ou et al., 2011). Ou et al. (Ou et al., 2011) showed that in the human genome, interchromosomal LCRs range in size from 5kb to over 50kb, all of which can act as substrates for reciprocal translocations. In addition, Hermetz et al. (2012) described a translocation occurring via homologous recombination between HERV elements on different chromosomes.

These different rearrangement events affecting genome organisation make it difficult to accurately determine the likelihood of a mechanism of origin of a duplicate. This is because it is difficult to determine from the organisation of the duplicate gene/segment(s) within a genome how many rearrangement events have happened since the origin of the duplicates. People have tried to address this problem by using the age of duplicates estimated by calculating the rates of synonymous substitutions (K_s) (Lynch and Conery, 2000, Ezawa et al., 2011, Katju and Lynch, 2003). Such duplicate age calculations have led to the conclusions that younger genes tend to be closer together in the genome, in particular being more represented in the duplicates in the intrachromosomal category rather than the duplicates in the interchromosomal category. An important caveat in the estimation of duplicate age is that it can be confounded by the process of gene conversion, which can homogenise gene sequence after the origin of duplicates (Lynch and Conery, 2000). Since gene conversion is more likely to occur between genes that are in close proximity then there will be a degree of misjudging the age of duplicates as inappropriately young, and this effect will be most prominent in the categories of closely linked genes such as tandem duplicates. Furthermore, the positive correlation between age and dispersal in the genome has recently been questioned with the proposal of a process named drift duplication (Ezawa et al., 2011). Ezawa et al. (2011) compared multiple animal genomes, from human, mouse, zebrafish, *C. elegans*, *Drosophila melanogaster*, and *D. pseudoobscura*, determining the age and genomic location of duplogs (see duplogs Appendix A for definition). This work showed a new pattern of high levels of interspersed intrachromsomal duplicates, which implies an interspersed intrachromosomal mode of duplication with a probability comparable to the observed rates of tandem duplication. This mode of duplication is named drift duplication.

The precise mechanism leading to drift duplication is not specified by Ezawa et al. (2011), and is likely to involve a combination of processes. One of these could well be the recently discovered process of duplication via circular DNA-based translocation. Durkin et al. (2012) recently found that in ‘lineback’ or ‘witrik’ cows a translocation of 492 kb occurred which was then followed by a repatriation of a 575 kb segment, including the KIT gene that is involved in the pigmentation patterning of the cows and their distinctive “lineback” phenotype. The intriguing aspect to these translocations is the order of sequences within the translocated segment, which is consistent with translocation via a circular DNA intermediate, which is opened up for reinsertion at a different point in the circle from the boundaries of the original excision. Also, duplication had happened as the repatriated segment is larger than the original excised fragment (Fig. 1.8).

Further examples of duplications via circular DNA intermediates are being found, such as the *vasa* genes of *Tilapia* (Fujimura et al., 2011). The difference between the cow and *Tilapia* examples, however, is that the cow circular DNA intermediate is repatriated into an ancestral locus, presumably due to homologous recombination, whereas the *Tilapia vasa* duplicates that arose via circular intermediates have gone to new locations. The *Tilapia vasa* example is thus more reminiscent of drift duplication, but it remains to be seen how widespread such circular DNA translocation events are and how the reintegration sites are selected (Fujimura et al., 2011).

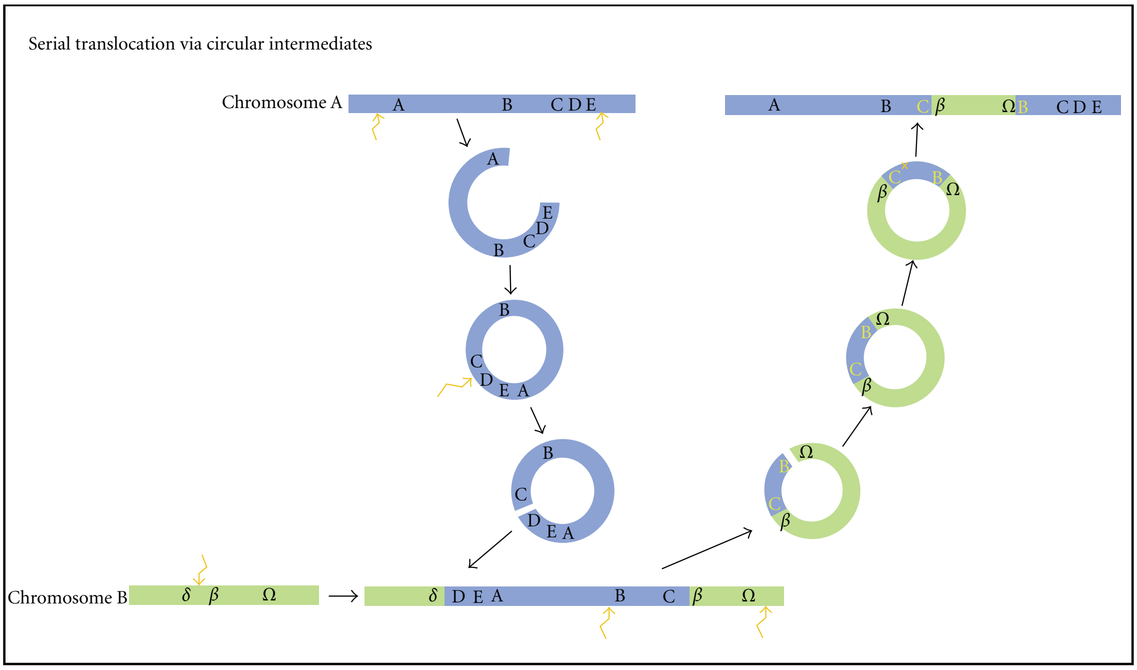


Figure 1.8.- Scheme of a serial translocation via circular DNA intermediates. *Two excisions create a fragment of chromosome A, delimited by genes A and E. This fragment circularizes. At reinsertion into a new genomic location, the circle is linearized by being opened between C and D and inserts between delta and beta of chromosome B. The subsequent translocation involves an excision delimited by genes B and omega. This fragment created circularizes and has sequence identity to the region on chromosome A between the C and B genes. This region of homology allows a repatriation of the segment of original genes from chromosome A, creating a duplication as well as translocating genes from chromosome B. Blue and green lines represent fragments of two different chromosomes. The capital and Greek letters represent genes within the chromosomes. The yellow capital letters denote the genes translocated from chromosome B (green line). The angled orange arrows represent excision points in the DNA. The orange cross represents a homologous recombination site. Adapted from Durkin et al.* ***(2012)*** *and Mendivil Ramos and Ferrier* ***(2012)***.

In light of the range of genomic rearrangement mechanisms and their apparent probabilities described above, it is surprising that syntenic arrangements can be conserved for vast evolutionary timespans. Such syntenic arrangements have been observed from humans to the origin of chordates and beyond, to some basal lineages of animals such as the cnidarian *Nematostella vectensis* and the placozoan *Trichoplax adhaerens* (Putnam et al., 2007, Srivastava et al., 2008). What is also striking is that this phenomenon of long-term general synteny conservation is not uniform across the animal kingdom. Some lineages and groups of animals seem to have particularly derived genome organisations relative to other animals (eg. *Oikopleura* and urochordates in general; *Drosophila* and other Diptera; nematodes like *C. elegans* (Adams, 2000, Seo et al., 2001, Stein et al., 2003, Stark et al., 2007)). One possibility is that this might be a reflection of different abundances of repetitive elements which can have a role in facilitating genomic rearrangements. Another possibility is that gene sizes, and perhaps more importantly gene densities within the chromosomes, vary significantly across the animal genomes. This variation might not just be the number of nucleotides spanned by the coding sequence, but also by the regulatory elements, which will influence how frequently rearrangement mutations can occur that are still compatible with organismal viability. Regardless of this, some animal genomes seem to be more tolerant of, or prone to, rearrangements than others.

With the increasing amounts of human genome sequence data, particularly in relation to disease and cancer genomics, a new phenomenon involving a catastrophic rearrangement of the genome has recently been described: chromothripsis (Stephens et al., 2011). Perhaps the process of chromothripsis has a relevance beyond the realms of cancer and disease biology and may be comparable to processes whereby some animal genomes become extensively rearranged relative to other lineages (see Fig. 1.9). A consideration of the general processes that govern the dynamics of animal genomes with particular attention to duplication and its distribution, are indispensable for understanding the potential augmentation and/or contraction of the developmental toolkit, such as the homeobox superfamily.

1.5 Homeobox genes

1.5.1 Classification of homeoboxes

An essential facet for understanding the evolution of the homeobox superfamily is its classification. A sensible and insightful classification makes it easier to compare structure, expression and function of orthologues when they are being compared between taxa (Ferrier, 2008). Determining and classifying complete homeobox complements across animal taxa sheds light on lineage-specific diversity, gene gains and losses.

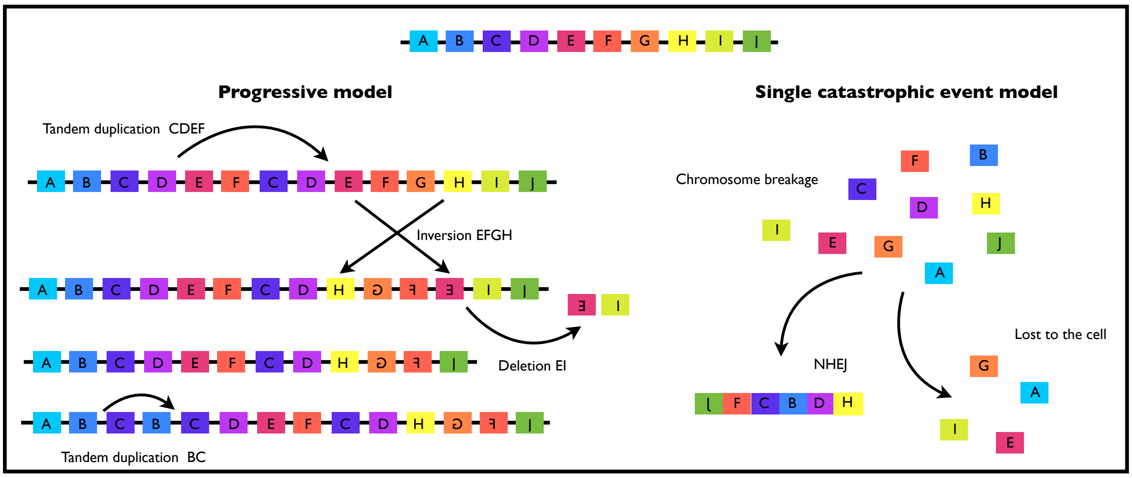


Figure 1.9.- Scheme of chromothripsis models of operation. *Chromothripsis has been described as a catastrophic rearrangement happening in a genome. The coloured squares with the capital letters represent genes within a chromosome. On the left-hand side the progressive model is represented. The chromosome undergoes a progressive series of rearrangements. On the right-hand side is the single catastrophic event model. The chromosome undergoes one single catastrophic event shattering the chromosome and subsequently rejoining some of the fragments via NHEJ and losing some others. Adapted from Liu et al.**(2011)**.*

The homeobox genes are classified mainly by the phylogenetic relationships of their homeodomain region (i.e. usually it is a 60 amino-acid motif that interacts with the DNA in a sequence-specific fashion) and secondarily by other motifs that usually enable protein-protein interactions during development (Bürglin, 1994, 2005). For the majority of the gene families across the animal kingdom this methodology is very robust at the family level (Ferrier, 2008). Based on phylogenies, the metazoan homeobox superfamily is composed of over 100 families and grouped into 11 classes (ANTP, PRD, ZF, TALE, CERS, POU, LIM, CUT, HNF, SINE and PROS) (Holland, 2007). In metazoans, the ANTP and PRD classes have greatly expanded and diversified, likely being instrumental in the animal radiation (Gellon and McGinnis, 1998, Ruiz-Trillo et al., 2008, Fonseca et al., 2008). Despite the importance of classification, some of the classification nomenclature in the ANTP class, which attempts to group genes into subclasses, has proved to be misleading when reconstructing the evolutionary relationships of the gene families (Ferrier, 2008). This is the case with the so-called Hox-like (HoxL) and NK-like (NKL) subclasses. The basis of this nomenclature is a combination of some ambiguous motifs, and poorly resolved family interrelations (i.e. if one takes a strict molecular phylogeny interpretation of reliable node support above 70%) (Fonseca et al., 2008, Ferrier, 2008). Different authors have disagreed in their classifications of genes within these subclasses (e.g. the Dlx gene (Howard-Ashby et al., 2006, Monteiro et al., 2006, Ryan et al., 2006, Larroux et al., 2007, Takatori et al., 2008, Ferrier, 2008)). Apart from creating classification problems, it also creates problems when trying to infer the evolutionary history of a subclass within the evolution of the ANTP class (see section below). Nevertheless, a solution has been put forward for this nomenclature (Ferrier, 2008, Hui et al., 2012). Instead of naming based on the ambiguous inter-family phylogenetic patterns that are prevalent in the ANTP-class, the new proposal uses the actual linkage of ANTP-class families (i.e. instead of Hox-like and NK-like it should be Hox-linked and NK-linked). In this way, the acronym does not change and its meaning is unambiguous and accurate, as outlined below.

1.5.2 Evolution of the ANTP class

Establishing the evolutionary relationships within the ANTP class has been the primary focus of many studies, with the aim of understanding the intriguing arrangement of some of its members in clusters within the genome. Hypotheses about the evolution of this class are heavily influenced by the choice of hypothesis for the phylogeny of the basal animal lineages and the continued isolation of new data from species at the base of the animal phylogeny.

1.5.2.1 The "Megacluster" hypothesis

The grounds of this idea stems from chordate genome data. In particular, from the recovery of clustered phylogenetic relationships of families of two large groups within the ANTP class, Hox and their relatives and NK and their relatives (Pollard and Holland, 2000, Garcia-Fernandez, 2005), and the observation of ANTP homeobox genes being linked in different chordate taxa (e.g. *Branchiostoma floridae* (Castro and Holland, 2003, Luke et al., 2003, Castro et al., 2006)). Based on these data, it has been postulated that before the origin of the Urmetazoan (the last common ancestor of all animals) the ancestral state of ANTP, the proto-ANTP gene, originated and underwent several tandem duplications leading to a cluster, the "Megacluster", of precursors of the different families (Hox, ParaHox, NK and other related families) that are currently observed ((Pollard and Holland, 2000, Hui et al., 2012); Fig. 1.10). Then, the "Megacluster" broke apart in several locations as lineages diverged from the ancestral state. These breaks presumably were at random positions of the "Megacluster" across the different lineages, apart from some functionally constrained clusters like the Hox. In this way, different lineages would be expected to contain distinct, but overlapping, remains of the "Megacluster" such that the evolutionary history of this cluster could be deduced from comparisons across the animal kingdom (Hui et al., 2012).

There are two alternative lines of thought that explain the origin and evolution of this cluster, the so-called "strong" and "weak" forms. The "strong" form of the "Megacluster" is the one mentioned above. The other, the "weak" form of the "Megacluster", proposes that the complete "Megacluster" never existed in its entirity. Under this premise significant portions of the ANTP class still did evolve in clusters, but not all of the families that are hypothesized to have been involved in the "Megacluster" actually ever existed all together in a single intact "Megacluster". There are two possible routes by which this could have occurred:

(i) A precursor cluster broke before all families evolved, and/or

(ii) Non-tandem duplications could have been involved in the origin of genes or groups of genes within the whole complement of families of the "Megacluster".

The two forms ("strong" and "weak") of the "Megacluster" will lead to similar but not identical remains of the ANTP-class gene linkage patterns in different animal lineages. The obvious way to distinguish between these two possibilities is by examining the patterns of linkage and chromosomal locations of the homeoboxes within different animal genomes outside the chordates. This type of analysis encounters three difficulties. One is the sub-chromosomal level of assembly of many draft genome sequences of the animals used in these analyses, and the second is the ambiguous phylogenetic resolution of some members of the ANTP-class. The third difficulty involves finding animal genomes that are not significantly derived and rearranged relative to deep animal ancestral states. Regarding this third difficulty, the choice of an animal genome to test the "Megacluster" hypothesis is crucial as it needs a chromosome number potentially similar to that of the protostome-deuterostome ancestor (or

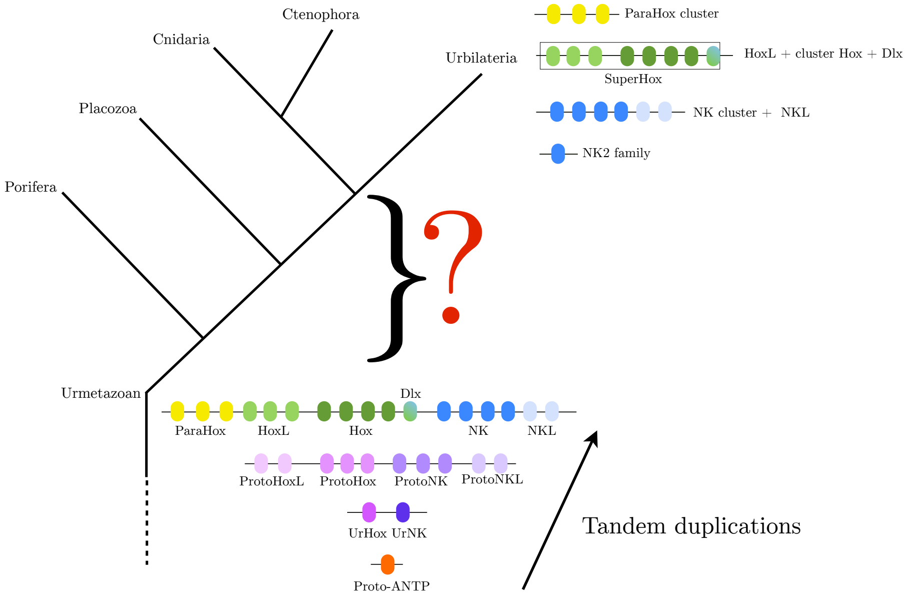


Figure 1.10.- The "Megacluster" hypothesis. *The ovals represent genes and the colouring represent different families of the ANTP-class genes. The orange, pink and purple ovals represents different precursors. Yellow ovals represent ParaHox genes, dark green ovals represent Hox genes, pale green ovals represent HoxL genes, mixed green, blue represent Dlx, dark blue represent NK genes and pale blue represent NKL genes. The "Megacluster" cluster existed before the Urmetazoan. Pressumably broke apart over the basal lineages in an unknown fashion and thus, the question mark. In the Urbilaterian the different families resided in different chromosomes according to Hui et al.* ***(2012)****. Also in the Urbilateria existed the "SuperHox" cluster ("EuHox" plus 8 Hox-linked genes)* ***(Butts et al., 2008)****.*

Urbilateria), thus reducing the chances of genome macrorearrangements (e.g. fissions and fussions). The second difficulty stems from a homoebox gene key in the "Megacluster" hypothesis, Dlx. Based on poor phylogenetic support, this gene has been hypothesized to be related to those genes in the NK cluster and traditionally named as an NK-like. However, in chordates this gene is linked to the Hox cluster and this linkage has been used as evidence for the NK and Hox linkage in the "Megacluster", with an interchromosomal translocation event supposedly separating all NK-like genes except Dlx from the Hox-like genes (Pollard and Holland, 2000, Garcia-Fernandez, 2005). However, the ambiguity in the phylogenetic placement of Dlx and the rather tenuous nature of its classification as a family with closer ties to the NK genes than the Hox genes casts serious doubt on the veracity of the "Megacluster" hypothesis, at least in its strong form.

In a recent study using chromosomal fluorescence *in situ* hybridisation in the lophochotrozoan *Platynereis durmerilii*, as a means to have an independent source of linkage data from the chordates that does not suffer from the third difficulty outlined above (which clearly affects other protostomes like fruit flies and nematodes whose genomes are highly rearranged and chromosome numbers reduced), showed similar patterns of breakage of the ANTP-class genes as those found in chordates. Thus, if the "Megacluster" ever existed in the Urmetazoan then it had already broken into four chromosomes in the Urbilaterian (i.e. the common ancestor of protostomes and deuterostomes)(Hui et al., 2012) (see Fig. 1.10).

1.5.2.2 The ProtoHox hypothesis

Before the formulation of the "Megacluster" hypothesis, the ProtoHox state (i.e. the precursor of Hox and ParaHox) was postulated following the discovery of the ParaHox cluster, the paralogous cluster of Hox, in *Branchiostoma floridae* (Brooke et al., 1998). The ParaHox cluster consists of the genes Gsx, Xlox and Cdx and molecular phylogenetic analysis shows that these genes are more similar to the Hox genes, consistent with the idea of being paralogues ((Brooke et al., 1998) see Fig 1.11). Recently, Osborne et al. (2009)

showed that the ParaHox cluster exhibits spatial and temporal colinearity (see Fig 1.11). Therefore, deep in animal ancestry after the hypothetical ProtoHox cluster duplicated and produced the actual Hox and ParaHox clusters. However, the exact timing of this duplication has been the subject of much debate, which will be examined later in this thesis.

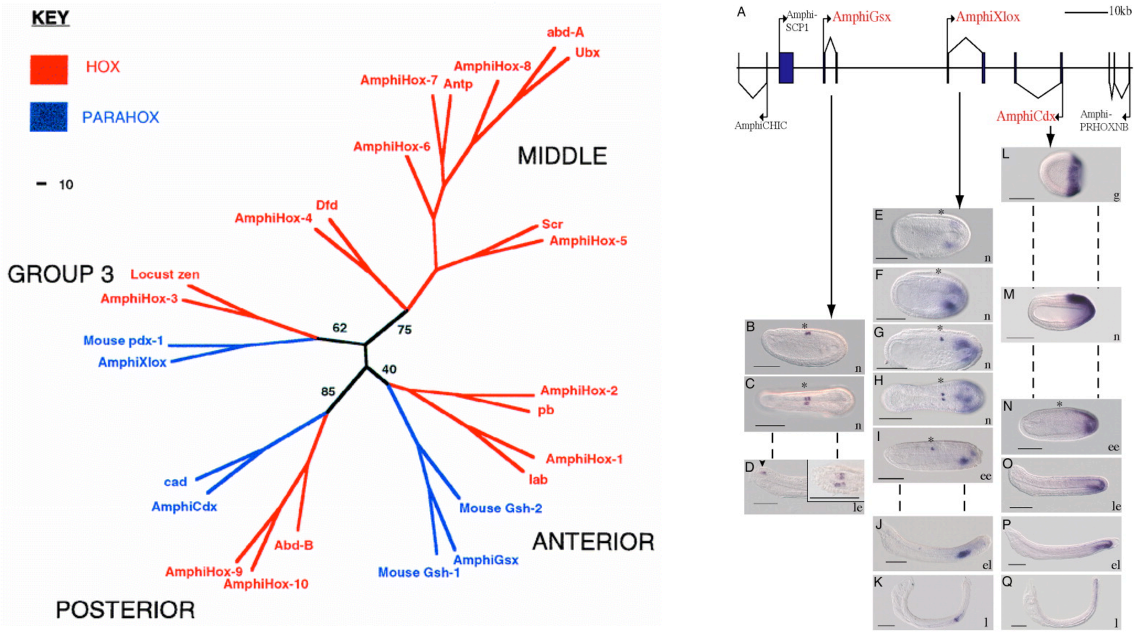


***Figure 1.11.- The ParaHox cluster in Branchiostoma floridae.** Left hand side phylogenetic relationships of the ParaHox genes (in blue) with the main groups of the Hox genes (in red) showing paralogous relationship adapted from Brooke et al. **(1998)**. Right hand side ParaHox cluster organisation and expression data of Branchiostoma floridae adapted from Osborne et al. **(2009)**.*

Recent whole genome sequence data from basal animal lineages, such as *Nematostella vectensis*, *Trichoplax adhaerens* and *Amphimedon queenslandica*, have allowed new genome-wide surveys of ANTP class genes and further reconsiderations of evolutionary models of the ANTP class in comparison with the information provided from bilaterians. In particular, this new data has led to competing views on different aspects of the ProtoHox: (i) where did the ProtoHox come from? (ii) when did it arise, in relation to which animal lineages? and (iii) was ProtoHox a gene or a cluster?

1.5.2.2.1 Where did the ProtoHox come from?

Regarding the aspect of where did the ProtoHox come from there are two lines of thought. First, Gauchat et al. (2000), hypothesized that the ProtoHox gene originated from another pre-existing ANTP class gene, Evx. This model is based upon the clustering observed in cnidarians and bilaterians of Evx and the Hox cluster (see Fig. 1.12). Second, there is the model proposed by Larroux et al. (2007), in which the ProtoHox cluster originated from the NK cluster. This follows two assumptions: *Amphimedon queenslandica*, a desmosponge, contains only NK genes (and no Hox/ParaHox gene families) in a cluster and Porifera is the most basal animal lineage ((Larroux et al., 2007), see Fig. 1.13).

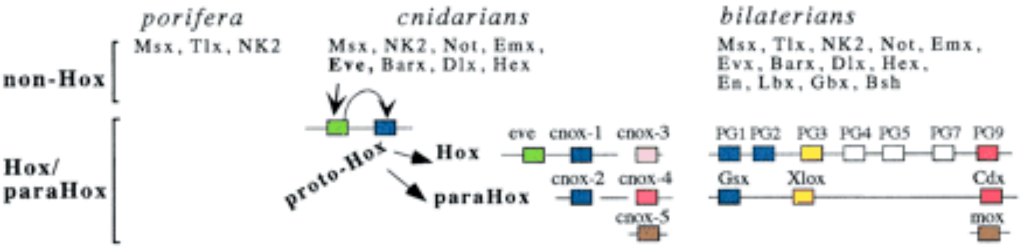


Figure 1.12.- Scheme describing the origin of the Hox and ParaHox clusters according Gauchat et al. (2000). *Adapted from* ***(Gauchat et al., 2000)****.*

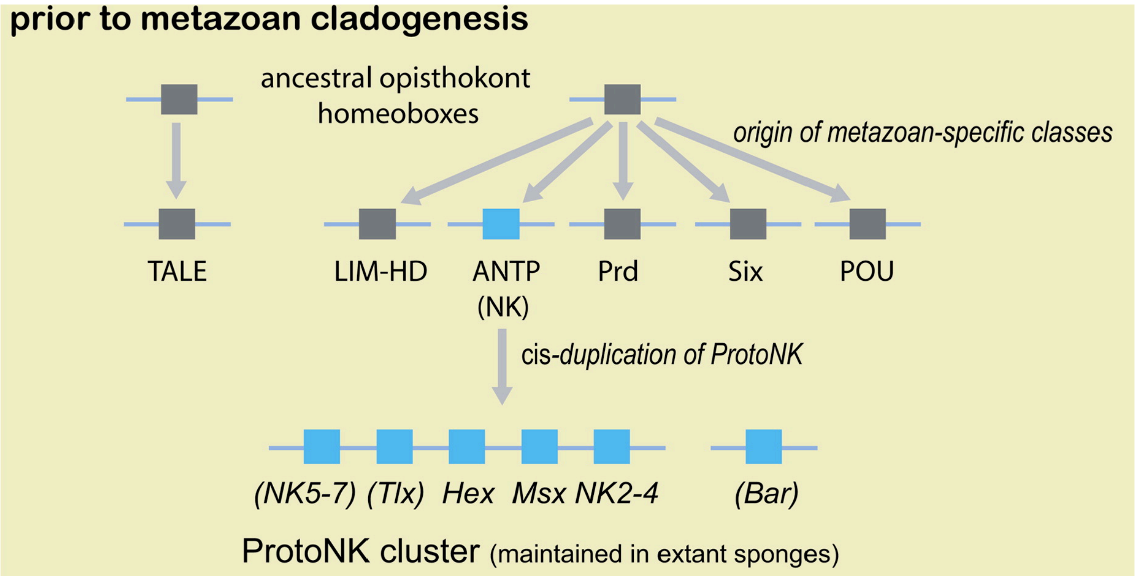

prior to metazoan cladogenesis
ancestral opisthokont
homeoboxes
origin of metazoan-specific classes
TALE
LIM-HD
ANTP
(NK)
Prd
Six
POU
cis-duplication of ProtoNK
(NK5-7)
(Tlx)
Hex
Msx
NK2-4
(Bar)
ProtoNK cluster (maintained in extant sponges)

Figure 1.13.- Scheme describing the origin of the Hox and ParaHox clusters according Larroux et al. (2007). *Adapted from Larroux et al.* ***(2007)****.*

1.5.2.2.2 When did the ProtoHox arise?

Regarding the aspect of when did the ProtoHox arise with respect to particular animal lineages there are various lines of thought. Peterson and Sperling (2007) based on strict interpretation of phylogenetic analyses of homeodomain genes from basal animal lineages, hypothesized that the ProtoHox originated before Porifera. The interpretation of these analyses assumed independent homeodomain gene losses in Placozoa and Porifera. This contradicts the view of Larroux et al. (2007) in which the ProtoHox arose after Porifera. Dellaporta et al. (2006) hypothesized that ProtoHox originated in the Placozoa. Furthermore, due to the lack of Hox/ParaHox genes in the ctenophore *Mnemiopsis leidyi* and the potential early branching of this lineage relative to the rest of the metazoans, Ryan snd colleagues proposed that the ProtoHox arose after this lineage, claiming Placozoa, Cnidaria and Bilateria now comprise the ParaHoxozoa (Ryan et al., 2010). A further viewpoint is that of Kamm et al. (2006), who hypothesized that the Hox-like genes of cnidarians arose from independent duplications from those that generated the Hox and ParaHox genes of bilaterians.

1.5.2.2.3 Was the ProtoHox a gene or a cluster?

Another aspect of the evolution of ProtoHox is whether the ProtoHox was a cluster or a gene. In the original hypothesis of the ProtoHox state a cluster was proposed (Brooke et al., 1998). This ProtoHox cluster was hypothesized to have consisted of four genes, which included precursors for Gsx/ Hox1-2, Xlox/Hox3, central Hox and Cdx/Hox9+. When this cluster duplicated to give rise to the Hox and ParaHox clusters, the Hox3 and Hox4-8 genes would have been lost from the cnidarian Hox cluster and a ParaHox gene paralogous to the central Hox genes would have been lost from the Cnidaria-Bilateria Ancestor. I will refer to this model as the four-gene model. Subsequent genomic data from the base of the animal phylogeny led to the two- and three-gene models, which I will examine in turn.

There are two two-gene ProtoHox models. Both of them were based upon cnidarian ANTP class sequence data. The two-gene model proposed by Garcia-Fernàndez et al. (2005) hypothesized that cnidarians possessed only anterior and posterior Hox and ParaHox genes, lacking orthologues of Xlox/Hox3 or central Hox families. Thus, the two genes that comprise the ProtoHox cluster were a precursor of Gsx and Hox1/2 and precursor of Cdx and Hox9+. The two-gene model of Chourrout et al. (2006) proposed that one of the ProtoHox genes was the precursor of central Hox/Xlox and the other one the precursor of Gsx and Hox1/2.

In the three-gene models of Finnerty and Martindale (1999) and Ferrier and Holland (2001), it is hypothesized that the central Hox genes evolved within the Hox cluster after the ProtoHox to Hox/ParaHox duplication.

An alternative to the whole cluster duplication models mentioned above is the cluster splitting model of Ryan et al. (2007). This model postulates a ProtoHox gene that underwent sequential tandem duplication, expanding to a cluster that contained the precursors for the different Hox and ParaHox families. Once this cluster existed it broke apart and produced the separate Hox and ParaHox clusters (see Fig. 1.14). I refer to this model as the one-gene model.

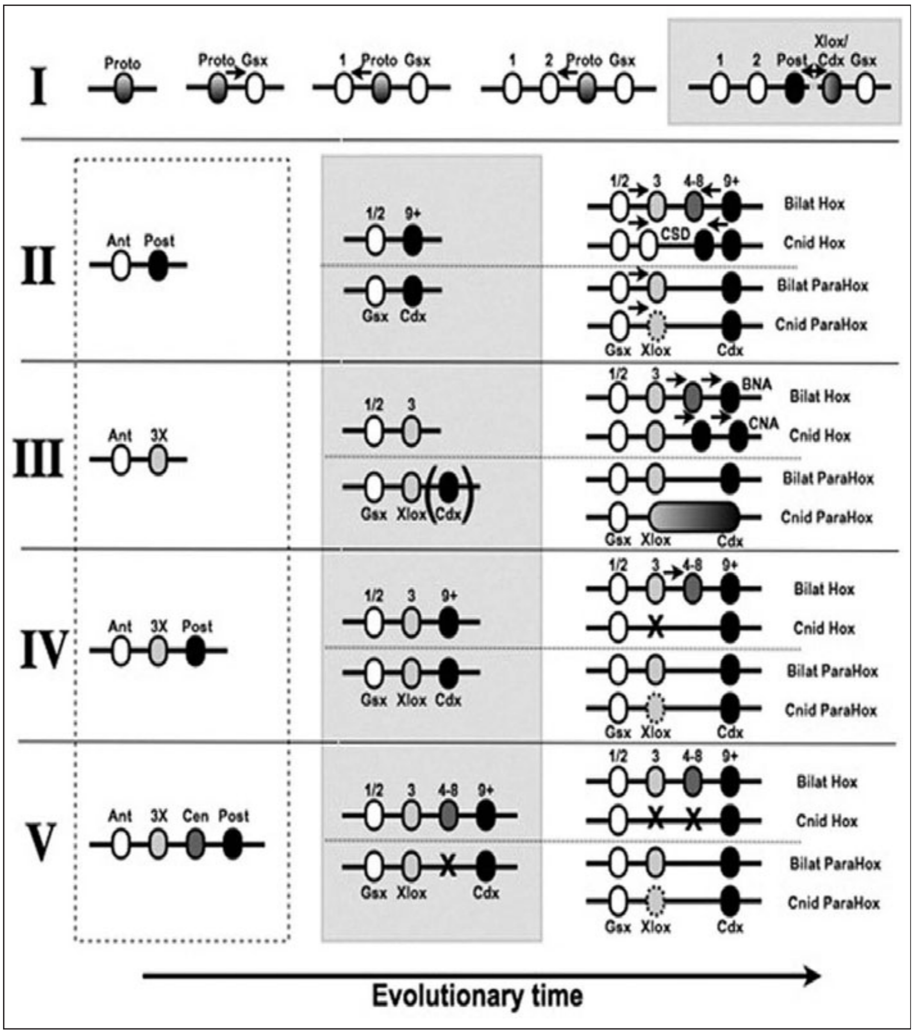


Figure 1.14.- Was the ProtoHox a gene or a cluster? *Figure adapted from Ferrier et al. (2010).*

This variety of models stems from a number of problems, one of which is the lack of robust resolution in many homeodomain phylogenies. Other problems include few basal animal lineages and relatively sparse sampling from these lineages, as well as a poor understanding of the dynamics of animal genome rearrangements and the probabilities of various rearrangement mechanisms across these lineages and in ancestral animal genomes (e.g. duplication rates in the basal lineages). Recently, all of these competing models have been statistically compared in a study by Lanfear and Bromham (2008). In particular, they compared the support for the different hypotheses using two statistical methods that contrast the Maximum-likelihood and Bayesian topologies from each one of the ProtoHox models. These tests favoured the three-gene and four-gene models and thus rejected the two-gene models and the one-gene model.

In concert with the statistical rejection of the two-gene and one-gene models, there are several pieces of ANTP sequence stemming from wider taxon sampling in Cnidaria that rejects the two-gene models. First, Cdx orthologues have been reported in two cnidarians: EdCnox4 in *Eleutheria dichotoma* and Anthox4 in *Metridium senile* (Kuhn et al., 1996, Finnerty and Martindale, 1997, Finnerty and Martindale, 1999). Thus, this data is inconsistent with the two-gene model of Chourrout et al. (2006) in regards of the origin of the Cdx gene. Second, Xlox has been proved to be present in some cnidarians (Quiquand et al., 2009). However, recently it has been disproved according to the data presented in Chiori et al. (2009). In this way, if one adopts the Quiquand et al. (2009) hypothesis the presence of Xlox in cnidarians refutes the two-gene model of Garcia-Fernàndez.

1.5.2.3 The "SuperHox" cluster

Butts et al. (2008) recently postulated the existence of a "SuperHox" cluster in the last common ancestor of the bilaterians, the Urbilaterian, from comparison of two bilaterians (*Tribolium* and *Branchiostoma*). They proposed that the "SuperHox" cluster is composed of the canonical Hox cluster or "EuHox" genes and eight other ANTP class homeobox genes (i.e. Mox, Hex, Ro, Mnx, En, Nedx, Dlx and Evx) (Fig. 1.10). It was unclear from the data available whether the "SuperHox" and the NK genes were linked in the Urbilateria.

1.5.3 Hox clustering diversity

In order to examine the different biological aspects of gene clustering, it is necessary to clarify the working definition of gene cluster that I will use for the rest of this thesis. The term “gene cluster” has a range of definitions in different research contexts, but for the purposes of this work, “gene cluster” refers to genes in the same family (operationally defined as having sequence similarity) that are clustered together in the genome. Gene clusters, defined as such, have been observed in diverse gene families within animal genomes. Apart from the homeobox gene families Hox, ParaHox and NK, other examples of developmental gene clusters have been described, such as: Wnt (Nusse, 2001, Sullivan et al., 2007), FGF (Itoh and Ornitz, 2004), forkhead (Mazet et al., 2006), bHLH (Simionato et al., 2007), Runx genes (Bao and Friedrich, 2008), GATA factors (Gillis et al., 2008), SP genes (Schaeper et al., 2010), and *achaete-scute* (Negre and Simpson, 2009).

The Hox cluster is one of the most renowned cases of developmental gene clustering and its organisation in an ordered cluster correlates in many cases to the spatial and temporal sequence of gene expression during embryo development. This facet of Hox biology is termed colinearity (i.e. the genes at the 3' end of this cluster control the differentiation of the anterior part of the embryo and the genes at the 5' end of this cluster control the diferentiation of the posterior part of the embryo). The organisation of this cluster and the position of the genes within it are correlated with the transcriptional regulation of these genes (Graham et al., 1989, Duboule and Dollé, 1989). Given its importance this cluster has been investigated across various animal lineages, and it has become clear that the integrity of the Hox cluster has not been maintained in many species. Describing the arrangement (broken or not) of this cluster in a wide range of bilaterian animals can provide an insight into the nature of an animal’s genome and its degree of conserved organization relative to ancestral states.

In the protostome *D. melanogaster* this cluster has been split into Antennapedia and Bithorax complexes, and the other recently sequenced drosophilids show that their Hox cluster has undergone different rearrangements with their Hox clusters splitting in different places (Negre and Ruiz, 2007). Early studies reported that in the Hox cluster of *Bombyx mori*, the labial gene is at a different location from the main Hox cluster in chromosome 6 (Yasukochi et al., 2004). More recently it was shown that this split of 12 Megabases took place between *labial* and *pb* (Chai et al., 2008). Not all insects have broken Hox clusters however, intact Hox clusters have been found in *Tribolium* (Shippy et al., 2008), *Schistocerca gregaria* (Ferrier and Akam, 1996), and *Apis mellifera* (Dearden et al., 2006). The Hox cluster of the crustacean *Daphnia pulex* is intact (supplementary information in (Colbourne et al., 2011)). So far, and despite the current explosion of genome-scale data, appropriate surveys of Hox clustering in arthropods have not yet been reported from outside the insects or crustaceans, prior to this thesis (see chapter 6). In other invertebrates like *C. elegans* the Hox cluster is dispersed and several genes have been lost (Aboobaker and Blaxter, 2003). Within the lophotrochozoans the Hox cluster surveys of *Lottia gigantea* showed a highly conserved cluster (Simakov et al., 2013) and *Capitella teleta* has a cluster broken towards the posterior end of the cluster (Fröbius et al., 2008). However, this is not the case of the other lophotrochozoan the leech, *Helobdella robusta*, whose cluster is completely disorganised (Simakov et al., 2013).

In deuterostomes the Hox clusters are highly variable in terms of gene order and clustering. The Hox cluster of the echinoderm *Strongylocentrotus purpuratus* has been reported to have four posterior Hox genes embedded in between three anterior genes and the four central genes, with the loss of *SpHox4* (Cameron et al., 2006). Recently, the organisation of the 12 gene Hox clusters of two hemichordates (*Saccoglossus kowalevski* and *Ptychodera flava*) has been reported (Freeman et al., 2012). Within these clusters the Hox1 to Hox 9/10 genes posses the same genomic organisation and transcriptional orientation as their orthologues in chordates and the 5'/posterior end of each cluster contains three posterior genes that are specific to Ambulacraria (the hemichordate clade). In the urochordates the Hox gene organization has been reported in *Ciona intestinalis* (Ikuta et al., 2004) and in the larvacean *Oikopleura dioica*

(Seo et al., 2004), showing in both cases a dispersed Hox cluster. In contrast to the urochordates, the cephalochordate amphioxus (Garcia-Fernàndez and Holland, 1994, Ferrier et al., 2000, Holland et al., 2008) has an intact Hox cluster, consisting of 15 Hox genes in their ancestral order and orientation. At the origin of the vertebrates two rounds of whole genome duplication occurred, and this has resulted in four Hox clusters in tetrapods, all of which have undergone some gene loss.

Some likely mechanisms contributing to functional constraints on clustering are known, largely from the work in mice. Vertebrates, other than teleosts, have four paralogous Hox clusters due to two whole genome duplications at the origin of the vertebrates (Dehal and Boore, 2005, Putnam et al., 2008). After the two rounds of whole genome duplication, the Hox cluster and other families underwent evolutionary innovation, neofunctionalization and subfunctionalization of their members. Evolution of mechanisms that conserve the clustered organisation present in vertebrates may have accompanied these innovations (Deschamps, 2007). Kmita and Duboule (Kmita and Duboule, 2003) outline three main mechanisms of cluster regulation: (i) local sharing of *cis* regulatory elements, (ii) long-range global enhancers (as shown in the globin gene cluster as well as the Hox cluster) and (iii) chromatin modulation, as this plays an important role in transcription with the decondensation proceeding from the anterior end to the posterior end of the cluster.

A specific example of local enhancer sharing in mice was provided by Sharpe et al. (Sharpe et al., 1998). Tarchini and Duboule (Tarchini and Duboule, 2006) showed the existence of global control regions at either end of the HoxD cluster in mouse. The correlation of the sequential opening of the chromatin with the collinear Hox expression in mouse and human has been demonstrated *in vivo* (Chambeyron and Bickmore, 2004) and *in vitro* (Chambeyron et al., 2005, Morey et al., 2007).

Besides these specific regulatory mechanisms that have been described in mice, it has been noted that in many species in which the Hox cluster has been broken, remnants of spatial colinearity have remained (Monteiro and Ferrier, 2006). However, in no case described to date has temporal 'colinearity' been found in a species with a broken cluster. This has led to the hypothesis that it is the mechanism(s) that produce temporal colinearity that is primarily responsible for the constraints that maintain an intact, ordered Hox cluster. Thus, some bilaterian Hox clusters have been accumulating viable rearrangements that break them meanwhile others have been constrained by the evolution of either ancestral or lineage-specific pan-cluster regulatory mechanisms (Monteiro and Ferrier, 2006). The nature of these pan-cluster regulatory mechanisms in invertebrates have yet to be determined.

1.6 Thesis structure

The structure of this thesis is as follows. First, in Chapter 2 I describe the general methods used in this thesis. In Chapters 3 and 4 it is addressed the long term debate about the origin of the Hox and ParaHox loci, making use of genome-wide comparisons and statistical analyses to reconstruct ancestral states and propose a point of Hox/ParaHox origin. Second, in Chapter 5 deals with my contribution to identifying possible ParaHox orthologues in calcareous sponges (*Sycon ciliatum* and *Leucosolenia sp.*), which provides an independent test of the new hypothesis formulated in Chapters 3 and 4 about the origin of these loci. Third, in Chapter 6 describes a survey of the homeobox complement of *Strigamia maritima,* compares it with the rest of the arthropods and describes instances of ancestral linkages and clustering within this new genomic sequence. In Chapter 7 I discuss the implications of my work for the field of 'evo-devo'.

Chapter 2

Materials and Methods

2.1 Orthologue analysis

2.1.1 Retrieving putative orthologues

To identify putative orthologues, I performed reciprocal BLAST (Altschul et al., 1990) searches against the genome. The general principle of the reciprocal best BLAST hit, also known as rbh, is that a "protein i in the genome I is the rbh of the protein j in the genome J if query of genome J with protein i yields as the top hit protein j and reciprocal query of genome I with protein j yields as a top hit protein i" (Wall et al., 2003). This approach is used in two ways in this thesis, depending on the nature of the analysis. If the analysis involves a large number of genes (over 40 genes) the rbhs is implemented in a Python script in order to retrieve them automatically based on an e-value of equal to or less than 10^{-05} and bit-score of equal to or over 70. Under the assumption that sequence similarity indicates homology, whatever protein that meets the combination of the thresholds of e-value and bit-score will be homologous sequences. If instead the analysis involves fewer than 40 genes, each one of these alignments are critically assessed by eye and further analysed using phylogentic trees (see Section 2.1.2). This type of screen helps to establish one-to-one, one-to-many or many-to-many relationships, or absence of orthologues, which is of importance for subsequent statistical analyses.

2.1.2 Identity of orthologues

Family members were aligned using MAFFT (v6.846b, default settings (Katoh et al., 2002)) and viewed in Jalview (v2.6.1, (Clamp et al., 2004)) to edit alignments for phylogenetic tree building. Alignment editing was refined by either cross-comparisons with multiple alignments built by GBLOCKS (v0.91b, (Castresana, 2000)) in order to remove saturated sites and uncertain columns or by eye (e.g. homeobox genes). An additional and complimentary way of looking for orthology is comparison of shared combinations of domains/motifs that can provide a distinctive signature for some orthology relationships. I used SMART (v7.0,(Schultz et al., 1998)) to help confirm these motifs and domains. In cases with family members with relationships of the form one-to-many or many-to-many, phylogenetic trees were constructed using Modelgenerator (v0.85, choosing the BIC criteria (Keane et al., 2006)) followed by Neighbour-Joining in PHYLIP (v3.69), Maximum Likelihood in PhyML (v3.0, (Guindon et al., 2010)) and Bayesian Markov chain Monte Carlo in MrBayes (v3.1.2, (Huelsenbeck and Ronquist, 2001)). Node support for NJ trees was estimated from 1000 bootstrap replicates, using the JTT model of sequence evolution. Node support for ML was estimated from 100 bootstrap replicates and for Bayesian trees I used 1000000 generations; 5000 for sample probability; burn-in of 50 samples; two runs of four chains each. This tree building helped to resolve some of the one-to-many and many-to-many relationships as one-to-one orthologies.

2.1.3 Orthologue location retrieval

Orthologous gene locations in the human genome were noted from MapViewer from the NCBI website (www.ncbi.nlm.nih.gov/mapview). For other genomes, orthologue scaffold locations were inferred from the relevant gff3 files retrieved from the ftp genome project site (see Section 2.3) via specific python scripts (specified in Chapters 3 and 4).

2.2 Synteny statistical analysis

Synteny conservation was examined statistically with two tests: the Exact Binomial test and Fisher's Exact test. These tests were conducted in R (v. 2.13.0) using specific codes (specified in Chapters 3 and 4). The derivation of the numbers of genes per chromosome or scaffold is different depending on the particular genome. If it is the human genome sequence in the synteny analysis, the gene numbers were taken as the number of protein coding genes (detailed in Appendix B, section B.1). For other genomes that are not human it is assumed that every annotated gene is a protein coding gene. These numbers are inferred via the respective genome sequence gff3 files. Once the number of genes per chromosome or scaffold was obtained, the expected probabilities of selecting a randomly chosen gene from a particular chromosome or scaffold were calculated.

The Exact Binomial test (Sokal and Rohlf, 1995) was used to estimate whether there is a statistical deviation from a theoretical expected distribution of observation into two categories (i.e.: location and non-location).

The Fisher's Exact test (Sokal and Rohlf, 1995) was used to estimate whether there is a statistically significant association (i.e. contingency) between the apparent concentration of orthologues within an animal genome scaffold(s) which is similar to a particular human chromosomal region. From the contingency tables the expected numbers are calculated as follows: for a particular cell by multiplying its row by its column totals and dividing the product by the grand total.

2.3 List of genomes used in this study

Species	Genome version	Source
Amphimedon queenslandica	Amphimedon queenslandica v1.0	http://spongezome.metazome.net/cgi-bin/gbrowse/amphimedon/
Trichoplax adhaerens	Trichoplax adhaerens Grell-BS-99 v1.0	http://www.ncbi.nlm.nih.gov/nuccore/ABGP00000000
Nematostella vectensis	Nematostella vectensis v1.0	http://www.ncbi.nlm.nih.gov/nuccore/ABAV00000000
Lottia gigantea	Lottia gigantea v1.0	http://genome.jgi-psf.org/Lotgi1/Lotgi1.info.html
Capitella teleta	Capitella teleta v1.0	http://genome.jgi-psf.org/Capca1/Capca1.home.html
Branchiostoma floridae	Branchiostoma floridae v2.0	http://genome.jgi-psf.org/Brafl1/Brafl1.home.html
Strigamia maritima	Strigamia maritima v1.0	http://www.strigamia-annotation.org/
Tribolium castaneum	Tribolium castaneum v1.0	http://beetlebase.org
Drosophila melanogaster	Drosophila melanogaster vX	http://flybase.org
Homo sapiens	Homo sapiens GRCh37.p2	http://www.ncbi.nlm.nih.gov/nuccore/ABGP00000000

Chapter 3

Reconstructing the ancestral condition of a cluster's locus. Insights from the placozoan lineage

(Adapted from Mendivil Ramos, O., Barker, D. & Ferrier, D. E. K. 2012. Ghost Loci Imply Hox and ParaHox Existence in the Last Common Ancestor of Animals. Current Biology, 22, 1951-1956)

Here I resolve the long debated origin and identity of the Hox-like gene in the placozoan *Trichoplax adhaerens*. I expose how an alternative methodology has helped to solve this controversy and, moreover, pushed back the origin of Hox and ParaHox in terms of evolutionary time and lineage.

3.1 Introduction

In recent years, there has been an increase in the number of hypotheses speculating about the origin and evolution of the Hox cluster (extensively reviewed in section 1.5.2.2; (Ferrier, 2010)). The common assumption of these hypotheses, which stems from the discovery of the Hox cluster and its genomic arrangement, is that this cluster originated early in animal evolution via tandem duplication and that extensive independent duplications occurred in major bilaterian lineages (Lemons and McGinnis, 2006)). This did occur in the case of the Hox cluster and it is extendable to the other families within the ANTP class. It also occurred in the paralogous sister of Hox, the ParaHox. The discovery of the ParaHox cluster in *Branchiostoma floridae* and its homologous features with the Hox cluster, including its arrangement in the genome, phylogenetic affinities and transcriptional collinearity, led to the hypothesis of an ancestral state of Hox and ParaHox, the ProtoHox (Brooke et al., 1998). At some point in animal evolution a ProtoHox state existed that eventually duplicated to give rise to the Hox genes and ParaHox genes. However, duplications have not been the only source of change in the composition and structure of these families during evolution; gene loss has also played a major role in shaping these families. This has led to different hypotheses about the nature of the ProtoHox duplication, and which animal lineages are descended from the different states (ProtoHox versus Hox/ParaHox, see chapter 1, section 1.5.2.2).

The foundations of many of these hypotheses are based on the analyses of Hox and ParaHox genes in basal animal lineages as proxies for evolutionary stages of these families. Starting from the sister group of bilaterians, the cnidarians contain a number of genes with Hox and ParaHox gene sequence affinities, however, their precise function and precise evolutionary relationships with their putative orthologues in bilaterians remains controversial (Ferrier, 2010). Beyond the phylogenetic ambiguities that tend to arise using homeodomain sequences, synteny analyses in *Nematostella vectensis* show that this cnidarian contains distinct Hox and ParaHox loci homologous to the bilaterian loci (Hui et al., 2008). Therefore, these loci evolved before the origin of Cnidaria. This case shows that the complement of the homeobox genes and their phylogenetic relationships with bilaterian sequences only provides limited resolution for dating the existence of these loci relative to the other basal animal lineages.

The placozoan lineage, represented by *Trichoplax adhaerens*, contains a single gene, Trox-2, with sequence similarity to the Hox-like genes. Different opinions exist about the orthology of this gene. Some authors believe that Trox-2 is orthologous to the ParaHox gene Gsx, and hence is an evolutionary sister to Hox genes (Schierwater et al., 2008). Others believe that this gene is a distinct descendant of the ProtoHox condition, which is hypothesized to have been the precursor to the Hox and ParaHox genes (Jakob, 2004).

Hox and ParaHox genes are absent from the genome sequences of the poriferan *Amphimedon queenslandica* and the ctenophore *Mnemiopsis leidyi*, and have not been found in any other members of these phyla (Larroux et al., 2007, Ryan et al., 2010).

An initial step in resolving the origin of Hox, and by extension the origin of ParaHox, involves deducing the orthology of the Hox-like gene of *T. adhaerens, Trox-2*. An alternative way of determining orthology, in addition to molecular phylogenetics using the homeodomain, is to perform synteny analysis as an independent means of sequence-based phylogeny reconstruction to determine the homology of gene loci. This approach has previously been used in *Nematostella vectensis*, *Platynereis dumerilii* and *Branchiostoma floridae* to infer the orthology of ParaHox clusters (Ferrier et al., 2005, Hui et al., 2008, Hui et al., 2012). I used a comparable approach to deduce the orthology of *Trox-2*.

I anticipate a number of possible scenarios with regards to *Trox-2*'s synteny:

1) *Trox-2* is in a ProtoHox locus, which implies that orthologues of the surrounding neighbours are typically found in both Hox and ParaHox loci of bilaterians and cnidarians. Therefore, *Trox-2* resides in a locus homologous to Hox and ParaHox loci in bilaterians and cnidarians and thus, is representative of the ProtoHox condition.

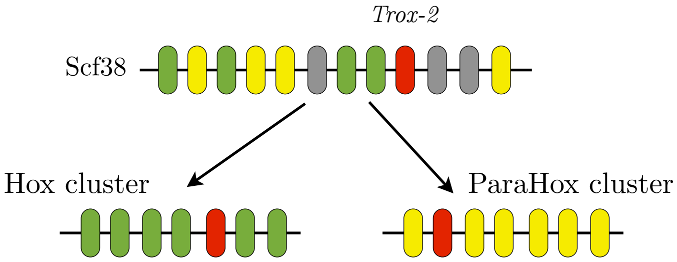


Figure 3.1.- ProtoHox synteny scenario. *Yellow ovals represent ParaHox neighbours. Green ovals represents Hox neighbours. Grey ovals represent non-Hox/ ParaHox neighbours. Red oval represents T. adhaerens Trox-2, and Hox/ParaHox cluster genes.*

2) *Trox-2* is in a ParaHox locus, which implies that orthologues of the surrounding neighbours are typically found in bilaterian ParaHox loci. This would be consistent with evidence from homeodomain molecular phylogenies, in which *Trox-2* is orthologous to Gsx. Since the ParaHox cluster is hypothesized to have arisen by the duplication of the ProtoHox cluster, then if *T. adhaerens* contains a ParaHox locus I would also expect it to contain a Hox locus. Taking into account the absence of other Hox-like sequences in this genome besides *Trox-2*, finding a locus in the genome sequence homologous to a Hox neighbourhood would suggest differential loss of genes.

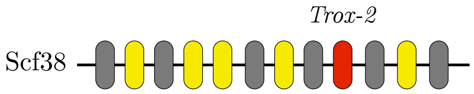


Figure 3.2.- ParaHox synteny scenario. *Yellow ovals represent ParaHox neighbours. Grey ovals represent non-Hox/ParaHox neighbours. Red oval represents T. adhaerens Trox-2.*

3) *Trox-2* is in a Hox locus, which implies that orthologues of the surrounding neighbours are typically found in bilaterian Hox loci. This would imply that despite this sequence's affinity with the ParaHox gene, Gsx surprisingly resides in a homologous locus to Hox.

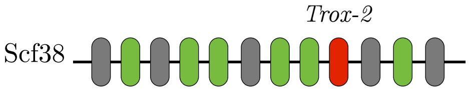


Figure 3.3.- Hox synteny scenario. *Green ovals represents Hox neighbours. Grey ovals represent non-Hox/ParaHox neighbours. Red oval represents T. adhaerens Trox-2.*

4) *Trox-2* is in neither a Hox or ParaHox locus, and orthologues of the surrounding neighbours are not found in either Hox or ParaHox loci of bilaterians as a result of any ancestral gene neighbourhoods having been broken apart along the placozoan lineage.

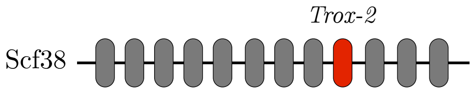


Figure 3.4.- Non-Hox/ParaHox synteny scenario. *Grey ovals represent non-Hox/ParaHox neighbours. Red oval represents T. adhaerens Trox-2.*

Here I have set out the basis on which to test whether synteny can actually give any further resolution and favour any of the current hypotheses regarding the evolution of this family within the placozoan lineage.

3.2 Materials and Methods

3.2.1 Analysis of *Trox-2*-containing scaffold 38

A search for repetitive elements was performed on scaffold 38 (GenBank accession number: DS985276.1) using RepeatMasker (www.repeatmasker.org), in order to clarify whether these repetitions are tandem duplications of coding sequences, assembly artefacts, transposons or mini/microsatellites. To visualize the repeat content of the scaffold, dot-plots comparing the nucleic acid sequence of the scaffold 38 against itself were computed by performing a BLASTn search in NCBI.

3.2.2 Orthologue analysis

Orthology assignment for each one of the 38 genes in scaffold 38 of *T. adhaerens* was performed as in Chapter 2, section 2.1 with the following modifications. Each gene within the scaffold was compared by rbh against the human genome (GRCh37.p2). This helped to establish whether each *T. adhaerens* gene had a one-to-one, one-to-many or many-to-many relationship with human genes, or no orthology at all (i.e. *Trichoplax* specific genes). *Trichoplax*'s sequences and their candidate human orthologues with their respective family members (if they had them) were aligned using MAFFT (see section 2.1.2 for specific details) and viewed in Jalview (see section 2.1.2 for specific details) to edit alignments for phylogenetic tree building. Alignment editing was refined by cross-comparison with multiple alignments post-processed by GBLOCKS (see section 2.1.2 for specific details). In cases without *T.adhaerens* or human family members or duplicates (i.e. putative one-to-one relationship), orthologue sequences of other chordates and *Nematostella vectensis* were included to help identify conserved domains and motifs and *T. adhaerens* gene identity. SMART (see section 2.1.2 for specific details) was used to help confirm these conserved domains and motifs (Appendix B, B.1). In cases with family members (one-to-many or many-to-many), phylogenetic trees were constructed using Modelgenerator followed by Neighbour-Joining, Maximum Likelihood and Bayesian Markov chain Monte Carlo. Node support for NJ trees was estimated from 1000 bootstrap replicates. Node support for ML were estimated from 100 bootstrap replicates and Bayesian trees there were estimated usings 1000000 generations; 5000 for sample probability; burn-in of 50; two runs of four chains each. See Appendix B, B.5 for multiple alignments and phylogenetic trees. This tree building helped to resolve some of the one-to-many and many-to-many relationships as one-to-one. Orthologous gene locations in the human genome were noted.

3.2.3 Orthologue statistical analysis

Orthologue statistical analysis was performed as specified in Chapter 2, section 2.2 with the following modifications. Once identified *T. adhaerens*-human orthologues were classified into Hox loci neighbour orthologues, ParaHox loci neighbour orthologues and Non-Hox/ParaHox loci neighbour orthologues. Hox loci neighbour orthologues are those *T. adhaerens* genes with human orthologues located on any of the human chromosomes bearing a Hox cluster (Chromosomes 2, 7, 12 and 17). ParaHox loci neighbour orthologues are those *T. adhaerens* genes with human orthologues located on any of the human chromosomes bearing a ParaHox 'cluster' (Chromosomes 4, 5, 13 and X). Non-Hox/ParaHox orthologues are those *T. adhaerens* genes with human orthologues located on chromosomes other than 2, 4, 5, 7, 12, 13, 17 and X. Also, two sets of tests were performed to accommodate tandem or segmental duplications on the human lineage which result in co-linkage of multiple members of a particular gene family. One version included the single location of each of the human orthologues and the second version included the collapsed location of the human paralogues (e.g. in the case of the torsins gene family four out of the five members are located in chromosome 9, and in this case we counted just one location in chromosome 9 within the second set of tests). These numbers were used to estimate observed probabilities of categories of orthologues. Expected probabilities of categories of the orthologues were inferred as mentioned in chapter 2 section 2.2. From these probabilities were calculated contingency tables (see probabilities for version human genome version GRCh37.p2 in Appendix B, B.2, B.3A and B.3.B). These probabilities were used to perform an Exact Binomial Test and a Fisher Exact Test in R (see codes in Appendix B, B. 4 and B.6).

3.2.4 Orthologue retrieval from Hox PAL

Orthologue retrieval was performed as specified in Chapter 2 section 2.1 with modifications. The Hox Putative Ancestral Linkage (PAL) gene list from *Nematostella vectensis* (Putnam et al., 2007) was used. The Hox PAL gene list (267 genes) accommodates orthologues into groups that have conserved linkage across bilaterian Hox-bearing chromosomes and *N. vectensis* scaffolds. This list was used as a query to perform rbh (BLASTp) against the *T. adherens* genome (see Appendix B, B.6).

3.2.5 Orthologue statistical analysis of Hox PAL

Once identified the *T. adhaerens*-human orthologues of scaffold 3 of *T. adhaerens* , the probabilities were calculated of a gene being in scaffold. These probabilities were used to perform an Exact Binomial Test in R (see Appendix B, B.7.).

3.3 Results

<u>3.3.1 *Trichoplax adhaerens Trox-2* is a ParaHox gene in a ParaHox locus...</u>

3.3.1.1 Analysis of scaffold 38

The scaffold that contains *Trox-2* is scaffold 38. This scaffold is built from contig ABGP01001092.1 (GenBank accession number): 1...229025, gap (50 bp), ABGP01001093.1 (GenBank accession number): 1...78836, gap (50 bp), ABGP01001094.1 (GenBank accession number): 1...83649, gap (15900 bp), ABGP01001095.1 (GenBank accession number): 1...3167. The dot-plot analysis of scaffold 38 shows that there are some repetitive elements, but no major duplications or large scale repetitions, which is consistent with this scaffold being well-assembled. The output of RepeatMasker analysis shows that these repetitive elements fall into the category of simple and low-complexity repeats (See Appendix X). From now on I will refer to scaffold 38 as the *Trox-2* scaffold.

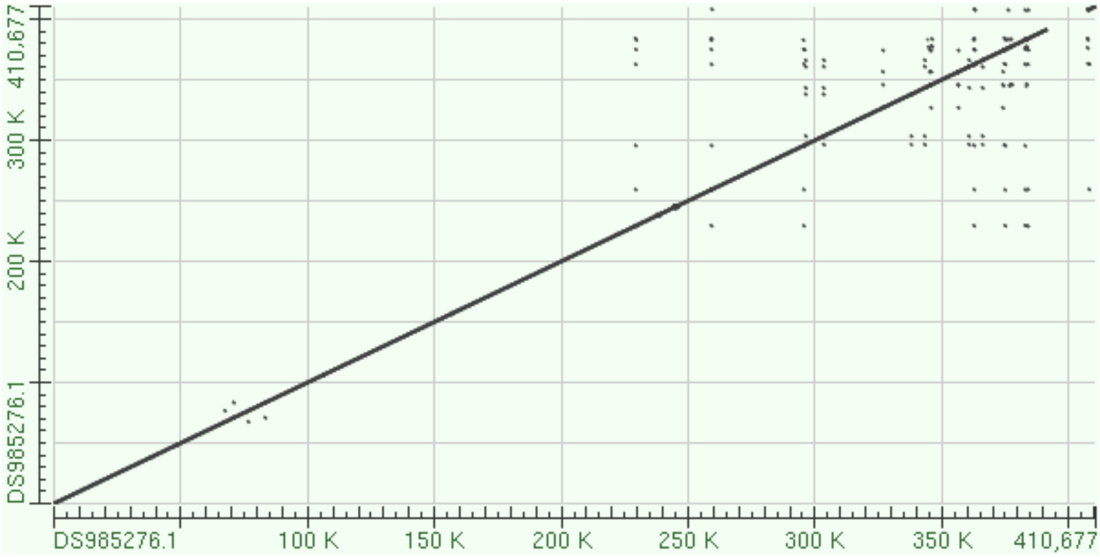


Figure 3.5.- Dot-plot of Trox-2 scaffold compared to itself.

3.3.1.2 Orthologue assignment of *Trox-2* scaffold

The details of each *T. adhaerens* gene on scaffold 38 and whether it can be assigned to a human orthologue or orthologues are as follows:

1) **TRIADDRAFT_62201** (Accession number: XP_002118187.1)

Reciprocal BLASTp searches show that this protein has similarity to the human kinesin-3 family. According to the current classification based on the Kinesin motor (KISc), human kinesins are comprised of 14 families (plus a collection of 'orphan' genes), divided into 28 subfamilies (Wickstead et al., 2010). The human kinesin-3 family is composed of eight members (KIF16B, StarD9, KIF1A, KIF1B, KIF1C, KIF13A, KIF13B and KIF14). Apart from the KISc motif, the FHA motif is characteristic of this family (Wickstead et al., 2010). *T. adhaerens* also contains four further kinesin-3 family sequences. A multiple alignment with all human kinesin genes and the putative *T. adhaerens* kinesin-3 genes showed no obvious affinity of TRIADDRAFT_62201 with a particular human kinesin subfamily. Moreover, the SMART analysis shows that TRIADDRAFT_62201 has no FHA motif, consistent with its very divergent nature. A neighbour-joining tree (JTT, 1000 bootstraps) show that this protein is a divergent member of the kinesin family. Due to the divergent nature of TRIADDRAFT_62201 I discard it from the synteny analysis.

2) **TRIADDRAFT_62202** (Accession number: XP_002118164.1)

Reciprocal BLASTp searches show that this protein is a putative orthologue to human pericentriolar material 1 or PCM1. BLASTp searches using the human PCM1 and TRIADDRAFT_62202 sequences against their own genomes revealed no other family members. Chordate PCM1 sequences have a GTP/ATP binding site motif with the consensus [A,G]-X4-G-K-[S,T] and various motifs rich in aspartic acid and glutamic acid (EDDEx6AEx3, DEx6QD and EDENEDEEMEEFEE) (Balczon et al., 1994). The *Trichoplax* orthologue does not show any of these motifs but does have extensive sequence similarity at the C-terminus end, which is also the case for the cnidarian putative PCM1 sequences from *Nematostella vectensis* and *Hydra magnipapillata*. Hence, I name this protein *Tad_PCM1*, and include this protein in the synteny analysis as a "one-to-one" orthologue relationship.

3) **TRIADDRAFT_62203** (Accession number: XP_002118165.1)

The results from the reciprocal BLASTp searches indicate that this protein has no significant match with any human protein. Consequently this protein is not informative for synteny analysis.

4) **TRIADDRAFT_33759** (Accession number: XP_002118188.1)

Reciprocal BLASTp searches indicate that this protein is a putative orthologue of human Torsin 1A. The human torsin family is composed of five members: TORSIN 1A, TORSIN 1B, TORSIN 2A, TORSIN3A and C9orf167. Also, the reciprocal BLASTp searches indicated another putative *Trichoplax* torsin, TRIADDRAFT_58752. The torsin family belongs to the superfamily AAA+. The torsins have four short motifs: Walker A, Walker B, SN, sensor IV. These motifs are all present in the *T. adhaerens* sequences. The ClpB heat shock protein family is closely related to the torsins (Ozelius et al., 1999, Zhu et al., 2008). Torsins and Clpbs are characterized by six conserved cysteines. In Torsin sequences the cysteine closest to the C-terminus is embedded in the motif GCK. In ClpB sequences the sequence is instead GAR (Ozelius et al., 1999, Zhu et al., 2008). A molecular phylogenetic analysis, including some ClpB genes as an outgroup, shows that TRIADDRAFT_58752 and TRIADDRAFT_33759 form a sister group to the torsins of humans and other animals. I thus classify the orthologue relationship as "many-to-many" and accommodate this in the statistical analyses as described below.

5) **TRIADDRAFT_64406** (Accession number: XP_002118166.1)

Reciprocal BLASTp searches indicate that this protein is a putative orthologue to human neurochondrin. Neurochondrin is a leucine-rich protein (Mochizuki et al., 1999). No further family members were found in *T. adhaerens* or in human. The multiple alignment shows extensive conservation of leucine-rich motifs in TRIADDRAFT_64406, confirmed by SMART. Hence, I name this protein *Tad_NCDN*, and include this protein in the synteny analysis as a "one-to-one" orthologue relationship.

6) **TRIADDRAFT_9204** (Accession number: XP_002118167.1)

Reciprocal BLASTp searches indicate that this protein is a putative orthologue to human matrilins, human fibrillins and human fibulins. These proteins share the following domains: epidermal growth factor-like domain (EGF), calcium-binding EGF-like domain (EGF_CA) and von Willebrand factor type A domain (VWA). The distinction amongst these families is by a characteristic combination of these domains (Deák et al., 1999, Handford et al., 2000, Kielty et al., 2002, Frank et al., 2002, Whittaker and Hynes, 2002, Timpl et al., 2003, Sicot et al., 2008). A multiple alignment of TRIADDRAFT_9204 with human matrillins, fibulins and fibrillins shows that single EGF and EGF_CA motifs are present in TRIADDRAFT_9204. The SMART analysis confirms this result. This *T. adhaerens* sequence is thus relatively short, with very few motifs, and its classification is consequently poorly resolved. Hence, I discard this protein from the analysis.

7) **TRIADDRAFT_51183** (Accession number: XP_002118168.1)

Reciprocal BLASTp searches identify TRIADDRAFFT_51183 as a putative orthologue to human Hydroxysteroid (17-beta) dehydrogenase 10. Other family members were retrieved from the *Trichoplax* and human genomes. The hydroxysteroid 17-beta dehydrogenase (HSD17B) family belongs to the short chain dehydrogenase/reductase (SDR) superfamily. The HSD17B family shares several amino acid sequence motifs with the SDR superfamily: TGXXXGXG (part of the Rossman fold), NAG (structural stabilization), YXXK (active centre) and PGXXXT (C-terminal to active site) (Baker, 2001, Kleiger and Eisenberg, 2002, Mindnich et al., 2004, Kavanagh et al., 2008). The multiple alignment and SMART analysis show the conservation of these motifs in TRIADDRAFFT_51183. I use a phylogenetic analysis to clarify specific orthology and paralogy relationships. For this analysis the rest of the members of the HSD17B family were used as an outgroup. From this analysis I identify a one-to-one relationship between TRIADDRAFFT_51183 and HSD17B10. Also, I identify a recent paralogue of this particular *T. adhaerens* sequence (TRIADDRAFT_22420) also orthologous to human HSD17B10, indicating a possible lineage-specific duplication in *T. adhaerens*. Hence, I name this protein *Tad_HSD17B10A*, and include this protein in the synteny analysis as a “one-to-one” orthologue relationship.

8) **TRIADDRAFT_33760** (Accession number: XP_002118168.1)

Reciprocal BLASTp searches indicate that this protein is a putative orthologue to a human fibrillin or fibulin protein. These two families share domains (see TRIADDRAFT_9204 above). A multiple alignment of TRIADDRAFT_33760 with human fibulins and human fibrillins, along with a SMART analysis, revealed conservation of some motifs. However, the TRIADDRAFT_33760 gene model is very short, which contributes to it lacking a clear affinity to a particular human gene(s). Therefore, due to this difficulty in identifying TRIADDRAFT_33760 orthology, I discard this gene from the synteny analysis.

9) **TRIADDRAFT_33711** (Accession number: XP_002118189.1)

Reciprocal BLASTp searches indicate that TRIADDRAFT_33711 is a putative orthologue to human vacuolar protein sorting 36 (VPS36). No further family members were found in the *T. adhaerens* or human genomes. The human protein is characterised by a split pleckstrin-homology domain Φ XKX(G/A/S/P)X...(K/R)...X(R/K)XRX(F/L) also known as the glue domain (Lemmon, 2001, Alam et al., 2006). Multiple alignment of TRIADDRAFT_33711 with chordate orthologues of VPS36, and SMART analysis, show that TRIADDRAFT_33711 also conserves this motif. Therefore, I name this protein *Tad_ VPS36*, and include this protein in the synteny analysis as a "one-to-one" orthologue relationship.

10) **TRIADDRAFT_33740** (Accession number: XP_002118170.1)

Reciprocal BLASTp searches indicate that TRIADDRAFT_33740 is a putative orthologue of human Aminoadipate-semialdehyde-dehydrogenase-phosphopantetheinyl transferase (AASDHPPT). No further family members are found in the *T. adhaerens* and human genomes. Human ASSDHPPT is characterised by the phosphopantetheinyl transferase motif GXD...E...(W/F/L)XX(K/R)E(A/S)XXK (Joshi et al., 2003). Multiple alignment of TRIADDRAFT_33740 with several other chordate orthologues, along with SMART analysis, show the conservation of the phosphopantetheinyl transferase motif in TRIADDRAFT_33740. Therefore, I name this protein

Tad_AASDHPPT, and include this protein in the synteny analysis as a "one-to-one" orthologue relationship.

11) **TRIADDRAFT_33763** (Accession number: XP_002118190.1)

Reciprocal BLASTp searches indicate that TRAIDDRAFT_33763 is a putative orthologue of human Apoptosis Induction Factors (AIFMs). The human AIFM family contains three members, characterised by a FAD or NAD(P) binding Rossmann motif ((V/I)XGX(1-2)GXXGXXX(G/A))(Susin et al., 1999). A second member of this family was retrieved from the *T. adhaerens* genome (TRIADRAFT_59728). Multiple alignment and SMART analysis of the human and *T. adhaerens* sequences show conservation of the Rossman motif. I investigated orthology and paralogy relationships with molecular phylogenetic analyses, and identify a one-to-one orthology relationship of TRAIDDRAFT_33763 with human AIFM1 (and a one-to-one orthology relationship between TRIADDRAFT_59728 and human AIFM3). Therefore I include TRAIDDRAFT_33763 in the synteny analysis.

12) **TRIADDRAFT_33726** (Accession number: XP_002118171.1)

Reciprocal BLASTp searches indicate that TRIADDRAFT_33726 is a putative orthologue to human tumour suppressor candidate 3 (TUSC3) and magnesium transporter protein (MAGT1). Both human proteins share the thioredoxin-like motif (CXXC) and oligosaccharyl transferase motif (Zhou and Clapham, 2009). A multiple alignment of the human proteins, other chordate orthologues and TRIADDRAFT_33726, along with a SMART analysis, confirm conservation of the motifs. The phylogenetic analysis shows that TRIADDRAFT_33726 is a proto-orthologue to both chordate proteins.

13) **TRIADDRAFT_62215** (Accession number: XP_002118191.1)

Reciprocal BLASTp searches indicate that TRIADDRAFT_62215 is a putative orthologue of a human G-protein couple receptor (GPCR). The current consensus based on the common functional unit, the seven α-helical transmembrane motif (7TM), divides the human GPCRs superfamily into five classes (Glutamate, Rhodopsin, Adhesion, Frizzled/Taste2 and Secretin (GRAFS)) (Fredriksson et al., 2003). Beyond this basic level of classification more fine-scale affinities are very poorly resolved in phylogenetic trees. Consistent with this, no unambiguous orthology of TRIADDRAFT_62215 with any particular human gene(s) could be determined. Therefore I discard this protein from the synteny analysis.

14) **TRIADDRAFT_62216** (Accession number: XP_002118172.1)

Reciprocal BLASTp searches indicate that this hypothetical protein is also a putative GPCR. For the same reasons as outlined for TRIADDRAFT_62215 I discard this protein from the synteny analysis.

15) **TRIADDRAFT_62217** (Accession number: XP_002118173.1)

Reciprocal BLASTp searches indicate that TRIADDRAFT_62217 is a putative orthologue of human thioredoxin. Thioredoxins are characterized by a cysteine-rich sequence motif (W-C-G-P-C-K followed by three cysteines). The thioredoxin superfamily is divided into families by a common thioredoxin fold encoded by the two residues in between the two cysteines of the active sites (thioredoxin: C-G-P-C; glutaredoxin: C-P-T-C; DbsA: C-P-H-C) (Martin, 1995, Carvalho et al., 2006, Atkinson and Babbitt, 2009). Further paralogues of TRIADDRAFT_62217 were retrieved from the *T. adhaerens* genome. Surprisingly three out of the four *T. adhaerens* family members were in scaffold 38 (TRIADDRAFT_62226 and TRIADDRAFT_62227). Multiple alignment of these three proteins demonstrated that they are unlikely to be recent duplicates as there are extensive differences between the sequences. Multiple alignment of the four putative *T. adhaerens* thioredoxins and the human thioredoxins showed the conserved motif W-C-G-P-C-K, but outside this motif no cysteine conservation was observed. Phylogenetic analyses with the entire coding sequences do not reveal a clear orthologue identification. Therefore, I exclude these proteins from the synteny analysis.

16) **TRIADDRAFT_33746** (Accession number: XP_002118192.1)

The results from the reciprocal BLASTp searches indicate that this hypothetical protein is an orthologue to human Yip family member 6 (YIPF6). The human Yip family is composed of seven members and characterized by the motif DLYGP and GY (Yang et al., 1998, Calero et al., 2002). Further members of the human Yip family as well as four putative members of this family from the *T. adhaerens* genome were retrieved. One of the *T. adhaerens* genes is also in scaffold 38 (TRIADDRAFT_5826). The SMART analysis confirms the presence of conserved motifs. The molecular phylogenetic analysis helped to identify a "one-to-one" orthologue relationship between human YIPF6 and TRIADDRAFT_33746, which I named *Tad_ YIPF6,* and a "one-to-many" orthologue relationship between TRIADDRAFT_5826 (which I named *Tad_ YIPF5/7*) and human YIPF5 and YIPF7. Therefore, I included both proteins in the synteny analysis.

17) **TRIADDRAFT_33724** (Accession number: XP_002118174.1)

The results from the reciprocal best-hit BLASTp search indicate that this hypothetical protein is a putative orthologue to Glucosamine-6-phosphate deaminase 2 (GNPDA2). The human GNPDA family is composed of two members (GNPDA1 and GNPDA2) and the motif that characterizes this family is (L/I/V/M)3XGX(L/I/T)X(L/I/V/M)XG(L/I/V/M)GX(D/E/I)3XGX(I)X(L)X(V)XG(I)GX(D)H (Wolosker et al., 1998). No further family members were found in the *T. adhaerens* genome. The multiple alignment with TRIADDRAFT_33724 and other chordate orthologues, along with SMART analysis, confirm conservation of the family-characterizing motif. The phylogenetic analysis helped to identify a "one-to-many" orthologue relationship between TRIADDRAFT_33724, GNPDA1 and GNPDA2. Therefore, I name this protein *Tad_ GNPDA* and included it in the synteny analysis.

18) **TRIADDRAFT_62220** (Accession number: XP_002118175.1)

The reciprocal best hit BLASTp search indicates that this hypothetical protein is a putative orthologue of the human FERM and PDZ domain containing proteins: collectively the FRMPDs. There are four FRMPDs (FRMPD1, FRMPD2, FRMPD3 and FRMPD4) in the human genome and they are distinguished by the order of appearance of the FERM and PDZ domains in their sequences (Stenzel et al., 2009). Another putative FRMPD member in the *T. adhaerens* genome was found. The multiple alignment and SMART analysis of TRIADRAFT_62220 with human FRMPD domain containing proteins show conservation of the FERM domain and conserved tryptophan motifs. Phylogenetic analysis helped to identify a “one-to-many” orthologue relationship between TRIADDRAFT_62220 and human FRMPD1, FRMPD3 and FRMPD4. The other *T. adhaerens* sequence, TRIADDRAFT_64201 shows affinity with human FRMPD2. Hence, I name TRIADDRAFT_62220 *Tad_FRMPD1/3/4* and include it in the synteny analysis.

19) **TRIADDRAFT_62221** (Accession number: XP_002118193.1)

The reciprocal best-hit BLASTp searches indicate that this hypothetical protein is a putative orthologue to the BTB/POZ domain-containing proteins. The human genome possesses 16 BTB/POZ domain-containing proteins. No further BTB/POZ domain-containing proteins were found in the *T. adhaerens* genome. The multiple alignment of TRIADDRAFT_62221 with the human BTB/POZ domain containing proteins show affinity with the human BTB/POZ domain containing protein 12 (Andersen et al., 2009). SMART analysis confirm the conserved motifs. Molecular phylogenetic analysis confirm a “one-to-one” orthologue relationship of TRIADDRAFT_62221 with human BTB/POZ domain containing protein 12. Hence, I name this protein *Tad_BTB/POZ12* and included it in the synteny analysis.

20) **TRIADDRAFT_62222** (Accession number: XP_002118176.1)

A GPCR, which is excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

21) **TRIADDRAFT_62223** (Accession number: XP_002118194.1)

The reciprocal best-hit BLASTp search indicate that this protein is a putative orthologue to members of the human intracellular membrane-associated calcium-independent phospholipase (PNPLA) family. The patatin domain characterizes the human PNPLA family, which is composed of nine members (Wilson et al., 2006). The human PNPLA genes differ in their motif content besides the patatin domain. Further putative *T. adhaerens* PNPLA genes were retrieved. The patatin motif is conserved in TRIADDRAFT_62223 according to the SMART analysis, and a multiple alignment shows affinity with human PNPLA8 and PNPLA9. However, molecular phylogenetic analysis revealed that

TRIADDRAFT_62223 does not have a clear affinity with any particular human PNPLA gene, whilst other *T. adhaerens* PNPLA genes do have clearer orthologue relationships. Therefore, I exclude it from the synteny analysis analysis.

22) **TRIADDRAFT_62224** (Accession number: XP_002118177.1)

The results from the reciprocal BLASTp searches indicate that this protein does not have a significant match with any human protein, and so it is excluded from the synteny analysis.

23) **TRIADDRAFT_33732** (Accession number: XP_002118195.1)

The results from the best-hit reciprocal BLASTp searches indicate that this hypothetical protein is a putative orthologue of human Chloride channel proteins. The human Chloride channel protein family is composed of seven members, characterized by seven very well conserved transmembrane helices (Mindell and Maduke, 2001). Further members of this family were retrieved from the *T. adhaerens* genome. The multiple alignment and SMART analysis show conservation of the transmembrane helices in TRIADDRAFT_33732 and affinity for human CLCN3, 4, 5 genes. Molecular phylogenetic analysis helped identify a "one-to-many" orthologue relationship between TRIADDRAFT_33732 and CLCN3, CLCN4 and CLCN5. Therefore, I name this protein *Tad_ CLCN3/4/5* and included it in the synteny analysis.

24) **TRIADDRAFT_62226** (Accession number: XP_002118196.1)

A putative thioredoxin, excluded from the analysis as discussed for TRIADDRAFT_62217.

25) **TRIADDRAFT_62227** (Accession number: XP_002118197.1)

A putative thioredoxin, excluded from the analysis as discussed for TRIADDRAFT_62217.

26) **TRIADDRAFT_5826** (Accession number: XP_002118178.1)

Discussed in 16) TRIADDRAFT_33746

27) **TRIADDRAFT_62229** (Accession number: XP_002118198.1)

Reciprocal BLASTp searches indicate that this protein has no significant match with any human protein, and so I exclude it from the synteny analysis.

28) **TRIADDRAFT_62230** (Accession number: XP_002118179.1)

A GPCR, which is excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

29) **TRIADDRAFT_7464** (Accession number: XP_002118199.1)

A GPCR, which is excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

30) **TRIADDRAFT_5463** (Accession number: XP_002118200.1)

A GPCR, which is excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

31) **TRIADDRAFT_64407** (Accession number: XP_002118180.1)

The BLASTp search indicate no BLAST hits at all.

32) **TRIADDRAFT_62233** (Accession number: XP_002118181.1)

The results from the best-hit reciprocal BLASTp searches indicate that this hypothetical protein is a putative orthologue of human sterol regulatory element-binding transcription factors (SREBF1). This family belongs to a higher-order group B of the basic helix-loop-helix (bHLH) superfamily (Simionato et al., 2007). The human sterol regulatory element-binding transcription factors family is composed of two members and as for other bHLH superfamily members is characterized by a DNA-binding basic region followed by two α-helices. No further members of this family were retrieved from the *T. adhaerens* genome. The multiple alignment shows conservation of the α-helices and DNA-binding basic region and affinity for the orthologues of SREBF1. The SMART analysis confirms the conserved motifs. Molecular phylogenetic analysis helped identify a "one-to-many" orthologue relationship between TRIADDRAFT_62233 and human SREBF1 and SREBF2. Thus, I name this protein *Tad_SREBF1/2* and included it in the synteny analysis.

33) **TRIADDRAFT_33728** (Accession number: XP_002118201.1)

Trox-2

34) **TRIADDRAFT_62235** (Accession number: XP_002118182.1)

A GPCR, which is excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

35) **TRIADDRAFT_62236** (Accession number: XP_002118183.1)

A GPCR, which is excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

36) **TRIADDRAFT_62237** (Accession number: XP_002118184.1)

A GPCR, which is excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

37) **TRIADDRAFT_62238** (Accession number: XP_002118185.1)

A GPCR, which is excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

38) **TRIADDRAFT_62239** (Accession number: XP_002118186.1)

A GPCR, which is excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

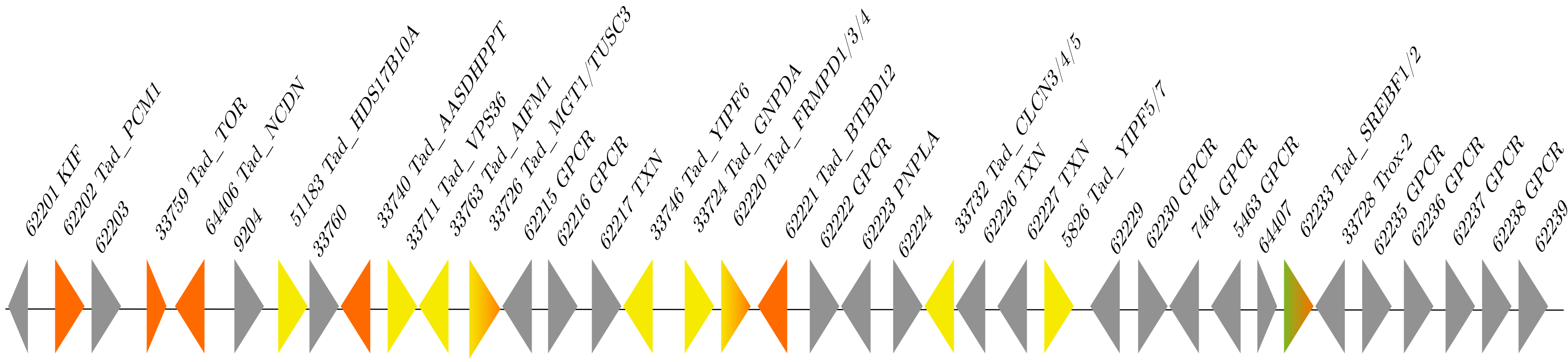


Figure 3.6.- Trox-2 scaffold orthology. *Triangles represents genes, with directionality representing gene orientation. ParaHox neighbour orthologues are defined as T. adhaerens genes with human orthologues located on any of the human chromosomes bearing ParaHox loci (Chromosomes 4, 5, 13 and X). Hox neighbour orthologues are defined as T. adhaerens genes with human orthologues located on any of the human chromosomes bearing a Hox cluster (Chromosomes 2, 7, 12, 17). Triangles in grey are genes with no detectable human orthology, triangles in orange are orthologues of human genes not linked to human Hox/ParaHox loci, triangles in yellow are orthologues of human ParaHox neighbours, triangles in yellow-orange are orthologues of human genes that are a mix of ParaHox and non-Hox/ParaHox neighbours, triangle in green-orange is a gene with human orthologues that are a mix of Hox and non-Hox/ParaHox neighbours.*

Here I summarize the classification of orthologous relationships between the genes found in Scaffold 38 of the *T. adhaerens* genome and human genes.

One-to-one orthologues	**One-to-many orthologues**	**Many-to-many**
TRIADDRAFT_62202 : PCM1	TRIADDRAFT_33726 : TUSC3 and MGT1	TRIADDRAFT_33759 and TRIADDRAFT_58752: TORSIN 1A, TORSIN 1B, TORSIN 2A, TORSIN 3A and C9orf167
TRIADDRAFT_64406 : NCDN	TRIADDRAFT_33724 : GNPDA1 and GNPDA	
TRIADDRAFT_51183 : HSD17B10	TRIADDRAFT_62220 : FRMPD1, FRMPD3 and FRMPD4	
TRIADDRAFT_33711 : VPS36	TRIADDRAFT_33732 : CLCN3, CLCN4 and CLCN5	
TRIADDRAFT_33740 : ASSDHPPT	TRIADDRAFT_5826 : YIP1 M5 and YIP1 M7	
TRIADDRAFT_33763 : AIFM1	TRIADDRAFT_62233 : SREBF1 and SREBF2	
TRIADDRAFT_33746 : YIP1 M6		
TRIADDRAFT_62221 : BTBD12		

Table 3.1.- Summary of the orthologue identities in scaffold 38. *The orthologue relationship is noted as Trichoplax gene(s): Human gene(s).*

3.3.1.2 Statistical significance of the observed synteny conservation of *Trox-2* scaffold

After identifying *T. adhaerens*-human orthologues I classified them into Hox loci neighbour orthologues, ParaHox loci neighbour orthologues and Non-Hox/ParaHox loci neighbour orthologues. Hox loci neighbour orthologues are those *T. adhaerens* genes with human orthologues located on any of the human chromosomes bearing a Hox cluster (Chromosomes 2, 7, 12 and 17). ParaHox loci neighbour orthologues are those *T. adhaerens* genes with human orthologues located on any of the human chromosomes bearing ParaHox loci (Chromosomes 4, 5, 13 and X). Non-Hox/ParaHox orthologues are those *T. adhaerens* genes with human orthologues located on chromosomes other than 2, 7, 12, 17, 4, 5, 13 or X. Also, I performed two sets of tests to accommodate tandem or segmental duplications on the human lineage which result in co-linkage of multiple members of a particular gene family. One version included the single location of each of the human orthologues and the second version included the collapsed location of the human paralogues (e.g., in the case of the torsins four out of the five members are located on human chromosome 9, and in this case we counted just one location on chromosome 9 within the second set of tests; Table 3.2 and Table 3.3).

***T. adhaerens* protein in Scaffold 38**	**Human orthologue**	**Human chromosomal location**
TRIADDRAFT_62202	PCM1	8
TRIADDRAFT_33759	TORSIN 1A	9
	TORSIN 1B	9
	TORSIN 2A	9
	TORSIN 3A	1
	C9orf167	9
TRIADDRAFT_64406	NCDN	1
TRIADDRAFT_51183	HSD17B10	X
TRIADDRAFT_33711	VPS36	13
TRIADDRAFT_33740	ASSDHPPT	11
TRIADDRAFT_33763	AIF1	X
TRIADDRAFT_33726	TUSC3	8
	MGT1	X
TRIADDRAFT_33746	YIP1 M6	X
TRIADDRAFT_33724	GNPDA1	5
	GNPDA2	4
TRIADDRAFT_62220	FRMPD1	9
	FRMPD3	X
	FRMPD4	X
TRIADDRAFT_62221	BTBD12	6
TRIADDRAFT_33732	CLCN3	4
	CLCN4	X
	CLCN5	X
TRIADDRAFT_5826	YIP1 M5	5
	YIP1 M7	4
TRIADDRAFT_62233	SREBF1	17
	SREBF2	22

Table 3.2.- Summary of orthologue identities with their single locations in the human genome for version 1 of the statistical tests.

***T. adhaerens* protein in Scaffold 38**	**Human orthologue**	**Human chromosomal location**
TRIADDRAFT_62202	PCM1	8
TRIADDRAFT_33759	TORSIN 1A, TORSIN 1B, TORSIN 2A and C9orf167	9
	TORSIN 3A	1
TRIADDRAFT_64406	NCDN	1
TRIADDRAFT_51183	HSD17B10	X
TRIADDRAFT_33711	VPS36	13
TRIADDRAFT_33740	ASSDHPPT	11
TRIADDRAFT_33763	AIF1	X
TRIADDRAFT_33726	TUSC3	8
	MGT1	X
TRIADDRAFT_33746	YIP1 M6	X
TRIADDRAFT_33724	GNPDA1	5
	GNPDA2	4
TRIADDRAFT_62220	FRMPD1	9
	FRMPD3 and FRMPD4	X
TRIADDRAFT_62221	BTBD12	6
TRIADDRAFT_33732	CLCN3	4
	CLCN4 and CLCN5	X
TRIADDRAFT_5826	YIP1 M5	5
	YIP1 M7	4
TRIADDRAFT_62233	SREBF1	17
	SREBF2	22

Table 3.3.- Summary of orthologue identities with their collapsed locations in the human genome for version 2 of the statistical tests.

The observed synteny conservation was statistically tested with two tests: Exact Binomial test and Fisher's Exact test. The numbers derived for these tests are based on human genome version 37 patch 2 and are as follows:

C1	number of genes on Hox Chromosomes	4489
C2	number of genes in the Hox clusters	39
C3	number of genes that are Hox loci neighbours	4450
C4	number of genes that are non-Hox loci neighbours	15997
C5	number of genes on ParaHox Chromosomes	2865
C6	number of genes in the ParaHox 'clusters'	6
C7	number of genes that are ParaHox loci neighbours	2859
C8	number of genes that are non-ParaHox loci neighbours	17588
C9	number of genes that are non-(Hox/ParaHox) loci neighbours	13093
C10	number of genes that are Hox/ParaHox loci neighbours	7309
C11	total number of genes in genome minus Hox and ParaHox clusters	20402
C12	total number of genes in genome	20447

***Table 3.4.- Summary of number of genes (protein coding genes) in the human genome version 37 patch 2.* See Appendix B, section B.1 for full derivation.**

From these numbers we calculated the probabilities of a randomly chosen human gene being a Hox locus neighbour, ParaHox locus neighbour and Non-Hox/ParaHox neighbour. The probabilities were as follows:

Probability of being a Hox locus neighbour Ph	0.217635839 (= C3 / C11)
Probability of not being a Hox locus neighbour Qh	0.782364161(= C4 / C11)
Total	1

Table 3.5.- Summary of the probabilities of being a Hox neighbour.

Probability of being a ParaHox locus neighbour Pph	0.139824913 (= C7 / C11)
Probability of not being a ParaHox locus neighbour Qph	0.860175087 (= C8 / C11)
Total	1

Table 3.6.- Summary of the probabilities of being a ParaHox neighbour.

Probability of being a Hox/ParaHox neighbour Pnhph	0.357460752 (= C10 / C11)
Probability of being a non-Hox/ParaHox neighbour Qnhph	0.641750809 (= C9 / C11)
Total	1

Table 3.7.- Summary of the probabilities of being a Non-Hox/ParaHox neighbour.

These probabilities were used to perform the Binomial Exact Test. The Exact Binomial Test was used to compare the observed number of Hox neighbour orthologues (or ParaHox neighbour orthologues or Hox/ParaHox neighbour orthologues) on scaffold 38 with those expected on the basis of the probability of Hox neighbours (or ParaHox neighbours or Hox/ParaHox neighbours) in the human genome.

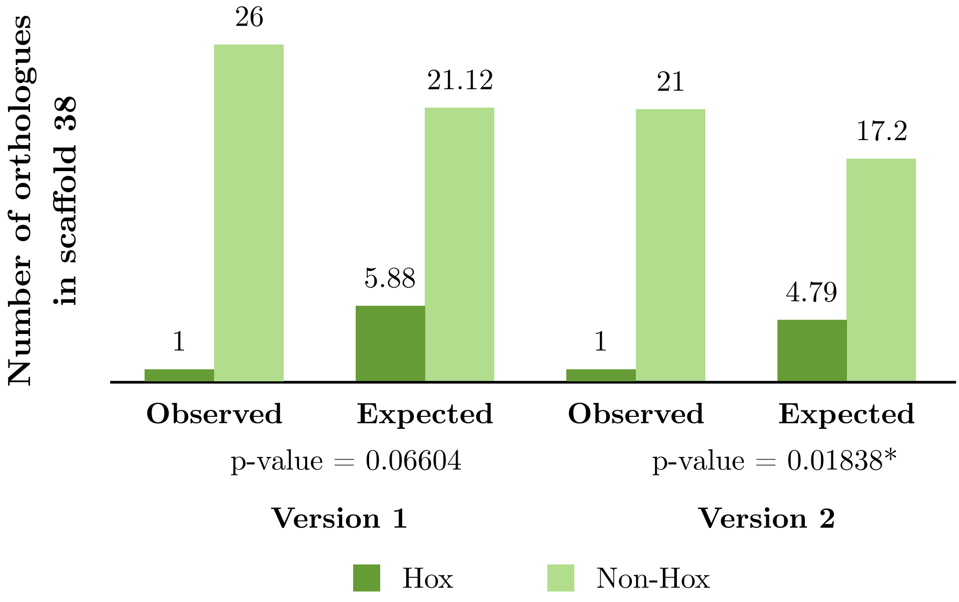


Figure 3.7.- Binomial exact test Hox case.

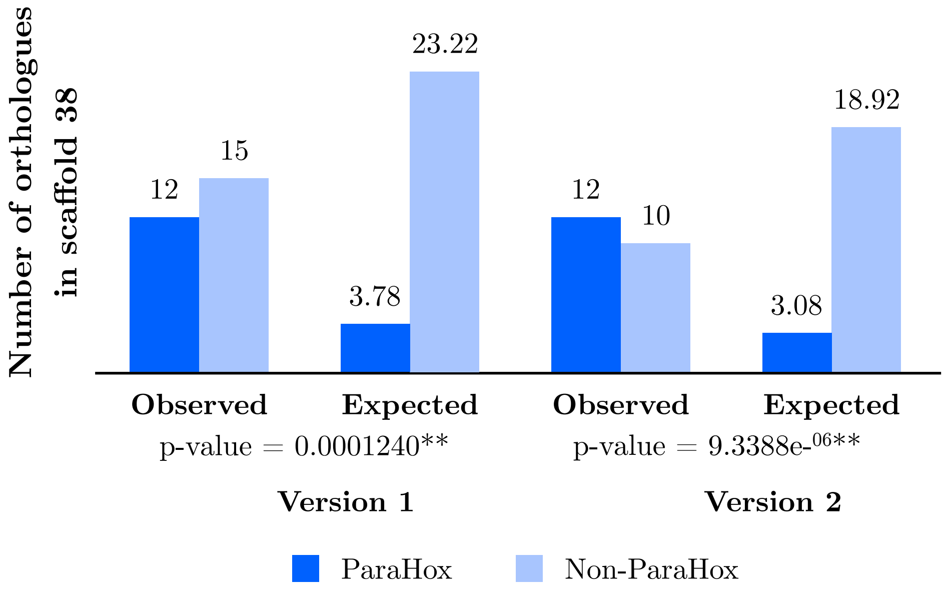


Figure 3.8.- Binomial exact test ParaHox case.

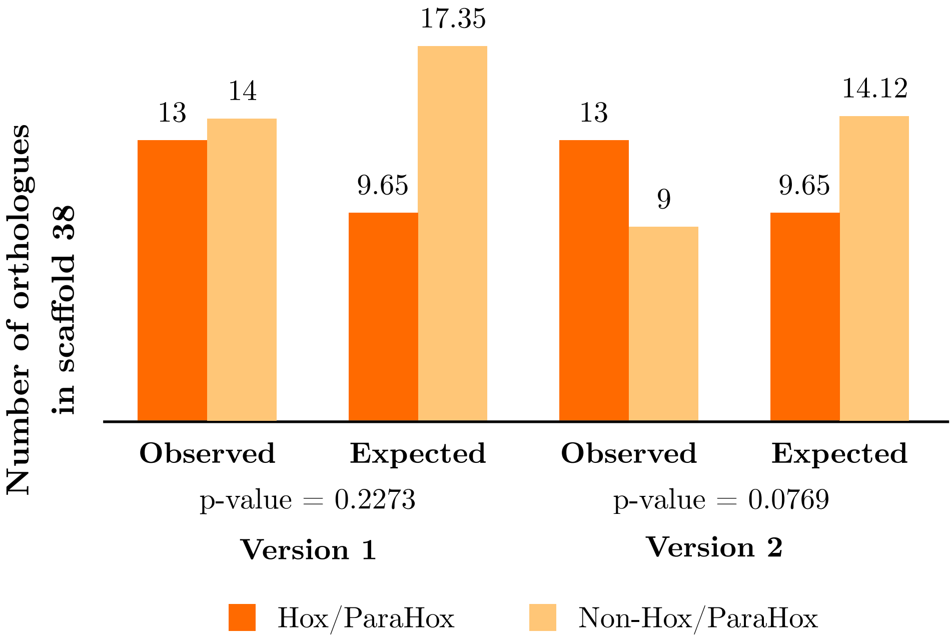
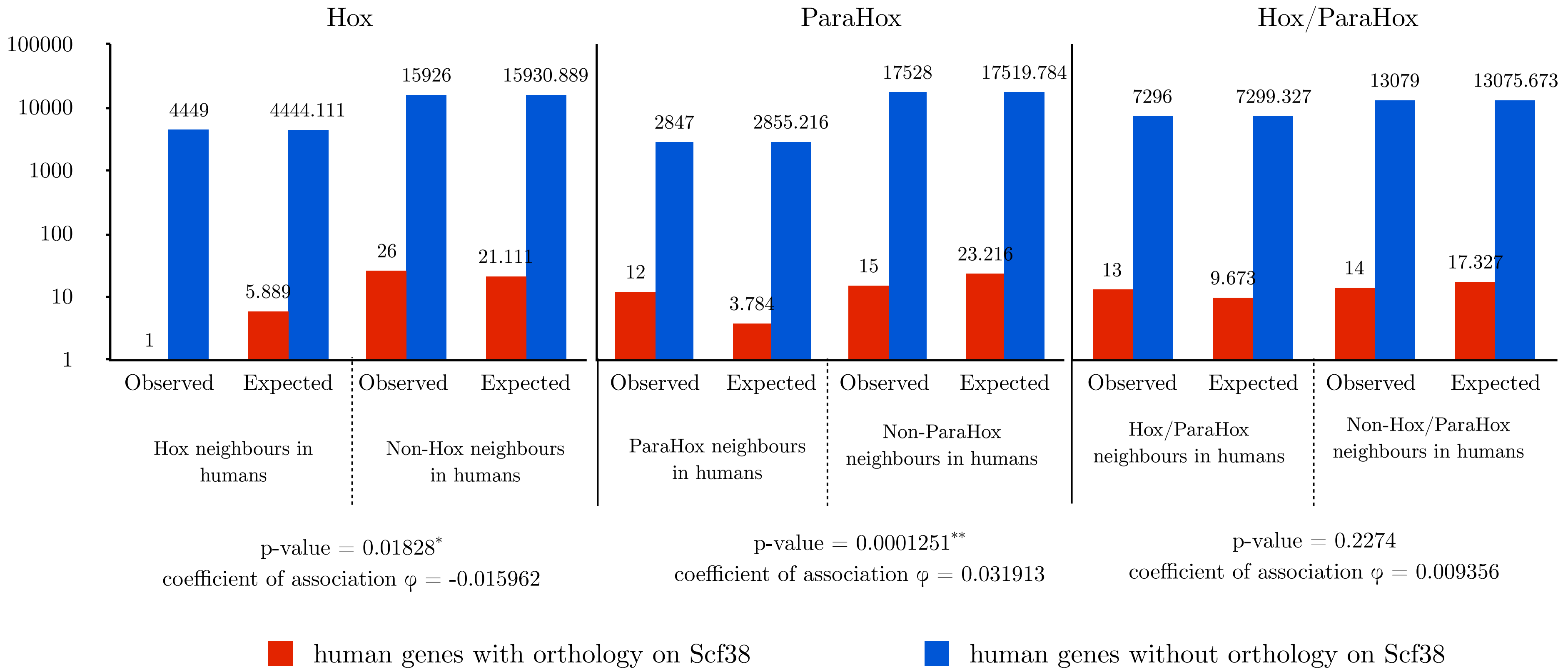


Figure 3.9.- Binomial exact test Hox-ParaHox case.

For the computation of the Fisher's Exact Tests I computed contingency tables which are based on the numbers derived in Table 3.4 for each version and are available in Appendix B, sections B.2 and B.3. In figures 3.10 and 3.11 I summarize the contingency tables and the results.



Version 1

Figure 3.10.- Version 1 Fisher exact tests for Hox, ParaHox and Hox/ParaHox cases. *Single asterisks denote statistical significance at 5%, and double asterisks denote statistical significance at 1%.*

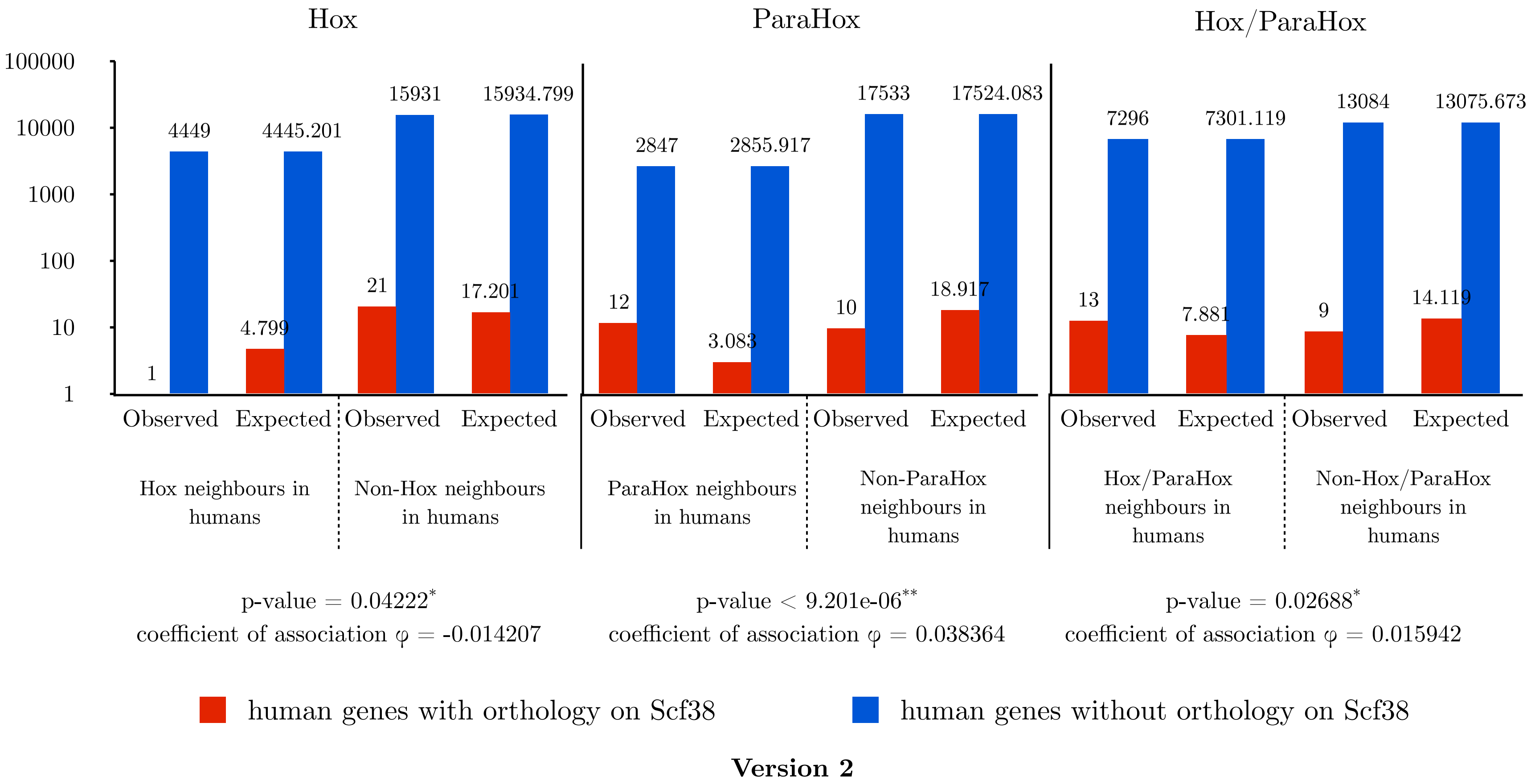


Figure 3.11.- Version 2 Fisher exact tests for Hox, ParaHox and Hox/ParaHox cases. *Single asterisks denote statistical significance at 5%, and double asterisks denote statistical significance at 1%.*

3.3.2 ...and Placozoa have a Ghost Hox locus

3.3.2.1 Identification of orthologues and synteny analysis of scaffold 3 of the *Trichoplax adhaerens* genome

Since *T. adhaerens* has a ParaHox locus with a ParaHox gene, I wanted to test whether there is a Hox locus in *T. adhaerens* that lacks a Hox gene, that is a "ghost" Hox locus. I used the Hox Putative Ancestral Linkage (PAL) gene list from *N. vectensis* (Putnam et al., 2007). The Hox PAL gene list arranges orthologues into groups that have conserved linkage across chordates and *Nematostella vectensis* Hox-bearing chromosomes and *N. vectensis* scaffolds. I used this gene list to perform BLASTp searches against the *T. adhaerens* genome, using the reciprocal best-hit criteria to compile the list of *Trichoplax* orthologues that could be part of the bilaterian-cnidarian-placozoan (BCP) Hox PAL (see Appendix B, section B.5). Starting from 267 *N. vectensis* genes in the list we found 222 orthologues in *T. adhaerens*. Of these 222 orthologues 114 are in *T. adhaerens* scaffold 3.

3.3.2.2 Statistical significance of the observed synteny conservation of scaffold 3 of the *Trichoplax adhaerens* genome

In order to test whether the apparent concentration of Hox loci neighbour orthologues found in scaffold 3 of *T. adhaerens* is significantly different from a random distribution in the *T. adhaerens* genome, I performed an Exact Binomial test (Figure 3.12). For this test I calculated the probability of a gene being in scaffold 3 of *T. adhaerens* by chance, which is the number of genes annotated in scaffold 3 (1071) divided by the total number of genes annotated in all *T. adhaerens* scaffolds (11520). The probability of a gene not being somewhere in scaffold 3 is one minus the probability of a gene being in scaffold 3.

Probability of being a Hox loci neighbour in Scaffold 3 PSc3	0.092968750
Probability of not being a Hox loci neighbour in Scaffold 3 QSc3	0.907031250
Total	1

Table 3.8.- Probabilities of a gene being in Scaffold 3 of the genome of Trichoplax adhaerens.

These probabilities are used to perform the Binomial Exact Test. The Exact Binomial Test was used to compare observed number of Hox loci neighbour orthologues (or non-Hox loci neighbours) on scaffold 3 from those expected on the basis of the probability of Hox neighbours (or ParaHox neighbours or Hox/ParaHox neighbours) in the human genome. In figure 3.12 I summarize the observed and expected numbers as well as the results.

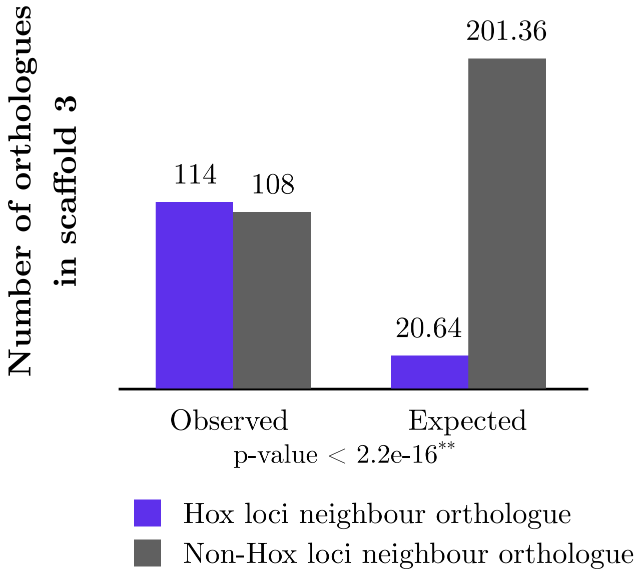


Figure 3.12.- Binomial exact test of Hox ghost loci.

I found that there are 222 *T. adhaerens* genes orthologous to cnidarian-bilaterian ancestral Hox neighbours. From those 222 genes, there are 114 genes residing in scaffold 3 of *T.adhaerens.* I found that there is a significant association of these genes residing in scaffold 3 with the cnidarian-bilaterian ancestral Hox neighbours.

3.4 Discussion

3.4.1 Trox-2 is in a placozoan ParaHox locus

To resolve whether the placozoan Hox-like gene, *Trox-2*, is a ParaHox gene or a direct ProtoHox gene descendant (Figures 3.1, 3.2, 3.3 and 3.4) I analysed the entire genomic scaffold containing *Trox-2* for conserved synteny with the human genome. First, I searched the *Trox-2* scaffold for genes with clear orthology to distinct human genes, to select genes that could be used in our statistical analyses. With this curated list of 27 *T. adhaerens* genes we tested whether the neighbours of *Trox-2* are significantly similar to the neighbours of human ParaHox loci, or instead are similar to the Hox neighbours, or lack significant synteny to human ParaHox and Hox loci. The *T. adhaerens Trox-2* scaffold shares significant synteny with the ParaHox loci of humans (Binomial and Fisher's exact tests, $P<0.0005$ Figures 3.7, 3.8, 3.9, 3.10 and 3.11). This is consistent with two scenarios. Either *Trox-2* is a ParaHox gene, in which case there should be no synteny with the human Hox loci because the ProtoHox neighbours would be expected to have distributed evenly between the descendant Hox and ParaHox loci; or *Trox-2* is a ProtoHox descendant, in which case the *Trox-2* scaffold should also have significant synteny with human Hox loci as well as the ParaHox loci, because the ProtoHox neighbours have not been split between the two loci (Hox and ParaHox). There is a significant lack of synteny with human Hox loci (Binomial and Fisher's exact tests, $P<0.02$). Synteny of *T. adhaerens Trox-2* neighbours with the human genome strongly supports a ParaHox identity for *Trox-2*. This is consistent with the topology of molecular phylogenetic trees including *Trox-2* and contradicts the hypothesis that *Trox-2* is a direct ProtoHox descendant.

3.4.2 A ghost Hox locus exists in placozoans

If *Trox-2* is indeed a ParaHox gene and an evolutionary sister (or paralogue) to Hox genes, then we would expect there to be a *T. adhaerens* locus with synteny to human Hox loci, but which lacks a Hox gene. To find this "ghost" Hox locus I used the Putative Ancestral Linkage (PAL) group information from the cnidarian *N. vectensis* genome (Putnam et al., 2007). By comparing the *N. vectensis* genome with those of chordates Putnam et al. (Putnam et al., 2007) deduced a list of 267 genes that were adjacent to the Hox genes in the cnidarian-bilaterian ancestor. I found 222 *T. adhaerens* orthologues of these cnidarian-bilaterian ancestral Hox neighbours. I found a highly significant association of these genes with *T. adhaerens* scaffold 3 (114 genes out of 222; *P*<2.2e-06; Figure 3.12). *T. adhaerens* thus has a ParaHox locus in which *Trox-2* resides, and a ghost Hox locus with synteny to cnidarian and bilaterian Hox loci but without a resident Hox gene. This implies that Hox gene(s) have been lost along the placozoan lineage and that both the Hox and ParaHox loci evolved before the origin of the Placozoa (Figure 3.13).

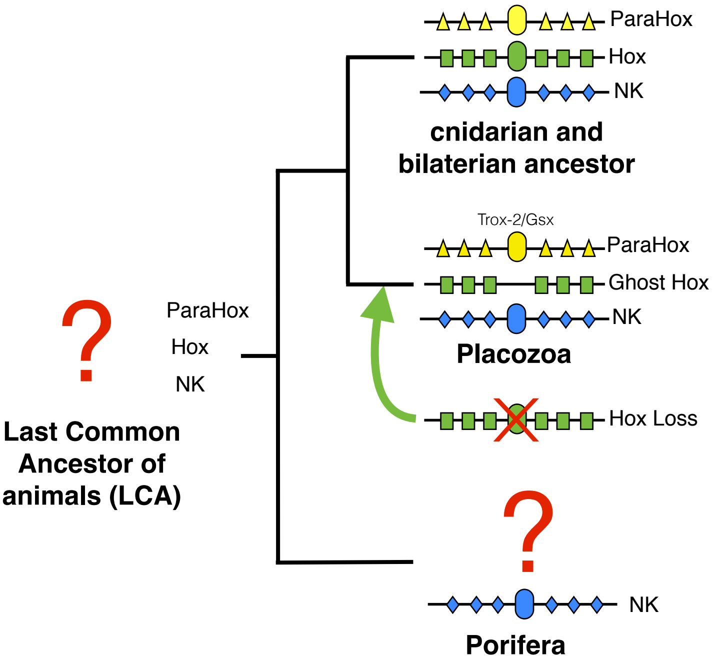


Figure 3.13.- Summary of the findings within the cnidarian and placozoan lineage. *Cnidarian and bilaterian ancestors had Hox, ParaHox and NK loci. Placozoans have lost their Hox gene(s) but retained a ghost Hox locus, and Trox-2 is a ParaHox gene in a ParaHox locus.*

Chapter 4

Reconstructing the ancestral condition of a cluster's locus. Insights from the poriferan lineage.

(Adapted from Mendivil Ramos, O., Barker, D. & Ferrier, D. E. K. 2012. Ghost Loci Imply Hox and ParaHox Existence in the Last Common Ancestor of Animals. Current Biology, 22, 1951-1956)

After pushing the origin of the Hox and ParaHox loci further back to before the placozoan lineage, I test whether these loci originated before the poriferan lineage. I use an extension of the same strategy applied in the previous chapter but here use it on a broader scale. Also, I check that the synteny signal is exclusive to metazoans and I propose a new hypothesis that pushes the origin of the Hox and ParaHox loci back to the last common ancestor of all animals.

4.1 Introduction

In the previous chapter I illustrated that both the Hox and ParaHox loci evolved before the origin of the Placozoa. The remaining lineages, Ctenophora and Porifera, are the next candidates for testing whether the ghost Hox and ParaHox loci are present, to see if the origin of the loci should be pushed even deeper in animal evolution.

The genome sequences of the poriferan *Amphimedon queenslandica* and the ctenophore *Mnemiopsis leidyi* do not possess any Hox or ParaHox genes (Larroux et al., 2007, Ryan et al., 2010). The absence of Hox and ParaHox genes from all sponges that have been examined so far, including the whole genome sequence of *A. queenslandica* (Larroux et al., 2007, Larroux et al., 2008), has led to conflicting hypotheses about whether Hox and ParaHox genes evolved before or after the origin of the poriferan lineage. Larroux et al. (2007) found a cluster of NK homeobox genes in the genome of *A. queenslandica*, which like Hox and ParaHox genes are members of the ANTP-class of genes. This combination of a cluster of genes with sequence affinity to Hox and ParaHox genes, with the lack of bona fide Hox and ParaHox genes, led Larroux et al. (2007) to propose that Hox/ParaHox genes arose from an NK gene cluster after divergence of the poriferan lineage (see Fig. 4.1B). Peterson and Sperling (2007) used phylogenetic trees to propose an alternative hypothesis, that several homeobox gene families, including the Hox and ParaHox families, were lost during poriferan evolution (see Fig. 4.1A). Poor inter-family support values within homeodomain phylogenies make it difficult to resolve between these two hypotheses with confidence.

In this case, in which there are no Hox or ParaHox genes in the genome sequence of *A. queenslandica*, testing any of the proposed hypotheses of the origin of Hox and ParaHox loci relies on inferring loci orthology. This entails looking at synteny on a large-scale in a genome with a sub-chromosomal level of assembly. I will use a comparable approach to deduce orthologous regions within the *Amphimedon queenslandica* genome to that performed in the previous chapter. However, in this case:

1) I will test whether there are the orthologues of Hox and ParaHox neighbours clustered.

2) I will test whether the Hox and ParaHox neighbour orthologues are clustered, is this clustering overlapping (ProtoHox), or not (Hox and ParaHox).

3) I will test whether the NK locus in *A. queenslandica* is distinct from the Hox and ParaHox loci, or instead the NK locus acted as the source of the ProtoHox/Hox/ParaHox loci as postulated by Larroux et al (Larroux et al., 2007).

4) I will check that the clustering of loci like those found in the metazoans analysed to date is exclusive to metazoans (or not) when compared to the sister group of metazoans, choanoflagellates, and in particular to the genome sequence of *Monosiga brevicollis.*

Here I set out the basis on which to test whether synteny can actually give any further resolution and favour any of the hypotheses regarding the evolution of Hox and ParaHox in the poriferan lineage.

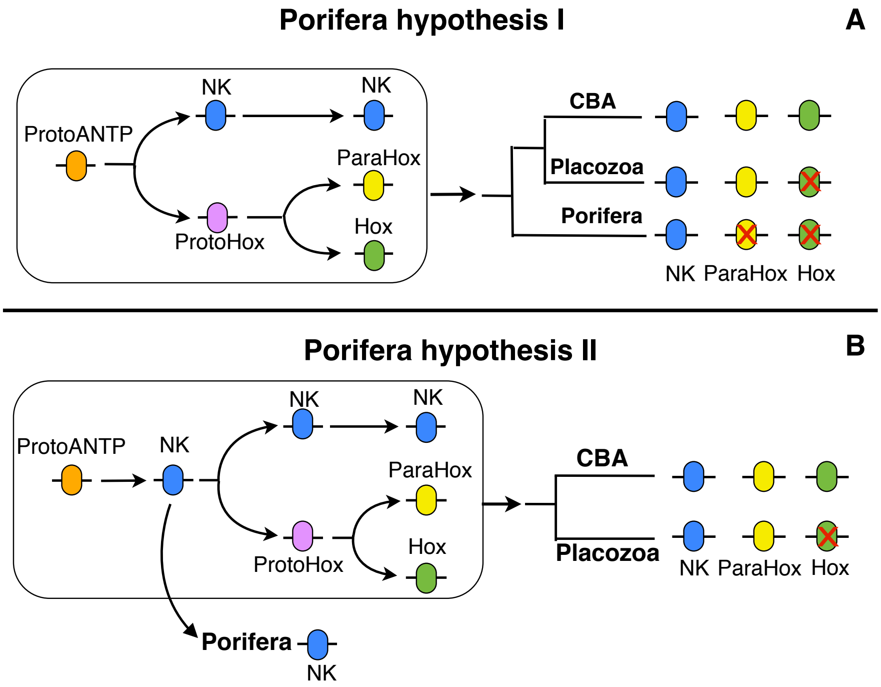


Figure 4.1 .- The ProtoHox hypothesis and alternative views of the poriferan condition. *(A) Porifera hypothesis I is that the Hox and ParaHox loci evolved before the origin of poriferans, but that these homeobox genes were lost in the sponge lineage. (B) Porifera hypothesis II is that the poriferan lineage arose before the evolution of the Hox and ParaHox loci, which evolved by duplication from the NK cluster locus.*

4.2 Materials and Methods

4.2.1 Orthologue retrieval from bilaterian-cnidarian Hox PAL gene list in *A. queenslandica.*

Orthologue retrieval was performed as specified in Chapter 2, section 2.1 but with the following modifications. The Hox Putative Ancestral Linkage (PAL) gene list from *Nematostella vectensis* (Putnam et al., 2007) was used. The Hox PAL gene list (267 genes) accommodates orthologues into groups that have conserved linkage across bilaterian Hox-bearing chromosomes and *N. vectensis* scaffolds. This list was used as a query to perform rbh (BLASTp) against the *A. queenslandica* genome-wide protein set (see Appendix C, C.1).

4.2.2 Construction of localized ParaHox PAL (l-ParaHox PAL).

There is no putative ancestral linkage gene list for the ParaHox loci in *N. vectensis*. This is due to the fact that the *N. vectensis* ParaHox synteny is more localized than the scale of analysis used by Putnam et al. (2007) (Hui et al., 2008). However, *T. adhaerens* scaffold 5 has significant synteny with the close, localized neighbourhoods of the ParaHox loci of humans. These close neighbourhoods were described by Srivastava et al. (2008) as chromosomal segments with particular coordinates.

Their annotation is from the version of the human genome corresponding to build 36. I checked whether the coordinates annotated for that genome build have changed in the current build used in this study (human genome version GRCh37.p2), and confirmed that no relevant changes had occurred. I thus used these segments to build up a localized-ParaHox PAL gene list from *T. adhaerens*. First, the number of genes (the protein coding genes, pcg) for each human segment were gathered. With each gene of the human segments a BLASTp search against the *T. adhaerens* genome-wide protein set was performed. A filter was applied to the BLASTp search outputs, retaining a gene if it is a top hit and has a bit score greater than 70 and an e-value less than 10^{-10} and is also located in *T. adhaerens* scaffold 5.

These *T. adhaerens* genes were next used for BLASTp searches against the human genome, filtering the outputs for genes that were a top hit and had a bit score greater than 70 and an e-value less than 10 and were located in the human chromosomal segments 5.4, X.6, 13.1 and 4.2. This resulted in 70 pairs of orthologues. Within these pairs were five GPCR pairs, which were discarded due to the ambiguity in their classification and the difficulty in assigning orthology with confidence (see chapter 3). This left 65 gene pairs in the PAL list (see Appendix C section C.2).

4.2.3 Orthologue retrieval from l-ParaHox PAL gene list in *A. queenslandica.*

Orthologue retrieval was performed as specified in Chapter 2 section 2.1 but with modifications. The l-ParaHox Putative Ancestral Linkage (PAL) gene list from *T. adhaerens* and the neighbouring genes of scaffold 38 of *T.adhaerens* was used. This list was used as a query to perform rbh (BLASTp) against the *A. queenslandica* genome-wide protein set (see Appendix C section C.3).

4.2.4 Monte Carlo-based test for synteny in *Amphimedon queenslandica.*

A Monte Carlo-based test for synteny was implemented and performed as follows. The genome of *A. queenslandica* is assembled to a sub-chromosomal level (i.e. scaffold level) (Srivastava et al., 2010). In order to test whether there is clustering of the Hox neighbour orthologues in a genome, I obtained an empirical null distribution of the number of scaffolds expected to be occupied by this number of genes, in absence of any conservation of synteny, based on 1000 simulations (Manly, 1991). In each simulation, all of the genes were randomly allocated to all of the scaffolds, with the scaffold randomly selected with replacement and with a probability of selection proportional to its observed gene content, with the locations of the Hox/ParaHox neighbour orthologues being recorded. This simulated genome is then compared to the actual genome scaffolds. The comparison being made is between the Hox/ParaHox neighbour orthologues placed at random and the expected probability of Hox/ParaHox neighbour orthologues for each scaffold. If the content of Hox/ParaHox neighbour orthologues observed in a scaffold exceeds the expected probability of Hox/ParaHox neighbour orthologues of that scaffold, as judged from the simulated genome, for that cycle the “exceeded probability” would increase by one. This comparison was performed for all simulated scaffolds. The cycle ends once this comparison is finished. Each cycle is repeated 1000 times. In practice the “exceeded probability” always equalled the number of scaffolds occupied by one or more of the Hox/ParaHox neighbour orthologues . See Appendix C, C.4.

The empirical *P* value for a one-tailed test of the alternative hypothesis of clustering may be calculated as the proportion of simulations in which the number of scaffolds occupied by a certain number of genes is less than or equal to the actual number observed.

The test for a ProtoHox scenario, with the results obtained from the simulation, was done as follows: for each cycle of both the Hox and ParaHox simulations the number of scaffolds with an overlap of at least one orthologue of a Hox neighbour and at least one orthologue of a ParaHox neighbour was recorded. The empirical *P* value for a test of the alternative hypothesis of clustering versus the null hypothesis of no clustering was calculated as the proportion of simulations in which the number of scaffolds with both kinds of orthologue was greater than or equal to the observation.

4.2.5 Synteny analysis of NK loci of *A. queenslandica* and statistical test

Orthologue retrieval was performed as specified in Chapter 2 section 2.1 but with the following modifications. The neighbouring genes of scaffold 13506 of *Amphimedon queenslandica*, in which the NK cluster resides, were used to perform orthologue retrieval via rbh (BLASTp) searches against the lophotrochozoan genomes of *Capitella teleta* and *Lottia gigantea.* (Codes are available from Appendix C, C.5)

4.2.6 Orthologue retrieval from BCP Hox, l-ParaHox PAL and *T. adhaerens* scaffold 38 gene list in *Monosiga brevicollis* genome.

Orthologue retrieval was performed as specified in Chapter 2 section 2.1 but with the following modifications. First, the Hox Putative Ancestral Linkage (PAL) gene list from *Nematostella vectensis* (Putnam et al., 2007) was used. The Hox PAL gene list (267 genes) accommodates orthologues into groups that have conserved linkage across bilaterian Hox-bearing chromosomes and *N. vectensis* scaffolds. This list was used as a query to perform rbh (BLASTp) against the *M. brevicollis* genome. Second, the l-ParaHox PAL (see previous section) and the neighbouring genes of scaffold 38 in *T. adhaerens* were used. These two lists were used as queries to perform rbh (BLASTp) against the *M. brevicollis* genome. See Appendix C, C.6 and C.7.

4.2.7 Monte Carlo-based test for synteny in *M. brevicollis* genome.

Performed as in section 4.2.4 but in this case using the *M. brevicollis* genomes and the orthologues retrieved.

4.3 Results

4.3.1 Sponges have distinct Hox and ParaHox loci...

4.3.1.1 Identification of orthologues in *Amphimedon queenslandica* using the bilaterian-cnidarian-placozoan (BCP) Hox PAL gene list

Using the same logic as I did for the *T. adhaerens* ghost Hox locus, I first wanted to determine whether there are orthologues of human Hox loci neighbours in the *A. queenslandica* genome and then deduce whether these orthologues are clustered.

In order to accomplish this, I used the BCP Hox PAL gene list to conduct BLASTp searches against the *A. queenlandica* genome. I followed the reciprocal best-hit criteria to find putative orthologues to the Hox loci neighbours in *A. queenslandica.* From here I produced a list of 187 *A. queenslandica* genes orthologous to the BCP Hox PAL genes (see Appendix C, section C.1). The 187 genes are distributed in the scaffolds as shown in Figure 4.2.

4.3.1.2 Monte Carlo-based test for synteny conservation of the BCP Hox PAL genes in the *A. queenslandica* genome

The *A. queenslandica* genome is assembled to a subchromosomal level (i.e. scaffold level) and therefore, chromosome-level linkage is not immediately

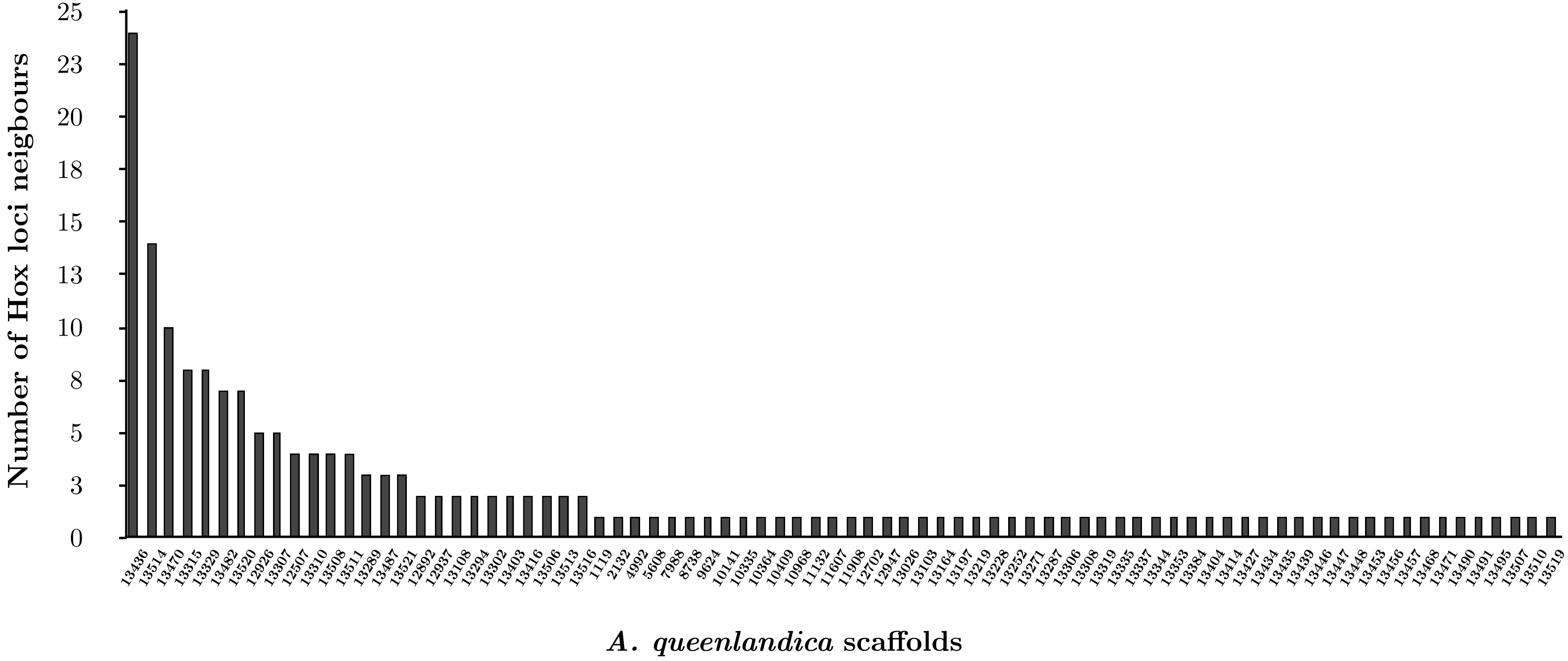


Figure 4.2.- Distribution of orthologues of Hox loci neighbours in A. queenslandica genome scaffolds.

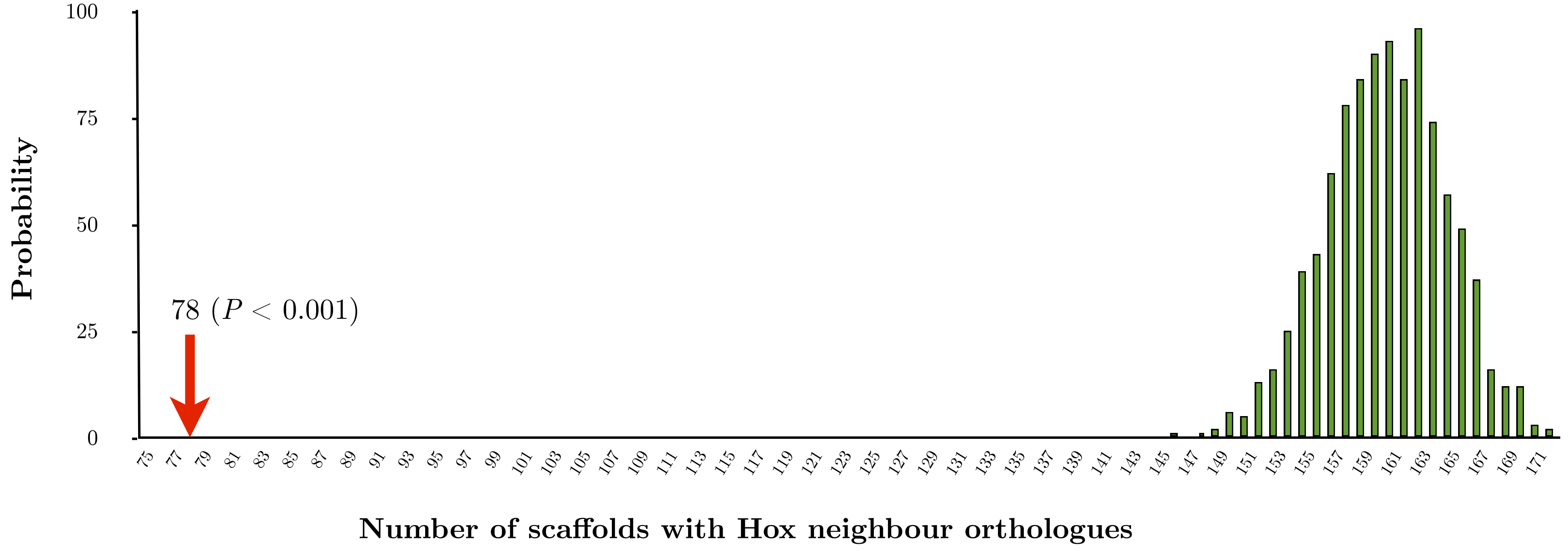


Figure 4.3.- Sponges have a distinct ghost Hox locus. *Simulation of randomized location of A. queenslandica orthologues of human Hox neighbours across the sponge scaffolds. The arrow indicates observed number of scaffolds in A. queenslandica that actually contain the 187 orthologues of Hox neighbours.*

apparent. This meant that I had to test whether the observed distribution of the Hox loci neighbour orthologues across the scaffolds are clustered. The test was designed on the basis of a Monte Carlo experiment, and entailed the generation of an empirical null distribution based upon 1000 simulations. Each simulation is the number of *A. queenslandica* scaffolds expected to be occupied by the 187 genes in the absence of any conservation of synteny. That is, the 187 genes are randomly scattered across the *A. queenslandica* scaffolds, and are not clustered. The empirical null distribution obtained after the Monte Carlo-based experiment is shown in Figure 4.3. The calculated empirical *P*-value for a one-tailed test of the alternative hypothesis of clustering versus the null hypothesis of no clustering is less than 0.001 as is indicated by the distribution to the left of the red arrow in Figure 4.3.

4.3.1.3 Creation of *T. adhaerens* localised-ParaHox PAL

There is no putative ancestral linkage gene list for the ParaHox loci in *N. vectensis.* This is due to the fact that the *N. vectensis* ParaHox synteny is more localized than the scale of analysis used by Putnam et al (2007) (Hui et al., 2008). However, *T. adhaerens* scaffold 5 has significant synteny with the close, localized neighbourhoods of the human ParaHox genes (see Tables S8.2 and S8.3 in(Srivastava et al., 2008)). These close neighbourhoods were described by Srivastava et al. (2008) as chromosomal segments with particular coordinates and are summarised in the following table with the add-on of number of genes (i.e. protein coding genes) for each segment:

Chromosome	Segment Name	Molecular Coordinates	Number of genes per segment
5	5.4	139835480-167951722	210
X	X.6	70406305-106924338	157
13	13.1	1-41837067	125
4	4.2	25986602-57101698	103

Table 4.1.- Human chromosomal segments containing the ParaHox "clusters". *Identified by Srivastava et al. (2008) with significant synteny to the T. adhaerens genome scaffold 5.*

Their annotation is dated for the version of the human genome corresponding to build 36. I checked whether the coordinates annotated for that genome build have changed in the current build used in this study (i.e. checking in the archive of ensembl and their web-based checker of build 36 versus the human build 37 patch 2). I confirmed that no relevant change had occurred and so used these segments to build up a localized-ParaHox PAL gene list from *T. adhaerens.* This list contains 70 pairs of *T. adhaerens*-human orthologues. Within these pairs are five GPCR pairs. I discarded these due to the ambiguity in their classification and the difficulty in assigning orthology with confidence, as discussed for TRIADDRAFT_62215 in Chapter 3, which left 65 gene pairs in our localised-ParaHox PAL list (see Appendix C, section C.2). This localized-ParaHox PAL gene list was used to test for a ghost ParaHox locus in the *A. queenslandica* genome.

It is noteworthy that scaffold 5 has the clear ParaHox neighbourhood synteny signal in the analyses of Srivastava *et al.* (2008), and not scaffold 38, which contains *Trox-2.* This is because scaffold 38 is too small, with too few genes, to be included in the *T.adhaerens* synteny analysis of Srivastava *et al.* (2008). I predict that *T. adhaerens* scaffold 5 and 38 are potentially closely linked in the placozoan genome.

4.3.1.4 Identification of orthologues in *A. queenslandica* genome using *T. adhaerens* localized-ParaHox PAL (l-ParaHox PAL) gene list

I used the same procedure as I did for finding the *T. adhaerens* ghost Hox locus to first determine whether there are orthologues of human ParaHox neighbours in *A. queenslandica,* and second, to deduce whether these orthologues are clustered. I found 44 l-ParaHox PAL orthologues in the *A. queenslandica* genome (see Appendix C, section C3), distributed in the manner shown in Figure 4.4.

4.3.1.5 Monte Carlo-based test for synteny conservation of the l-ParaHox PAL genes in the *A. queenslandica* genome

I performed the same simulations as for the Hox loci neighbours, but incorporating the number of ParaHox neighbour orthologues determined in the previous section. The empirical null distribution obtained after the Monte Carlo-based test is shown in Figure 4.5. The calculated empirical *P*-value for a one-tailed test of the alternative hypothesis of clustering versus the null hypothesis of no clustering is less than 0.001 as is indicated by the distribution to the left of the red arrow in Fig. 4.5.

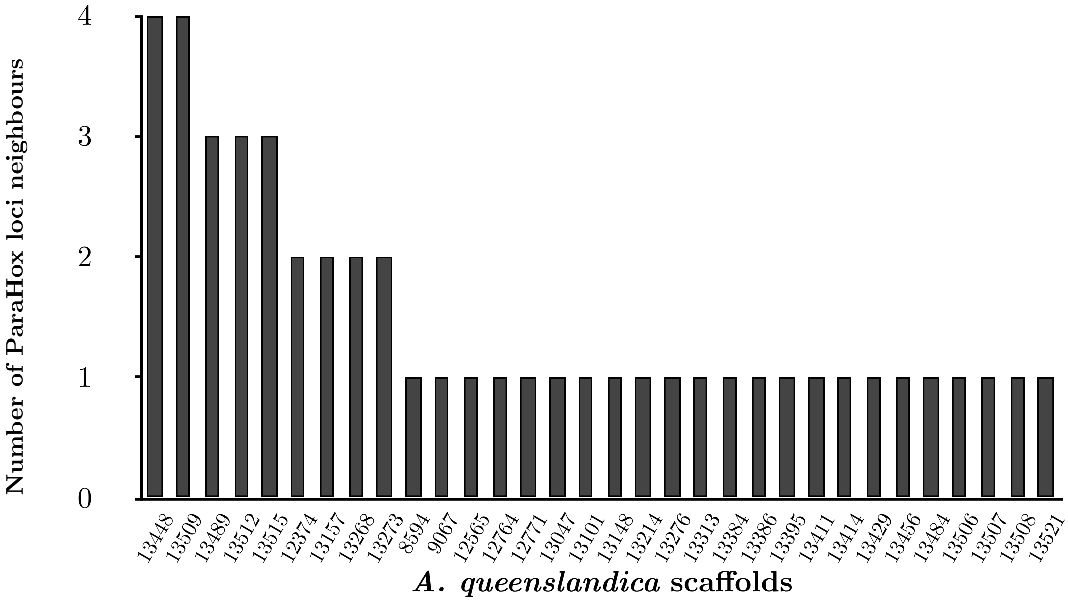


Figure 4.4.- Distribution of ParaHox locus neighbour orthologue in A. queenslandica genome scaffolds.

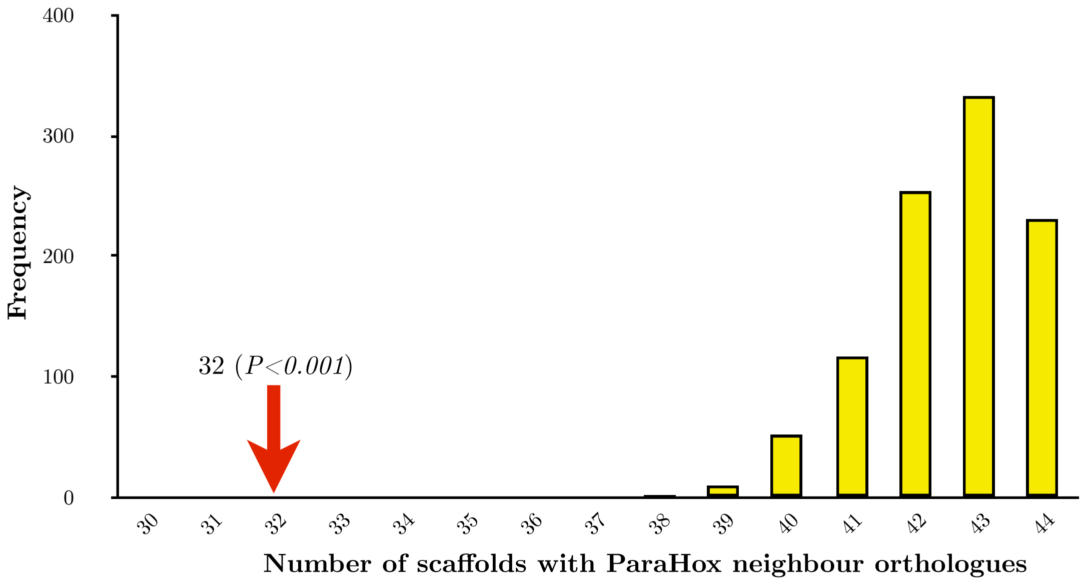


Figure 4.5.- Sponges have a distinct ghost ParaHox locus. *Simulation for ParaHox neighbour orthologues. The arrow indicates observed number of scaffolds with orthologues of ParaHox neighbours in A. queenslandica.*

4.3.1.6 Determining whether the *A. queenslandica* genome has a ghost ProtoHox locus or ghost Hox and ParaHox loci.

In order to infer whether the clustered Hox and ParaHox neighbour orthologues in *A. queenslandica* are coincident, as would be expected for a ProtoHox locus, or whether they are distinct, independent ghost loci, I used the output of both Hox and ParaHox simulations from above. For each cycle of both experiments I recorded how many scaffolds had an overlap, with at least one orthologue of a Hox neighbour and at least one orthologue of a ParaHox neighbour. The distribution is shown in Figure 4.6.

The empirical *P* value for the test of the alternative hypothesis of coincident Hox and ParaHox neighbour clustering versus the null hypothesis of random co-occurrence of Hox and ParaHox neighbours is 0.316 and is represented by the arrow in Figure 4.6. This implies that the overlap is not significantly different from random, and that *A. queenslandica* has separate ghost Hox and ParaHox loci, as opposed to a ProtoHox condition which would have entailed the overlap of Hox and ParaHox neighbours occurring with a probability beyond the upper tail of the empirical null distribution.

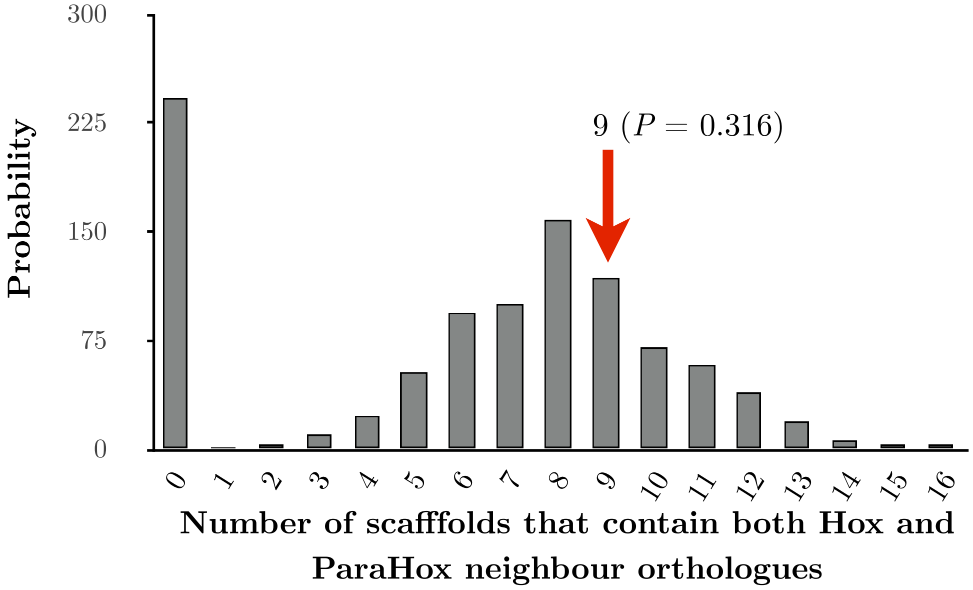


Figure 4.6.-Sponges have distinct ghost Hox and ParaHox loci. *Overlap plot of both simulations to distinguish whether the Hox and ParaHox neighbour clustering is coincident or distinct. The arrow indicates observed number of scaffolds with co-localization of Hox and ParaHox neighbour orthologues in A. queenslandica.*

4.3.2 Sponges also have a distinct NK locus

4.3.2.1 Synteny analysis of NK loci of *A. queenslandica* and statistical significance of observed synteny

As a further test of whether the Hox and ParaHox loci are already distinct from the NK locus in *A. queenslandica* (as implied above) or whether the NK locus acted as the source of the ProtoHox/Hox/ParaHox loci (as inferred by Larroux et al. (2007)), I analysed the neighbouring genes of the NK cluster-bearing scaffold in *A. queenslandica* (scaffold 13506). I performed orthologue retrieval by BLASTp searches against the lophotrochozoan genomes of *Capitella teleta* and *Lottia gigantea*. I did not use ecdysozoan genomes due to their extensive genome rearrangements, particularly with respect to the linkage patterns of the ANTP-class genes (Larroux et al., 2007, Wotton et al., 2009). Also, vertebrate genomes cannot be used for this particular NK-versus-ParaHox/Hox linkage analysis because in vertebrates some NK clusters have become secondarily linked with some ParaHox loci. It is known that these linkages do not reflect the ancestral chordate condition from the data from amphioxus and *Platynereis durmerilii* (Hui et al., 2012).

I used the reciprocal BLAST best-hit criteria to identify orthologues of the *A. queenslandica* NK cluster neighbours. Then, I determined which of these genes localised to either NK cluster gene-bearing scaffolds, Hox gene-bearing scaffolds, or ParaHox gene-bearing scaffolds in both *C. teleta* and *L. gigantea*. In *C. teleta* nine orthologues are located on NK-cluster gene scaffolds, which themselves have a total number of 239 genes (excluding the homeobox genes themselves). In *L. gigantea* 35 orthologues are on NK-cluster gene scaffolds, which contain a total of 1,246 genes. For the ParaHox scaffolds *C. teleta* has zero orthologues of sponge NK neighbours from a total of 28 genes, whilst *L. gigantea* has four out of 167. For the Hox scaffolds *C. teleta* has one orthologue out of 104 genes, and *L. gigantea* has one orthologue out of 360 genes.

I used an Exact Binomial Test to test whether this distribution of orthologues of sponge NK neighbours in lophocotrozoan NK, ParaHox and Hox scaffolds represents statistically significant synteny with the *A. queeslandica* NK cluster scaffold. I calculated the probability of a gene being on a Hox scaffold as the total number of annotated genes in the Hox scaffolds (*C. teleta* 104, *L. gigantea* 360), divided by the total number of annotated genes for the genome (*C. teleta* 32415, *L. gigantea* 23851). The probability of a gene being on a ParaHox scaffold is the total of the annotated genes on ParaHox scaffolds (*C. teleta* 28, *L. gigantea* 167), divided by the total number of annotated genes for the genome (*C. teleta* 32415, *L. gigantea* 23851). Finally, the probability of a gene being on an NK scaffold is the total of the annotated genes in the NK scaffolds (*C. teleta* 239, *L. gigantea* 1246), divided by the total number of annotated genes for the genome (*C. teleta* 32415, *L. gigantea* 23851) (see Tables 4.2, 4.3 and 4.4). In order to test whether the apparent concentration of NK loci neighbours found in *C.teleta* and *L. gigantea* genomes is similar to the one in the *A. queenslandica* NK-bearing scaffold (Contig13506) I performed a Binomial Exact Test. The same test was conducted to test whether the apparent concentration of Hox and ParaHox loci neighbours found in *C. teleta* and *L.*

gigantea is significantly different to that in the *A. queenslandica* NK-bearing scaffold (Tables 4.3 and 4.4).

	Capitella teleta	***Lottia gigantea***
Probability of a gene being in a Hox scaffold P_H	0.00320839118	0.0150937068
Probability of a gene being in a ParaHox scaffold P_{PH}	0.00086379762	0.0070018029
Probability of a gene being in a NK scaffold P_H	0.00737312972	0.0522409962

Table 4.2.- Probabilities of a gene being in Hox, ParaHox or NK scaffolds in Capitella teleta and Lottia gigantea genomes.

Capitella teleta	scaffold	capacity	neighbouring orthologues
NK p-value = 2.501e-5 **	815	14	0
	493	16	0
	315	26	2
	725	20	0
	95	63	2
	33020	1	0
	31	89	3
	694	10	2
ParaHox p-value = 0.4809	70	29	0
	292	13	0
	33	62	1
Hox p-value = 1	760	10	0
	444	18	0

Table 4.3.- Summary of gene numbers of Hox-, ParaHox- and NK-bearing scaffolds and p-values of Binomial Exact Test in Capitella teleta.

Lottia gigantea	scaffold	capacity	neighbouring orthologues
NK p-value = 2.3e-10 **	122	44	0
	19	277	9
	72	97	0
	40	168	4
	88	88	5
	21	245	8
	9	321	9
	263	6	0
ParaHox p-value = 0.0508	85	82	3
	80	85	1
Hox p-value = 0.3789	12	360	1

Table 4.4.- Summary of gene numbers of Hox-, ParaHox- and NK-bearing scaffolds and p-values of Binomial Exact Test in Lottia gigantea.

The results from these tests show there is a significant association of the lophotrochozoan orthologues of the *Amphimedon* NK cluster neighbours with the lophotrochozoan NK gene-containing scaffolds. Also, there is no significant association with either the lophotrochozoan Hox or ParaHox-containing scaffolds. This implies that the NK cluster locus of *Amphimedon* is orthologous with the NK loci of the lophotrochozoans, but that there is no association with the ParaHox or Hox loci and thus, no synteny-based evidence for the Larroux *et al.* (2007) hypothesis of the ProtoHox/Hox/ParaHox cluster evolving from an NK cluster.

4.3.3 Hox and ParaHox loci are metazoan-specific

4.3.3.1 Identification of orthologues in *Monosiga brevicollis* using the bilaterian-cnidarian Hox PAL gene list and l-ParaHox PAL and *T. adhaerens* scaffold 38 gene list

I wanted to test whether the clustering of Hox and ParaHox neighbour orthologues is exclusive to metazoans, as might be predicted from the complete lack of ANTP-class homeobox genes from non-metazoan lineages, and whether the ProtoHox condition evolved with the origin of the Metazoa. For this purpose I used the genome of *Monosiga brevicollis*, as a representative from the choanoflagellate sister group of metazoans (King et al., 2008). Using the same logic as used for the *T. adhaerens* and *A. queenslandica* ghost Hox loci, I wanted to first find whether there are orthologues of bilaterian-cnidarian Hox loci neighbours in *M. brevicollis* and then infer whether these orthologues are clustered or not. Also, I wanted to determine whether there is clustering of orthologues of the l-ParaHox PAL genes that I deduced from the comparions between *T. adhaerens*, *N. vectensis* and humans. I also included a search for *Monosiga* orthologues of the genes in *T. adhaerens* scaffold 38, which contains the ParaHox gene *Trox-2.*

In order to accomplish the first aim I used the BC Hox PAL gene list to perform BLASTp searches against the *M. brevicollis* genome. I followed the reciprocal best-hit criteria to find putative orthologues to the Hox loci neighbours in *M. brevicollis.* This produced a list of 139 *M. brevicollis* genes (see Appendix C, section C.4). Similarly, the search for putative orthologues to the ParaHox loci neighbours in *M. brevicollis* produced a list of 52 *M. brevicollis* genes orthologous to the l-ParaHox PAL list (41 orthologues) and to *T. adhaerens* scaffold 38 (the *Trox-2* scaffold) (11 orthologues) (see Appendix C, section C.5).

4.3.3.2 Monte Carlo-based test for synteny conservation of the Hox and ParaHox loci neighbours in the *Monosiga brevicollis* genome.

I performed the same simulations as for the Hox and ParaHox loci neighbour analyses in *A. queenslandica*, but incorporating the number of Hox (139) and ParaHox (52) neighbour orthologues in *M. brevicollis.* Also, I used the total number of genes for *M. brevicollis*, 9196, and the total number of scaffolds, 218, with their respective gene densities. The empirical null distributions obtained after the Monte Carlo-based tests are shown in Figures 4.7 and 4.8.

The calculated empirical P-value for a one-tailed test of the alternative hypothesis of clustering versus the null hypothesis of no clustering is 0.703 for Hox and 0.903 for ParaHox and is indicated by the red arrows.

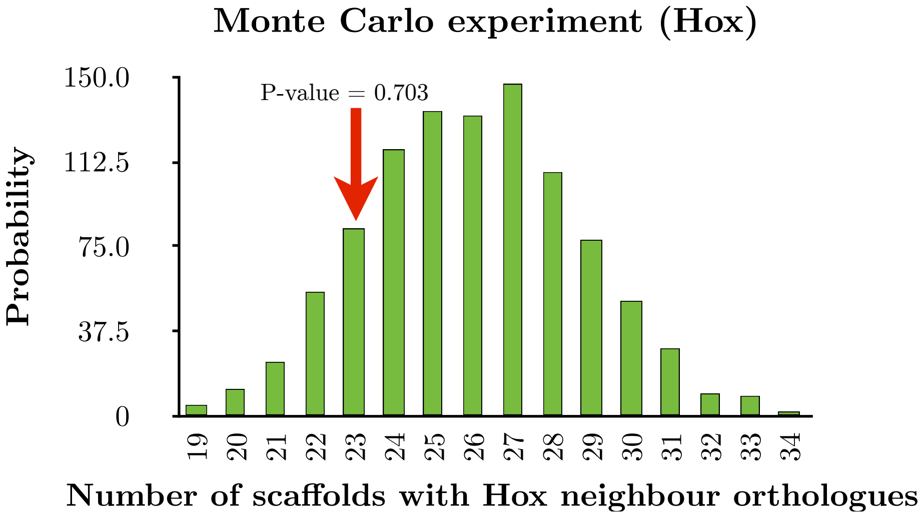


Figure 4.7.- Histogram of the Monte Carlo experiments of Hox PAL genes found in M. brevicollis*. Simulation of randomized location of M. brevicollis orthologues of bilaterian-cnidarian Hox neighbours across M. brevicollis scaffolds. Red arrow indicates observed number of scaffolds with Hox neighbour orthologues.*

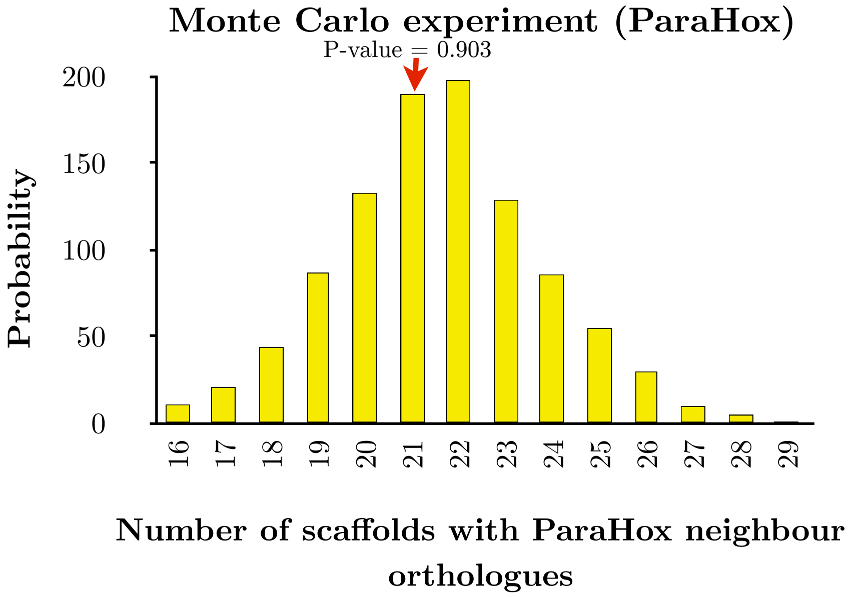


Figure 4.8.- Histogram of the Monte Carlo experiments of ParaHox PAL genes found in M. brevicollis*. Simulation of randomized location of M. brevicollis orthologues of bilaterian-cnidarian ParaHox neighbours across M. brevicollis scaffolds. Red arrow indicates observed number of scaffolds with ParaHox neighbour orthologues.*

The observed distribution of Hox and ParaHox neighbour orthologues in *M. brevicollis* does not differ from the null simulated distributions that represent random distributions of these genes across the choanoflagellate genome (see Figs. 4.7 and 4.8). This lack of clustering of these genes reveals that there are no ghost Hox and ParaHox loci in *M. brevicollis*. As expected, the Hox and ParaHox loci thus appear to be specific to the Metazoa.

4.4 Discussion

4.4.1 Sponges have distinct Hox and ParaHox loci

Here I have described how I tested whether Hox and ParaHox loci can be detected even earlier in animal evolution. Porifera constitute the lineage most commonly considered to be more basal than Placozoa and Cnidaria within the animal phylogeny (Philippe et al., 2009, Pick et al., 2010), and a whole genome sequence from a sponge is available, from *A. queenslandica*.

Using the Hox PAL gene list derived from *N. vectensis*-bilaterian comparisons I found 187 orthologues in *A. queenslandica*. I then tested whether these 187 sponge genes are clustered in the *A. queenslandica* genome as a ghost Hox (or ProtoHox) locus, or are randomly scattered throughout the genome, as might be the case if the Hox locus did not evolve before the origin of poriferans. This last scenario could also, alternatively, be interpreted as the *A. queenslandica* genome having become rearranged to the extent that synteny with other phyla has been largely lost. According to simulations, the 187 *A. queenslandica* genes show significant evidence of clustering onto a small number of scaffolds (one-tailed test of clustering, $P < 0.001$, Fig. 4.3).

This clustering of cnidarian-bilaterian Hox neighbour orthologues in this sponge can reflect one of two possibilities: either *A. queenslandica* has a ghost Hox locus, or this animal has a ghost ProtoHox locus. To distinguish between these two possibilities we determined whether *A. queenslandica* has a ghost ParaHox locus that is distinct from the ghost Hox locus, as would be expected if the origin of the Hox and ParaHox loci occurred before the origin of the Porifera. If instead sponge orthologues of ParaHox gene neighbours cluster in a fashion co-localized with the above Hox neighbour clustering, then this would imply the existence of a ghost ProtoHox locus, with the duplication into Hox and ParaHox loci occurring after the divergence of poriferans. To determine whether orthologues of ParaHox neighbours are clustered in *A. queenslandica* I first constructed a list of human ParaHox neighbouring genes that are also neighbours in the placozoan *T. adhaerens,* and hence form a ParaHox PAL in the placozoan-cnidarian-bilaterian ancestor. I used the synteny information of Srivastava et al. (Srivastava et al., 2008), which matched human genome segments containing the human ParaHox loci with a single scaffold in the *T. adhaerens* genome (scaffold 5). From the 595 genes in these human genomic segments I found 167 genes on *T. adhaerens* scaffold 5, which when filtered for reciprocal best BLAST hits back to specific human ParaHox segments resulted in 65 genes in the localized-ParaHox PAL list (l-ParaHox PAL). Using this l-ParaHox PAL list I detected 44 *A. queenslandica* genes. These 44 sponge genes cluster together on significantly fewer scaffolds than expected for randomly distributed genes (one-tailed test for clustering $P<0.001$, see Fig. 4.5).

Furthermore, I tested whether these clustered ParaHox and Hox PAL orthologues co-localise representing the ProtoHox condition, or whether they instead form two distinct loci representing the Hox and ParaHox condition. The observed number of *A. queenslandica* scaffolds containing both Hox and ParaHox PAL orthologues is nine, which does not differ significantly from the null expectation of random co-localization (one-tailed, $P = 0.316$, Fig. 4.6), providing no significant evidence for the ProtoHox hypothesis. I conclude that the clustering of Hox PAL orthologues is distinct from the ParaHox PAL orthologue clustering in *A. queenslandica*, which implies that distinct Hox and ParaHox ghost loci exist in this poriferan. This is consistent with the gene loss hypothesis explaining the absence of Hox and ParaHox genes in sponges (Fig. 4.1 (A)), and is inconsistent with the hypothesis of Hox/ParaHox (or ProtoHox) genes arising from an NK gene cluster (Fig. 4.1 (B)). I found further evidence against the NK-ProtoHox hypothesis (Fig. 4.1 (B)) from an analysis of the genes neighbouring the *A. queenslandica* NK cluster, which show no significant linkage with the Hox or ParaHox loci of bilaterians, in contrast to what might have been expected if the Hox/ParaHox/ProtoHox genes had evolved from duplication of the NK locus (Fig. 4.1 (B)). I also found that the existence of ghost Hox and ParaHox loci is restricted to the animals. Analysis of the genome of a choanoflagellate, *M. brevicollis*, from the sister group to the Metazoa revealed no clustering of the orthologues of the metazoan Hox and ParaHox neighbours (Figs. 4.7 and 4.8).

4.4.2 A last common ancestor with Hox and ParaHox was followed by gene loss

The assumption underlying all the analyses is that the Hox and ParaHox loci evolved by duplication of a ProtoHox locus such that neighbours of the ProtoHox cluster distributed relatively equally with the post-duplication Hox and ParaHox loci (Fig. 4.9).

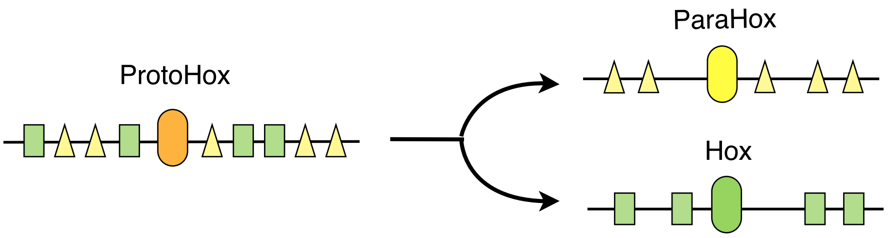


Figure 4.9.- The ProtoHox hypothesis.

If instead the Hox/ParaHox genes evolved by some mechanism like a retrotransposition or a small-scale DNA-based transposition, then the daughter gene would have inserted into a distinct genomic location without necessarily taking neighbours from the parent (ProtoHox) locus. I consider this less likely than the ghost loci hypothesis (see Chapter 7), which merely implies duplication and gene loss, a phenomenon that is known to be common (Hughes and Friedman, 2004, Danchin, 2006, Miller et al., 2007, Wyder et al., 2007, Takahashi et al., 2009), and which is consistent with gene phylogeny topologies (Peterson and Sperling, 2007).

The discovery of ghost Hox and ParaHox loci in a sponge, and a ParaHox locus containing *Trox-2* alongside a ghost Hox locus in a placozoan, implies that the last common ancestor of animals possessed distinct Hox and ParaHox loci (Fig 4.9). This, in turn, implies loss of these homeobox genes during the evolution of some basal animal lineages, which, in terms of these developmental control genes, have been simplified relative to the last common ancestor of animals (Figure 4.10).

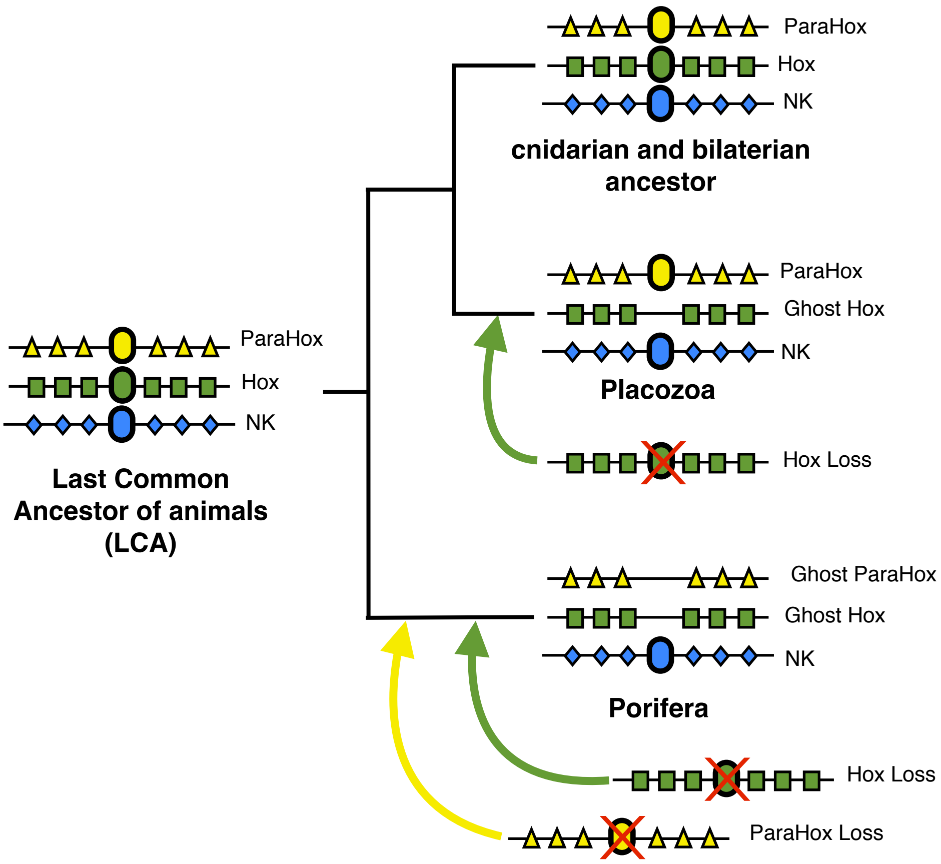


Figure 4.10.- Last Common Ancestor of animals had Hox, ParaHox and NK loci. *Placozoans have lost their Hox gene(s) but retained a ghost Hox locus, and Trox-2 is a ParaHox gene in a ParaHox locus. Poriferans have lost Hox and ParaHox genes but retained distinct ghost Hox and ParaHox loci. Cnidarian and bilaterian ancestors had Hox, ParaHox and NK loci as did the Last Common Ancestor of animals.*

Chapter 5

Are there ParaHox genes in the calcareous poriferans *Sycon ciliatum* and *Leucosolenia sp.*?

(Adapted from Fortunato S. et al. “The ANTP complement of calcareous sponges” in preparation)

This chapter describes my contribution to determining the orthology assignment of potential ParaHox genes in the sponges *S. ciliatum* and *Leucosolenia sp.*. I also describe the gene neighbours surrounding this *Sycon ciliatum* gene as an alternative means to give resolution in the orthology assignment. Finally, I describe how this gene supports the new hypothesis that Hox and ParaHox genes existed in the last common ancestor of the animals.

5.1 Introduction

To date, separate ANTP-class gene surveys have not identified any Hox or ParaHox genes in a variety of sponges. This has led to competing views about the origin of Hox and ParaHox genes and disagreement as to whether the ancestor of sponges did or did not have Hox and ParaHox (or ProtoHox) genes. Chapters 3 and 4 demonstrated the existence of ghost Hox and ParaHox loci in a sponge, *Amphimedon queenslandica*, implying that these homeobox genes were lost during the evolution of the sponge lineage (Peterson and Sperling, 2007, Mendivil Ramos et al., 2012). However, a precise timing of the loss of Hox and ParaHox is still unclear. That is, whether these gene losses occurred early in sponge evolution, before the various classes arose and diverged, or instead the gene losses happened multiple times independently in distinct poriferan lineages. This can be tested by investigating further poriferan lineages in addition to the ones already examined.

The group of Dr. Maja Adamska (Sars Institute, Bergen) recently sequenced the whole genome sequence of *Sycon ciliatum* and they have been cataloguing the homeobox complement of this sponge, with their current focus on the NK families. In parallel they are also sequencing the genome of another calcareous sponge *Leucosolenia sp.* from which its homeobox complement has been isolated. In collaboration with the Adamska group, I have been analysing a particular ANTP-class homeobox gene that may have some affinity with a ParaHox gene. If a ParaHox gene is present in a sponge this could verify the ghost loci hypothesis and be an independent proof of the results and conclusions presented in Chapter 3 and 4.

Sycon ciliatum and *Leucosolenia sp.* are calcareous sponge and their genome sequence are the only representatives of this lineage. To date, the phylogenetic relationship amongst sponge lineages (Demospongiae, Hexactinellida, Calcarea, and Homoscleromorpha) is unclear (Wörheide et al., 2012). This is especially challenging for the Calcarea lineage as classical and molecular systematics are largely in disagreement as to its “correct” phylogenetic position (Wörheide et al., 2012). Furthermore, there are currently two competing views, one proposing the monophyletic relationship of all sponge lineages and the other proposing sponge paraphyly ((Wörheide et al., 2012) see Fig. 5.7). These two competing hypotheses imply that I will have to carefully consider how to frame my comparative analyses regarding this newly found homeobox gene (see later Section 5.4).

The first task is to assign orthology to this homeobox gene. If this newly found gene is indeed a ParaHox gene, this will require a reassessment of the current understanding of the Hox and ParaHox complement. It would also be interesting to determine whether the immediate surrounding neighbours of this candidate gene reveal conserved synteny to human loci and to the PALs that I described in Chapters 3 and 4. I attempt to identify the orthology of this homeobox gene by multiple sequence and motif comparisons, phylogenetic analyses and by an examination of synteny in the case of *Sycon*.

5.2 Materials and Methods

5.2.1 Genome sequencing and annotation of *Sycon ciliatum* and isolation of ANTP-class genes

Performed by the Adamska group.

5.2.2 Orthologue analysis of 34059 of *Sycon ciliatum* and 70333 of *Leucosolenia sp.*

Homeodomain sequences from *S. ciliatum* and *Leucosolenia sp.* were kindly provided by Sofia Fortunato from Dr. Maja Adamska's research group at Sars Institute (Bergen) available in Appendix D, D.1. The orthologue analysis was performed as specified in Chapter 2, Section 2.1. ANTP-class and PRD-class homeodomain and homeobox gene sequences were downloaded from HomeoDB and/or GenBank and are available in Appendix D, D.1. PRD-class sequences were included as an outgroup. The acronyms of species used in multiple alignment and phylogenetic trees are Hsa (human), Bfl (*Branchiostoma floridae*), Cte (*Capitella teleta*), Lgi (*Lottia gigantea*), Nve (*Nematostella vectensis)*, Tad (*Trichoplax adhaerens*), Tca (*Tribolium castaneum*), Sci (*Sycon ciliatum*), Lsp (*Leucosolenia sp.*), Edi (*Eleutheria dichotoma),* Nv (*Nereis virens*), Pdu (*Platynereis durmerilii*) and Aqu (*Amphimedon queenslandica*).

5.2.3 Synteny analysis of scaffold 34095 of *Sycon ciliatum*

The sequence of scaffold 34095 from *S. ciliatum* was kindly provided by Sofia Fortunato from Dr. Maja Adamska's research group at SARS (Norway) (see Appendix D, D.2 and D.3). The synteny analysis was performed as specified in Chapter 2, section 2.2. Each one of the genes within this scaffold was used as a query to perform rbh (BLASTp) against the Human, *Amphimedon queenslandica, Trichoplax adhaerens, Nematostella vectensis*, *Lottia gigantea* and *Capitella teleta* genomes.

5.3 Results

5.3.1 Orthology analysis of gene 34059 of *Sycon ciliatum* and gene 70333 of *Leucosolenia sp.*

The initial phylogenetic analyses of genes 34059 of *Sycon ciliatum* and 70333 of *Leucosolenia sp.* performed by the Adamska group were unable to distinguish whether this gene was an NK gene, like Hex, or a ParaHox gene, like Cdx (Sofia Fortunato personal communication). I first constructed a Neighbour-Joining phylogenetic tree of the homedomain sequences of the ANTP-class genes of *Tribolium castaneum* and *Branchiostoma floridae*, the genes 34059 of *S. ciliatum* and 70333 of *Leucosolenia sp.*. This revealed some affinity of *Sycon* 34059 and Leucosolenia 70333 with the Cdx/Cad genes of amphioxus and *T. castaneum.* However, it is noteworthy that the support value for this association is very low (40.9%) and the long branch associated with this gene is indicative of its divergent nature, such that caution that must be exercised when deducing its orthology (see Fig. 5.1).

As mentioned above, there is the possibility of this *Sycon* homeobox gene having affinities with an NK gene family, Hex. The lack of robust resolution of this homeodomain sequence led me to next examine a multiple alignment of a selection of ANTP-class protein sequences to check whether there are any motifs outside of the homeodomain that could assist with identifying the orthology of the *Sycon* and *Leucosolenia* genes (34095 and 70333), as well as make comparisons to *Sycon* and *Leucosolenia's* closest available relative *Amphimedon queenslandica.*

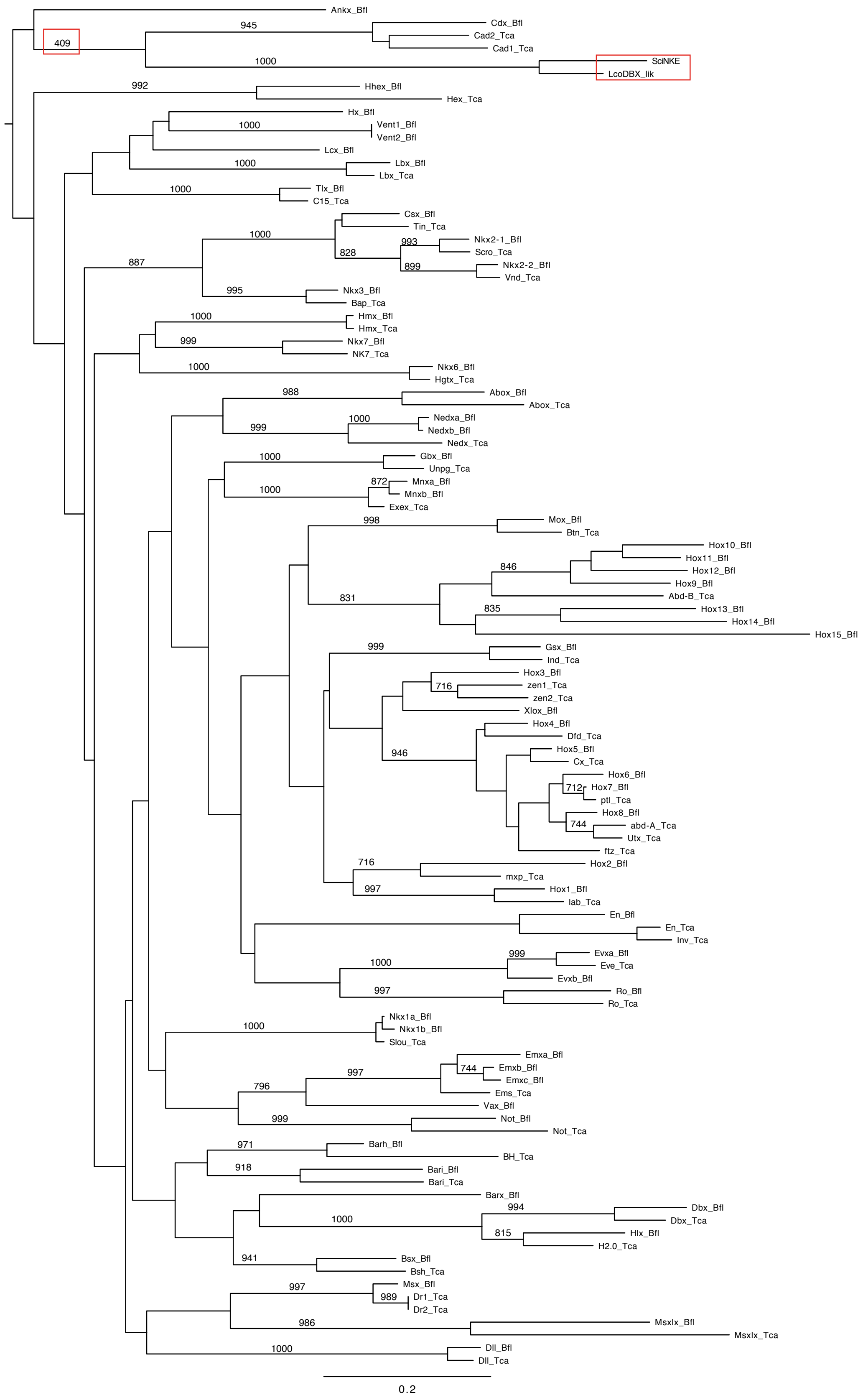
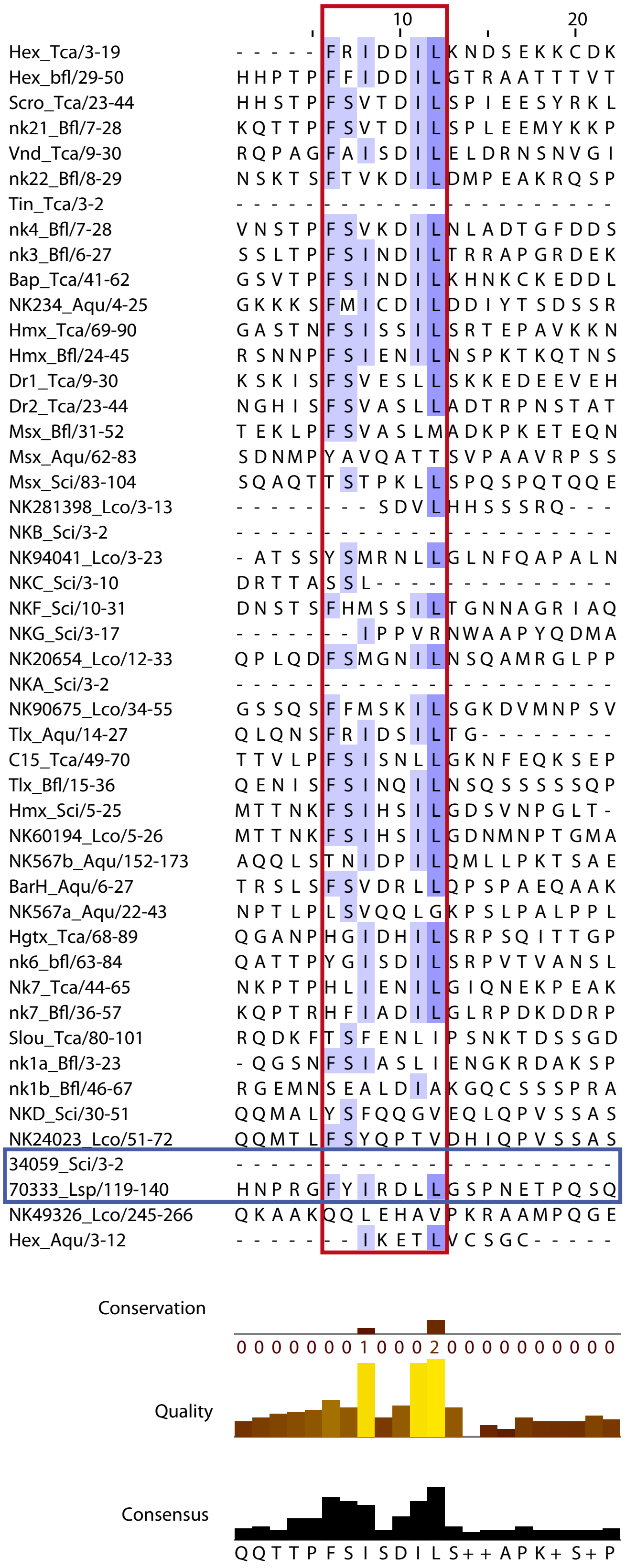


Figure 5.1.- Neighbour-joining tree of ANTP-class genes from B. floridae and T. castaneum and gene 34059 from S. ciliatum and 70333 Leucosolenia (indicated by a red box). *This phylogenetic tree was constructed using the JTT model and 1000 bootstrap replicates. The bootstrap support values equal or above 500 are shown in black and in red the support values for Sycon 34059 and Leucosolenia 70333 protein.*

A. queenslandica only possesses the following NK genes: Hex, Msx, NK5/6/7, NK2/3/4, Tlx and BarH (Larroux et al., 2007). I used these homeobox genes and their corresponding orthologues from *B. floridae* and *T.castaneum* to look for the most similar regions of these sequences and characterise possible motifs outside the homeodomain regions (see below and MA in Appendix D, D.1). No further motifs were found across all NK sequences, although many of the NK cluster proteins do contain the conserved region shown in Fig. 5.2. This region is, however, not universally found in NK proteins and is not therefore a reliable diagnostic, and since it is not in the *Sycon* 34059 sequence anyway, it does not help with the identification. The *Leucosolenia* 70333 present some similarities in this region, but not a clear match with other motifs. Thus, no reliably informative motifs outside of the homeodomain were found in the NK genes. In a similar fashion the motifs outside of the Cdx/Caudal homeodomain did not help with the identification of the *Sycon* and *Leucosolenia* genes either.

Due to the lack of informative motifs outside of the homeodomain, I examined the residues of the homeodomain itself, to see if there were particular residues that could be diagnostic for either NK or Cdx genes and whether any of these are shared with *Sycon* 34059 and *Leucosolenia* 70333. I constructed a multiple alignment of all the NK sequences and a wide range of Cdx/Cad genes from a variety of bilaterians and cnidarians, and included the *Sycon* and *Leucosolenia* genes (MA in Appendix D, D.1). The multiple alignment reveals a combination of amino acids within the second helix of the homeodomain that is restricted to Cdx/Caudal and only one or two other sequences. This motif has the sequence Y-I-T (see Fig. 5.3). The Engrailed (En) and developing brain homeobox (Dbx) families are the other ANTP-class genes that share some similarity with this motif. These observations led to refined phylogenetic analyses, focusing only on the homeodomain of the NK sequences present in the *Amphimedon* NK cluster, the other ANTP-class families that also have the YIT motif and a wide range of Cdx/Cad genes from sponges and a range of bilaterians. The trees were rooted with some members of the PRD class.



```
Hex_Tca/3-19          - - - - - F R I D D I L K N D S E K K C D K
Hex_bfl/29-50         H H P T P F F I D D I L G T R A A T T T V T
Scro_Tca/23-44        H H S T P F S V T D I L S P I E E S Y R K L
nk21_Bfl/7-28         K Q T T P F S V T D I L S P L E E M Y K K P
Vnd_Tca/9-30          R Q P A G F A I S D I L E L D R N S N V G I
nk22_Bfl/8-29         N S K T S F T V K D I L D M P E A K R Q S P
Tin_Tca/3-2           - - - - - - - - - - - - - - - - - - - - - -
nk4_Bfl/7-28          V N S T P F S V K D I L N L A D T G F D D S
nk3_Bfl/6-27          S S L T P F S I N D I L T R R A P G R D E K
Bap_Tca/41-62         G S V T P F S I N D I L K H N K C K E D D L
NK234_Aqu/4-25        G K K K S F M I C D I L D D I Y T S D S S R
Hmx_Tca/69-90         G A S T N F S I S S I L S R T E P A V K K N
Hmx_Bfl/24-45         R S N N P F S I E N I L N S P K T K Q T N S
Dr1_Tca/9-30          K S K I S F S V E S L L S K K E D E E V E H
Dr2_Tca/23-44         N G H I S F S V A S L L A D T R P N S T A T
Msx_Bfl/31-52         T E K L P F S V A S L M A D K P K E T E Q N
Msx_Aqu/62-83         S D N M P Y A V Q A T T S V P A A V R P S S
Msx_Sci/83-104        S Q A Q T T S T P K L L S P Q S P Q T Q Q E
NK281398_Lco/3-13     - - - - - - - - S D V L H H S S S R Q - - -
NKB_Sci/3-2           - - - - - - - - - - - - - - - - - - - - - -
NK94041_Lco/3-23      - A T S S Y S M R N L L G L N F Q A P A L N
NKC_Sci/3-10          D R T T A S S L - - - - - - - - - - - - - -
NKF_Sci/10-31         D N S T S F H M S S I L T G N N A G R I A Q
NKG_Sci/3-17          - - - - - - - I P P V R N W A A P Y Q D M A
NK20654_Lco/12-33     Q P L Q D F S M G N I L N S Q A M R G L P P
NKA_Sci/3-2           - - - - - - - - - - - - - - - - - - - - - -
NK90675_Lco/34-55     G S S Q S F F M S K I L S G K D V M N P S V
Tlx_Aqu/14-27         Q L Q N S F R I D S I L T G - - - - - - - -
C15_Tca/49-70         T T V L P F S I S N L L G K N F E Q K S E P
Tlx_Bfl/15-36         Q E N I S F S I N Q I L N S Q S S S S S Q P
Hmx_Sci/5-25          M T T N K F S I H S I L G D S V N P G L T -
NK60194_Lco/5-26      M T T N K F S I H S I L G D N M N P T G M A
NK567b_Aqu/152-173    A Q Q L S T N I D P I L Q M L L P K T S A E
BarH_Aqu/6-27         T R S L S F S V D R L L Q P S P A E Q A A K
NK567a_Aqu/22-43      N P T L P L S V Q Q L G K P S L P A L P P L
Hgtx_Tca/68-89        Q G A N P H G I D H I L S R P S Q I T T G P
nk6_bfl/63-84         Q A T T P Y G I S D I L S R P V T V A N S L
Nk7_Tca/44-65         N K P T P H L I E N I L G I Q N E K P E A K
nk7_Bfl/36-57         K Q P T R H F I A D I L G L R P D K D D R P
Slou_Tca/80-101       R Q D K F T S F E N L I P S N K T D S S G D
nk1a_Bfl/3-23         - Q G S N F S I A S L I E N G K R D A K S P
nk1b_Bfl/46-67        R G E M N S E A L D I A K G Q C S S S P R A
NKD_Sci/30-51         Q Q M A L Y S F Q Q G V E Q L Q P V S S A S
NK24023_Lco/51-72     Q Q M T L F S Y Q P T V D H I Q P V S S A S
34059_Sci/3-2         - - - - - - - - - - - - - - - - - - - - - -
70333_Lsp/119-140     H N P R G F Y I R D L L G S P N E T P Q S Q
NK49326_Lco/245-266   Q K A A K Q Q L E H A V P K R A A M P Q G E
Hex_Aqu/3-12          - - - - - - - I K E T L V C S G C - - - - -

Conservation          0 0 0 0 0 0 0 1 0 0 0 2 0 0 0 0 0 0 0 0 0 0
Consensus             Q Q T T P F S I S D I L S + + A P K + S + P
```

Figure 5.2.- Section of the multiple alignment of the NK family of bilaterians and sponges. *The red rectangle delineates a potential motif of this family showing that it is not universal. The blue rectangle indicates the Sycon 34059 and Leucosolenia 70333*

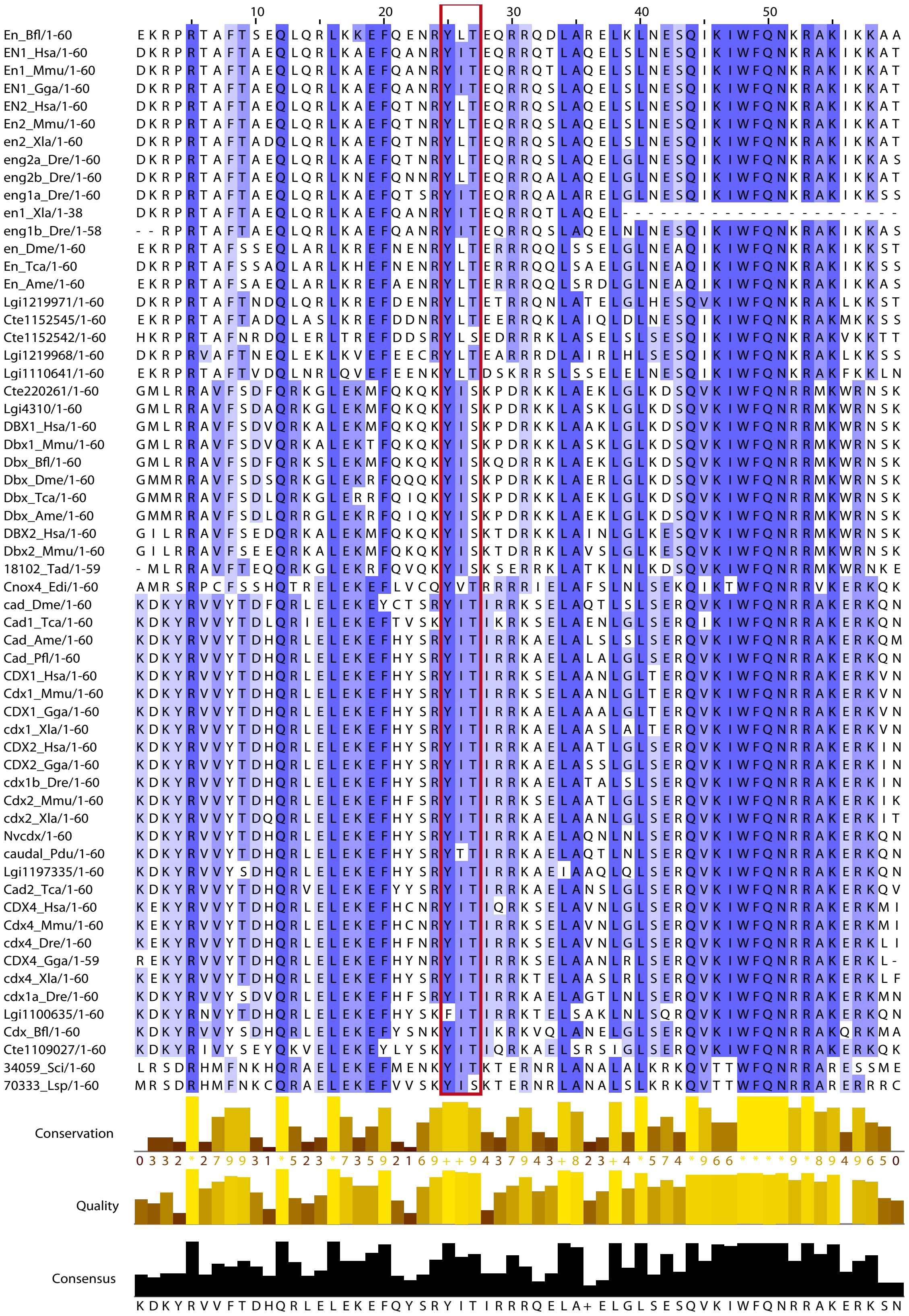

10
20
30
40
50
Conservation
Quality
Consensus
K D K Y R V V F T D H Q R L E L E K E F Q Y S R Y I T I R R R Q E L A + E L G L S E S Q V K I W F Q N R R A K E R K S N

Figure 5.3.- Variability with the Cdx/Cad, En and Dbx genes in bilaterians, cnidarians, placozoan and sponges.

The phylogenetic trees (Figures 5.4, 5.5 and 5.6) all show the same pattern of clustering, grouping Sycon 34059 and *Leucosolenia* 70333 with Cdx from bilaterians and cnidarian. It is noted that the support values vary (85.7%, 64.9% and 53.6%). These support values may well be low due to the long span of evolutionary time that separates this sponge sequence from its putative bilaterian orthologues. Nevertheless, given the consistent grouping of *Sycon* 34059 and *Leucosolenia* 70333 with the Cdx family in a variety of trees incorporating different combinations of ANTP-class families (Figures 5.4, 5.5 and 5.6), I suggest that these *Sycon* and *Leucosolenia* genes are indeed Cdx genes and as such should be re-name as 'SciCdx' and 'LeuCdx'.

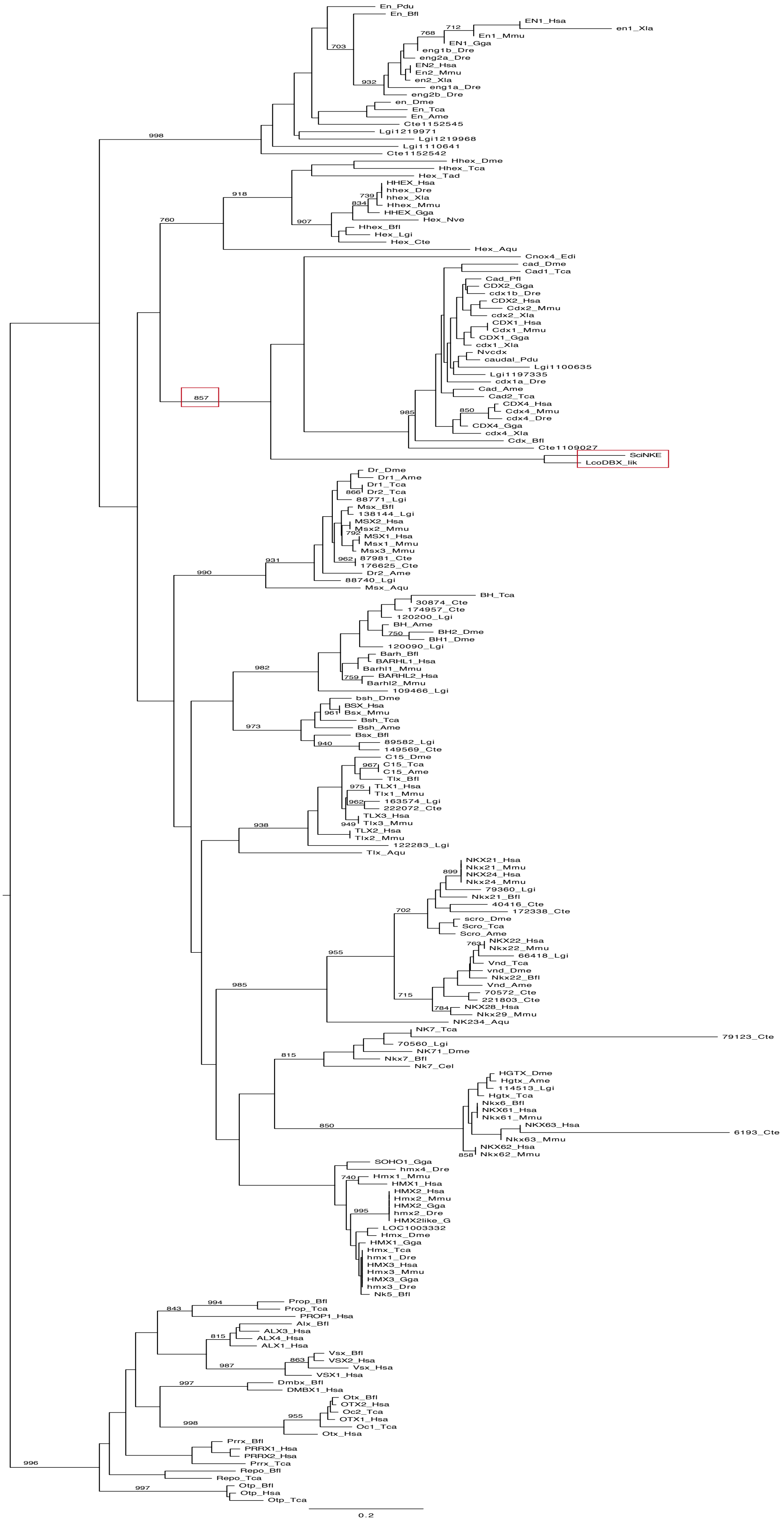


Figure 5.4.- Phylogenetic tree of Sycon 34059 and Leucosolenia 70333 (indicated by red boxes). *Dbx group and A. queenslandica NK5/6/7a/b excluded. NJ (1000) The bootstrap support values equal to or above 500 are shown in black.*

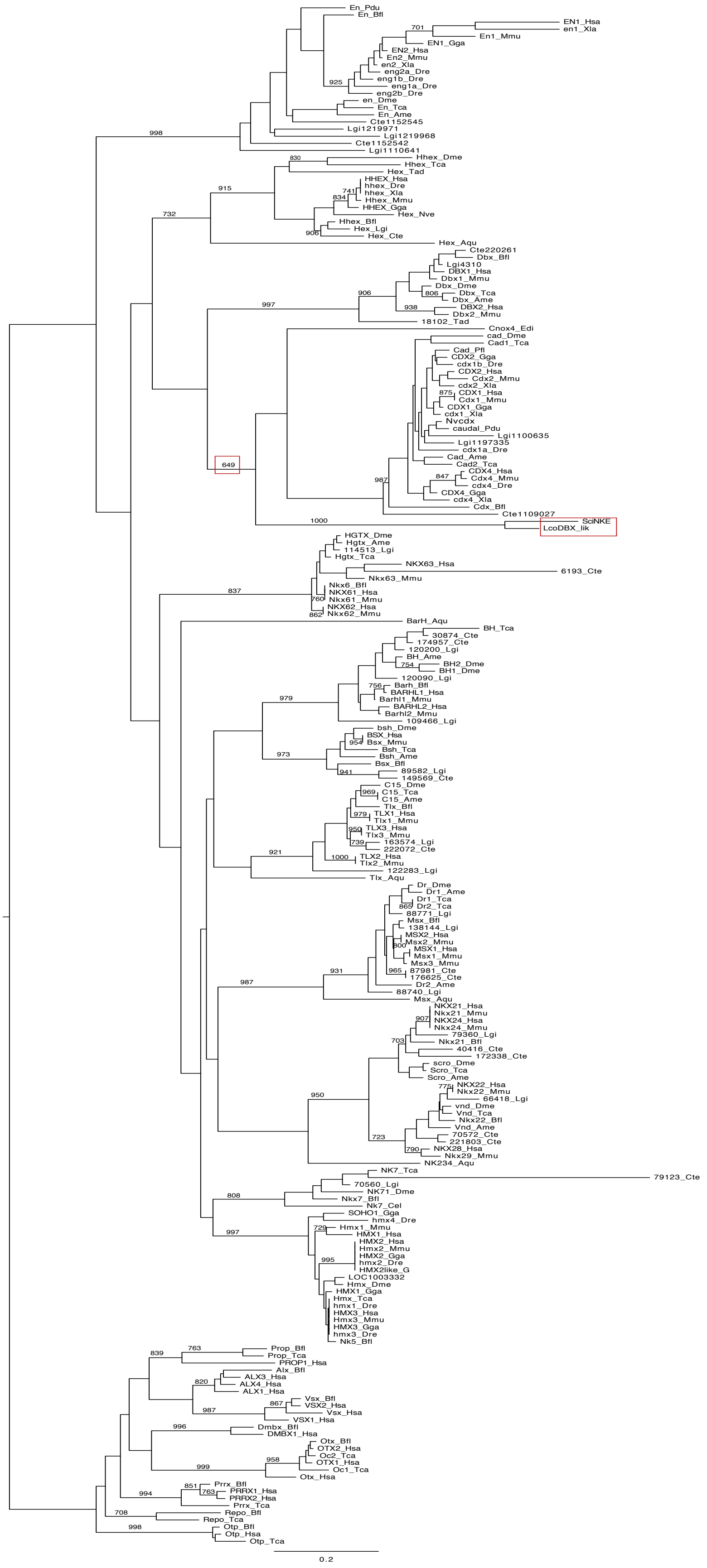


Figure 5.5.- Phylogenetic tree of Sycon 34059 and Leucosolenia 70333 (indicated by red boxes). *A. queenslandica NK5/6/7a/b excluded. NJ (1000) The bootstrap support values equal to or above 500 are shown in black*

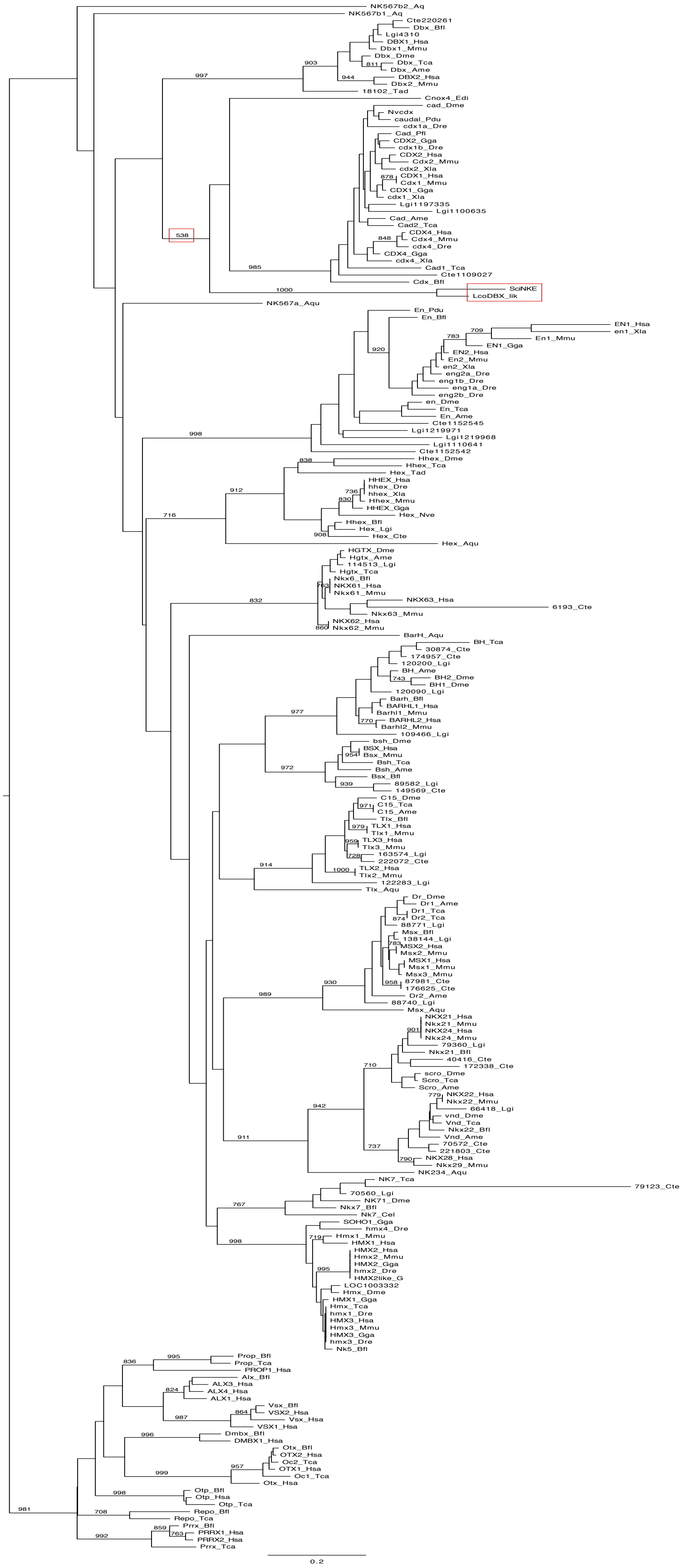


Figure 5.6.- Phylogenetic tree of Sycon 34059 and Leucosolenia 70333 (indicated by red boxes). *Dbx group and A. queenslandica NK5/6/7a/b included. NJ (1000) The bootstrap support values equal to or above 500 are shown in black*

5.3.2 Synteny of *Sycon* scaffold 34059/SciCdx

Alongside the analyses of the sequence of the *Sycon* gene itself, as a means to identify its orthology I also examined the neighbours of the *Sycon* gene to assess whether there is any synteny conserved with bilaterian loci.

This scaffold is 86441 bp long and contains 7 genes, excluding the 34059 *Sycon* gene, which is located towards one end of the scaffold. Orthologue retrieval was performed in the same way as for *Trichoplax* scaffold 38 (Chapter 3; Section 3.2.2). For the retrieval of orthologues and for the purpose of comparing this scaffold with Hox, ParaHox and NK loci I used a variety of animal genomes ranging from bilaterians (human, *Branchiostoma floridae*, *Capitella teleta* and *Lottia gigantea*) to the basal animal lineages (*Nematostella vectensis*, *Trichoplax adhaerens* and *Amphimedon queenslandica*).

The analysis of each protein in this scaffold is summarised in the following table:

Protein number	**Transcript support**	**Corresponding human family (via BLASTp)**	**SMART**	**Corresponding human genes (via phylogeny)**	**Chromosomal locations (in humans)**
42087	236760	zinc and double PHD fingers family (DPF)	ZnFC2H2, PHD and RINGDPF motifs	members 1, 2 and 3	19, 14 and 11
25811	281809	SAR family	SAR1 motif	members A and B	5 and 10
2815	196056, 270307 and 97250	Histone family	H2A motif	Histone 2A	1 and 6
42474	137474	DNAJA family	DNAJ motifs	members 1, 2 and 4	1
22551	200395, 200396, 200397, 200398, 200399, 200400 and 200401	This sequence is relatively short, with very few motifs, and its classification is consequently poorly resolved. Hence, I discard this protein from the analysis.			
24615	307466, 307467, 307468, 307469, 307470, 307471, 307472, 307473, 307474, 307475, 307476, 307477	MACRO domain containing (MACROD) family	Aipp motifs	MacroD1 and D2	11 and 20
13732		This protein does not have a significant match with any human protein, and so it is excluded from the synteny analysis.			

Table 5.1.-Synteny analysis of Sycon 34059/SciCdx scaffold.

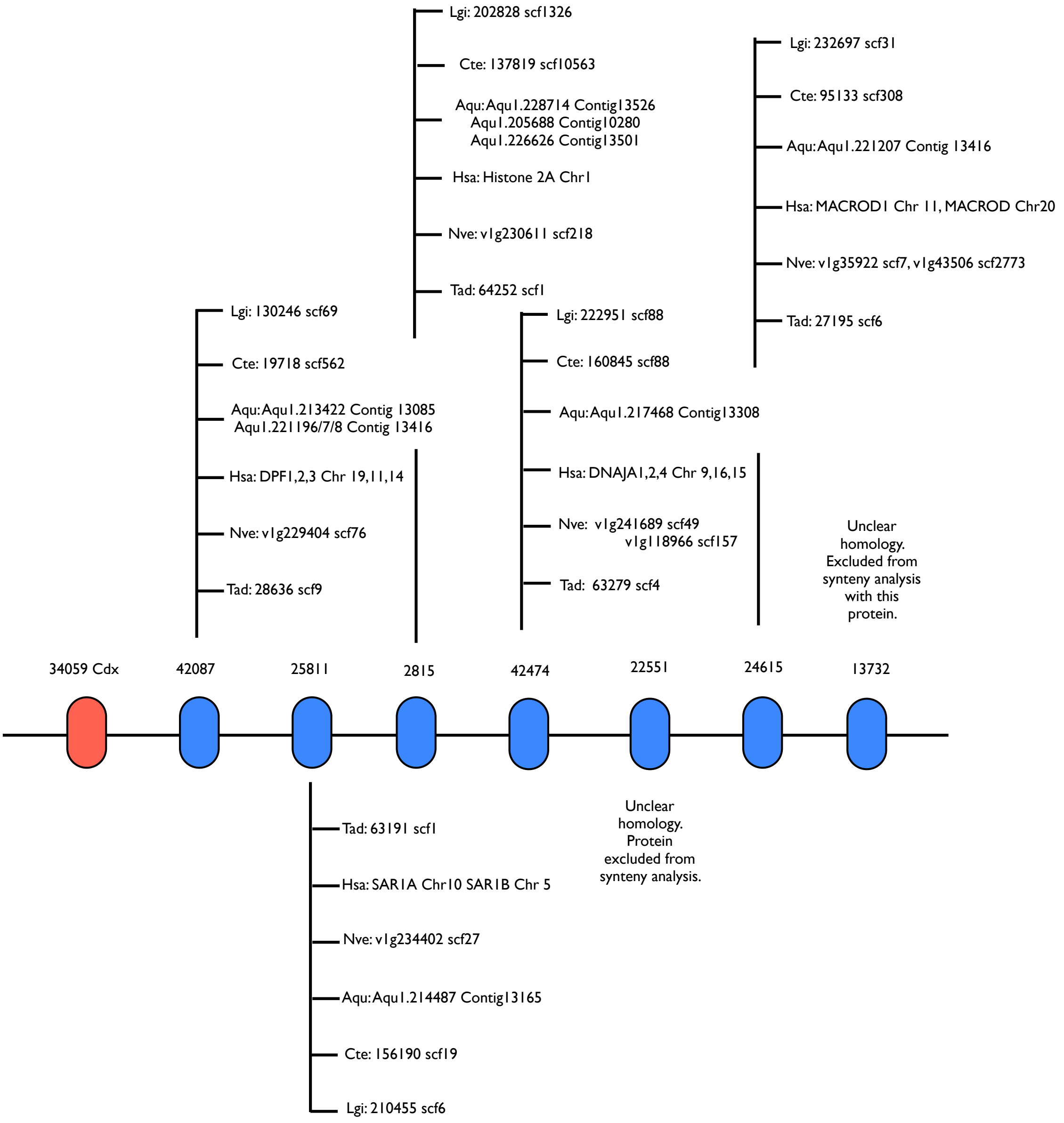


Figure 5.7.- Synteny analysis of Sycon* 34059*/SciCdx scaffold.

The synteny analysis unfortunately did not reveal a robust homologous signal with any of the chromosomes bearing human ParaHox loci (chromosomes 4, 5, 13 and X) or any of the homologous sponge, placozoan or cnidarian ParaHox PAL regions (see Chapters 3 and 4). However, the protein *Sycon* 28511 does provide some support for a ParaHox association. The human orthologues of Sci 28511 reside on chromosomes 10 and 5 and the *Nematostella* orthologue is next to the ParaHox genes NVHD065 and Anthox2. In order to appropriately test these observations statistically I would need an estimate of the complete number in genes as well as scaffold sizes and their gene content of the *Sycon* genome to perform a power analysis, i.e. estimation of the sample size (in this case minimum number of orthologues) needed to perform the test. This is currently not available in a reliable form until the annotation of the *Sycon* genome is complete.

5.4 Discussion

The *Sycon* gene 34059 and *Leucosolenia* gene 70333 appear to be an orthologues of Cdx/Cad, and hence represent the first instances of poriferan ParaHox (or Hox-like) gene. Regardless of the poor support values that unite this gene with the Cdx family that are encountered in some of the molecular phylogenies, this homeodomain still clusters with the Cdx/Caudal family in a wide variety of trees that contain various different ANTP-class members. As an independent route to resolving the orthology of this gene the synteny analysis shows some support in favour of this region being homologous to the ParaHox loci in bilaterians or the ParaHox PAL regions in the basal lineages. This support is however only modest, as it stems from one gene sequence, 25811, which is a bilaterian and *Nemastostella* ParaHox neighbour orthologue. The Adamska group has developed the whole mount *in situ* hybridisation technique for *Sycon* and so interesting future work would involve obtaining expression data for *Sycon* 34059/SciCdx and *Leucosolenia* 70333/LeuCdx, to reveal what role(s) these genes might be playing in these sponges and whether this expression can be related to the function of Cdx genes in other animals.

The presence of a ParaHox gene in this lineage is an independent corroboration of the predictions from the ghost loci hypothesis (Chapters 3 and 4). The ghost loci hypothesis proposes differential gene losses happening in the placozoan and poriferan lineages that affected the Hox and ParaHox genes and their loci, but whilst leaving the broad landscape of these loci intact (i.e. as ghost loci). The discovery of this ParaHox gene within this sponge lineage confirms that the last common ancestor of all animals is likely to have possessed ParaHox (and Hox) genes, and contrary to all previous indications not all sponges have lost all of these Hox/ParaHox genes. The Hox/ParaHox genes thus provide an example of differential losses of developmental control genes across different sponge lineages, which is a phenomenon that seems to be widespread across this phylum (M. Adamska personal communication).

Given the limited synteny signal from the *Sycon* 34059/SciCdx scaffold and the absence of a full genome sequence assembly and gene annotation, it remains to be resolved whether there is another region(s) homologous to bilaterian/cnidarian ParaHox loci which could be linked to this scaffold. Also, an important future avenue of research would be to resolve whether there is a ghost Hox locus in the *Sycon* genome.

Another calcarean sponge, *Leucosolenia sp.* is in the pipeline for assembly and annotation and public release by the Adamska group. This will provide an important further point of reference for resolving the scale of differential gene losses across sponges, particularly with regards to the Hox/ParaHox genes. In addition, the genome of the homoscleromorphan sponge, *Oscarella carmela* has recently be published (Feuda et al., 2012). I performed a preliminary *in silico* homeobox screen in *O. carmela,* and found no indication of Hox or ParaHox genes, but an analysis of synteny and search for ghost Hox/ParaHox loci would be an important avenue of future research in this species as well.

With this data in hand one must consider the alternative possible interpretations that relate to the differing poriferan phylogeny topologies currently being debated (i.e. monophyly versus paraphyly). These are schematized in Fig. 5.8 and Fig. 5.9 and are as follows:

(a) Poriferan monophyly: starting from an Urmetazoan/last common ancestor of animals (LCAA) with distinct Hox, ParaHox and NK loci containing each of these groups of homeobox gene, loss of the Hox gene(s) is most parsimoniously explained by loss from the last common ancestor of Porifera (LCAP) after the divergence from the lineage leading to the Eumetazoa and before the divergence of the various poriferan lineages. In contrast, ParaHox loss occurred at some point between the split into the two main clades of poriferans, (Demospongiae + Hexactinellida) and (Calcarea + Homoscleromorpha) and the origin of the Demospongiae.

(b) Poriferan paraphyly: starting from an Urmetazoan/LCAA with Hox, ParaHox and NK genes and loci at least two independent cases of Hox loss must have occurred; one after the divergence of the (Demospongiae + Hexactinellida) lineage and one in the Calcarea lineage. In contrast, ParaHox loss has occurred either in the (Demospongiae + Hexactinellida) prior to the divergence of these two classes or has occurred in the Demospongiae lineage.

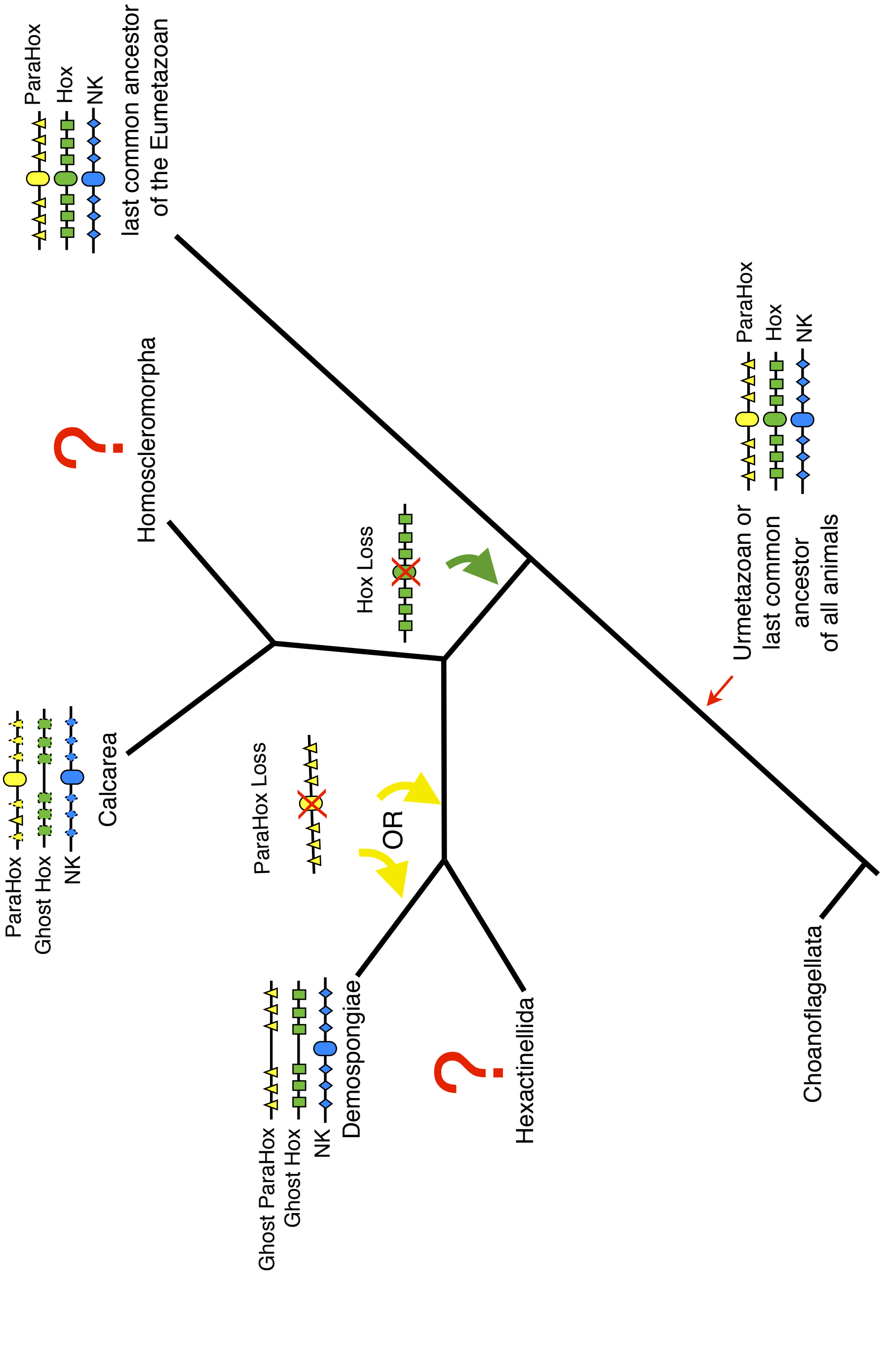


Figure 5.8 .- Alternative scenarios for interpretation of ghost loci and the presence of a ParaHox gene in the calcareid sponges Sycon cilliatum and Leucosolenia sp. in a monophyletic scenario. The red question marks denote unknown ANTP-class genes.

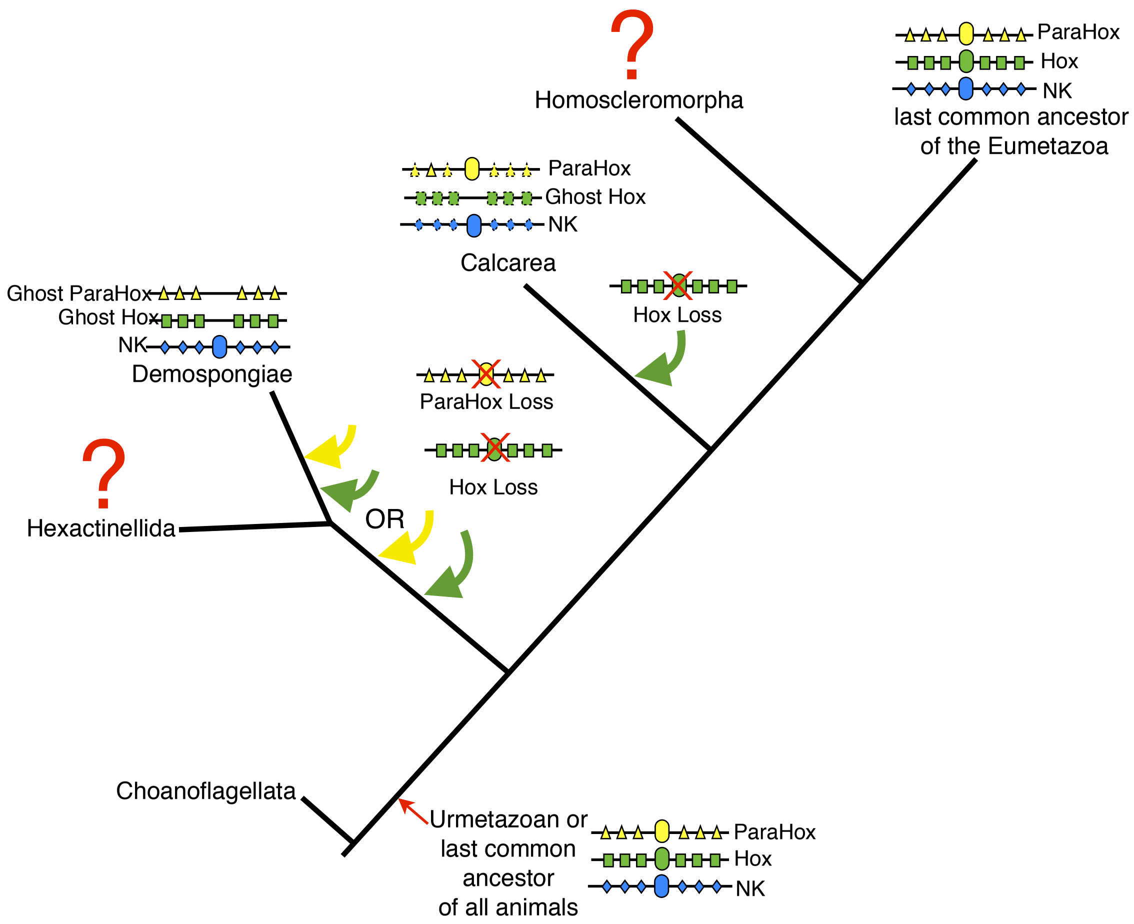


Figure 5.9 .- Alternative scenarios for interpretation of ghost loci and the presence of a ParaHox gene in the calcareid sponges Sycon cilliatum and Leucosolenia sp. in a paraphyletic scenario. The red question marks denote unknown ANTP-class genes.

Chapter 6

The homeobox complement of *Strigamia maritima.*

(Adapted from Strigamia consortium "A myriapod genome: Insights into arthropod evolution" in preparation)

In this chapter I describe how I searched for the homeobox complement of the myriapod *Strigamia maritima*, and curated, classified and annotated it. Also, I describe the clustering and linkage of some members and how this can be used to help reconstruct the evolution of this superfamily.

6.1 Introduction

At the moment, arthropod genome sequences are perhaps one of the most highly represented in the animal kingdom. However, the sequencing efforts in this group have been focused mainly in the holometabolous insects, especially drosophilids, and thus the taxonomic sampling diversity within the whole group of arthropods is limited. Within the drosophilids, the genome of *Drosophila melanogaster* is by far the most studied. As new genome sequences from other arthropods and other invertebrates are released, it is becoming ever more apparent that *Drosophila*'s genome is actually a poor representation of other arthropods and invertebrates. *Drosophila* genomes have lost a significant portion of the bilaterian gene complement and have undergone extensive rearrangements relative to other animal genomes (Stark et al., 2007), such as chordates and sea anemones (Putnam et al., 2007). To further understand when the unique characteristics of higher insects appeared and to depict the diversification of this clade there is a need for wider sampling of other arthropod genomes.

The genome sequence of the centipede, *Strigamia maritima*, respresents one of the four major extant lineages of arthropods, the Myriapoda, which is not represented by any other genome sequence to date. Recently, myriapods have been recognised as the living sister group to the clade that encompasses all insects and crustaceans (Regier et al., 2010). Thus, this genome represents a well-placed phylogenetic anchor to compare and determine ancestral character states for the arthropods and, moreover, help to resolve where particular evolutionary changes in either the insects or crustaceans occurred.

Surveying for the homeobox complement of the *S. maritima* genome will not only provide a descriptive catalogue of the homeoboxes in this genome, but will also provide insights into the evolutionary dynamics and ancestral states of this superfamily. In addition, as mentioned previously, the homeobox superfamily can act as a proxy with which to understand the biological nature of genome rearrangements via the clustering of some of its members. The challenges in this survey are identifying the potential gene losses and correct annotation of the putative orthologues. Great care must be taken when concluding that gene losses have occurred, to avoid making incorrect inferences about loss due to sequencing artefacts (e.g.sequencing errors, misassembly and/ or not enough coverage). One possible way to overcome such sequencing artefacts is to complement the genome assembly information with transcriptome data and with searches of the unassembled reads, both of which are available for this genome project.

In this chapter I will describe a method that I have developed to retrieve a list of putative homeobox gene candidates from an assembled whole genome sequence, and how I phylogenetically classify these candidates. I describe the instances of clustering and linkage of some of the members of this superfamily in *S. maritima.* Finally, I put these results into context with some of the hypotheses regarding the origin and evolution of the homeoboxes.

6.2 Materials and Methods

6.2.1 Survey and construction of a saturated list of putative candidate homeoboxes genes

6.2.1.1 Large-scale survey for candidate homeobox genes in a newly sequenced arthropod genome

In order to retrieve all homeobox genes from a genome sequence such as *S. maritima,* a Python script that parses a tBLASTn output was designed (see Appendix D). The query batches used for surveying were the homeodomain sequences of *T. castaneum* and *B. floridae.* The beetle and amphioxus homeodomains were obtained from HomeoDB (http://homeodb.cbi.pku.edu.cn/ (Zhong et al., 2008)), and the beetle and amphioxus searches were performed independently. The Python script retrieves the scaffold in which a candidate is located. The same process is performed on the transcriptome sequence data as well as the unassembled sequence reads of *S. maritima,* in order to perform as thorough a search as possible and distinguish those candidates that are supported by expression data. See Appendix E, E.1 and E.2.

6.2.1.2 Classification of the candidates

From this initial search a list of candidate genes located in particular scaffolds was obtained, which was then manually curated using the program

Apollo (v1.11.8, (Lewis et al., 2002)) to check for appropriate exon-intron boundaries and potential UTRs (i.e. untranslated regions). Once curated, the classification was performed using multiple alignments of the candidate homeobox genes with their potential orthologues in order to check for similarities within the homeodomain and other domains outside the homeodomain. From these alignments a neighbour-joining phylogenetic tree was built (1000 bootstrap replicates), using the whole set of homeodomains of the *Strigamia* candidates with the whole sets of homeodomains from *T. castaneum* and *B. floridae*. The membership of each class and family was then checked from this tree, and the *Strigamia maritima* condition for each category noted (Appendix E, E.3). A table of *Strigamia* homeoboxes with their orthologues of *Tribolium castaneum* and *Branchiotosma* f*loridae* is provided in Appendix E, E. 4.

Independent phylogenetic trees of the classes of ANTP, PRD, TALE, HNF and Xlox/Hox3 were reconstructed. Modelgenerator was used with each alignment to retrieve the appropriate model of sequence evolution to use for the inference of maximum-likelihood and bayesian phylogenetic trees. For each class tree a neighbour-joining (1000 bootstrap replicates), maximum-likelihood (100 bootstrap replicates) and bayesian trees (1000000 generations; 5000 for sample probability; burn-in of 50 samples; two runs of four chains each) were constructed.

The homeodomain genes other than those from *Tribolium* and *Branchiostoma* were retrieved from HomeoDB, NCBI and JGI. The species acronyms used for the phylogenetic trees were Ame (*Apis mellifera*), Bfl (*Branchiostoma floridae*), Cte (*Capitella teleta*), Dme (*Drosophila melanogaster*), Hsa (*Homo sapiens*), Lgi (*Lottia gigantea*), Nve (*Nematostella vectensis*), Sma (*Strigamia maritima*) and Tca (*Tribolium castaneum*). All the sequences, alignment and Newick format trees are available in Appendix E, E.5.

6.2.2 Synteny analysis of the scaffold 48457 and statistical test

Orthologue retrieval was performed as described in Chapter 2, Section 2.1. Each one of the genes within this scaffold was used as a query to perform rbh (BLASTp) against the Human genome. The statistical analyses were performed as specified in Chapter 2 Section 2.2 with the following modifications. Once identified *S. maritima*-human orthologues were classified into Hox loci neighbour orthologues, ParaHox loci neighbour orthologues and Non-Hox/ ParaHox loci neighbour orthologues. Expected probabilities of categories of the orthologues were inferred as described in Chapter 2 Section 2.2. From these probabilities were calculated contingency tables (see probabilities for version 64 human genome version in Appendix E, E.6). These probabilities were used to perform an Exact Binomial Test and a Fisher Exact Test in R (see codes as in Appendix B).

6.2.3 Clustering and linkage inference

The clustering and linkage distances of the homeobox genes were inferred based on exon boundaries.

6.3 Results

6.3.1 The homeobox complement of *Strigamia maritima*

I used the complete homeobox catalogues of an insect and chordate (*Tribolium castaneum* and *Branchiostoma floridae* respectively) as queries for a saturated search (i.e. it will not retrieve more homeobox candidates) of the whole genome assembly, as well as the unassembled reads and the transcriptome data of the *Strigamia maritima* genome sequencing project. I found 112 homeobox-containing genes, based upon phylogenetic analysis of the homeodomain (see Appendix E, E.4 for the complete list of *Strigamia* homeoboxes). This compares to 133 homeobox genes in the chordate amphioxus and 104, 103, and 93 in the insects *Drosophila melanogaster, Tribolium castaneum* and *Apis mellifera* respectively.

Of these 112 *Strigamia* homeobox genes, seven are very divergent and it was initially difficult to determine their orthology precisely. However, with a combination of molecular phylogenetics with Neighbour-Joining, Maximum-likelihood and Bayesian approaches, and using additional information from domains or sequence conservation outside of the homeodomain, I was able to place three of the seven genes in the ANTP class (two) and PRD class (one).

Besides the remaining four unclassified sequences, I found 54 ANTP-class genes, 25 PRD-class genes (see Appendix E, E.4) and 29 distributed amongst the nine remaining classes that are usually recognized (see Table 6.1). I found two genes with more than one homeobox, one in the Zinc Finger (ZF) class (containing four homeoboxes) and one in the Cut class (containing two homeoboxes).

Homeobox class	*Strigamia maritima*	*Branchiostoma floridae*	*Tribolium castaneum*
ANTP	54	60	45
PRD	25	28	25
TALE	8	9	8
SINE	3	3	3
LIM	6	7	7
POU	4	8	6
HNF	1	4	0
CUT	3	4	3
PROS	1	1	1
ZF	2	5	2
CERS	2	1	1
others	4	3	2

Table 6.1.- Summary of numbers of homeobox genes in each class in Strigamia maritima, Branchiostoma floridae and Tribolium castaneum.

The number of *Strigamia* homeobox genes is slightly larger than the numbers found in most other arthropods analysed to date. This, at least in part, may be due to several instances of lineage-specific duplications alongside a distinct lack of homeobox gene loss in *Strigamia*.

6.3.1.1 ANTP Class

I found multiple copies (usually two to three) of Eve, Not, Vnd, BarH, Btn, Cad, and Ind. I also found a duplication of a potential Hox3 gene (see discussion below). A further distinctive feature of the *Strigamia* ANTP homeobox complement is the presence of Vax, which has not previously been found in an arthropod genome. Thus, this gene can no longer be thought of as lost from the Arthropoda as a whole.

6.3.1.2 PRD Class

I found 2 copies of Unc4 and Otd. A further distinctive feature of the *Strigamia* PRD homeobox complement is the presence of Dmbx, which has not previously been found in an arthropod genome and so, as for Vax, this gene can no longer be thought of as lost from the Arthropoda as a whole.

6.3.1.3 Other classes

I found multiple copies of the Irq gene, which is a member of the TALE-class, which provides an interesting case of independent duplication within a homeobox cluster (see discussion below). Also, I found a *Strigamia* Hmbox gene, which is a member of the HNF-class. This is interesting on two counts. Firstly, the HNF class as a whole is missing from other arthropod genomes like those of the insects, and so this represents the first example of an arthropod HNF class gene described to date. Secondly, Hmbox genes have previously been proposed as chordate-specific, in contrast to more ancient members of the HNF class like HNF1/Tcf (a gene present in diploblasts as well as several bilaterians) (Ryan et al., 2006). Thus, this *Strigamia* Hmbox gene (which posseses a POU-like domain, the typical insertion for HNF-class genes of 15-20 amino acids between the second and third helix in the homeodomain, and bootstrap support of 92.6% for a grouping with chordate Hmbox genes in a HNF-class tree (see Appendix E, E.5 for multiple alignments and Fig. 6.1) implies that Hmbox genes are not chordate-specific but have been widely lost in multiple lineages of the animal kingdom. Also, the ancient HNF1/Tcf family has instead been lost from *Strigamia*.

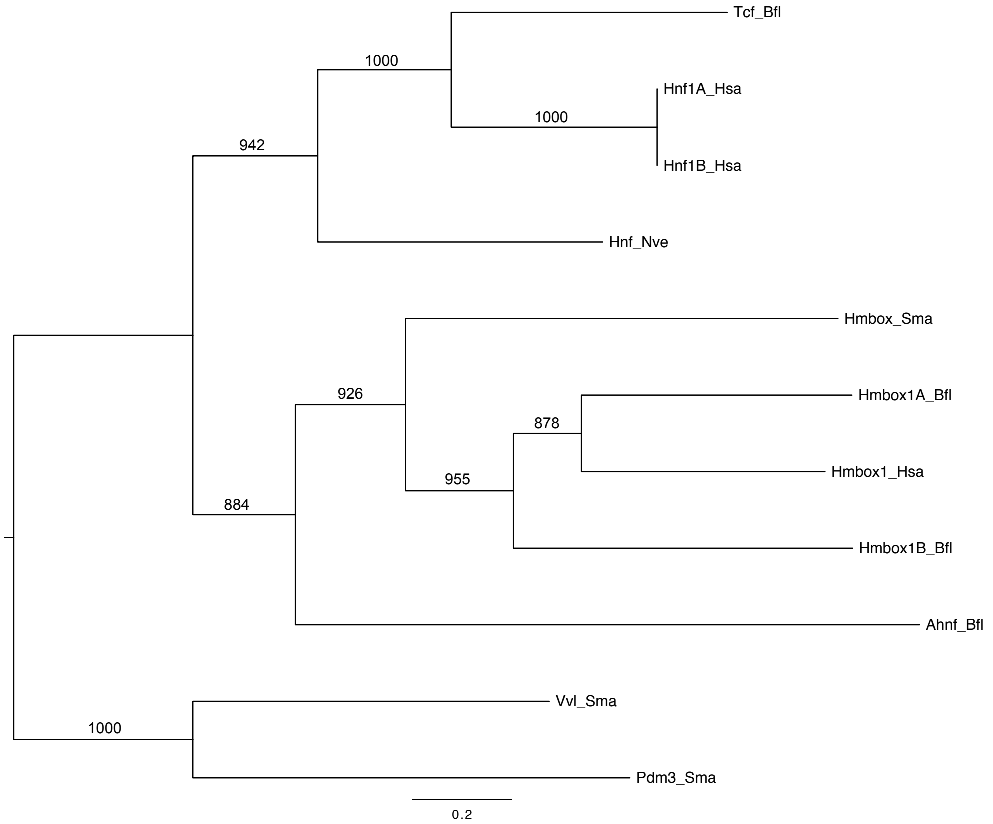


Figure 6.1.- Phylogenetic analysis of the HNF-class gene of Strigamia using different HNF genes from chordates and a cnidarian. *This phylogenetic analysis was constructed using neighbour-joining with the JTT distance matrix and 1000 bootstrap replicates. A multiple alignment of the entire coding sequences was used as a basis for the phylogenetic analysis. Two POU class genes (Vvl and Pdm3) were used as an outgroup to root the tree.*

6.3.2 Clustering of homeobox genes

The clustering and linkage of homeobox genes is often of functional significance (e.g. the Hox genes) or provides an important insight into the origins of this gene family as well as a useful proxy for the degree of genome rearrangement relative to other species. There is an intact Hox cluster in *S. maritima*. Closely linked to the posterior side of the Hox cluster is clustered Evxb (see Fig. 6.2). This clustering is also found in cnidarians and chordates (Gauchat et al., 2000, Minguillòn and Garcia-Fernàndez, 2003). The Hox cluster has been annotated by the research group of Michael Akam (University of Cambridge) and so it is not described in detail here. However, I note the absence of Hox3 from the cluster, the close linkage of only one of the Evx genes, and several potential non-homeobox gene models within the cluster.

In contrast to the intact Hox cluster, its evolutionary sister the ParaHox gene cluster is not intact, which reflects the situation found in other ecdyzosoans (Ferrier and Minguillon, 2003). In addition to the break-up of the ParaHox cluster, the ParaHox genes of *Strigamia* have undergone duplications, producing two copies of Ind and a third Ind-like gene and three of Cad, which is likely to have implications for their roles in early development of the ectoderm, nervous system and gut. No ecdysozoan Xlox, which is the third ParaHox gene, has been described to date. The counterpart to the Xlox ParaHox gene from the Hox cluster (following the ProtoHox to Hox/ParaHox model of Brooke et al. (Brooke et al., 1998)) is Hox3. In *Strigamia* Hox3 is absent from the Hox cluster, but elsewhere within the genome there are two genes with sequence affinities to Hox3/Xlox. It is thus interesting to try to determine whether these two *Strigamia* Hox3/Xlox genes are either Hox genes that have somehow translocated out of the Hox cluster (and Xlox is absent from *Strigamia* as with other ecdysozoans), or instead these genes are the first examples of edysozoan Xlox genes (and Hox3 has been deleted from the *Strigamia* Hox cluster and genome).

A Neighbour-Joining phylogenetic tree of the entire coding sequences of these *Strigamia* Hox3/Xlox genes along with a selection of Hox1, Hox2, Hox3, Hox4 and Xlox genes reveals some affinity of the *Strigamia* genes with the Xlox genes of amphioxus, *Lottia* and *Capitella*. However, it is noteworthy that the bootstrap support value for this association is very low (only 33%) and so the grouping of the *Strigamia* genes with the Xlox genes of other species cannot be considered as robust (see Fig. 6.3).

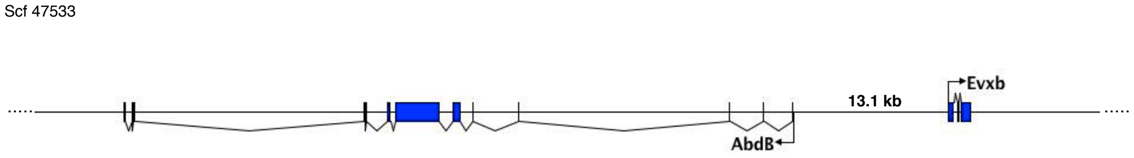


Figure 6.2.- Cluster of the posterior side of the Hox cluster (AbdB) and Evxb in S. maritima. *The rectangles linked by lines represent genes and the lines scaffolds. The colouring of rectangles represents the class that these genes belong to, in this case ANTP-class. The small arrows represent the transcriptional orientation.*

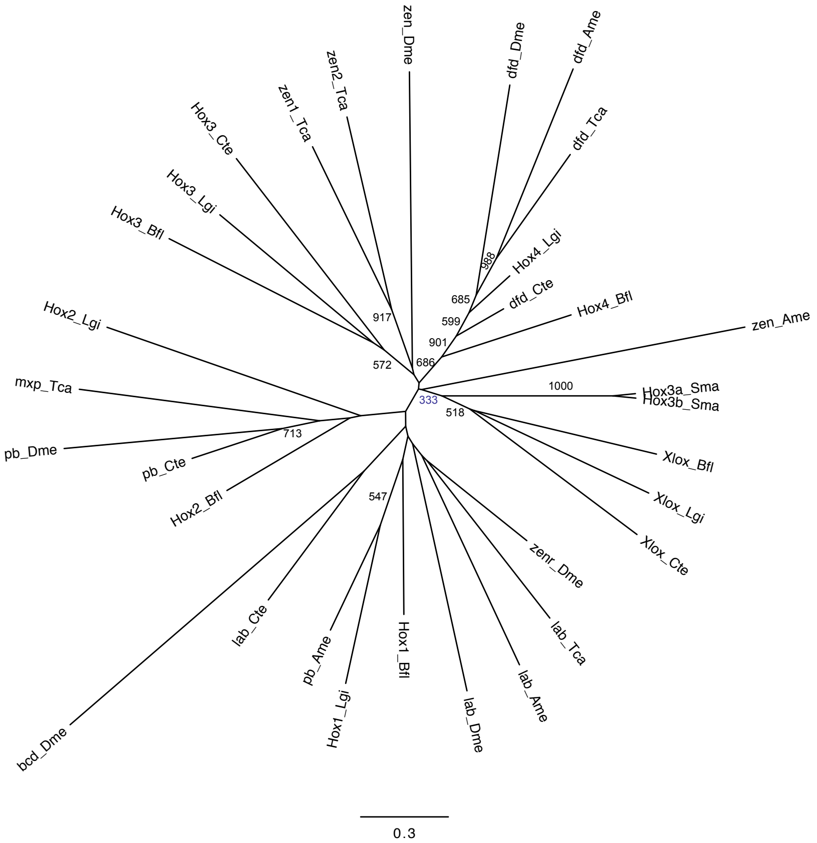


Figure 6.3.- Phylogenetic analysis of Xlox/Hox3 genes of Strigamia using a selection of Hox1, Hox2, Hox3, Hox4 and Xlox sequences. *This analysis was based upon the whole coding sequence of the genes. This phylogenetic analysis was constructed using neighbour-joining with a JTT distance matrix and 1000 bootstrap replicates . The blue support value (of 333) is the node that reveals the affinity between Xlox/Hox3 from Strigamia and Xlox sequences.*

Further phylogenetic analysis, focusing on the most similar regions of the Xlox and Hox sequences, including the hexapeptide and homeodomain regions (see Fig. 6.4) and rooting the trees with some members of the PRD class, now reveals a possible affinity with Hox3 genes rather than Xlox (see Fig. 6.5). But again there are no significant support values for this Hox3 grouping (the 42.9% support value is not shown in the tree as the threshold is 50%).

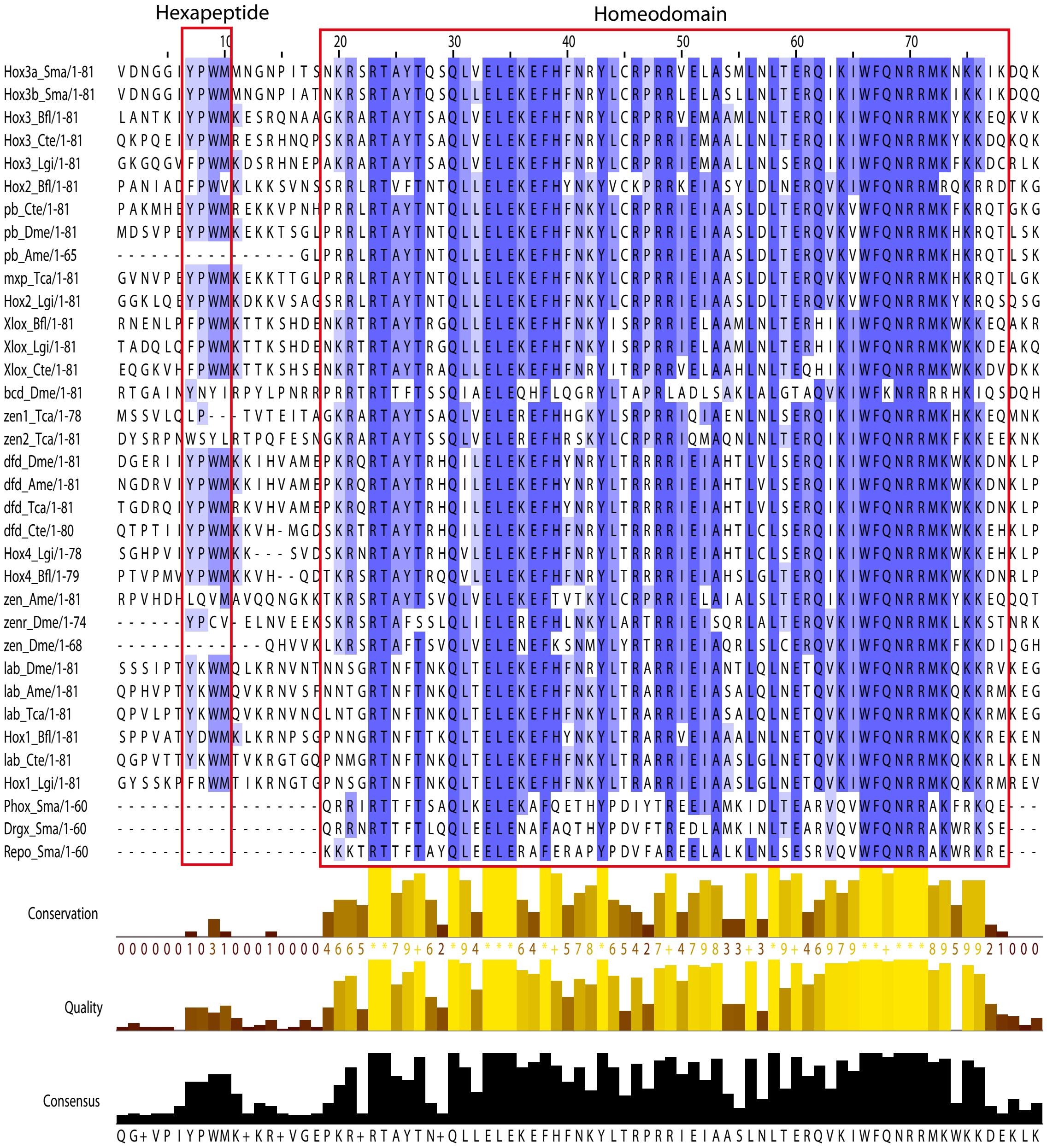

Hexapeptide
Homeodomain
10
20
30
40
50
60
70
Hox3a_Sma/1-81 VDNGGIYPWMMNGNPITSNKRSRTAYTQSQLVELEKEFHFNRYLCRPRRVELASMLNLTERQIKIWFQNRRMKNKKIKDQK
Hox3b_Sma/1-81 VDNGGIYPWMMNGNPIATNKRSRTAYTQSQLLELEKEFHFNRYLCRPRRLELASLLNLTERQIKIWFQNRRMKIKKIKDQQ
Hox3_Bfl/1-81 LANTKIYPWMKESRQNAAGKRARTAYTSAQLVELEKEFHFNRYLCRPRRVEMAAMLNLTERQIKIWFQNRRMKYKKEQKVK
Hox3_Cte/1-81 QKPQEIYPWMRESRHNQPSKRARTAYTSAQLVELEKEFHFNRYLCRPRRIEMAALLNLTERQIKIWFQNRRMKYKKDQKQK
Hox3_Lgi/1-81 GKGQGVFPWMKDSRHNEPAKRARTAYTSAQLVELEKEFHFNRYLCRPRRIEMAALLNLSERQIKIWFQNRRMKFKKDCRLK
Hox2_Bfl/1-81 PANIADFPWVKLKKSVNSSRRLRTVFTNTQLLELEKEFHYNKYVCKPRRKEIASYLDLNERQVKIWFQNRRMRQKRRDTKG
pb_Cte/1-81 PAKMHEYPWMREKKVPNHPRRLRTAYTNTQLLELEKEFHFNKYLCRPRRIEIAASLDLTERQVKVWFQNRRMKFKRQTGKG
pb_Dme/1-81 MDSVPEYPWMKEKKTSGLPRRLRTAYTNTQLLELEKEFHFNKYLCRPRRIEIAASLDLTERQVKVWFQNRRMKHKRQTLSK
pb_Ame/1-65 ----------------GLPRRLRTAYTNTQLLELEKEFHFNKYLCRPRRIEIAASLDLTERQVKVWFQNRRMKHKRQTLSK
mxp_Tca/1-81 GVNVPEYPWMKEKKTTGLPRRLRTAYTNTQLLELEKEFHFNKYLCRPRRIEIAASLDLTERQVKVWFQNRRMKHKRQTLGK
Hox2_Lgi/1-81 GGKLQEYPWMKDKKVSAGSRRLRTAYTNTQLLELEKEFHFNKYLCRPRRIEIAASLDLTERQVKVWFQNRRMKYKRQSQSG
Xlox_Bfl/1-81 RNENLPFPWMKTTKSHDENKRTRTAYTRGQLLELEKEFHFNKYISRPRRIELAAMLNLTERHIKIWFQNRRMKWKKEQAKR
Xlox_Lgi/1-81 TADQLQFPWMKTTKSHDENKRTRTAYTRGQLLELEKEFHFNKYISRPRRIELAAMLNLTERHIKIWFQNRRMKWKKDEAKQ
Xlox_Cte/1-81 EQGKVHFPWMKTTKSHSENKRTRTAYTRAQLLELEKEFHFNRYITRPRRVELAAHLNLTEQHIKIWFQNRRMKWKKDVDKK
bcd_Dme/1-81 RTGAINYNYIRPYLPNRRPRRTRTTFTSSQIAELEQHFLQGRYLTAPRLADLSAKLALGTAQVKIWFKNRRRRHKIQSDQH
zen1_Tca/1-78 MSSVLQLP---TVTEITAGKRARTAYTSAQLVELEREFHHGKYLSRPRRIQIAENLNLSERQIKIWFQNRRMKHKKEQMNK
zen2_Tca/1-81 DYSRPNWSYLRTPQFESNGKRARTAYTSSQLVELEREFHRSKYLCRPRRIQMAQNLNLTERQIKIWFQNRRMKFKKEEKNK
dfd_Dme/1-81 DGERIIYPWMKKIHVAMEPKRQRTAYTRHQILELEKEFHYNRYLTRRRRIEIAHTLVLSERQIKIWFQNRRMKWKKDNKLP
dfd_Ame/1-81 NGDRVIYPWMKKIHVAMEPKRQRTAYTRHQILELEKEFHYNRYLTRRRRIEIAHTLVLSERQIKIWFQNRRMKWKKDNKLP
dfd_Tca/1-81 TGDRQIYPWMRKVHVAMEPKRQRTAYTRHQILELEKEFHYNRYLTRRRRIEIAHTLVLSERQIKIWFQNRRMKWKKDNKLP
dfd_Cte/1-80 QTPTIIYPWMKKVH-MGDSKRTRTAYTRHQILELEKEFHFNRYLTRRRRIEIAHTLCLSERQIKIWFQNRRMKWKKEHKLP
Hox4_Lgi/1-78 SGHPVIYPWMKK---SVDSKRNRTAYTRHQVLELEKEFHFNRYLTRRRRIEIAHTLCLSERQIKIWFQNRRMKWKKEHKLP
Hox4_Bfl/1-79 PTVPMVYPWMKKVH--QDTKRSRTAYTRQQVLELEKEFHFNRYLTRRRRIEIAHSLGLTERQIKIWFQNRRMKWKKDNRLP
zen_Ame/1-81 RPVHDHLQVMAVQQNGKKTKRSRTAYTSVQLVELEKEFTVTKYLCRPRRIELAIALSLTERQIKIWFQNRRMKYKKEQQQT
zenr_Dme/1-74 ------YPCV-ELNVEEKSKRSRTAFSSLQLIELEREFHLNKYLARTRRIEISQRLALTERQVKIWFQNRRMKLKKSTNRK
zen_Dme/1-68 -------------QHVVKLKRSRTAFTSVQLVELENEFKSNMYLYRTRRIEIAQRLSLCERQVKIWFQNRRMKFKKDIQGH
lab_Dme/1-81 SSSIPTYKWMQLKRNVNTNNSGRTNFTNKQLTELEKEFHFNRYLTRARRIEIANTLQLNETQVKIWFQNRRMKQKKRVKEG
lab_Ame/1-81 QPHVPTYKWMQVKRNVSFNNTGRTNFTNKQLTELEKEFHFNKYLTRARRIEIASALQLNETQVKIWFQNRRMKQKKRMKEG
lab_Tca/1-81 QPVLPTYKWMQVKRNVNCLNTGRTNFTNKQLTELEKEFHFNKYLTRARRIEIASALQLNETQVKIWFQNRRMKQKKRMKEG
Hox1_Bfl/1-81 SPPVATYDWMKLKRNPSGPNNGRTNFTTKQLTELEKEFHYNKYLTRARRVEIAAALNLNETQVKIWFQNRRMKQKKREKEN
lab_Cte/1-81 QGPVTTYKWMTVKRGTGQPNMGRTNFTNKQLTELEKEFHFNKYLTRARRIEIAASLGLNETQVKIWFQNRRMKQKKRLKEN
Hox1_Lgi/1-81 GYSSKPFRWMTIKRNGTGPNSGRTNFTNKQLTELEKEFHFNKYLTRARRIEIAASLGLNETQVKIWFQNRRMKQKKRMREV
Phox_Sma/1-60 ------------------QRRIRTTFTSAQLKELEKAFQETHYPDIYTREEIAMKIDLTEARVQVWFQNRRAKFRKQE---
Drgx_Sma/1-60 ------------------QRRNRTTFTLQQLEELENAFAQTHYPDVFTREDLAMKINLTEARVQVWFQNRRAKWRKSE---
Repo_Sma/1-60 ------------------KKKTRTTFTAYQLEELERAFERAPYPDVFAREELALKLNLSESRVQVWFQNRRAKWRKRE---
Conservation
0000001031000100004665**79+62*94***64*+578*65427+479833+3*9+46979**+***8959921000
Quality
Consensus
QG+VPIYPWMK+KR+VGEPKR+RTAYTN+QLLELEKEFHFNKYLTRPRRIEIAASLNLTERQVKIWFQNRRMKWKKDEKLK

Figure 6.4.- Multiple alignment of relevant residues of the Hox1, Hox2, Hox3, Hox4 and Xlox sequences of different lineages. *Three Paired class genes are included as an outgroup. The grading of purple colouring of the amino acids shows the identity level of these sequences. The red rectangles in the multiple alignment delimit the core of the hexapeptide motif and the homeodomain.*

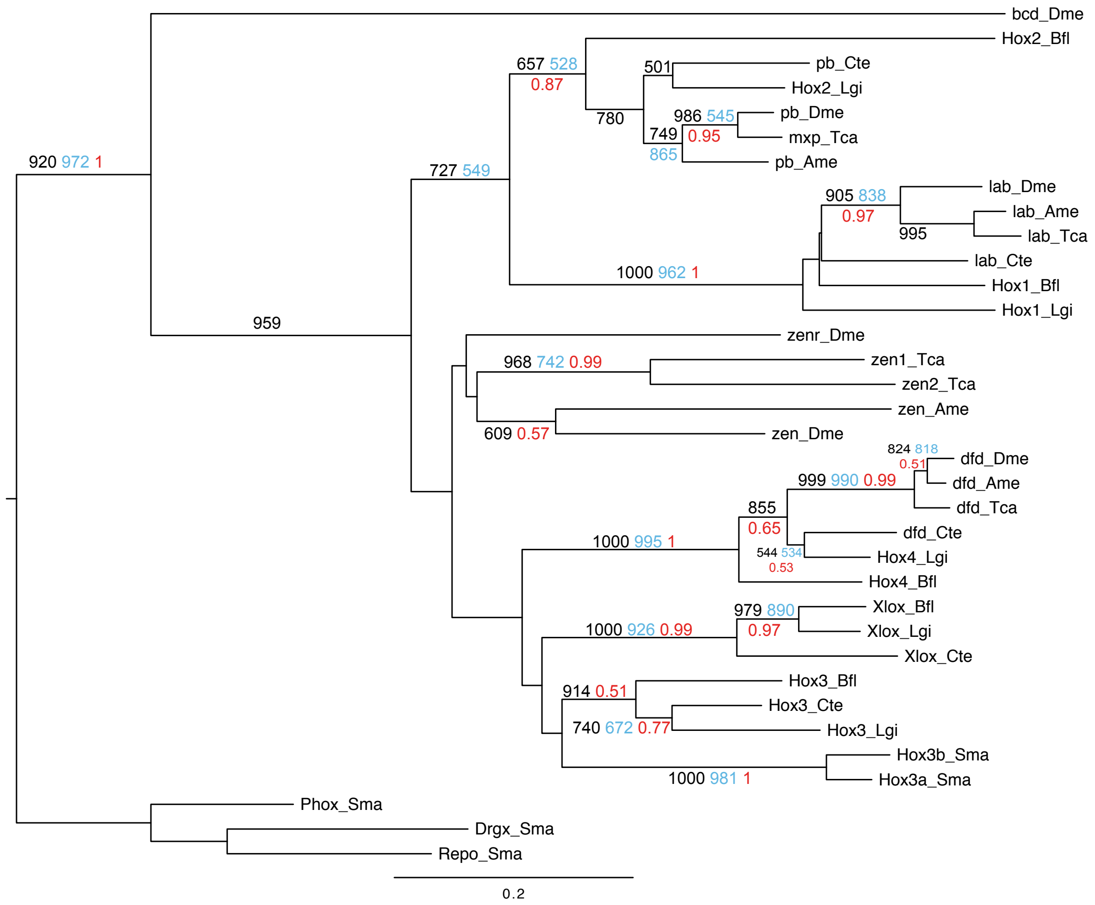


Figure 6.5.- Phylogenetic analysis of Strigamia Xlox/Hox3 homeodomain and hexapeptide motifs using a selection of Hox1, Hox2, Hox3, Hox4 and Xlox sequences. *This analysis used a section of the coding sequence including the hexapeptide and some flanking residues plus the homeodomain (alignment in Fig. 6.3). Three Paired class genes are included as an outgroup. This phylogeny was constructed using Neighbor-Joining with the JTT distance matrix and 1000 bootstrap replicates. Maximum Likelihood support values are shown in blue and Bayesian posterior probabilities in red.*

An alternative approach to phylogenetic trees that can sometimes help with resolving gene orthology is comparison of synteny (Hui et al., 2008). One of the *Strigamia* Hox3/Xlox genes (*Hox3b_ Sma*) is on a small scaffold with no neighbours and so comparative synteny cannot be analysed. However, the second gene (*Hox3a_ Sma*) is on a scaffold with 94 other genes (scaffold 48457). I found that by reciprocal best BLAST searches against the human genome (v68 from ENSEMBL) I retrieved 24 one-to-one *Strigamia* to human orthologues (see Table 6.2).

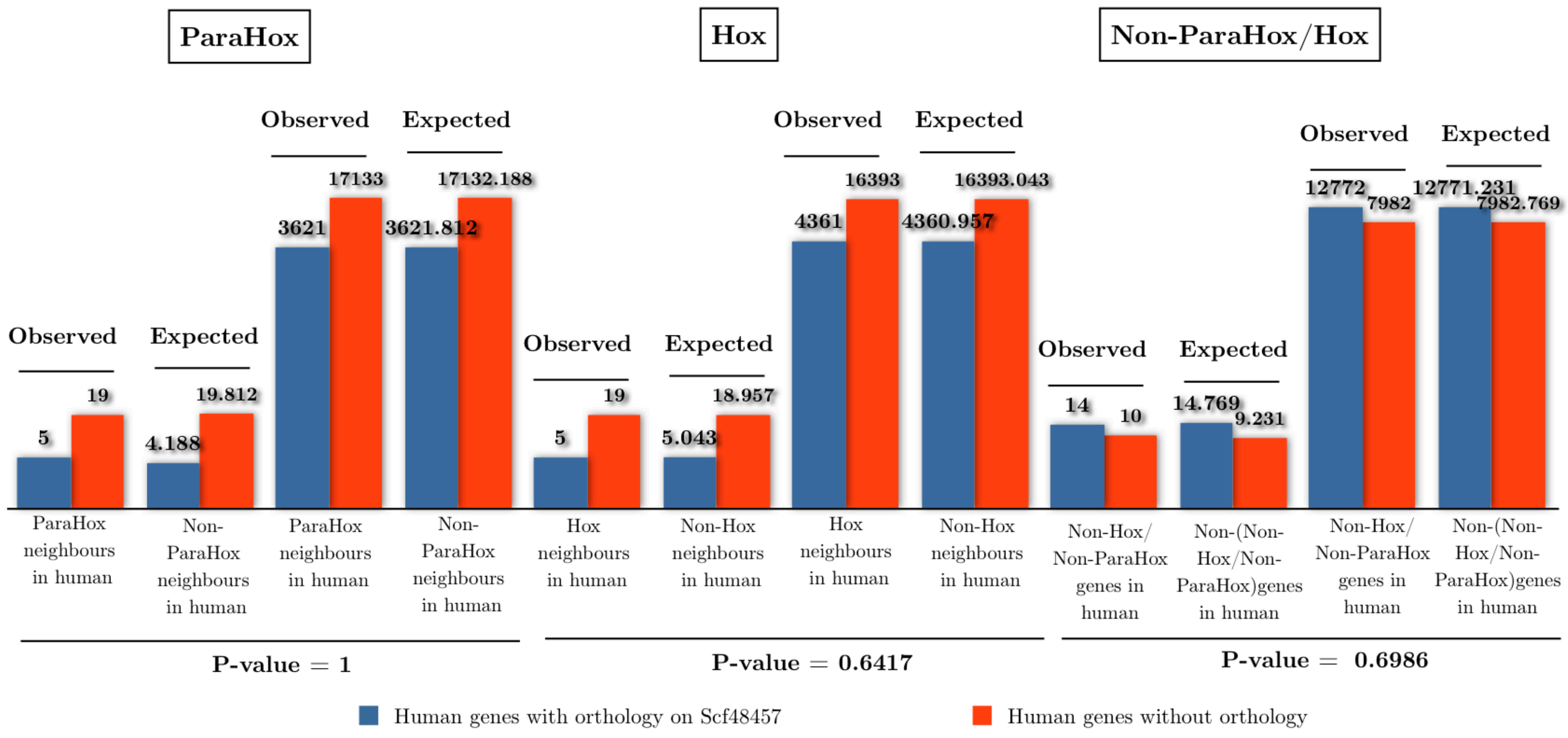


Figure 6.6.- Fisher's Exact Test to distinguish whether Strigamia scaffold 48457 has significant synteny conservation with ParaHox or Hox of humans.

S. maritima gene	Human gene	Chromosome location
Smar_temp_008046	ENSP0000362704	1
Smar_temp_007985	ENSP00000350967	18
Smar_temp_008065	ENSP00000279206	11
Smar_temp_008014	ENSP00000357973	6
Smar_temp_008000	ENSP00000404030	2
Smar_temp_008072	ENSP00000270517	19
Smar_temp_008026	ENSP00000329137	1
Smar_temp_007995	ENSP00000216862	20
Smar_temp_007986	ENSP00000260983	2
Smar_temp_008066	ENSP00000361236	6
Smar_temp_008018	ENSP00000254190	15
Smar_temp_008023	ENSP00000306340	4
Smar_temp_008004	ENSP00000339918	11
Smar_temp_008073	ENSP00000365014	9
Smar_temp_008048	ENSP00000445955	12
Smar_temp_008058	ENSP00000438978	22
Smar_temp_008029	ENSP00000312397	5
Smar_temp_008013	ENSP00000394071	X
Smar_temp_008051	ENSP00000439188	7
Smar_temp_008009	ENSP00000229270	12
Smar_temp_008017	ENSP00000454828	15
Smar_temp_008024	ENSP00000421488	4
Smar_temp_008019	ENSP00000355877	1
Smar_temp_008008	ENSP00000303525	4

Table 6.2.- One-to-one Strigamia to human orthologues starting from genes on Strigamia scaffold 48457. *The third column is the chromosomal location of the human orthologue.*

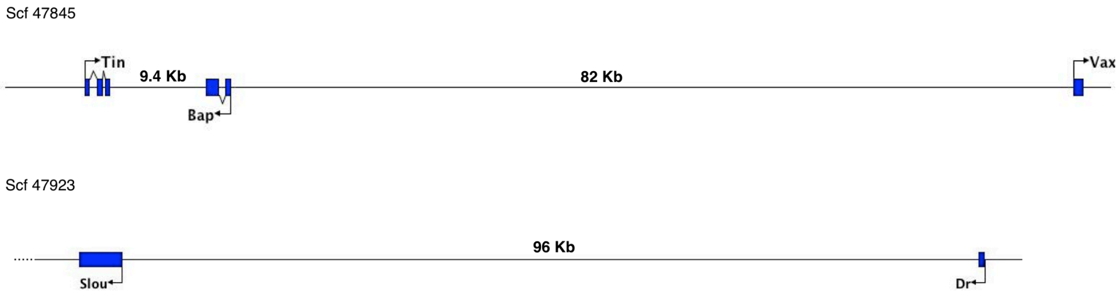


Figure 6.7.- NK cluster remains in S. maritima. *The rectangles linked by lines represent genes and the lines scaffolds. The colouring of rectangles represent the class that these genes belong to, in this case ANTP-class. The small arrows represent the transcriptional orientation.*

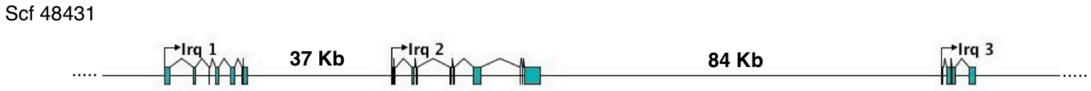


Figure 6.8.- The Iroquois cluster in S. maritima. *The rectangles linked by lines represents genes and the lines scaffolds. The colouring of the ovals represent the class that these genes belong to, in this case TALE-class. The small arrows represent the transcriptional orientation.*

Examining the locations in the human genome of these 24 genes revealed that five genes are located within chromosomes bearing human Hox clusters, five within chromosomes bearing human ParaHox loci and 14 in chromosomes with neither a Hox or ParaHox association (non-Hox/ParaHox chromosomes). Using Fisher's Exact Test I found no significant associations with Hox, ParaHox or non-Hox/ParaHox chromosomes (all tests p>=0.6, see Fig. 6.6). As with the phylogenetic analyses, the synteny analyses also unfortunately did not resolve whether these *Strigamia* genes are orthologues of Hox3 or Xlox.

In addition to the clustering of Hox and ParaHox genes some arthropods also contain an NK gene cluster. This cluster is involved in mesoderm development and provides an additional example of ANTP-class gene clustering. The clustering is likely due to the regulatory mechanisms operating on the genes, which so far are poorly characterized (Jagla et al., 2001, Cande et al., 2009). *S. maritima* does not possess an intact NK cluster, but does have some gene pairs that are remains from the ancestral cluster, potentially reflecting the retention of some shared regulatory mechanism(s). These pairs are tinman and bagpipe, often known as NK4 and NK3 in chordates, and slouch (NK1) and Drop (Msx) (Fig. 6.7). In addition, the NK cluster remnant of bagpipe and tinman is linked with Vax (Fig. 6.7), this linkage being relatively tight as there are only seven intervening genes. This linkage is also conserved in the mollusc *Lottia gigantea*. However, the number of intervening genes in *Lottia* is larger as well as the distance between bap and Vax (747 Kb). Thus, the linkage of Vax with the NK cluster is likely an ancient aspect of the organisation of these genes, dating to at least the divergence of the Ecdysozoa and Lophotrochozoa. Vax can thus be included as a new member of the ancestral ANTP-class Mega-homeobox cluster that arose deep in animal ancestry (Pollard and Holland, 2000, Garcia-Fernandez, 2005).

There is also a cluster of three TALE-class genes of the Irx/Iroquois family in *S. maritima* (Fig. 6.8). The three-gene clusters of insects and chordates are independently derived (Takatori et al., 2008, Irimia et al., 2008, Kerner et al., 2009). The three-gene cluster of *D. melanogaster* arose from an ancestral state (still present in most other insects) of two genes, one being orthologous to mirror and the second being pro-orthologous (Sharman, 1999) to araucan and caupolican. Two of the *S. maritima* Irq genes have affinity with the insect mirror gene in phylogenetic trees (Fig. 6.9). This may indicate that the three-gene cluster of this myriapod arose by duplication of the mirror gene rather than the araucan/caupolican gene in contrast to the route to the three gene cluster of *D. melanogaster*. The *S. maritima* Irq/Irx cluster thus represents a further example of the repeated independent expansion of this gene cluster in multiple lineages of the animal kingdom which intriguingly seems to settle on the three-gene state in each expanded case (Takatori et al., 2008, Irimia et al., 2008, Kerner et al., 2009).

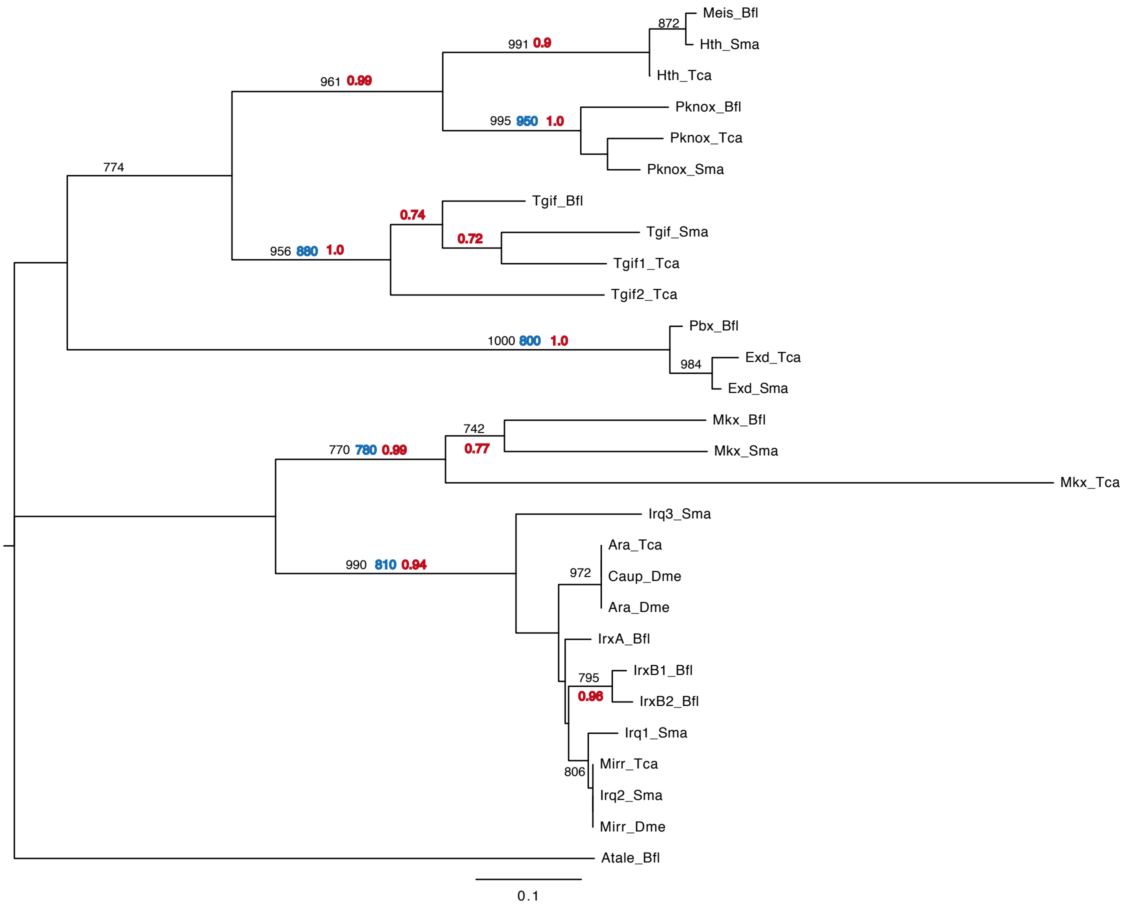


Figure 6.9.- Phylogenetic analysis of TALE class homeodomains of S. maritima using T. castaneum, D. melanogaster and B. floridae genes for comparison. *The phylogenetic tree was constructed using neighbour-joining with a JTT distance matrix and 1000 bootstrap replicates (support value in black). Nodes with support equal to or above 500 with Maximum-Likelihood (LG+G) analysis are in blue, and nodes with posterior probabilities equal to or above 0.5 with Bayesian (LG+G) are in red.*

An additional example of a homeobox gene cluster involving genes from outside the ANTP-class is the PRD-class cluster involving Orthopedia (Otp), Rax (Rx) and Homeobrain (Hbn). This cluster, which is present in *S. maritima* (Fig. 6.10), is also found in cnidarians, insects and molluscs (Mazza et al., 2010).

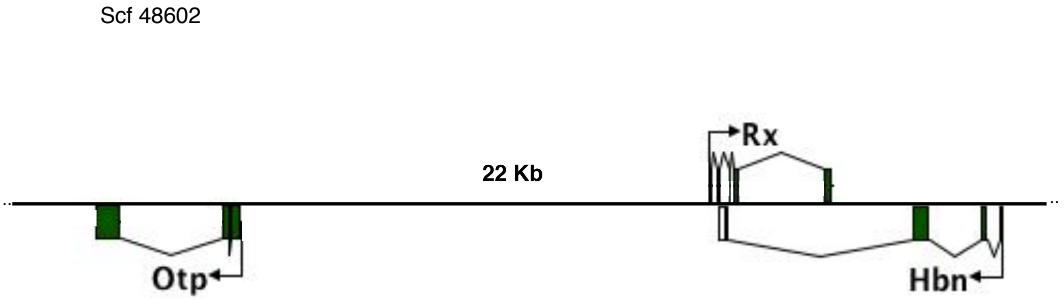


Figure 6.10.- The PRD cluster in the S. maritima genome. *The rectangles linked by lines represents genes and the line the scaffold. The colouring of the rectangles represents the class to which these genes belong, in this case the PRD-class. The small arrows denote the transcriptional orientation. The intergenic distances are indicated in kbp, except in the case of Rx-Hbn as these genes are overlapping but with opposite transcriptional orientations.*

6.3.3 Remains of ancestral homeobox clusters: the Megacluster and SuperHox

The ANTP-class of genes, including the Hox, ParaHox and NK genes, evolved very early in animal evolution, probably via states in which many of the genes were clustered into a Mega-homeobox cluster before the origin of the bilaterians and a SuperHox cluster in the Urbilaterian (Pollard and Holland, 2000, Garcia-Fernandez, 2005, Butts et al., 2008). I have found some remains of this SuperHox cluster in *S. maritima* (Fig. 6.11). SuperHox remains are represented by the linkage of Exex(Mnx)-Nedx-BtnA(Mox) in scaffold 48238 and the linkage of BtnB(Mox) with En in scaffold 48511. The Hmbox gene is linked to the Exex-Nedx-BtnA SuperHox remnant in *S. maritima* (Fig. 6.11). It remains to be seen, following future, more widespread genome sequencing, whether such a linkage represents a remnant of an ancestral state or not. The tight linkage of Ems with the IndB gene is potentially revealing with regards to the evolution of the Mega-homeobox cluster. Ems/Emx is a member of the ancestral NK linkage group (Garcia-Fernandez, 2005, Hui et al., 2012), whilst IndB is a ParaHox gene.

This tight linkage of these two genes in *S. maritima* may thus be a remnant from early animal evolution of their existence in the Mega-cluster. It should be noted that NK and ParaHox genes have become secondarily linked again in vertebrates, as Hui et al. have hypothesized that NK-cluster and ParaHox genes were on distinct chromosomes in the chordate and lophotrochozoan ancestors (Hui et al., 2012). Whilst this tight Ems-IndB linkage is intriguing, further, more taxonomically widespread examination of ANTP-class homeobox linkage patterns is certainly required to establish the veracity (or otherwise) of the Mega-cluster hypothesis. Similarly, the linkage of the ParaHox-like gene, Ind-like, with the NK gene scro may also be indicative of an ancestral linkage in the Mega cluster. However, this Ind-like - scro linkage in *S. maritima* is looser than the linkage of Ems - IndB (273 kb versus 10 kb (Fig. 6.12)) and so a secondary association cannot presently be excluded.

Finally, the linkage of the SINE class gene, sine oculis (so), with the ANTP-class genes Ems is not unique to *S. maritima*. Humans have two semi-orthologues of so, namely six1 and six2, and two semi-orthologues of ems, namely emx1 and emx2. Six2 is linked with emx1 on human chromosome 2, a linkage that is also echoed on zebrafish linkage group 13. A linkage of these SINE and ANTP-class genes at least as old as the bilaterian ancestor thus seems likely.

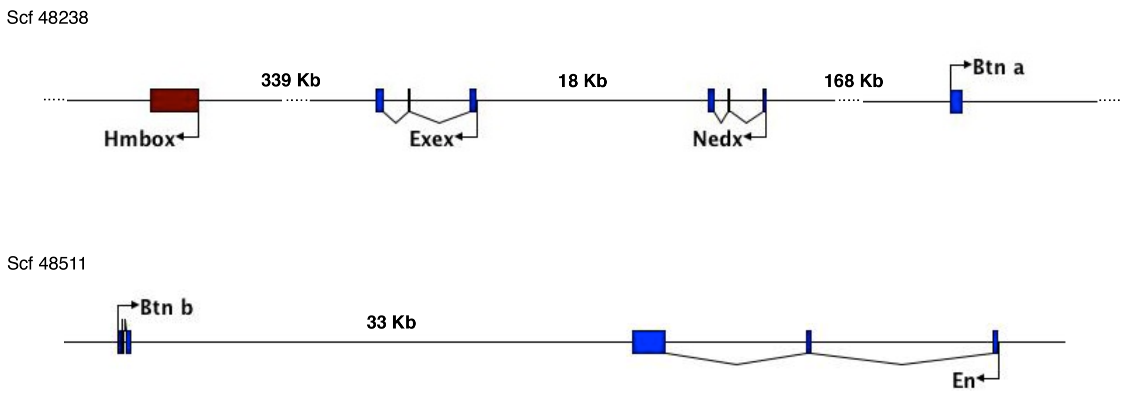


Figure 6.11.- SuperHox remains in the S. maritima genome. *The blue rectangles linked by lines are genes belonging to the ANTP-class and the brown rectangle linked by lines is a gene belonging to the HNF-class. The intergenic distances are indicated in kbp.*

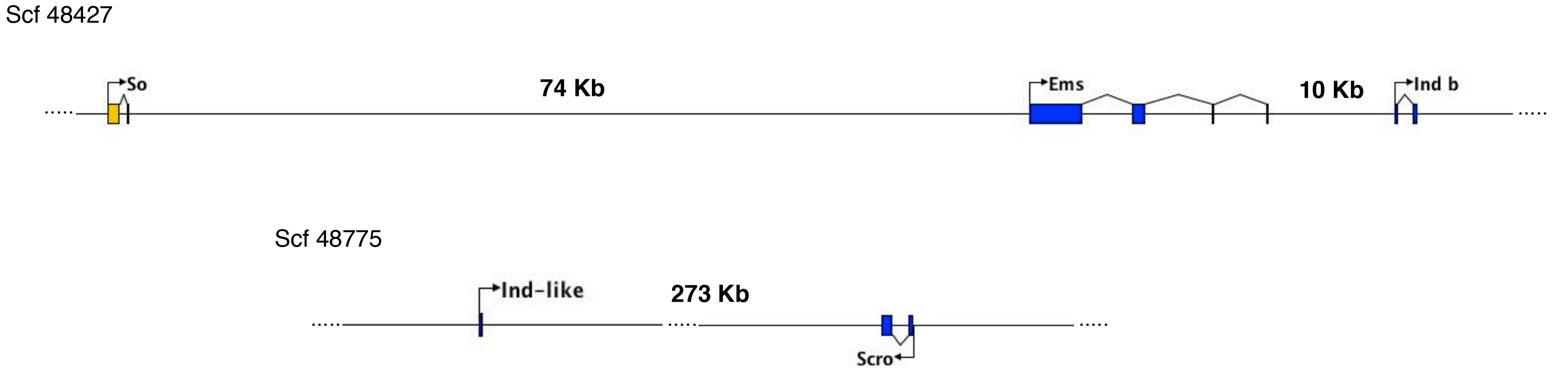


Figure 6.12.- Megacluster remains in the S. maritima genome. *The blue rectangles linked with lines are genes belonging to the ANTP-class and the yellow rectangles linked with lines is a gene belonging to the SINE-class. The intergenic distances are indicated in kbp.*

6.4 Discussion

The *S. maritima* genome represents the first genome sequenced from the myriapod lineage. This genome sequence contributes to the expansion of resources available to understand arthropod genome diversity, which in the past has been focused on other lineages. Interestingly, the *S. maritima* genome sequence retains significant traces of the large-scale genome organisation present before the divergence of protostomes and deuterosomes (Nik Putnam, personal communication). There is sufficient data available from the linkage of genes within scaffolds to reveal clear retained synteny between amphioxus and *S. maritima*. This implies that the last common ancestor of the arthropods retained significant synteny with the genomes of other animal phyla.

I have described over 9 examples of homeobox gene clustering in this myriapod genome. The remains of clustering and instances of linkage in this genome are another reflection of the retention of synteny from deep ancestors, and are a further indication of the relatively conservative nature of the *S. maritima* genome which should make it an excellent point of reference for further comparative genomics research.

As mentioned above, the number of genes in the homeobox complement of *S. maritima* is slightly larger than the numbers found in most other arthropods analysed to date. The ANTP and PRD classes have gone through several independent gene expansions. Also, these two classes include two genes, Vax and Dmbx, which have not been found before in any other arthropod genome. Also, there is the presence of a member of the HNF-class, the Hmbox gene. This gene represents the first example of a HNF-class gene in the arthropod, but a gene from the family from which the class takes its name, Hnf, has potentially been lost in this phylum. The loss of this Hnf family is potentially the only example of homeobox family loss in *S. maritima*. The greater retention of ancestral synteny relative to other arthropods so far analysed thus seems to be matched by the greater retention of gene family complements, at least if the homeobox genes are indicative.

Chapter 7

General discussion

This thesis has examined two aspects of the homeobox gene superfamily. One of these aspects deals with the origin of the Hox and ParaHox loci. Using comparative genomic approaches, such as large-scale synteny, I have compared genomes from basal lineages of the animal tree with bilaterians, allowing the identification of ancestral homologous genomic regions of the Hox and ParaHox loci. This has not only led to the formulation of a new hypothesis of how these loci originated and evolved, contradicting various other hypotheses dealing with this question, but has also been an important step in the reconstruction of a part of one of the most important genomes in the evolutionary history of the animals, i.e. the last common ancestor of all animals, the Urmetazoan. Furthermore, an independent means of corroboration stemmed from the ParaHox orthologue identification in calcareous sponges (*Sycon ciliatum* and *Leucosolenia sp.*). The second main aspect of this thesis involved cataloguing the diversity of the homeobox complement of the newly sequenced genome of the coastal centipede, *Strigamia maritima.* This catalogue improved the understanding of the evolution of homeobox gene clustering arrangements and provided further evidence of important ancestral states, such as the SuperHox and the Megacluster, of the homeobox superfamily.

This section summarises the findings of this PhD thesis, putting them into a wider context and highlighting their impacts on our understanding of some of the key aspects of animal evolution.

7.1 Macrosyntenic regions of basal animal genomes imply simplification events at the genome level that explain the origin and evolution of the Hox and ParaHox loci

The hypotheses dealing with the origin and evolution of the ANTP-class of genes have been based on the presence and absence of particular ANTP-class family members. Within the ANTP-class resides the paralogous Hox and ParaHox gene families, which have previously been hypothesized to have evolved via duplication from a common ancestral state: the ProtoHox. The timing and the origin of ProtoHox, Hox and ParaHox genes relative to particular animal lineages has led to conflicting hypotheses (examined in Chapter 1, Section 1.5.2). The common approach of these hypotheses focused only on the mode of duplication of the family members, ignoring the dynamics affecting these families at the whole genome level. In combination with this, the absence of these genes or cases of ambiguous phylogenetic resolution of family members has required the use an alternative approach, the inference of macrosynteny. In Chapters 3 and 4 it has been showed that there is a significant amount of very ancient genomic architecture, at least in the Hox and ParaHox loci of bilaterians and cnidarians (Putnam et al., 2007, Hui et al., 2008), and I showed that these loci are also present in the placozoans and poriferans (Chapters 3 and 4; (Mendivil Ramos et al., 2012)). The new term defining these regions is "ghost" loci. "Ghost" loci denotes a macrosyntenic region of the Porifera and Placozoa genomes homologous to the Hox and ParaHox loci in bilaterians, but this region does not actually contain the homeobox genes themselves.

The underlying general hypothesis of the ghost loci is that the ancestral ProtoHox locus, containing one or more homeobox genes along with a variety of neighbouring non-homeobox genes, duplicated to generate two loci that became the Hox and ParaHox loci. The evidence that the duplication of the ProtoHox was a large-scale event involving multiple genes stems from two sources:

(i) Collagen and tyrosine kinase receptor genes flank the Hox and ParaHox clusters and thus, the ProtoHox duplication included homeobox and neighbouring genes (Minguillón and Garcia-Fernàndez, 2003).

(ii) It has been postulated that at some point in evolution the ProtoHox gene/cluster was linked with a pro-orthologue of Mox and Evx, then the whole block must have duplicated in tandem, giving rise to the Hox cluster linked with Evx and the ParaHox cluster linked with Mox. Eventually, the ParaHox cluster translocated, leaving Mox distantly linked to the Hox-Evx block (Minguillón and Garcia-Fernàndez, 2003).

(iii) Lanfear and Bromham (Lanfear and Bromham, 2008) statistically tested the likelihoods of the alternative ProtoHox models (individual genes through to 2-, 3- and 4-gene clusters) and found support for either of the 3- or 4-gene models.

Since the Hox and ParaHox clusters are on separate chromosomes the implication is that the large duplication could have happened in one of two ways:

(i) One possibility is that this duplication stems from a whole genome (WGD) or whole chromosome duplication, i.e. the ProtoHox cluster and its neighbours located on one chromosome duplicated, giving rise to two distinct chromosomes. Following this event extensive loss occurred along the daughter chromosomes so that distinctive sets of Hox and ParaHox neighbours remained. Differential gene loss after such a large duplication has been observed in the human genome. Following the 2R WGD at the origin of the vertebrates, the human genome has retained less than 30% of the ohnologues generated in this event (Makino and McLysaght, 2010). Also, WGDs and polyploidy in animals are frequent events, with further examples being regularly identified (extensively reviewed in chapter 1, section 1.4.1).

(ii) The second possibility is that this duplication stems from a large, multi-gene segmental duplication within a single chromosome. This would have entailed an intra-chromosomal duplication of the ProtoHox cluster, followed by a translocation to another chromosome via a chromosome arm exchange or chromosome fission. This scenario would imply a differential distribution of neighbouring genes of the original ProtoHox to the descendant Hox and ParaHox loci.

Alternatively, one could argue that the ProtoHox duplication into Hox and ParaHox did not involve neighbouring genes. This would have been possible through retrotransposition or a small scale inter-chromosomal DNA-based transposition (Fig. 7.1).

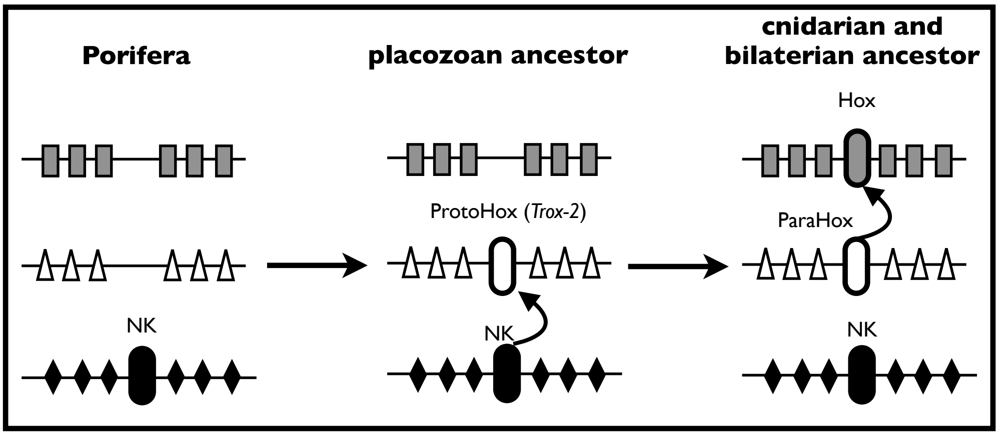


Figure 7.1.- Less parsimonious alternative to the Ghost Locus hypothesis. *Summary of duplication events occurring via retrotransposition or DNA-based transposition in the basal animal lineages. In the placozoan ancestor the duplication of the NK gene(s) via a transposition event that did not include non-homeobox neighbour genes gives rise to the ProtoHox (of which Trox-2 is a direct descendant). In the cnidarian and bilaterian ancestor the ProtoHox duplicates via another transposition event that did not include neighbouring genes. One of the copies evolves into ParaHox gene(s) and the other gives rise to the Hox gene(s). Note, this scenario also requires asymmetrical evolution after the ProtoHox/Trox-2 state, such that the ParaHox descendant gene Gsx retains greater similarity with the ProtoHox/Trox-2 gene than do any other descendant Hox and ParaHox genes.*

This event is unlikely, for two reasons:

(i) A retrotransposition involves a single coding sequence and thus, clashes with the fact that the ProtoHox duplication would most likely have entailed a cluster of genes (Brooke et al., 1998, Ferrier and Holland, 2001, Minguillón and Garcia-Fernàndez, 2003, Lanfear and Bromham, 2008).

(ii) Retrogenes would need to adopt the regulatory elements of the locus into which they were inserted, and thus adopt the expression profile of this region and not the parental one. This seems an unlikely explanation for the Hox/ParaHox duplication, since both gene clusters have comparable, complex patterns of expression involving anterior-posteriorly staggered expression patterns in the nervous system and other tissues in bilaterians. This is consistent with the ancestral ProtoHox duplication involving coding sequences and regulatory elements.

Thus, retrotrasposition is unlikely to have had a role in the duplication of ProtoHox into Hox and ParaHox. Futhermore, interchromosomal DNA-based transposition is also unlikely as these rarely involve a whole coding sequence and by extension is unlikely to include regulatory elements. These small-scale transpositions can occur either during the process of segmental duplication (SD) or only when a gene transposes without duplicating, known as a Positionally Relocated gene (PRG (Bhutkar, 2007)). Regarding segmental duplications, it has been observed that the median size of the duplication is much smaller than the average size of a gene in nematodes, human and flies (Katju and Lynch, 2003, Zhang et al., 2005, Meisel, 2009a). Furthermore, these sizes are inferred solely from the exon boundaries, as there is currently a lack of information about regulatory elements across whole animal genomes. In summary, even though PRGs or SDs could include a whole coding sequence in a new chromosomal location they most likely lack an ancestral regulatory region, and the new duplicated and transposed fragment or coding sequence will acquire a new regulatory input and a novel expression pattern relative to the ancestral locus. Once more, this clashes with the similar patterns of expression that the Hox and ParaHox genes possess.

As these events (PRGs and SDs) do happen in any genome, and thus cannot be discarded as a possible mechanism of separation of the Hox and ParaHox loci, their frequency is, however, rather low. To explain the separation of the Hox/ParaHox situation in terms of an SD-like event it would be rare, as they are required to be interchromosomal (which is even rarer than intrachromosomal (Katju and Lynch, 2003, Zhang et al., 2005, Bhutkar, 2007, Meisel, 2009a)), multi-genic (i.e. homeobox genes and a few non-homeobox neighbours) and all their regulatory elements. Therefore, I favour the hypothesis of a whole locus split or duplication including homeobox genes and non-homeobox neighbours, followed by differential gene loss of descendant neighbours (gene loss being a common occurrence (Hughes and Friedman, 2004, Danchin, 2006, Miller et al., 2007, Wyder et al., 2007, Takahashi et al., 2009, Makino and McLysaght, 2010)) and in some cases loss of the homeobox genes themselves, resulting in ghost loci. In opposition to the common view that evolutionary events contribute lead to increasing complexity, the hypothesis that I favour implies a simplification from the complexity of the last common ancestor's genome. Moreover, the implications of these findings contradicts the ParaHoxozoa nomenclature proposed by Ryan et al. (2010) that has been proposed to denote Placozoa, Cnidaria and Bilateria, but now will actually comprise all metazoans as the Urmetazoan had Hox and ParaHox loci.

In Chapter 3 I predicted that scaffold 5 and 38 of the *T. adhaerens* genome are linked, together composing the ParaHox locus. While it is outside of the scope of the current work, this prediction provides an interesting avenue for future work. For example, it could be verified by using genome walking techniques and/or using fluorescence *in situ* hybridisation to locate some of the gene positions within the chromosome, or by using an *in silico* approach involving sequencing reads of the genome and trying to retrieve an enlarged version of scaffold 5 and/or 38. Likewise, the Monte-Carlo simulations of the *A. queenslandica* genome sequence performed in Chapter 4 predicted that the Hox neighbour orthologue genes are clustered separately and independently from the cluster of the ParaHox neighbour orthologue genes. The distinct clustering arrangement of Hox and ParaHox could also be verified by chromosomal fluorescence *in situ* hybridisation. Lastly, the increase in taxon sampling of sponge genomes sequences or ANTP-surveys might reveal over new homeoboxes and independently verify this hypothesis as explained next.

7.2 ParaHox genes in calcareous sponges support the "ghost" loci hypothesis?

The increase of the taxon sampling of sponges possibly will provide new insights that could verify the ghost loci hypothesis. New calcareous sponge genomes sequences, *Sycon cilliatum* (in late stages of assembly) and *Leucosolenia sp.* (in the pipeline for genome sequencing and assembly), are being the focus of the study of the Adamska lab (SARS, Norway). As has been discussed above, I have demonstrated the existence of ghost Hox and ParaHox loci in a sponge, *Amphimedon queenslandica*, implying that these homeobox genes were lost during the evolution of the sponge lineage (Chapters 3 and 4 (Mendivil-Ramos et al., 2012)). In collaboration with the Adamska lab it was possible to identify two ParaHox sequences of the aforementioned calcareous sponges and thus, further investigate the precise timing of the loss of Hox and ParaHox.

In Chapter 5 I helped to identify the orthology of two ANTP-class genes of the calcareous sponges *Sycon* and *Leucosolenia.* The sequence analyses performed indicate that very likely these sequences are very likely orthologues of the bilaterian ParaHox gene, Cdx. This is taking into account the length of the phylogenetic tree branches, and the retention of some informative combinations of amino acids and the persistent clustering of these genes with the bilaterian and cnidarian Cdx sequences in a variety of phylogenies. Furthermore, I recovered very weak synteny signal stemming from one of the surroundings genes of *Sycon* Cdx that associate this scaffold with the bilaterian and cnidarian ParaHox loci. Both findings indicate the ParaHox genes are present in this lineage and were likely to be present in the last common ancestor of all animals. These ParaHox genes are the first ever identified in sponges and contradict all the previous indications that all sponges have lost all of the Hox/ParaHox genes. Also, the implications of these findings again contradicts the ParaHoxozoa nomenclature, that now effectively becomes synonymous with "Metazoa". Moreover, the general views from these findings are that the Hox/ParaHox genes have undergone differential loss across the different sponge lineages.

These general views can be further explored in the near future. Given the limited synteny signal from the *Sycon* Cdx scaffold and the absence of a full genome sequence assembly and gene annotation, it remains to be resolved whether there is another region(s) homologous to bilaterian/cnidarian ParaHox loci which could be linked to this scaffold. Also, resolving whether there is a "ghost" Hox locus in the *Sycon* genome, should be an immediate avenue of research to pursue.

7.3 Diversity of the homeobox complement and synteny conservation in *Strigamia maritima* contributes to further reconstruction of ancestral states in Ecdysozoa and to bilaterians (the Urbilaterian)

The genome sequence of *S. maritima* has increased the diversity of available arthropod genome sequences, as it is the first ever myriapod species sequenced. This genome is noteworthy for the numerous instances of retained ancestral gene families (Michael Akam personal communication) and the significant traces of large-scale conservation of genome organisation relative to the genomes of other animals. The homeobox gene superfamily catalogued herein is consistent with the general pattern of conserved synteny of this genome. In Chapter 6 I described nine cases of clustering of this superfamily within this genome and the presence of several homeobox genes not previously described in the arthropods (e.g.: Dmbx, Vax and Hmbox and the retention of remnants of the Megacluster and SuperHox clusters). Thus, one could say that this genome represents an arthropod genome that is less derived from the ancestral bilaterian state than other available arthropod (an ecdysozoan) genome sequences. Quite possibly, the general retention of micro- and macro-synteny of this genome is also accompanied by ancestral cis-regulatory regions and thus, ancestral expression patterns and functions. Future work will further characterise these expression patterns.

In this vein, it would be interesting to examine the ambiguous orthology of the genes (Smar_temp_SM33002 and Smar_temp_SM33003 with their tentative names *Hox3a* and *Hox3b*). In Chapter 6, I demonstrated that these genes show weak affinity to bilaterian Hox 3 and Xlox sequences and thus are of unclear orthology, in contrast to the conclusions of Panfilo and Akam (2007). Given these results, the expression data of these genes could perhaps shed light on their orthology. However, expression data would have to be interpreted with great care because Hox 3 gene orthologues in arthropods have gone through changes in expression, function and copy number. The expression of Hox3 in the crustacean *Daphnia pulex* has been shown to be representative of a canonical Hox gene involved in anterior-posterior patterning (Papillon and Telford, 2007). On the other hand, this gene has duplicated and diverged in function (i.e.: becoming involved in katatrepsis) in the lineages of *D. melanogaster* and *T. castaneum.* In the case of *D. melanogaster* this gene duplicated, giving rise to *zen1* and *zen2* (Rushlow et al., 1987). In the case of *T. castaneum* this gene also duplicated and the duplicates subsequently evolved by subfunctionalization (van der Zee et al., 2005). To date, no Xlox gene orthologue has been identified in the ecdyzosoans and thus, the only expression data that would be comparable come from outside the ecdyzosoans. Lophotrochozoan and deuterostome Xlox genes are expressed in regions of the CNS and midgut (Hui et al., 2009). Peculiarly, the Hox cluster in *S. maritima* lacks Hox 3. Similarly, it is worth noting that there is a dispersed ParaHox cluster that potentially lacks Xlox. Consequently it remains to be seen whether the functional studies of these genes can actually shed light on their orthology relationships.

7.4 General conclusions and future directions

This PhD project has studied the origin and evolution of the homeobox gene superfamily in animals. This superfamily is one of the most distinctive groups of genes involved in the evolution of developmental processes. This work provides an example of the way in which comparative genomics can enhance the resolution of classical 'evo-devo' questions and general evolutionary biology questions. In particular, the approach undertaken here has provided insights at a genome-scale of evolutionary events within this superfamily. This has led to the formulation of a new hypothesis of the origin of the Hox and ParaHox loci and the genetic complexity of the last common ancestor of all animals. Likewise, the comparative approach used to build up a platform for classification of the homeobox gene complement of *S. maritima* has provided a foundation for investigating the biological significance of ancestral clustering of this superfamily and insights into the reconstruction of ancestral states of this superfamily.

Many of the limitations faced during this project have been due to the limitations of the comparative genomics field. Perhaps the most important constraint comes from the quality of the genome sequences. The great repertoire of the ever-increasing animal genome sequences have a high level of variability in ‘quality’ (i.e. the number of gaps in the sequence, and independent mapping data used to confirm the assembly (Mendivil Ramos and Ferrier, 2012)). In that way, careful considerations should be made when handling some of these sequences. Furthermore, quality is fundamental for gene annotation, orthologue identification and eventually genome assembly that goes further than subchromosomal assembly. High quality of genomes will enable the deduction of more accurate conclusions about macro-mutations that constitute major forces in evolutionary innovation (e.g. duplication). This should be complemented by investigations of the cis-regulatory architecture governing the underlying structure of a genome.

Moreover, the increased taxon sampling of animal is allowing increased resolution of the animal tree and helping to indicate key nodes within the animal tree that represent important evolutionary transitions. The identification of key nodes within the animal phylogeny, in combination with higher quality genome assemblies, will enable the provision of alternatives to human genome with which to deduce genome-scale evolutionary processes across the animal kingdom. The continued efforts to develop *in silico* tools and theoretical models to estimate the rearrangement rates of these macro-mutations needs to be applied across a range of animals in order to distinguish general processes from lineage-specific peculiarities. The gradual and continual expansion of the comparative genomics field will greatly contribute to our understanding of the evolution of developmental mechanisms in a much greater detail than known to date, especially with regards to one of the most important groups of developmental genes in “evo-devo”: the homeobox genes.

Bibliography

Abbasi, A. A. 2010. Unraveling ancient segmental duplication events in human genome by phylogenetic analysis of multigene families residing on HOX-cluster paralogons. *Mol Phylogenet Evol,* 57, 836-48.

Aboobaker, A. & Blaxter, M. 2003. Hox gene evolution in nematodes: novelty conserved. *Current Opinion in Genetics & Development,* 13, 593-598.

Adams, M. D. 2000. The Genome Sequence of Drosophila melanogaster. *Science,* 287, 2185-2195.

Adoutte, A., Balavoine, G., Lartillot, N., Lespinet, O., Prud'homme, B. & De Rosa, R. 2000. The new animal phylogeny: Reliability and implications. *Proceedings of the National Academy of Sciences,* 97, 4453-4456.

Aguinaldo, A. M. A., And & Lake, J. A. 1998. Evolution of the Multicellular Animals. *American Zoologist,* 38, 878-887.

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

Akam, M. 1989. Hox and HOM: Homologous gene clusters in insects and vertebrates. *Cell,* 57, 347-349.

Alam, S. L., Langelier, C., Whitby, F. G., Koirala, S., Robinson, H., Hill, C. P., And & Sundquist , W. I. 2006. Structural basis for ubiquitin recognition by the human ESCRT-II EAP45 GLUE domain. *Nature Structural & Molecular Biology,* 13, 1029-1030.

Allendoft, F. T., Gh 1984. *Tetraploidy and the evolution of salmonids,* New York, Plenum Press.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology,* 215, 403-410.

Andersen, S. L., Bergstralh, D. T., Kohl, K. P., Larocque, J. R., Moore, C. B. & Sekelsky, J. 2009. Drosophila MUS312 and the Vertebrate Ortholog BTBD12 Interact with DNA Structure-Specific Endonucleases in DNA Repair and Recombination. *Molecular Cell,* 35, 128-135.

Anderson, F. E., Córdoba, A. J. & Thollesson, M. 2004. Bilaterian Phylogeny Based on Analyses of a Region of the Sodium–Potassium ATPase β-Subunit Gene. *Journal of Molecular Evolution,* 58, 252-268.

Arai, K. M., K; Suzuki, R 1993. Production of polyploids and viable gynogens using spontaneously occurring tetraploid load, *misgurnus anguillocaudatus. Aquaculture,* 117, 227-235.

Atkinson, H. J. & Babbitt, P. C. 2009. An Atlas of the Thioredoxin Fold Class Reveals the Complexity of Function-Enabling Adaptations. *PloS Comput Biol,* 5, e1000541.

Ax, P. 1996. *Multicellular animals: A new approach to the phylogenetic order in nature.*, Springer Verlag, Berlin.

Babushok, D. V. & Kazazian, H. H., Jr. 2007. Progress in understanding the biology of the human mutagen LINE-1. *Hum Mutat,* 28, 527-39.

Bailey, J., Liu, G. & Eichler, E. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *The American Journal of Human Genetics,* 73, 823-834.

Bailey, J. A. & Eichler, E. E. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet,* 7, 552-64.

Baker, M. E. 2001. Evolution of 17B-hydroxysteroid dehydrogenases and their role in androgen, estrogen and retinoid action. *Molecular and Cell Endocrinology,* 171, 211-215.

Balczon, R., Bao, L. & Zimmer, W. E. 1994. PCM-1, A 228-kD Centrosome Autoantigen with a Distinct Cell Cycle Distribution. *Journal Cell Biology,* 124, 783-793.

Bao, R. & Friedrich, M. 2008. Conserved cluster organization of insect Runx genes. *Development Genes and Evolution,* 218, 567-574.

Bhutkar, A., Russo, S.M., Smith, T.F., and Gelbart, W.M., 2007. Genome-scale analysis of positionally relocated genes. *Genome Research,* 17, 1880-1887.

Böger, H. 1983. Versuch über das phylogenetische System der Porifera. *Meyniana,* 40, 143-154.

Brooke, N. M., Garcia-Fernàndez, J. & Holland, P. W. H. 1998. The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature,* 392, 920-922.

Brooke, N. M., Garcia-Fernàndez J. & Holland P.W. H. 1998. The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature,* 392, 920-922.

Bürglin, T. R. 1994. A Comprehensive Classification of Homeobox Genes *In:* DUBOULE, D. (ed.) *Guidebook to the Homeobox Genes.* Oxford University Press.

Bürglin, T. R. 2005. *Homeodomain proteins,* Wiley-VCH Verlag GmbH & Co. , Weinham.

Butts, T., Holland, P. W. H. & Ferrier, D. E. K. 2008. The Urbilaterian Super-Hox cluster. *TRENDS in Genetics,* 24, 259-262.

Calero, M., Winand, N. J. & Collins, R. N. 2002. Identification of the novel proteins Yip4p and Yip5p as Rab GTPase interacting factors. *FEBS Lett,* 515, 89-98.

Cameron, R. A., Rowen, L., Nesbitt, R., Bloom, S., Rast, J. P., Berney, K., Arenas-Mena, C., Martinez, P., Lucas, S., Richardson, P. M., Davidson, E. H., Peterson, K. J. & Hood, L. 2006. Unusual gene order and organization of the sea urchin hox cluster. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution,* 306B, 45-58.

Cande, J. D., Chopra, V. S. & Levine, M. 2009. Evolving enhancer-promoter interactions within the tinman complex of the flour beetle, Tribolium castaneum. *Development,* 136, 3153-3160.

Cañestro, C., Yokoi, H. & Postlethwait, J. H. 2007. Evolutionary developmental biology and genomics. *Nat Rev Genet,* 8, 932-42.

Caravas, J. & Friedrich, M. 2010. Of mites and millipedes: Recent progress in resolving the base of the arthropod tree. *BioEssays,* 32, 488-495.

Carvalho, A. P., Fernandes, P. A. & Ramos, M. J. 2006. Similarities and differences in the thioredoxin superfamily. *Prog Biophys Mol Biol,* 91, 229-248.

Carvalho, C. M., Ramocki, M. B., Pehlivan, D., Franco, L. M., Gonzaga-Jauregui, C., Fang, P., Mccall, A., Pivnick, E. K., Hines-Dowell, S., Seaver, L. H., Friehling, L., Lee, S., Smith, R., Del Gaudio, D., Withers, M., Liu, P., Cheung, S. W., Belmont, J. W., Zoghbi, H. Y., Hastings, P. J. & Lupski, J. R. 2011. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat Genet,* 43, 1074-81.

Castresana, J. 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol Biol Evol,* 17, 540-552.

Castro, L. F. C. & Holland, P. W. H. 2003. Chromosomal mapping of ANTP class homeobox genes in amphioxus: piecing together ancestral genomes. *Evolution & Development,* 5, 459-465.

Castro, L. F. C., Rasmussen, S. L. K., Holland, P. W. H., Holland, N. D. & Holland, L. Z. 2006. A Gbx homeobox gene in amphioxus: Insights into ancestry of the ANTP class and evolution of the midbrain/hindbrain boundary. *Developmental Biology,* 295, 40-51.

Chai, C.-L., Zhang, Z., Huang, F.-F., Wang, X.-Y., Yu, Q.-Y., Liu, B.-B., Tian, T., Xia, Q.-Y., Lu, C. & Xiang, Z.-H. 2008. A genomewide survey of homeobox genes and identification of novel structure of the Hox cluster in the silkworm, Bombyx mori. *Insect Biochemistry and Molecular Biology,* 38, 1111-1120.

Chambeyron, S. & Bickmore, W. A. 2004. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes & Development,* 18, 1119-1130.

Chambeyron, S., Da Silva, N. R., Lawson, K. A. & Bickmore, W. A. 2005. Nuclear re-organisation of the Hoxb complex during mouse embryonic development. *Development,* 132, 2215-2223.

Chang, E. M., Han, J. E., Kwak, I. P., Lee, W. S., Yoon, T. K. & Shim, S. H. 2012. Preimplantation genetic diagnosis for couples with a Robertsonian translocation: practical information for genetic counseling. *J Assist Reprod Genet,* 29, 67-75.

Chenuil, A. G., N; Berrebi, P 1999. A test of the hypothesis of an autopolyploid vs. allopolyploid origin for tetraploid lineage: application to the genus *Barbus* (Cyprinidae). *Heredity,* 82, 373-380.

Chiori, R., Jager, M., Denker, E., Wincker, P., Da Silva, C., Le Guyader, H., Manuel, M. & Quéinnec, E. 2009. Are Hox Genes Ancestrally Involved in Axial Patterning? Evidence from the Hydrozoan <italic>Clytia hemisphaerica</italic> (Cnidaria). *PLoS One,* 4, e4231.

Chourrout, D., Delsuc, F., Chourrout, P., Edvardsen, R. B., Rentzsch, F., Renfer, E., Jensen, M. F., Zhu, B., De Jong, P., Steele, R. E. & Technau, U. 2006. Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements. *Nature,* 442**,** 684-687.

Clamp, M., Cuff, J., Searle, S. M. & Barton, G. J. 2004. The Jalview Java alignment editor. *Bioinformatics,* 20**,** 426-427.

Colbourne, J. K., Pfrender, M. E., Gilbert, D., Thomas, W. K., Tucker, A., Oakley, T. H., Tokishita, S., Aerts, A., Arnold, G. J., Basu, M. K., Bauer, D. J., Cáceres, C. E., Carmel, L., Casola, C., Choi, J.-H., Detter, J. C., Dong, Q., Dusheyko, S., Eads, B. D., Fröhlich, T., Geiler-Samerotte, K. A., Gerlach, D., Hatcher, P., Jogdeo, S., Krijgsveld, J., Kriventseva, E. V., Kültz, D., Laforsch, C., Lindquist, E., Lopez, J., Manak, J. R., Muller, J., Pangilinan, J., Patwardhan, R. P., Pitluck, S., Pritham, E. J., Rechtsteiner, A., Rho, M., Rogozin, I. B., Sakarya, O., Salamov, A., Schaack, S., Shapiro, H., Shiga, Y., Skalitzky, C., Smith, Z., Souvorov, A., Sung, W., Tang, Z., Tsuchiya, D., Tu, H., Vos, H., Wang, M., Wolf, Y. I., Yamagata, H., Yamada, T., Ye, Y., Shaw, J. R., Andrews, J., Crease, T. J., Tang, H., Lucas, S. M., Robertson, H. M., Bork, P., Koonin, E. V., Zdobnov, E. M., Grigoriev, I. V., Lynch, M. & Boore, J. L. 2011. The Ecoresponsive Genome of Daphnia pulex. *Science,* 331**,** 555-561.

Collares-Pereira, M., Madeira, J. & Rab, P. 1995. Spontaneous triploidy in the stone loach *Noemacheilus barbatulus* (Balitoridae). *Copeia,* 2**,** 483-484.

Conant, G. C. & Wagner, A. 2003. Asymmetric sequence divergence of duplicate genes. *Genome Res,* 13**,** 2052-8.

Consortium, I. H. G. S. 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

Consortium, S. U. G. S., Sodergren, E., Weinstock, G. M., Davidson, E. H., Cameron, R. A., Gibbs, R. A., Angerer, R. C., Angerer, L. M., Arnone, M. I., Burgess, D. R., Burke, R. D., Coffman, J. A., Dean, M., Elphick, M. R., Ettensohn, C. A., Foltz, K. R., Hamdoun, A., Hynes, R. O., Klein, W. H., Marzluff, W., Mcclay, D. R., Morris, R. L., Mushegian, A., Rast, J. P., Smith, L. C., Thorndyke, M. C., Vacquier, V. D., Wessel, G. M., Wray, G., Zhang, L., Elsik, C. G., Ermolaeva, O., Hlavina, W., Hofmann, G., Kitts, P., Landrum, M. J., Mackey, A. J., Maglott, D., Panopoulou, G., Poustka, A. J., Pruitt, K., Sapojnikov, V., Song, X., Souvorov, A., Solovyev, V., Wei, Z., Whittaker, C. A., Worley, K., Durbin, K. J., Shen, Y., Fedrigo, O., Garfield, D., Haygood, R., Primus, A., Satija, R., Severson, T., Gonzalez-Garay, M. L., Jackson, A. R., Milosavljevic, A., Tong, M., Killian, C. E., Livingston, B. T., Wilt, F. H., Adams, N., Bellé, R., Carbonneau, S., Cheung, R., Cormier, P., Cosson, B., Croce, J., Fernandez-Guerra, A., Genevière, A.-M., Goel, M., Kelkar, H., Morales, J., Mulner-Lorillon, O., Robertson, A. J., Goldstone, J. V., Cole, B., Epel, D., Gold, B., Hahn, M. E., Howard-Ashby, M., Scally, M., Stegeman, J. J., Allgood, E. L., Cool, J., Judkins, K. M., Mccafferty, S.

S., Musante, A. M., Obar, R. A., Rawson, A. P., Rossetti, B. J., Gibbons, I. R., Hoffman, M. P., Leone, A., Istrail, S., Materna, S. C., Samanta, M. P., Stolc, V., et al. 2006. The Genome of the Sea Urchin Strongylocentrotus purpuratus. *Science,* 314**,** 941-952.

Copley, R. R., Aloy, P., Russell, R. B. & Telford, M. J. 2004. Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of Caenorhabditis elegans. *Evolution & Development,* 6**,** 164-169.

D'souza, T. G., Storhas, M., Schulenburg, H., Beukeboom, L. W. & Michiels, N. K. 2004. Occasional sex in an 'asexual' polyploid hermaphrodite. *Proc Biol Sci,* 271**,** 1001-7.

Danchin, E. G. J., Gouret, P., and Pontarotti, P., 2006. Eleven ancestral gene families lost in mammals and vertebrates while otherwise universally conserved in animals. *BMCEvolutionary Biology,* 6.

De Rosa, R., Grenier, J. K., Andreeva, T., Cook, C. E., Adoutte, A., Akam, M., Carroll, S. B. & Balavoine, G. 1999. Hox genes in brachiopods and priapulids and protostome evolution. *Nature,* 399**,** 772-776.

Deák, F., Wagener, R., Kiss, I. & Paulsson, M. 1999. The matrilins: a novel family of oligomeric extracellular matrix proteins. *Matrix Biology,* 18**,** 55-64.

Dearden, P. K., Wilson, M. J., Sablan, L., Osborne, P. W., Havler, M., Mcnaughton, E., Kimura, K., Milshina, N. V., Hasselmann, M., Gempe, T., Schioett, M., Brown, S. J., Elsik, C. G., Holland, P. W. H., Kadowaki, T. & Beye, M. 2006. Patterns of conservation and change in honey bee developmental genes. *Genome Res,* 16**,** 1376-1384.

Dehal, P. & Boore, J. L. 2005. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLoS Biol,* 3**,** e314.

Dehal, P., Satou, Y., Campbell, R. K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D. M., Harafuji, N., Hastings, K. E. M., Ho, I., Hotta, K., Huang, W., Kawashima, T., Lemaire, P., Martinez, D., Meinertzhagen, I. A., Necula, S., Nonaka, M., Putnam, N., Rash, S., Saiga, H., Satake, M., Terry, A., Yamada, L., Wang, H.-G., Awazu, S., Azumi, K., Boore, J., Branno, M., Chin-Bow, S., Desantis, R., Doyle, S., Francino, P., Keys, D. N., Haga, S., Hayashi, H., Hino, K., Imai, K. S., Inaba, K., Kano, S., Kobayashi, K., Kobayashi, M., Lee, B.-I., Makabe, K. W., Manohar, C., Matassi, G., Medina, M., Mochizuki, Y., Mount, S., Morishita, T., Miura, S., Nakayama, A., Nishizaka, S., Nomoto, H., Ohta, F., Oishi, K., Rigoutsos, I., Sano, M., Sasaki, A., Sasakura, Y., Shoguchi, E., Shin-I, T., Spagnuolo, A., Stainier, D., Suzuki, M. M., Tassy, O., Takatori, N., Tokuoka, M., Yagi, K., Yoshizaki, F., Wada, S., Zhang, C., Hyatt, P. D., Larimer, F., Detter, C., Doggett, N., Glavina, T., Hawkins, T., Richardson, P., Lucas, S., Kohara, Y., Levine, M., Satoh, N. & Rokhsar, D. S. 2002. The Draft Genome of Ciona intestinalis: Insights into Chordate and Vertebrate Origins. *Science,* 298**,** 2157-2167.

Dellaporta, S. L., Xu, A., Sagasser, S., Jakob, W., Moreno, M. A., Buss, L. W. & Schierwater, B. 2006. Mitochondrial genome of Trichoplax adhaerens supports Placozoa as the basal lower metazoan phylum. *Proceedings of the National Academy of Sciences,* 103, 8751-8756.

Delsuc, F., Brinkmann, H. & Philippe, H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet,* 6, 361-375.

Deschamps, J. 2007. Ancestral and recently recruited global control of the Hox genes in development. *Current Opinion in Genetics & Development,* 17, 422-427.

Dewey, C. N. 2011. Positional orthology: putting genomic evolutionary relationships into context. *Brief Bioinform,* 12, 401-12.

Dohrmann, M., Janussen, D., Reitner, J., Collins, A. G. & Wörheide, G. 2008. Phylogeny and evolution of glass sponges (Porifera, Hexactinellida). *Systematic Biology,* 57, 388-405.

Duboule, D. & Dollé, P. 1989. The structural and functional organization of the murine HOX gene family resembles that of Drosophila homeotic genes. *EMBO J,* 8, 1497-1505.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sorensen, M. V., Haddock, S. H. D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q., And & Giribet, G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature,* 452, 745-749.

Durkin, K., Coppieters, W., Drogemuller, C., Ahariz, N., Cambisano, N., Druet, T., Fasquelle, C., Haile, A., Horin, P., Huang, L., Kamatani, Y., Karim, L., Lathrop, M., Moser, S., Oldenbroek, K., Rieder, S., Sartelet, A., Solkner, J., Stalhammar, H., Zelenika, D., Zhang, Z., Leeb, T., Georges, M. & Charlier, C. 2012. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature,* 482, 81-4.

Echelle, A. & Mosier, D. 1981. All-female fish: a criptic species of *Menidia* (Atherinidae). *Science,* 212, 1411-1413.

Erwin, P. M. & Thacker, R. W. 2007. Phylogenetic analyses of marine sponges within the order Verongida: A comparison of morphological and molecular data. *Invertebrate Biology,* 126, 220-234.

Ewing, R. S., Cg; Evenson, Dp 1991. Flow cytometric identification of larval triploid walleyes. *Progressive Fish Culturist,* 53, 177-180.

Ezawa, K., Ikeo, K., Gojobori, T. & Saitou, N. 2011. Evolutionary patterns of recently emerged animal duplogs. *Genome Biol Evol,* 3, 1119-35.

Ferrier, D. E. K. 2008. When is a Hox gene not a Hox gene? The importance of nomenclature. *In:* MINELLI, A. & FUSCO, G. (eds.) *Evolving pathways: key themes in evolutionary developmental biology.* Cambridge (MA): Cambridge University Press.

Ferrier, D. E. K. 2010. Evolution of Hox Complexes. *In:* DEUTSCH, J. S. (ed.) *Hox Genes: Studies from the 20th to the 21st Century.* Landes Bioscience and Springer Science + Business Media.

Ferrier, D. E. K. & Akam, M. 1996. Organization of the Hox gene cluster in the grasshopper, Schistocerca gregaria. *Proceedings of the National Academy of Sciences,* 93, 13024-13029.

Ferrier, D. E. K., Dewar, K., Cook, A., Chang, J. L., Hill-Force, A. & Amemiya, C. 2005. The chordate ParaHox cluster. *Current Biology,* 15, R820-R822.

Ferrier, D. E. K. & Holland, P. W. H. 2001. Ancient Origin of the Hox Gene Cluster. *Nature Reviews,* 2, 34-38.

Ferrier, D. E. K. & Minguillon, C. 2003. Evolution of the Hox/ParaHox gene clusters. *Int J Dev Biol,* 47, 605-611.

Ferrier, D. E. K., Minguillón, C., Holland, P. W. H. & Garcia-Fernàndez, J. 2000. The amphioxus Hox cluster: deuterostome posterior flexibility and Hox14. *Evolution & Development,* 2, 284-293.

Feuda, R., Hamilton, S. C., Mcinerney, J. O. & Pisani, D. 2012. Metazoan opsin evolution reveals a simple route to animal vision. *Proceedings of the National Academy of Sciences.*

Field, K. G., Olsen, G. J., Lane, D. J., Giovannoni, S. J., Ghiselin, M. T., Raff, E. C., Pace, N. R. & Raff, R. A. 1988. Molecular phylogeny of the animal kingdom. *Science,* 239, 748-53.

Finnerty, J. R. & Martindale, M. Q. 1997. Homeoboxes in Sea Anemones (Cnidaria; Anthozoa): A PCR-Based Survey of Nematostella vectensis and Metridium senile. *The Biological Bulletin,* 193, 62-76.

Finnerty, J. R. & Martindale, M. Q. 1999. Ancient origins of axial patterning genes: Hox genes and ParaHox genes in the Cnidaria. *Evolution & Development,* 1, 16-23.

Fiston-Lavier, A. S., Anxolabehere, D. & Quesneville, H. 2007. A model of segmental duplication formation in *Drosophila melanogaster. Genome Res,* 17, 1458-70.

Fitch, W. 1970. Distinguishing homologues from analogous proteins. *Systematic Zoology,* 19, 99-113.

Fonseca, N., Vieira, C., Holland, P. & Vieira, J. 2008. Protein evolution of ANTP and PRD homeobox genes. *BMC Evol Biol,* 8, 200.

Fontana, F., Congiu, L., Mudrak, V. A., Quattro, J. M., Smith, T. I., Ware, K. & Doroshov, S. I. 2008. Evidence of hexaploid karyotype in shortnose sturgeon. *Genome,* 51, 113-9.

Frank, S., Schulthess, T., Landwehr, R., Lustig, A., Mini, T., Jeno, P., Engel, J., And & Kammerer, R. A. 2002. Characterization of the Matrilin Coiled-coil Domains Reveals Seven Novel Isoforms. *The Journal of Biological Chemistry,* 277, 19071-19079.

Fredriksson, R., Lagerström, M. C., Lundin, L. & Schiöth, H. B. 2003. The G-Protein-Coupled Receptors in the Human Genome Form Five Main

Families. Phylogenetic Analysis, Paralogon Groups, and Fingerprints. *Molecular Pharmacology,* 63, 1256-1272.

Freeman, R., Ikuta, T., Wu, M., Koyanagi, R., Kawashima, T., Tagawa, K., Humphreys, T., Fang, G.-C., Fujiyama, A., Saiga, H., Lowe, C., Worley, K., Jenkins, J., Schmutz, J., Kirschner, M., Rokhsar, D., Satoh, N. & Gerhart, J. 2012. Identical Genomic Organization of Two Hemichordate Hox Clusters. *Current biology : CB,* 22, 2053-2058.

Fröbius, A. C., Matus, D. Q. & Seaver, E. C. 2008. Genomic Organization and Expression Demonstrate Spatial and Temporal <italic>Hox</italic> Gene Colinearity in the Lophotrochozoan <italic>Capitella</italic> sp. I. *PLoS One,* 3, e4004.

Fujimura, K., Conte, M. & Kocher, T. 2011. Circular DNA intermediate in the duplication of Nile Tilapia *vasa* genes. *PLoS One,* 6, e29477.

Furlong, R. & Holland, P. 2004. Polyploidy in vertebrate ancestry: Ohno and beyond. *Biological Journal of the Linnean Society,* 82, 425-430.

Furlong, R. F. & Holland, P. W. 2002. Were vertebrates octoploid? *Philos Trans R Soc Lond B Biol Sci,* 357, 531-44.

Gallardo, M., Bickham, J. W., Honeycutt, R., Ojeda, R. & Köhler, N. 1999. Discovery of tetraploidy in a mammal. *Nature,* 401, 341.

Garcia-Fernandez, J. 2005. The genesis and evolution of homeobox gene clusters. *Nat Rev Genet,* 6, 881-892.

Garcia-Fernàndez, J. & Holland, P. W. H. 1994. Archetypal organization of the amphioxus Hox gene cluster. *Nature,* 370, 563-566.

Gauchat, D., Mazet, F., Berney, C., Schummer, M., Kreger, S., Pawlowski, J. & Galliot, B. 2000. Evolution of Antp-class genes and differential expression of Hydra Hox/paraHox genes in anterior patterning. *Proceedings of the National Academy of Sciences,* 97, 4493-4498.

Gellon, G. & Mcginnis, W. 1998. Shaping animal body plans in development and evolution by modulation of Hox expression patterns. *BioEssays,* 20, 116-125.

Gillis, W., Bowerman, B. & Schneider, S. 2008. The evolution of protostome GATA factors: Molecular phylogenetics, synteny, and intron/exon structure reveal orthologous relationships. *BMC Evol Biol,* 8, 112.

Giribet, G. 2003. Molecules, development and fossils in the study of metazoan evolution; Articulata versus Ecdysozoa revisited. *Zoology,* 106, 303-326.

Giribet, G. & Edgecombe, G. D. 2012. Reevaluating the Arthropod Tree of Life. *Annual Reviews Entomology,* 57, 167-186.

Giribet, G., Edgecombe, G. D. & Wheeler, W. C. 2001. Arthropod phylogeny based on eight molecular loci and morphology. *Nature,* 413, 157-161.

Giribet, G. & Ribera, C. 2000. A Review of Arthropod Phylogeny: New Data Based on Ribosomal DNA Sequences and Direct Character Optimization. *Cladistics,* 16, 204-231.

Giribet, G., Richter, S., Edgecombe, G. D. & Wheeler, W. C. 2005. The position of crustaceans within the Arthropoda - evidence from nine molecular loci and morphology. *In:* KOENEMANN, S. & JENNER, R. A. (eds.) *Crustacean Issues 16: Crustacea and Arthropod Relationships. Festschrift for Frederick R. Schram.* Boca Raton: Taylor & Francis.

Graham, A., Papalopulu, N. & Krumlauf, R. 1989. The murine and Drosophila homeobox gene complexes have common features of organization and expression. *Cell,* 57**,** 367-378.

Gui, J. L., Y; Li, K; Hong, Y; Zhou, T; 1985. Studies on the karyotypes of Chinese cyprinid dishes 6: karyotypes of three tetraploid species in Barbinae and one tetraploid species in Cyprininae. *Acta Genetica Sinica,* 12**,** 202-208.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology,* 59**,** 307-321.

Haase, A., Stern, M., Wächtler, K. & Bicker, G. 2001. A tissue-specific marker of Ecdysozoa. *Development Genes and Evolution,* 211**,** 428-433.

Halanych, K. M. 2004. THE NEW VIEW OF ANIMAL PHYLOGENY. *Annual Review of Ecology, Evolution, and Systematics,* 35**,** 229-256.

Halanych, K. M., Bacheller, J. D., Aguinaldo, A. M., Liva, S. M., Hillis, D. M., And & Lake, J. A. 1995. Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. *Science,* 267**,** 1641-3.

Handford, P. A., Downing, A. K., Reinhardt, D. P. & Sakai, L. Y. 2000. Fibrillin: from domain structure to supramolecular assembly. *Matrix Biology,* 19**,** 457-470.

Hatschek, B. 1888. *Lebrbuch der Zologie,* Jena.

Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G. W., Edgecombe, G. D., Martinez, P., Baguñà, J., Bailly, X., Jondelius, U., Wiens, M., Müller, W. E. G., Seaver, E., Wheeler, W. C., Martindale, M. Q., Giribet, G. & Dunn, C. W. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings of the Royal Society B: Biological Sciences.*

Hermetz, K. E., Surti, U., Cody, J. D. & Rudd, M. K. 2012. A recurrent translocation is mediated by homologous recombination between HERV-H elements. *Mol Cytogenet,* 5**,** 6.

Holland, L. Z., Albalat, R., Azumi, K., Benito-Gutierrez, E., Blow, M. J., Bronner-Fraser, M., Brunet, F., Butts, T., Candiani, S., Dishaw, L. J., Ferrier, D. E., Garcia-Fernandez, J., Gibson-Brown, J. J., Gissi, C., Godzik, A., Hallbook, F., Hirose, D., Hosomichi, K., Ikuta, T., Inoko, H., Kasahara, M., Kasamatsu, J., Kawashima, T., Kimura, A., Kobayashi, M., Kozmik, Z., Kubokawa, K., Laudet, V., Litman, G. W., Mchardy, A. C., Meulemans, D., Nonaka, M., Olinski, R. P., Pancer, Z., Pennacchio, L. A., Pestarino, M., Rast, J. P., Rigoutsos, I., Robinson-Rechavi, M., Roch, G., Saiga, H., Sasakura, Y., Satake, M., Satou, Y., Schubert, M., Sherwood, N., Shiina, T., Takatori, N., Tello, J., Vopalensky, P., Wada, S., Xu, A., Ye, Y., Yoshida, K., Yoshizaki, F., Yu, J. K., Zhang, Q., Zmasek, C. M., De Jong, P. J., Osoegawa, K., Putnam, N. H., Rokhsar, D. S., Satoh, N. & Holland, P. W. 2008. The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res,* 18, 1100-11.

Holland, P. W. H., Booth, H.A.F., and Bruford, E.A., 2007. Classification and nomenclature of all human homeobox genes. *BMC Biology,* 5.

Howard-Ashby, M., Materna, S. C., Brown, C. T., Chen, L., Cameron, R. A. & Davidson, E. H. 2006. Identification and characterization of homeobox transcription factor genes in Strongylocentrotus purpuratus, and their expression in embryonic development. *Developmental Biology,* 300, 74-89.

Huelsenbeck, J. P. & Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics,* 17, 754-755.

Hughes, A. 1999. Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *Journal of Molecular Evolution,* 48, 565-576.

Hughes, A. L. & Friedman, R. 2004. Differential loss of ancestral gene families as a source of genomic divergence in animals. *Proc Biol Sci,* 271 Suppl 3, S107-9.

Hui, J., Raible, F., Korchagina, N., Dray, N., Samain, S., Magdelenat, G., Jubin, C., Segurens, B., Balavoine, G., Arendt, D. & Ferrier, D. E. K. 2009. Features of the ancestral bilaterian inferred from Platynereis dumerilii ParaHox genes. *BMC Biology,* 7, 43.

Hui, J. H. L., Holland, P. W. H. & Ferrier, D. E. K. 2008. Do cnidarians have a ParaHox cluster? Analysis of synteny around a Nematostella homeobox gene cluster. *Evolution and Development,* 10, 725-730.

Hui, J. H. L., Mcdougall, C., Monteiro, A. S., Holland, P. W. H., Arendt, D., Balavoine, G., And & Ferrier, D. E. K. 2012. Extensive Chordate and Annelid Macrosynteny Reveals Ancestral Homeobox Gene Organization. *Molecular Biology and Evolution,* 29, 157-165.

Hyman, L. H. 1940. The Invertebrates. New York: McGraw-Hill.

Ikuta, T., Yoshida, N., Satoh, N. & Saiga, H. 2004. Ciona intestinalis Hox gene cluster: Its dispersed structure and residual colinear expression in development. *Proc Natl Acad Sci U S A,* 101, 15118-15123.

Irimia, M., Maeso, I. & Garcia-Fernàndez, J. 2008. Convergent Evolution of Clustering of Iroquois Homeobox Genes across Metazoans. *Mol Biol Evol,* 25, 1521-1525.

Irimia, M., Maeso, I., Penny, D., Garcia-Fernàndez, J. & Roy, S. W. 2007. Rare Coding Sequence Changes are Consistent with Ecdysozoa, not Coelomata. *Mol Biol Evol,* 24, 1604-1607.

Irimia, M., Tena, J. J., Alexis, M. S., Fernandez-Miñan, A., Maeso, I., Bogdanović, O., De La Calle-Mustienes, E., Roy, S. W., Gómez-Skarmeta, J. & Fraser, H. B. 2012. Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Res,* 22, 2356-2367.

Itoh, N. & Ornitz, D. M. 2004. Evolution of the Fgf and Fgfr gene families. *TRENDS in Genetics,* 20, 563-569.

Jagla, K., Bellard, M. & Frasch, M. 2001. A cluster of Drosophila homeobox genes involved in mesoderm differentiation programs. *BioEssays,* 23, 125-133.

Jakob, W., Sagasser, S., Dellaporta, S., Holland, P.W.H,, Kuhn, K., and Schierwater, B., 2004. The Trox-2 Hox/ParaHox gene of Trichoplax (Placozoa) marks an epithelial boundary. *Dev Genes Evol,* 214, 170-175.

Janko, K., Bohlen, J., Lamatsch, D., Flajshans, M., Epplen, J. T., Rab, P., Kotlik, P. & Slechtova, V. 2007. The gynogenetic reproduction of diploid and triploid hybrid spined loaches (Cobitis: Teleostei), and their ability to establish successful clonal lineages--on the evolution of polyploidy in asexual vertebrates. *Genetica,* 131, 185-94.

Johnson, A. D. & Herskowitz, I. 1985. A repressor (MATα2 product) and its operator control expression of a set of cell type specific genes in yeast. *Cell,* 42, 237-247.

Joshi, A. K., Zhang, L., Rangan, V. S. & Smith, S. 2003. Cloning, Expression, and Characterization of a Human 4'- Phosphopantetheinyl Transferase with Broad Substrate Specificity. *278,* 35.

Kamm, K., Schierwater, B., Jakob, W., Dellaporta, S.L., and Miller, D.J., 2006. Axial Patterning and Diversification in the Cnidaria Predate the Hox System. *Current Biology,* 16, 920-926.

Katju, V. & Lynch, M. 2003. The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. *Genetica,* 165, 1793-1803.

Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research,* 30, 3059-3066.

Kavanagh, K. L., Jörnvall, H., Persson, B. & Oppermann, U. 2008. The SDR superfamily: functional and structural diversity within a family of metabolic and regulatory enzymes. *Cellular and Molecular Life Sciences,* 65, 3895-3906.

Keane, T., Creevey, C., Pentony, M., Naughton, T. & Mclnerney, J. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol,* 6, 29.

Kerner, P., Ikmi, A., Coen, D. & Vervoort, M. 2009. Evolutionary history of the iroquois/Irx genes in metazoans. *BMC Evol Biol,* 9, 74.

Kielty, C. M., Baldock, C., Lee, D., Rock, M. J., Ashworth, J. L., And & Shuttleworth, C. A. 2002. Fibrillin: from microfibril assembly to biomechanical function. *Philos Trans R Soc Lond B Biol Sci*, 207-217.

Kim, Y. & Nirenberg, M. 1989. Drosophila NK-homeobox genes. *Proceedings of the National Academy of Sciences,* 86, 7716-7720.

King, N., Westbrook, M. J., Young, S. L., Kuo, A., Abedin, M., Chapman, J., Fairclough, S., Hellsten, U., Isogai, Y., Letunic, I., Marr, M., Pincus, D., Putnam, N., Rokas, A., Wright, K. J., Zuzow, R., Dirks, W., Good, M., Goodstein, D., Lemons, D., Li, W., Lyons, J. B., Morris, A., Nichols, S., Richter, D. J., Salamov, A., Sequencing Jgi, Bork, P., Lim, W. A., Manning, G., Miller, W. T., Mcginnis, W., Shapiro, H., Tjian, R., Grigoriev, I. V. & Rokhsar, D. 2008. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature,* 451**,** 783-788.

Kleiger, G. & Eisenberg, D. 2002. GXXXG and GXXXA Motifs Stabilize FAD and NAD(P)-binding Rossmann Folds Through Ca – H· · ·O Hydrogen Bonds and van der Waals Interactions. *Journal Molecular Biology,* 323**,** 69-76.

Kmita, M. & Duboule, D. 2003. Organizing Axes in Time and Space; 25 Years of Colinear Tinkering. *Science,* 301**,** 331-333.

Koonin, E. 2005. Orthologs, paralogs, and evolutionary genomics. *Annual Reviews Genetics,* 39**,** 309-338.

Kuhn, K., Streit, B. & Schierwater, B. 1996. Homeobox Genes in the CnidarianEleutheria dichotoma:Evolutionary Implications for the Origin ofAntennapedia-Class (HOM/Hox) Genes. *Mol Phylogenet Evol,* 6**,** 30-38.

Lander, E. S. 2011. Initial impact of the sequencing of the human genome. *Nature,* 470**,** 187-97.

Lanfear, R. & Bromham, L. 2008. Statistical Tests between Competing Hypotheses of Hox Cluster Evolution. *Systematic Biology,* 557**,** 708-718.

Larroux, C., Fahey, B., Degnan, S. M., Adamski, M., Roksar, D. S. & Degnan, B. M. 2007. The NK homeobox gene cluster predates the origin of Hox genes *Current Biology,* 17**,** 706-710.

Larroux, C., Fahey, B., Degnan, S.M., Adamski, M., Roksar, D.S., and Degnan, B.M., 2007. The NK homeobox gene cluster predates the origin of Hox genes *Current Biology,* 17**,** 706-710.

Larroux, C., Luke, G. N., Koopman, P., Rokhsar, D. S., Shimeld, S. M. & Degnan, B. M. 2008. Genesis and Expansion of Metazoan Transcription Factor Gene Classes. *Mol Biol Evol,* 25**,** 980-996.

Lartillot, N. & Philippe, H. 2004. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Mol Biol Evol,* 21**,** 1095-1109.

Lartillot, N. & Philippe, H. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 363**,** 1463-1472.

Laubichler, M. D. & Maienschein, J. 2007. *From Embryology to Evo-Devo: A History of Developmental Evolution,* Cambridge, MA, MIT Press.

Le Comber, S. & Smith, C. 2004. Polyploidy in fishes: patterns and processes. *Biological Journal of the Linnean Society,* 82**,** 431-442.

Lemmon, M. A., And Ferguson, K.M., 2001. Molecular determinants in pleckstrin homology domains that allow specific recognition of phosphoinositides. *Biochem. Soc. Trans.,* 29, 377-384.

Lemons, D. & Mcginnis, W. 2006. Genomic Evolution of Hox Gene Clusters. *Science,* 313, 1918-1922.

Lewis, E. B. 1978. A gene complex controlling segmentation in *Drosophila. Nature,* 276, 565-578.

Lewis, S., Searle, S., Harris, N., Gibson, M., Iyer, V., Richter, J., Wiel, C., Bayraktaroglu, L., Birney, E., Crosby, M., Kaminker, J., Matthews, B., Prochnik, S., Smith, C., Tupy, J., Rubin, G., Misra, S., Mungall, C. & Clamp, M. 2002. Apollo: a sequence annotation editor. *Genome Biol,* 3, research0082.1 - 0082.14.

Liu, G. E., Ventura, M., Cellamare, A., Chen, L., Cheng, Z., Zhu, B., Li, C., Song, J. & Eichler, E. E. 2009. Analysis of recent segmental duplications in the bovine genome. *BMC Genomics,* 10, 571.

Liu, P., Erez, A., Nagamani, S. C., Dhar, S. U., Kolodziejska, K. E., Dharmadhikari, A. V., Cooper, M. L., Wiszniewska, J., Zhang, F., Withers, M. A., Bacino, C. A., Campos-Acevedo, L. D., Delgado, M. R., Freedenberg, D., Garnica, A., Grebe, T. A., Hernandez-Almaguer, D., Immken, L., Lalani, S. R., Mclean, S. D., Northrup, H., Scaglia, F., Strathearn, L., Trapane, P., Kang, S. H., Patel, A., Cheung, S. W., Hastings, P. J., Stankiewicz, P., Lupski, J. R. & Bi, W. 2011. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell,* 146, 889-903.

Lorenzen, M. D., Gnirke, A., Margolis, J., Garnes, J., Campbell, M., Stuart, J. J., Aggarwal, R., Richards, S., Park, Y. & Beeman, R. W. 2008. The maternal-effect, selfish genetic element Medea is associated with a composite Tc1 transposon. *Proc Natl Acad Sci U S A,* 105, 10085-9.

Luke, G. N., Castro, L. F. C., Mclay, K., Bird, C., Coulson, A. & Holland, P. W. H. 2003. Dispersal of NK homeobox gene clusters in amphioxus and humans. *Proceedings of the National Academy of Sciences,* 100, 5292-5295.

Lv, J., Havlak, P. & Putnam, N. H. 2011. Constraints on genes shape long-term conservation of macro-synteny in metazoan genomes. *BMC Bioinformatics,* 12 Suppl 9, S11.

Lynch, M. & Conery, J. 2000. The evolutionary fate and consequences of duplicate genes. *Science,* 290, 1151-1155.

Mable, B. K. 2004. Why polyploidy is rarer in animals that in plants: myths and mechanisms. *Biological Journal of the Linnean Society,* 82, 453-466.

Makino, T. & Mclysaght, A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A,* 107, 9270-4.

Mallatt, J. & Winchell, C. J. 2002. Testing the New Animal Phylogeny: First Use of Combined Large-Subunit and Small-Subunit rRNA Gene Sequences to Classify the Protostomes. *Mol Biol Evol,* 19, 289-301.

Mallatt, J. M., Garey, J. R. & Shultz, J. W. 2004. Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. *Mol Phylogenet Evol,* 31, 178-191.

Manly, B. F. J. 1991. *Randomization and Monte Carlo Methods in Biology*, Chapman and Hall.

Martin, J. L. 1995. Thioredoxin - a fold for all reasons. *Structure,* 3, 245-250.

Mazet, F., Amemiya, C. T. & Shimeld, S. M. 2006. An ancient Fox gene cluster in bilaterian animals. *Current Biology,* 16, R314-R316.

Mazik, E. T., At; Rab, P 1989. Karyotype study of four species of the genus *Diptychus* (Pisces, Cyprinidae) with remarks on polyploidy of Scizothoracine fishes. *Folia Zoologica,* 38, 325-332.

Mazza, M., Pang, K., Reitzel, A., Martindale, M. & Finnerty, J. R. 2010. A conserved cluster of three PRD-class homeobox genes (homeobrain, rx and orthopedia) in the Cnidaria and Protostomia. *Evodevo,* 1, 3.

Mcginnis, W., Garber, R. L., Wirz, J., Kuroiwa, A. & Gehring, W. J. 1984. A homologous protein-coding sequence in drosophila homeotic genes and its conservation in other metazoans. *Cell,* 37, 403-408.

Meisel, R. P. 2009a. Evolutionary Dynamics of Recently Duplicated Genes: Selective Constraints on Diverging Paralogs in the Drosophila pseudoobscura Genome. *J Mol Evol,* 69, 81-93.

Meisel, R. P. 2009b. Repeat mediated gene duplication in the Drosophila pseudoobscura genome. *Gene,* 438, 1-7.

Mendivil Ramos, O., Barker, D. & Ferrier, D. E. K. 2012. Ghost Loci Imply Hox and ParaHox Existence in the Last Common Ancestor of Animals. *Current Biology,* 22, 1951-1956.

Mendivil Ramos, O. & Ferrier, D. E. K. 2012. Mechanisms of Gene Duplication and Translocation and Progress towards Understanding Their Relative Contributions to Animal Genome Evolution. *International Journal of Evolutionary Biology,* 2012.

Mendivil-Ramos, O., Barker, D. & Ferrier, D. E. K. 2012. Ghost Loci Imply Hox and ParaHox Existence in the Last Common Ancestor of Animals. *Current Biology,* 22, 1951-1956.

Miller, D. J., Hemmrich, G., Ball, E. E., Hayward, D. C., Khalturin, K., Funayama, N., Agata, K., And & Bosch, T. C. 2007. The innate immune repertoire in Cnidaria - ancestral complexity and stochastic gene loss. *Genome Biology,* 8.

Mindell, J. A. & Maduke, M. 2001. ClC chloride channels. *Genome Biology,* 2, 3003.1-3003.6.

Mindnich, R., Möller, G. & Adamski, J. 2004. The role of 17 beta-hydroxysteroid dehydrogenases. *Molecular and Cell Endocrinology,* 218, 7-20.

Minguillón, C. & Garcia-Fernàndez, J. 2003. Genesis and evolution of the Evx and Mox genes and the extended Hox and ParaHox gene clusters. *Genome Biology,* 4, R12.1-R12.8.

Mochizuki, R., Ishizuka, Y., Yanai, K., Koga, Y., And & Fukamizu, A. 1999. Molecular cloning and expression of human neurochondrin-1 and -2. *Biochem. Biophys. Acta,* 1446, 397-402.

Monteiro, A. S. & Ferrier, D. E. 2006. Hox genes are not always Colinear. *International journal of biological sciences,* 2, 95-103.

Monteiro, A. S., Schierwater, B., Dellaporta, S. L. & Holland, P. W. H. 2006. A low diversity of ANTP class homeobox genes in Placozoa. *Evolution & Development,* 8, 174-182.

Morey, C., Da Silva, N. R., Perry, P. & Bickmore, W. A. 2007. Nuclear reorganisation and chromatin decondensation are conserved, but distinct, mechanisms linked to Hox gene activation. *Development,* 134, 909-919.

Morin, R. D., Chang, E., Petrescu, A., Liao, N., Griffith, M., Chow, W., Kirkpatrick, R., Butterfield, Y. S., Young, A. C., Stott, J., Barber, S., Babakaiff, R., Dickson, M. C., Matsuo, C., Wong, D., Yang, G. S., Smailus, D. E., Wetherby, K. D., Kwong, P. N., Grimwood, J., Brinkley, C. P., 3rd, Brown-John, M., Reddix-Dugue, N. D., Mayo, M., Schmutz, J., Beland, J., Park, M., Gibson, S., Olson, T., Bouffard, G. G., Tsai, M., Featherstone, R., Chand, S., Siddiqui, A. S., Jang, W., Lee, E., Klein, S. L., Blakesley, R. W., Zeeberg, B. R., Narasimhan, S., Weinstein, J. N., Pennacchio, C. P., Myers, R. M., Green, E. D., Wagner, L., Gerhard, D. S., Marra, M. A., Jones, S. J. & Holt, R. A. 2006. Sequencing and analysis of 10,967 full-length cDNA clones from Xenopus laevis and Xenopus tropicalis reveals post-tetraploidization transcriptome remodeling. *Genome Res,* 16, 796-803.

Muller, G. B. 2007. Evo-devo: extending the evolutionary synthesis. *Nat Rev Genet,* 8, 943-949.

Negre, B. & Ruiz, A. 2007. HOM-C evolution in Drosophila: is there a need for Hox gene clustering? *TRENDS in Genetics,* 23, 55-59.

Negre, B. & Simpson, P. 2009. Evolution of the achaete-scute complex in insects: convergent duplication of proneural genes. *TRENDS in Genetics,* 25, 147-152.

Nusse, R. 2001. An ancient cluster of Wnt paralogues. *TRENDS in Genetics,* 17, 443.

Ogilvie, M. C. & Scriven, P. N. 2002. Meiotic outcomes in reciprocal translocation carriers ascertained in 3-day human embryos. *Eur J Hum Genet,* 10, 801-6.

Ohno, S. 1970. *Evolution by Gene Duplication,* New York, Springer-Verlag.

Olivier-Bonet, M., Navarro, J., Carrera, M., Egozcue, J. & Benet, J. 2002. Aneuploid and unbalanced sperm in two tranlocation carriers: evalutation of the genetic risk. *Molecular Human Reproduction,* 8, 958-963.

Osborne, P. W., Benoit, G., Laudet, V., Schubert, M. & Ferrier, D. E. K. 2009. Differential regulation of ParaHox genes by retinoic acid in the invertebrate chordate amphioxus (Branchiostoma floridae). *Developmental Biology,* 327, 252-262.

Ou, Z., Stankiewicz, P., Xia, Z., Breman, A. M., Dawson, B., Wiszniewska, J., Szafranski, P., Cooper, M. L., Rao, M., Shao, L., South, S. T., Coleman, K., Fernhoff, P. M., Deray, M. J., Rosengren, S., Roeder, E. R., Enciso, V. B., Chinault, A. C., Patel, A., Kang, S. H., Shaw, C. A., Lupski, J. R. & Cheung, S. W. 2011. Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. *Genome Res,* 21, 33-46.

Ozelius, L. J., Page, C. E., Klein, C., Hewett, J. W., Mineta, M., Leung, J., Shalish, C., Bressman, S. B., De Leon, D., Brin, M. F., Fahn, S., Corey, D. P., And & Breakefield, X. O. 1999. The TOR1A (DYT1) Gene Family and Its Role in Early Onset Torsion Dystonia. *Genomics,* 62, 377-384.

Pan, D. & Zhang, L. 2007. Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: a novel strategy to estimate gene duplication rates. *Genome Biol,* 8, R158.

Pandey, N. & Lakra, W. 1997. Evidence of female heterogamety, B-chromosome and natural tetraploidy in the Asian catfish, *Clarias batrachus*, used in aquaculture. *Aquaculture,* 149, 1-2.

Pandian, T. K., R 1999. Natural occurrence of monoploids and polyploids in the Inidan catfish, *Heteropneustes fossilis. Current Science,* 76, 1134-1137.

Panfilio, K. A. & Akam, M. 2007. A comparison of Hox3 and Zen protein coding sequences in taxa that span the Hox3/zen divergence. *Development Genes and Evolution,* 217, 323-329.

Papillon, D. & Telford, M. J. 2007. Evolution of Hox3 and ftz in arthropods: insights from the crustacean Daphnia pulex. *Development Genes and Evolution,* 217, 315-322.

Peterson, K. J. & Sperling, E. A. 2007. Poriferan ANTP genes: primitively simple or secondarily reduced? *Evolution and Development,* 9, 405-408.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Worheide, G. & Baurain, D. 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLoS Biol,* 9, e1000602.

Philippe, H., Brinkmann, H., Martinez, P., Riutort, M. & Baguñà, J. 2007. Acoel Flatworms Are Not Platyhelminthes: Evidence from Phylogenomics. *PLoS One,* 2, e717.

Philippe, H., Delsuc, F., Brinkmann, H. & Lartillot, N. 2005a. Phylogenomics. *Annual Review of Ecology, Evolution, and Systematics,* 36, 541-562.

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Quènnec, E., Da Silva, C., Wincker, P., Le Guyader, H., Leys, S., Jackson, D. J., Schreiber, F., Erpenbeck, D., Morgenstern, B., Wörheide, G. & Manuel, M. 2009. Phylogenomics Revives Traditional Views on Deep Animal Relationships. *Current Biology,* 19**,** 706-712.

Philippe, H., Lartillot, N. & Brinkmann, H. 2005b. Multigene Analyses of Bilaterian Animals Corroborate the Monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol Biol Evol,* 22**,** 1246-1253.

Philippe, H. & Telford, M. J. 2006. Large-scale sequencing and the new animal phylogeny. *Trends Ecol Evol,* 21**,** 614-20.

Pick, K. S., Philippe, H., Schreiber, F., Erpenbeck, D., Jackson, D. J., Wrede, P., Wiens, M., Alie, A., Morgenstern, B., Manuel, M., And & Worheide, G. 2010. Improved Phylogenomic Taxon Sampling Noticeably Affects Nonbilaterian Relationships. *Molecular Biology and Evolution,* 27**,** 1983-1987.

Pisani, D., Poling, L., Lyons-Weiler, M. & Hedges, S. 2004. The colonization of land by animals: molecular phylogeny and divergence times among arthropods. *BMC Biology,* 2**,** 1.

Pollard, S. L. & Holland, P. W. H. 2000. Evidence for 14 homeobox gene clusters in human genome ancestry. *Current Biology,* 10**,** 1059-1062.

Putnam, N. H., Butts, T., Ferrier, D. E. K., Furlong, R. F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J. K., Benito-Gutierrez, E. L., Dubchak, I., Garcia-Fernandez, J., Gibson-Brown, J. J., Grigoriev, I. V., Horton, A. C., De Jong, P. J., Jurka, J., Kapitonov, V. V., Kohara, Y., Kuroki, Y., Lindquist, E., Lucas, S., Osoegawa, K., Pennacchio, L. A., Salamov, A. A., Satou, Y., Sauka-Spengler, T., Schmutz, J., Shin, I. T., Toyoda, A., Bronner-Fraser, M., Fujiyama, A., Holland, L. Z., Holland, P. W., Satoh, N. & Rokhsar, D. S. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature,* 453**,** 1064-71.

Putnam, N. H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V. V., Jurka, J., Genikhovich, G., Grigoriev, I. V., Lucas, S. M., Steele, R. E., Finnerty, J. R., Technau, U., Martindale, M. Q. & Rokhsar, D. S. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science,* 317**,** 86-94.

Qian, W. & Zhang, J. 2008. Gene Dosage and Gene Duplicability. *Genetics,* 179**,** 2319-2324.

Quiquand, M., Yanze, N., Schmich, J., Schmid, V., Galliot, B. & Piraino, S. 2009. More constraint on ParaHox than Hox gene families in early metazoan evolution. *Developmental Biology,* 328**,** 173-187.

Raicu, P. T., E 1972. *Misgurnus fossilis*, a tetraploid fish species. *Journal of Heredity,* 63**,** 92-94.

Ranz, J. M., Casals, F. & Ruiz, A. 2001. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila. *Genome Res,* 11, 230-9.

Regier, J. C., Shultz, J. W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J. W. & Cunningham, C. W. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature,* 463, 1079-1083.

Reid, R. G. B. 2007. *Biological Emergences: Evolution by Natural Experiment,* Cambridge, MA, MIT Press.

Reitner, J. & Mehl, D. 1996. Monophyly of Porifera. *Verhandlungen des natrurwisenschaftlichen Vereins Hamburg,* 36, 5-32.

Richter, S., Moller, O. S. & Wirkner, C. S. 2009. Advances in crustancean phylogenetics. *Arthopod Syst. Phylogenet.,* 67, 275-286.

Rishi, K. S. R., S 1998. Karyotype study on six Indian hill-stream fishes. *Chromosome Science,* 2, 9-13.

Rogozin, I. B., Wolf, Y. I., Carmel, L. & Koonin, E. V. 2007. Analysis of Rare Amino Acid Replacements Supports the Coelomata Clade. *Mol Biol Evol,* 24, 2594-2597.

Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G. D., Longhorn, S. J., Peterson, K. J., Pisani, D., Philippe, H. & Telford, M. J. 2011. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proceedings of the Royal Society B: Biological Sciences,* 278, 298-306.

Ruiz-Trillo, I., Paps, J., Loukota, M., Ribera, C., Jondelius, U., Baguñà, J. & Riutort, M. 2002. A phylogenetic analysis of myosin heavy chain type II sequences corroborates that Acoela and Nemertodermatida are basal bilaterians. *Proceedings of the National Academy of Sciences,* 99, 11246-11251.

Ruiz-Trillo, I., Roger, A. J., Burger, G., Gray, M. W. & Lang, B. F. 2008. A Phylogenomic Investigation into the Origin of Metazoa. *Mol Biol Evol,* 25, 664-672.

Rushlow, C., Doyle, H., Hoey, T. & Levine, M. 1987. Molecular characterization of the zerknüllt region of the Antennapedia gene complex in Drosophila. *Genes & Development,* 1, 1268-1279.

Ryan, J. F., Burton, P., Mazza, M., Kwong, G., Mullikin, J. & Finnerty, J. 2006. The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes: evidence from the starlet sea anemone, Nematostella vectensis. *Genome Biol,* 7, R64.

Ryan, J. F., Mazza, M. E., Pang, K., Matus, D. Q., Baxevanis, A. D., Martindale, M. Q., And & Finnerty, J. R. 2007. Pre-Bilaterian Origins of the Hox Cluster and the Hox

Code: Evidence from the Sea Anemone, Nematostella

vectensis. *PLoS ONE,* 2.

Ryan, J. F., Pang, K., Nisc Comparative Sequencing Program, Mullikin, J. C., Martindale, M. Q., And & Baxevanis, A. D. 2010. The homeodomain complement of the ctenophore Mnemiopsis leidyi suggests that Ctenophora and Porifera diverged prior to the ParaHoxozoa. *EvoDevo,* 1.

Schaeper, N. D., Prpic, N. M. & Wimmer, E. A. 2010. A clustered set of three Sp-family genes is ancestral in the Metazoa: evidence from sequence analysis, protein domain structure, developmental expression patterns and chromosomal location. *BMC Evol Biol,* 10, 88.

Schierwater, B., Eitel, M., Jakob, W., Osigus, H.-J., Hadrys, H., Dellaporta, S. L., Kolokotronis, S.-O. & Desalle, R. 2009. Concatenated Analysis Sheds Light on Early Metazoan Evolution and Fuels a Modern "Urmetazoon" Hypothesis. *PLoS Biol,* 7, e1000020.

Schierwater, B., Kamm, K., Srivastava, M., Rokhsar, D. S., Rosengarten, R. D. & Dellaporta, S. L. 2008. The Early ANTP Gene Repertoire: Insights from the Placozoan Genome. *PLoS ONE,* 3, e2457-e2457.

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

Scholtz, G. 2002. The Articulata hypothesis – or what is a segment? *Organisms Diversity & Evolution,* 2, 197-215.

Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. 1998. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences,* 95, 5857-5864.

Schultz, R. 1980. *The role of polyploidy in the evolution of fishes,* New York, Plenum Press.

Seo, H.-C., Edvardsen, R. B., Maeland, A. D., Bjordal, M., Jensen, M. F., Hansen, A., Flaat, M., Weissenbach, J., Lehrach, H., Wincker, P., Reinhardt, R. & Chourrout, D. 2004. Hox cluster disintegration with persistent anteroposterior order of expression in Oikopleura dioica. *Nature,* 431, 67-71.

Seo, H.-C., Kube, M., Edvardsen, R. B., Jensen, M. F., Beck, A., Spriet, E., Gorsky, G., Thompson, E. M., Lehrach, H., Reinhardt, R. & Chourrout, D. 2001. Miniature Genome in the Marine Chordate Oikopleura dioica. *Science,* 294, 2506.

Sharman, A. 1999. Some new terms for duplicated genes. *Seminars in Cell & Developmental Biology* 10, 561-563.

Sharpe, J., Nonchev, S., Gould, A., Whiting, J. & Krumlauf, R. 1998. Selectivity, sharing and competitive interactions in the regulation of Hoxb genes. *EMBO J,* 17, 1788-1798.

She, X., Cheng, Z., Zollner, S., Church, D. M. & Eichler, E. E. 2008. Mouse segmental duplication and copy number variation. *Nat Genet,* 40, 909-14.

Shimizu Y, O. T., Sakaizumi M 1993. Electrophoretic studies of diploid, triploid and tetraploid forms of Japanese silver crucian carp, Carassius auratus. *Japanese Journal of Ichthyology,* 40, 65-75.

Shippy, T. D., Ronshaugen, M., Cande, J., He, J. P., Beeman, R. W., Levine, M., Brown, S. J. & Denell, R. E. 2008. Analysis of the Tribolium homeotic complex: insights into mechanisms constraining insect Hox clusters. *Development Genes and Evolution,* 218, 127-139.

Sicot, F. X., Tsuda, T., Markova, D., Klement, J. F., Arita, M., Zhang, R. Z., Pan, T. C., Mecham, R. P., Birk, D. E., And & Chu, M. L. 2008. Fibulin-2 Is Dispensable for Mouse Development and Elastic Fiber Formation. *Molecular and Cellular Biology,* 28, 1061-1067.

Simakov, O., Marletaz, F., Cho, S.-J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo, D.-H., Larsson, T., Lv, J., Arendt, D., Savage, R., Osoegawa, K., De Jong, P., Grimwood, J., Chapman, J. A., Shapiro, H., Aerts, A., Otillar, R. P., Terry, A. Y., Boore, J. L., Grigoriev, I. V., Lindberg, D. R., Seaver, E. C., Weisblat, D. A., Putnam, N. H. & Rokhsar, D. S. 2013. Insights into bilaterian evolution from three spiralian genomes. *Nature,* 493, 526-531.

Simionato, E., Ledent, V., Richards, G., Thomas-Chollier, M., Kerner, P., Coornaert, D., Degnan, B. & Vervoort, M. 2007. Origin and diversification of the basic helix-loop-helix gene family in metazoans: insights from comparative genomics. *BMC Evol Biol,* 7, 33.

Slack, J. M., Holland, P. W. H. & Graham, C. F. 1993. The zootype and the phylotypic stage. *Nature,* 361, 490-492.

Small, K., Brudno, M., Hill, M. & Sidow, A. 2007. A haplome alignment and reference sequence of the highly polymorphic Ciona savignyi genome. *Genome Biol,* 8, R41.

Sokal, R. R. & Rohlf, F. J. 1995. *Biometry: The Principles and Practices of Statistics in Biological Research,* New York, W.H. Freeman and Co.

Sonnhammer, E. K., Ev 2002. Orthology, paralogy and proposed classification for paralog subtypes. *TRENDS in Genetics,* 18, 619-620.

Srivastava, M., Begovic, E., Chapman, J., Putnam, N. H., Hellsten, U., Kawashima, T., Kuo, A., Mitros, T., Salamov, A., Carpenter, M. L., Signorovitch, A. Y., Moreno, M. A., Kamm, K., Grimwood, J., Schmutz, J., Shapiro, H., Grigoriev, I. V., Buss, L. W., Schierwater, B., Dellaporta, S. L. & Rokhsar, D. S. 2008. The Trichoplax genome and the nature of placozoans. *Nature,* 454, 955-60.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M. E., Mitros, T., Richards, G. S., Conaco, C., Dacre, M., Hellsten, U., Larroux, C., Putnam, N. H., Stanke, M., Adamska, M., Darling, A., Degnan, S. M., Oakley, T. H., Plachetzki, D. C., Zhai, Y., Adamski, M., Calcino, A., Cummins, S. F., Goodstein, D. M., Harris, C., Jackson, D. J., Leys, S. P., Shu, S., Woodcroft, B. J., Vervoort, M., Kosik, K. S., Manning, G., Degnan, B. M. & Rokhsar, D. S. 2010. The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature,* 466, 720-6.

Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N., Ruby, J. G., Brennecke, J., Hodges, E., Hinrichs, A. S., Caspi, A., Paten, B., Park, S.-

W., Han, M. V., Maeder, M. L., Polansky, B. J., Robson, B. E., Aerts, S., Van Helden, J., Hassan, B., Gilbert, D. G., Eastman, D. A., Rice, M., Weir, M., Hahn, M. W., Park, Y., Dewey, C. N., Pachter, L., Kent, W. J., Haussler, D., Lai, E. C., Bartel, D. P., Hannon, G. J., Kaufman, T. C., Eisen, M. B., Clark, A. G., Smith, D., Celniker, S. E., Gelbart, W. M. & Kellis, M. 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature,* 450, 219-232.

Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., Coulson, A., D'eustachio, P., Fitch, D. H. A., Fulton, L. A., Fulton, R. E., Griffiths-Jones, S., Harris, T. W., Hillier, L. W., Kamath, R., Kuwabara, P. E., Mardis, E. R., Marra, M. A., Miner, T. L., Minx, P., Mullikin, J. C., Plumb, R. W., Rogers, J., Schein, J. E., Sohrmann, M., Spieth, J., Stajich, J. E., Wei, C., Willey, D., Wilson, R. K., Durbin, R. & Waterston, R. H. 2003. The Genome Sequence of <italic>Caenorhabditis briggsae:</italic> A Platform for Comparative Genomics. *PLoS Biol,* 1, e45.

Stenzel, N., Fetzer, C. P., Heumann, R. & Erdmann, K. S. 2009. PDZ-domain-directed basolateral targeting of the peripheral membrane protein FRMPD2 in epithelial cells. *Journal of Cell Science,* 122, 3374-3384.

Stephens, P., Greenman, C., Fu, B., Yang, F. & Al, E. 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell,* 144, 27-40.

Sullivan, J. C., Ryan, J. F., Mullikin, J. C. & Finnerty, J. R. 2007. Conserved and novel Wnt clusters in the basal eumetazoan Nematostella vectensis. *Development Genes and Evolution,* 217, 235-239.

Susin, S. A., Lorenzo, H. K., Zamzami, N., Marzo, I., Snow, B. E., Brothers, G. M., Mangion, J., Jacoto, E., Costantini, P., Loef ¯Er, M., Larochette, N., Goodle, D. R., Aebersold, R., Siderovski, D. P., Penninger, J. M. & Kroemer, G. 1999. Molecular characterization of mitocondrial apoptosis-inducing factor. *Nature,* 397, 441-448.

Suzuki , A. T., Y 1981. Karyotype of tetraploid origin in a tropical Asian cyprinid, *Acrossocheilus sumatranus. Japanese Journal of Ichthyology,* 28, 173-176.

Szamlek, J., Cooper, D., Schempp, W., Minich, P., Kohn, M., Hoegel, J., Goidts, V., Hameister, H. & Kehrer-Sawtzki, H. 2006. Polymorphic micro-inversions contribute to genomic variability of humans and chimpazees. *Hum Genet,* 119, 103-112.

Takahashi, T., Mcdougall, C., Troscianko, J., Chen, W. C., Jayaraman-Nagarajan, A., Shimeld, S. M. & Ferrier, D. E. 2009. An EST screen from the annelid Pomatoceros lamarckii reveals patterns of gene loss and gain in animals. *BMC Evol Biol,* 9, 240.

Takatori, N., Butts, T., Candiani, S., Pestarino, M., Ferrier, D. E. K., Saiga, H. & Holland, P. W. H. 2008. Comprehensive survey and classification of homeobox genes in the genome of amphioxus, Branchiostoma floridae. *Development Genes and Evolution,* 218, 579-590.

Tarchini, B. & Duboule, D. 2006. Control of Hoxd Genes' Collinearity during Early Limb Development. *Developmental Cell,* 10, 93-103.

Telford, M. J. 2008. Resolving Animal Phylogeny: A Sledgehammer for a Tough Nut? . *Developmental Cell,* 14, 457-459.

Telford, M. J. 2013. Field et al. Redux. *Evodevo,* 4, 5.

Timpl, R., Sasaki, T., Kostka, G. & Chu, M. 2003. Fibulins: A versatile family of extracellular matrix proteins. *Nature Reviews Molecular Cell Biology,* 4, 479-489.

Van Der Zee, M., Berns, N. & Roth, S. 2005. Distinct Functions of the Tribolium zerknu®llt Genes in Serosa Specification and Dorsal Closure. *Current biology : CB,* 15, 624-636.

Vergilino, R., Belzile, C. & Dufresne, F. 2009. Genome size evolution and polyploidy in the *Daphnia pulex* complex (Cladocera: Daphniidae). *Biological Journal of the Linnean Society,* 97, 68-79.

Vervoort, A. 1980. Tetraploidy in *Protopterus* (Dipnoi). *Experentia,* 36, 294-296.

Vrijenhoek, R., Dawley, R., Cole, C. & Bogart, J. 1989. *A list of known unisexual vertebrates,* New York, The State University of New York.

Wall, D. P., Fraser, H. B. & Hirsh, A. E. 2003. Detecting putative orthologs. *Bioinformatics,* 19, 1710-1711.

Wang, J. T., Li, J. T., Zhang, X. F. & Sun, X. W. 2012. Transcriptome analysis reveals the time of the fourth round of genome duplication in common carp (Cyprinus carpio). *BMC Genomics,* 13, 96.

Whittaker, C. A. & Hynes, R. O. 2002. Distribution and Evolution of von Willebrand/Integrin A Domains: Widely Dispersed Domains with Roles in Cell Adhesion and Elsewhere. *Mol Biol Cell,* 13, 3369-3387.

Wickstead, B., Gull, K. & Richards, T. A. 2010. Patterns of kinesin evolution reveal a complex ancestral eukaryote with a multifunctional cytoskeleton. *BMC Evolutionary Biology,* 10.

Wilson, P. A., Gardner, S. D., Lambie, N. M., Commans, S. A. & Crowther, D. J. 2006. Characterization of the human patatin-like phospholipase family. *Journal of Lipid Research,* 47, 1940-1949.

Wolfe, K. 2000. Robustness-it's not where you think it is. *Nature,* 25, 3-4.

Wolosker, H., Kline, D., Bian, Y., Blackshaw, S., Cameron, A. M., Fralich, T. J., Schnaar, R. L., And & Snyder, S. H. 1998. Molecularly cloned mammalian glucosamine-6- phosphate deaminase localizes to transporting epithelium and lacks oscillin activity. *The FASEB Journal,* 12, 91-99.

Wörheide, G., Dohrmann, M., Erpenbeck, D., Larroux, C., Maldonado, M., Voigt, O., Borchiellini, C. & Lavrov, D. V. 2012. Chapter One - Deep Phylogeny and Evolution of Sponges (Phylum Porifera). *In:* MIKEL A. BECERRO, M. J. U. M. M. & XAVIER, T. (eds.) *Advances in Marine Biology.* Academic Press.

Wotton, K. R., Weierud, F. K., Juárez-Morales, J. L., Alvares, L. E., Dietrich, S. & Lewis, K. E. 2009. Conservation of gene linkage in dispersed vertebrate NK homeobox clusters. *Development Genes and Evolution,* 219, 481-496.

Wyder, S., Kriventseva, E. V., Schroder, R., Kadowaki, T. & Zdobnov, E. M. 2007. Quantification of ortholog losses in insects and vertebrates. *Genome Biol,* 8, R242.

Yang, X., Matern, H. T. & Gallwitz, D. 1998. Specific binding to a novel and essential Golgi membrane protein (Yip1p) functionally links the transport GTPases Ypt1p and Ypt31p. *EMBO J,* 17, 4954-4963.

Yasukochi, Y., Ashakumary, L. A., Wu, C., Yoshido, A., Nohata, J., Mita, K. & Sahara, K. 2004. Organization of the Hox gene cluster of the silkworm, Bombyx mori: a split of the Hox cluster in a non-Drosophila insect. *Development Genes and Evolution,* 214, 606-614.

Yu, X., Zhou, T., Li, K., Li, Y. & Zhou, M. 1987. On the karyosystematics of cyprinid fishes and a summary of fish chromosome studies in China. *Genetica,* 72, 225-236.

Zhang, L., Lu, H. H., Chung, W. Y., Yang, J. & Li, W. H. 2005. Patterns of segmental duplication in the human genome. *Mol Biol Evol,* 22, 135-41.

Zheng, J., Rogozin, I. B., Koonin, E. V. & Przytycka, T. M. 2007. A Rigorous Analysis of the Pattern of Intron Conservation Supports the Coelomata Clade of Animals. *In:* TESLER, G. & DURAND, D. (eds.) *Comparative Genomics.* Springer Berlin Heidelberg.

Zhong, Y., Butts, T. & Holland, P. W. H. 2008. HomeoDB: a database of homeobox gene diversity. *Evolution & Development,* 10, 516-518.

Zhou, H. & Clapham, D. E. 2009. Mammalian MagT1 and TUSC3 are required for cellular magnesium uptake and vertebrate embryonic development. *Proceeding of the National Academy of Sciences,* 106, 15750-15755.

Zhu, L., Wrabl, J. O., Hayashi, A. P., Rose, L. S. & Thomas, P. J. 2008. The Torsin-family AAA+ Protein OOC-5 Contains a Critical Disulfide Adjacent to Sensor-II That Couples Redox State to Nucleotide Binding. *Mol Biol Cell,* 19, 3599-3612.

Zrzavý, J., Hypša, V. & Vlášková, M. 1998. Arthropod phylogeny: taxonomic congruence, total evidence and conditional combination approaches to morphological and molecular data sets. *In:* FORTEY, R. A. & THOMAS, R. H. (eds.) *Arthropod Relationships.* Springer Netherlands.

Appendices

Appendix A

Duplicate genes nomenclature adapted from Mendivil Ramos, O. & Ferrier, D. E. K. 2012. Mechanisms of Gene Duplication and Translocation and Progress towards Understanding Their Relative Contributions to Animal Genome Evolution. International Journal of Evolutionary Biology, 2012, 10.

The terminology used to define the evolutionary relationships between duplicated genes has become increasingly detailed. The precise inference of the evolutionary relationships between duplicated genes is fundamental for most comparative genomic studies, but it can be complicated because duplication is often combined with speciation and subsequent gene loss (Sharman, 1999).

The most widely-used terms for describing evolutionary relationships between genes are homologous, orthologous and paralogous. Fitch (1970) defined homologous genes as those that share a common ancestor. A subset of homologous genes are orthologous, these being the genes separated only by speciation and not by a duplication event (Figure 1.A). Another subset of homologous genes are paralogous, which are those resulting from a duplication event (Figure 1.B). Sharman (1999) defined additional terms to describe the relationships amongst paralogues. Pro- orthology denotes the relationship of a gene to one of the descendants of its orthologue after duplication of that orthologue (Figure 1.C). Conversely, semi-orthology is the relationship of one of a set of duplicated genes to a gene that is orthologous to the ancestor of the whole set (Figure 1.D). Sharman (1999) also proposed the term trans-homology to describe members of the same gene family descendant from an ancestral gene via two independent gene duplication events. A further important term connected with paralogy is the one proposed by Wolfe (2000), who coined the term ohnologue for those paralogues stemming from a whole genome duplication

(Figure 1.F). Two years later, Sonnhammer and Koonin (2002) highlighted that the definition of a paralogous relationship can be related to a speciation event. Thus, they coined the terms inparalogues and outparalogues. Inparalogues are paralogues in a given lineage that all evolved by gene duplications that happened after a speciation event that separated the given lineage from the other lineage under consideration (Figure 1.E). Outparalogues are paralogues in a given lineage that evolved by gene duplications that happened before a speciation event (Figure 1.E). Careful consideration must be taken when using the terms such as inparalogues, outparalogues and ohnologues. The specification of the relation of the duplication event to the speciation event must be included when these terms are used, otherwise evolutionary interpretations and use of terminology can easily be confused. Finally, a new umbrella term, duplogs (Ezawa et al., 2011), has been thrown into the duplication terminology pool to define intraspecies paralogues. This term amalgamates all the types of paralogues within a species, including inparalogues, outparalogues and ohnologues.

Sonnhammer and Koonin (2002) also defined co-orthologues, which are synonymous with Sharman's (1999) definition of trans-homologues, and are inparalogues of one lineage which are homologous to another set of inparalogues in a second lineage. Artefacts stemming from phylogenetic inference, such as lineage-specific gene loss, can mislead the deduction of the evolutionary relationship of genes. For this purpose, Koonin (2005) devised the term pseudoorthologue to accommodate those genes that are essentially paralogues but appear to be orthologues due to differential, lineage-specific gene loss (Figure 1.G). Further useful terms are xenologue and pseudoparalogue. Xenologues are homologues acquired through horizontal gene transfer by one or both species that are being compared, but appearing to be orthologues when pairwise comparison of the genomes is performed (Figure 1.H) (Koonin, 2005). Pseudoparalogues are homologues that through the analysis in a single genome are interpreted as paralogues, however, these homologues originated by a combination of vertical inheritance and horizontal gene transfer (Figure 1.H) (Koonin, 2005).

Recently a new term, toporthology, has been specified, which aims to include another aspect of the concept of orthology, that of positional orthology (Dewey, 2011). Toporthology describes the evolutionary relationship of orthologues that retain their ancestral genomic positions. In the context of gene duplications, a duplication event is said to be 'symmetric' if deletion of either of the copies of the duplicated sequences would return the gene order to the original, ancestral state. Thus, tandem duplicates and whole-chromosome/ genome duplication are symmetrical duplications. A duplication event is 'asymmetric' if deleting only one of the copies could return the gene order to its original, ancestral state. Consequently, dispersed segmental duplications and retrotranspositions are asymmetrical duplications. From these definitions two genes are positionally homologous, topohomologous, if they are homologous and neither gene comes from an asymmetric duplication since the time of their common ancestor. The contrast to this case is atopohomologous. Furthermore, toporthologous genes would be those genes that are topohomologues and orthologues, topoparalogous genes would be those genes that are topohomologues and paralogues, atoporthologues genes would be those genes that are atopohomologues and orthologues and atopoparalogues genes would be those genes that are atopohomologues and paralogues.

The term toporthology and its associated derivations need to be used with extreme caution (Dewey, 2011). The value, and aim, of distinguishing toporthologues/topoparalogues is to distinguish those genes (which are not necessarily one-to-one orthologues) that are most comparable in terms of their evolutionary history. However, being able to distinguish toporthology obviously requires reliable, accurate genome assemblies and hinges on distinguishing parent/source locations from daughter/target locations of duplicated regions. Also, the distinction of toporthology can obviously be complicated by genomic rearrangements that occur after the duplication event and which can obscure whether a duplication was symmetric or asymmetric. Currently, the complications introduced by such post-duplication genomic rearrangements lead to some counterintuitive uses of the terminology. One might assume that toporthology simply refers to orthologues that are both in the ancestral locations, and conversely that atoporthology simply describes the situation in which at least one of the genes is no longer in the ancestral location. Similarly, the prefixes can be used with paralogues, to give topoparalogues and atopoparalogues and might be assumed to simply be used when paralogues are both in the ancestral location or one or other has moved respectively. The use of the terminology is not so straight-forward, however, as can be seen by a close inspection of Figure 2 in (Dewey, 2011), in which YA1 and YA2 are topoparalogues rather than atopoparalogues despite YA2 no longer being in the ancestral location. The classification of YA1 and YA2 as topoparalogues arises because they were not produced by an asymmetric duplication, but then the subsequent change of position of YA2 has obscured this. Consequently the precision of the data (taxonomic sampling and quality of genome assembly) severely compromises the utility of this terminology. Despite the apparent use of the terms to reflect relationships relative to ancestral locations within the genome, in fact the movement of genes to new, non-ancestral locations subsequent to the duplication event is not accommodated. Consequently toporthologues are not necessarily both in the ancestral genomic position. This terminology thus risks being counterintuitive and confusing in its present form.

The above summary of duplicate terminology serves to illustrate two things. Firstly, there is the complexity of the evolutionary processes involved in production of duplicates and the care that must thus be exercised when comparing genes between species. Secondly, there is currently an over-abundance of terminology, some of which is redundant and some of which is counterintuitive. It is to be hoped that with time the terminology will settle on a consensus of selected terms and those that are impractical or potentially misleading will be abandoned.

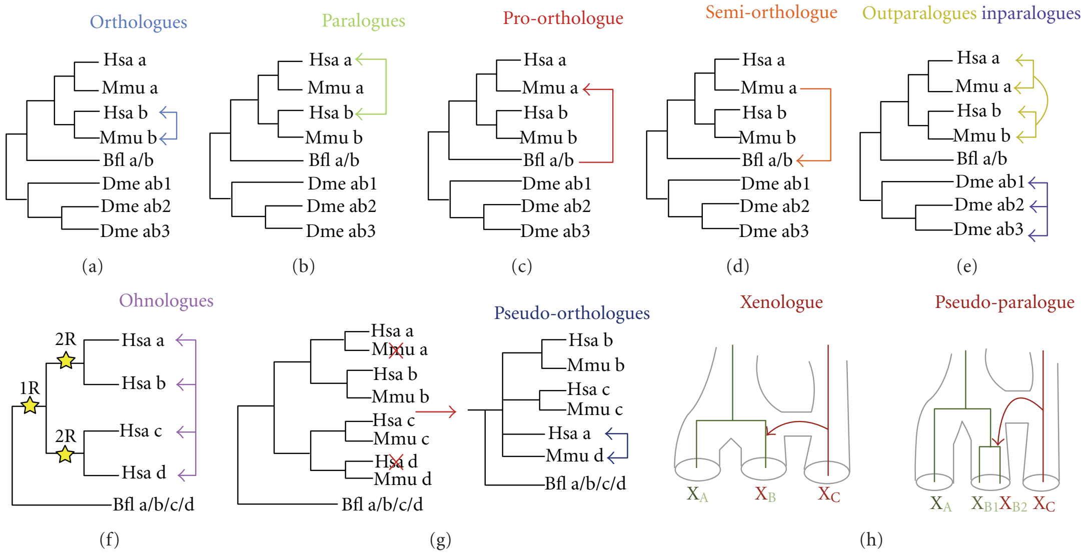


Figure A1.- Overview of the current terminology. *The different panels represent term(s) for duplicated genes. (a) Orthologues. The square blue arrows represent an orthologous relationship between the two genes. (b) Paralogues. The square green arrows represent paralogous relationships between the genes. (c) Proto-orthologue. The square red arrow represents the pro-orthologue relationship of gene a/b from Branchiostoma floridae to gene a from Mus musculus. (d) Semi-orthologue. The square orange arrow represents the semi-orthologous relationship of gene a of Mus musculus to gene a/b from Branchiostoma floridae. (e) Inparalogues and Outparalogues. The square yellow arrows represent the outparalogous relationship in which human and mouse a genes are outparalogous to human and mouse b genes. As a set, genes a and b from mouse and human represents coorthologues. The square purple arrows represent the inparalogous relationship between the genes which duplicated within this lineage. (f) Ohnologues. The square pink arrows delimit all the paralogues coming from WGD and the stars represent the duplication events. (g) Pseudo-orthologues. The square navy arrows represent the pseudo-orthologues. The red Xs represent lineage-specific gene losses. (h) Xenologues and Pseudo-paralogues. Species are represented by subindices A, B, and C, and the Xs represent the orthologous genes with their colouring designating the species of origin. All of the figures are adapted from Sharman* ***(1999)*** *and Koonin* ***(2005)*** *Bfl: Branchiostoma floridae, Dme: Drosophila melanogaster, Hsa: Homo sapiens, and Mmu: Mus musculus.*

Bibliography

Dewey, C. N. 2011. Positional orthology: putting genomic evolutionary relationships into context. *Brief Bioinform,* 12, 401-12.

Ezawa, K., Ikeo, K., Gojobori, T. and Saitou, N. 2011. Evolutionary patterns of recently emerged animal duplogs. *Genome Biol Evol,* 3, 1119-35.

Fitch, W. 1970. Distinguishing homologues from analogous proteins. *Systematic Zoology,* 19, 99-113.

Koonin, E. 2005. Orthologs, paralogs, and evolutionary genomics. *Annual Reviews Genetics,* 39, 309-338.

Sharman, A. 1999. Some new terms for duplicated genes. *Seminars in Cell and Developmental Biology* 10, 561-563.

Sonnhammer, E. K., Ev 2002. Orthology, paralogy and proposed classification for paralog subtypes. *TRENDS in Genetics,* 18, 619-620.

Wolfe, K. 2000. Robustness-it's not where you think it is. *Nature,* 25, 3-4.

Appendix B

B.1 SMART motifs of the orthologues in scaffold 38

This can be found in the CD folder Appendix B>B1>SMART_ACC.xls

B.2 Derivation of numbers for statistical tests

This can be found in the CD folder Appendix B>B2>sts.xls

The numbers used in our test are based on human genome version GRCh37.p2 are derived as follows:

1) The total number of protein-coding genes (pcg) in chromosomes 1 to 23 and X (20447 pcg, Table S4 C12). From this number we subtracted the number of protein-coding genes in the Hox clusters (39 pcg, Table S4, C2) and ParaHox 'clusters' (6 pcg, Table S4, C6), leaving a total number of protein-coding genes without Hox and ParaHox genes (20402 pcg, Table S4, C11).
2) We made the distinction of type of orthologues according to their location in the human genome. Hox loci neighbours (4450 pcg, Table S4, C3) include the total number of protein-coding genes on chromosomes 2 (1275 pcg), 7 (942 pcg), 12 (1055 pcg) and 17 (1217 pcg), excluding the genes from the Hox clusters (39 pcg). ParaHox loci neighbours (2859 pcg, Table S4, C7) include the total number of protein-coding genes on chromosomes 4 (781 pcg), 5 (899 pcg), 13 (333 pcg) and X (852 pcg) excluding the genes from the ParaHox 'clusters' (6 pcg). Non-Hox loci neighbours (15952 pcg, Table S4, C4) are those excluding the Hox loci neighbours (4450 pcg), Hox clusters (39 pcg) and ParaHox 'clusters' (6 pcg) from the total number of protein-coding genes, and non-ParaHox loci neighbours (17543 pcg, Table S4, C8) are those excluding the ParaHox loci neighbours (2859 pcg) from the total number of protein-

coding genes without Hox and ParaHox genes (20402 pcg). Hox/ParaHox loci neighbours (7309 pcg, Table S4, C10) are the sum of Hox loci neighbours (4450 pcg) and ParaHox loci neighbours (2859 pcg), and non-Hox/ParaHox loci neighbours (13093 pcg, Table S4, C9) are those excluding ParaHox (2859 pcg) and Hox (4450 pcg) loci neighbours as well as Hox (39 pcg) and ParaHox (6 pcg) from the total number of protein-coding genes (20447 pcg).

From these numbers we calculated the probabilities of a randomly chosen human gene being a Hox locus neighbour, ParaHox locus neighbour and Non-Hox/ParaHox neighbour. These probabilities are used to perform the Binomial Exact Test (Table S5, Table S6 and Table S7).

The Exact Binomial Test was used to test departure of observed numbers of Hox neighbour orthologues (or ParaHox neighbour orthologues or Hox/ ParaHox neighbour orthologues) on scaffold 38 from those expected on the basis of the probability of Hox neighbours (or ParaHox neighbours or Hox/ParaHox neighbours) in the human genome. We plotted the observed and expected number of genes in scaffold 38 for each one of the tests. For all the plots the expected number of orthologues is calculated by multiplying the total observed number of orthologues (i.e. 27 genes in version 1 and 22 genes in version 2) by the category probabilities from Table S5, Table S6, Table S7. (Fig. S1, Fig. S2 and Fig. S3).

B.3A Contingency tables for Fisher's Exact test version 1

OBSERVED	Hox loci neighbours in human	Non-Hox loci neighbours in human	Row total
human genes with orthology on Scf38	1	26	27
human genes without orthology on Scf38	4449	15926	20375
Column total	4450	15952	20402
EXPECTED	**Hox loci neighbours in human**	**Non-Hox loci neighbours in human**	**Row total**
human genes with orthology on Scf38	5.889128517 (= (27 *4450)/ 20402)	21.11087148 (=(15952*27)/ 20402)	27
human genes without orthology on Scf38	4444.110871 (= (20375*4450)/ 20402)	15930.88913 (=(20375*15952/20402))	20375
Column total	4450	15952	20402

Table B1 .- Fisher's Exact Test Hox contingency table (version 1)

OBSERVED	ParaHox loci neighbours in human	Non-ParaHox loci neighbours in human	Row total
human genes with orthology on Scf38	12	15	27
human genes without orthology on Scf38	2847	17528	20375
Column total	2859	17543	20402
EXPECTED	**ParaHox loci neighbours in human**	**Non-ParaHox loci neighbours in human**	**Row total**
human genes with orthology on Scf38	3.783599647 (=(27*2859)/20402)	23.21640035 (=(27*17543)/20402)	27
human genes without orthology on Scf38	2855.2164 (=(20375*2859)/20402)	17519.7836 (=(20375*17543)/20402)	20375
Column total	2859	17543	20402

Table B2 .- Fisher's Exact Test ParaHox contingency table (version 1)

OBSERVED	**Hox/ParaHox loci neighbours in human**	**Non-Hox/ParaHox loci neighbours in human**	**Row total**
human genes with orthology on Scf38	13	14	27
human genes without orthology on Scf38	7296	13079	20375
Column total	7309	13093	20402
EXPECTED	**Hox/ParaHox loci neighbours in human**	**Non-Hox/ParaHox loci neighbours in human**	**Row total**
human genes with orthology on Scf38	9.672728164 (=(27*7309)/20402)	17.32727184 (=(27*13093)/20402)	27
human genes without orthology on Scf38	7299.327272 (=(20375*7309)/20402)	13075.67273 (=(20375*13093)/20402)	20375
Column total	7309	13093	20402

Table B3 .- Fisher's Exact Test Hox/ParaHox contingency table (version 1)

B.3B Contingency tables for Fisher's Exact test version 2

OBSERVED	**Hox loci neighbours in human**	**Non-Hox loci neighbours in human**	**Row total**
human genes with orthology on Scf38	1	21	22
human genes without orthology on Scf38	4449	15931	20380
Column total	4450	15952	20402
EXPECTED	**Hox loci neighbours in human**	**Non-Hox loci neighbours in human**	**Row total**
human genes with orthology on Scf38	4.798549162 (= (22*4450)/20402)	17.20145084 (= (22*15952)/20402)	22
human genes without orthology on Scf38	4445.201451 (= (20380*4450)/20402)	15934.79855 (=(20380*15952)/20402)	20380
Column total	4450	15952	20402

Table B4.- Fisher's Exact Test Hox contingency table (version 2)

OBSERVED	ParaHox loci neighbours in human	Non-ParaHox loci neighbours in human	Row total
human genes with orthology on Scf38	12	10	22
human genes without orthology on Scf38	2847	17533	20380
Column total	2859	17543	20402
EXPECTED	ParaHox loci neighbours in human	Non-ParaHox loci neighbours in human	Row total
human genes with orthology on Scf38	3.082933046 (=(22*2859)/20402)	18.91706695 (=(22*17543)/20402)	22
human genes without orthology on Scf38	2855.917067 (=(20380*2859)/20402)	17524.08293 (=(20380*17543)/20402)	20380
Column total	2859	17543	20402

Table B5.- Fisher's Exact Test ParaHox contingency table (version 2)

OBSERVED	Hox/ParaHox loci neighbours in human	Non-Hox/ParaHox loci neighbours in human	Row total
human genes with orthology on Scf38	13	9	22
human genes without orthology on Scf38	7296	13084	20380
Column total	7309	13093	20402
EXPECTED	Hox/ParaHox loci neighbour in human	Non-Hox/ParaHox loci neighbours in human	Row total
human genes with orthology on Scf38	7.881482208 (=(22*7309)/20402)	14.11851779 (=(22*13093)/20402)	7.881482208
human genes without orthology on Scf38	7301.118518 (=(20380*7309)/20402)	13078.88148 (=(20380*13093)/20402)	7301.118518
Column total	7309	13093	7309

Table B6.- Fisher's Exact Test Hox/,ParaHox contingency table (version 2)

B.4 R codes for Fisher's Exact Test, Binomial Exact Test and coefficient of association

The source for the FET codes are in CD Appendix B>B4>*.R or *.dat.

The binomial extact test is executed as following:

```
R>binom.test(x, n, p=? )
```

The parameters are as following: x being the number of successes, n being the number of trials and p = ? hypothesized probability.

B.5 Multiple alignments and phylogenies

The source for all the alignments and phylogenies are in CD Appendix B>B5>MA_phylogenies.txt.

B.6 Bilaterian-cnidarian-placozoan (BCP) Hox PAL list

Human	Nematostella	Trichoplax	Trichoplax scaffold
NP_005010.2, NP_000915.1	v1g117138	TRIADDRAFT_23032, TRIADDRAFT_50458	scaffold_3, scaffold_7
NP_002872.1, NP_005393.2	v1g181274	TRIDDRAFT_22031	scaffold_3
NP_002482.1	v1g165533	N/A	N/A
NP_056085.1, NP_899200.1	v1g234384	TRIADDRAFT_21914	scaffold_3
NP_057287.2, NP_059127.2, NP_001193638.1	v1g163878	TRIADDRAFT_50086	scaffold_3
NP_001120793.1, NP_009207.2, NP_006798.1	v1g158808	N/A	N/A
NP_001025167.2	v1g60162	TRIADDRAFT_54019	scaffold_3
NP_036360.3, NP_001026849.1	v1g96919, v1g90236	TRIADDRAFT_63115	scaffold_5
NP_004932.1, NP_003578.2	v1g158380, v1g177626	TRIADDRAFT_50235, TRIADDRAFT_61750	scaffold_4, scaffold_28
NP_000979.1	v1g234712	TRIADDRAFT_63098	scaffold_3
NP_803190.2, NP_113622.1	v1g241475	TRIADDRAFT_53805	scaffold_2
NP_689953.1	v1g178116	TRIADDRAFT_50898	scaffold_16
NP_036423.4	v1g241520	N/A	N/A
NP_060559.2	v1g204669, v1g61248	TRIADDRAFT_54018	scaffold_3
NP_060229.3, NP_060292.3	v1g39783, v1g97651	TRIADDRAFT_60586	scaffold_16
NP_001020.2	v1g177484	TRIADDRAFT_37138	scaffold_2
NP_065811.1, NP_055867.3	v1g33527	TRIADDRAFT_22516	scaffold_3
NP_001193998.1, NP_054757.1	v1g241916	TRIADDRAFT_63688	scaffold_2
NP_115729.1	v1g161954	TRIADDRAFT_30973	scaffold_15
NP_079457.2	v1g238113	TRIADDRAFT_54142	scaffold_3

Human	Nematostella	Trichoplax	Trichoplax scaffold
NP_954587.2	v1g97664	TRIADDRAFT_55076	scaffold_3
NP_659478.1, NP_060844.2, NP_076961.1	v1g238404	N/A	N/A
NP_079029.3	v1g80474	TRIADDRAFT_53605	scaffold_2
NP_006384.1, NP_006166.3, NP_001157979.1, NP_006139.1	v1g81835, v1g178189	N/A	N/A
NP_055126.1	v1g158158	N/A	N/A
NP_001005209.1	v1g97345, v1g225955	TRIADDRAFT_63705	scaffold_3
NP_001017957.1	v1g117135	TRIADDRAFT_60534	scaffold_16
NP_005585.1, NP_954984.1	v1g240229	TRIADDRAFT_38286	scaffold_16
NP_001028217.1	v1g241468	N/A	N/A
NP_055078.1	v1g158178	TRIADDRAFT_31113	scaffold_16
NP_003911.2	v1g81863	TRIADDRAFT_53559	scaffold_2
NP_659447.1	v1g164783	N/A	N/A
NP_055400.1	v1g97531	TRIADDRAFT_1773	scaffold_41
NP_612405.2, NP_001096032.1, NP_065865.1	v1g189518	N/A	N/A
NP_001030022.1, NP_835227.1	v1g238446	N/A	N/A
NP_079178.2	v1g234616	TRIADDRAFT_23261	scaffold_3
NP_001120863.1	v1g240206	TRIADDRAFT_54232	scaffold_3
NP_066024.1, NP_997221.2, NP_002290.2, NP_004786.2	v1g86012, v1g201226	N/A	N/A
NP_005799.2, NP_872580.1	v1g61841	TRIADDRAFT_15923	scaffold_3
NP_524146.1, NP_524147.2, NP_002467.1	v1g189473, v1g177611	TRIADDRAFT_60842, TRIADDRAFT_37105	scaffold_18, scaffold_1
NP_036232.2	v1g238321	TRIADDRAFT_37272	scaffold_2
NP_037473.3	v1g164749	TRIADDRAFT_30808	scaffold_15
NP_054859.2	v1g80671	TRIADDRAFT_27056	scaffold_6
NP_002786.2	v1g99390	TRIADDRAFT_50102	scaffold_3
NP_057399.1	v1g80887	N/A	N/A

Human	Nematostella	Trichoplax	Trichoplax scaffold
NP_001106178.1	v1g117131	TRIADDRAFT_23353	scaffold_3
NP_036365.1, NP_001165906.1	v1g80293	TRIADDRAFT_23084	scaffold_3
NP_848930.1, NP_036229.1	v1g195794, v1g235432	TRIADDRAFT_63238	scaffold_3
NP_009228.2, NP_689484.3, NP_079054.3, NP_940863.3	v1g238046, v1g106903, v1g177096	TRIADDRAFT_54061	scaffold_3
NP_064527.1, NP_002480.1	v1g197216	N/A	N/A
NP_001097.2, NP_001607.1	v1g80560	TRIADDRAFT_3190	scaffold_16
NP_001194.1, NP_004320.2, NP_001096.1	v1g165860, v1g178197	TRIADDRAFT_27560, TRIADDRAFT_22452, TRIADDRAFT_22033	scaffold_7, scaffold_3, scaffold_3
NP_060599.1	v1g80869	TRIADDRAFT_53373	scaffold_2
NP_001231.2, NP_001232.1	v1g99661	TRIADDRAFT_31077, TRIADDRAFT_27302	scaffold_16, scaffold_6
NP_036565.2	v1g241911	TRIADDRAFT_54034	scaffold_3
NP_689609.2, NP_859076.3, NP_060866.2, NP_683759.1	v1g234398, v1g240471, v1g197432	TRIADDRAFT_22960, TRIADDRAFT_22540, TRIADDRAFT_52388	scaffold_3, scaffold_1
NP_003133.1, NP_742067.3	v1g236578, v1g25031, v1g99557	TRIADDRAFT_54211, TRIADDRAFT_57299	scaffold_3, scaffold_6
NP_005792.1	v1g241461	TRIADDRAFT_63244	scaffold_3
NP_001001550.1, NP_004481.2, NP_005301.2	v1g205661	TRIADDRAFT_54042, TRIADDRAFT_54043	scaffold_3
NP_003066.2, NP_003065.3	v1g30373	TRIADDRAFT_31063	scaffold_16
NP_001247.3, NP_001107563.1	v1g30881	TRIADDRAFT_21924	scaffold_3
NP_004513.1, NP_004975.2, NP_004512.1	v1g227907	TRIADDRAFT_54045	scaffold_3
NP_982288.1, NP_671723.1, NP_071358.1, NP_078828.2	v1g96861	TRIADDRAFT_58575	scaffold_8
NP_065970.2, NP_667340.2	v1g164111	TRIADDRAFT_60444	scaffold_15

Human	Nematostella	Trichoplax	Trichoplax scaffold
NP_840101.1, NP_003059.1, NP_005976.2	v1g236363	TRIADDRAFT_7674	scaffold_1
NP_075559.2, NP_001093894.1	v1g238337	TRIADDRAFT_54404	scaffold_3
NP_060621.3	v1g47548, v1g57220	TRIADDRAFT_54011	scaffold_3
NP_079265.2	v1g161986	TRIADDRAFT_53016	scaffold_2
NP_037409.2	v1g238395	TRIADDRAFT_55060	scaffold_3
NP_061167.1, NP_006046.1	v1g184757	TRIADDRAFT_54924, TRIADDRAFT_54925	scaffold_3
NP_079095.3	v1g229212	TRIADDRAFT_55032	scaffold_3
NP_002586.2, NP_148978.2, NP_002587.2	v1g84454, v1g241402	TRIADDRAFT_20204, TRIADDRAFT_50028	scaffold_1, scaffold_3
NP_003496.1, NP_003498.1, NP_001457.1, NP_003459.2, NP_114072.1	v1g171640, v1g183962	TRIADDRAFT_12196	scaffold_3
NP_066564.2, NP_003875.3	v1g96946	TRIADDRAFT_22087	scaffold_3
NP_057551.1, NP_644809.1, NP_054901.1	v1g241507	TRIADDRAFT_23050	scaffold_3
NP_001019839.1	v1g241397	TRIADDRAFT_63756	scaffold_3
NP_001926.2, NP_004451.2, NP_001171507.1, NP_001927.3	v1g197301	TRIADDRAFT_64278, TRIADDRAFT_60461	scaffold_15, scaffold_15
NP_002889.1, NP_002888.1, NP_001171182.1	v1g80463	TRIADDRAFT_28658	scaffold_9
NP_002147.2	v1g178049	TRIADDRAFT_54071	scaffold_3
NP_835455.1	v1g97005	TRIADDRAFT_9370	scaffold_3
NP_004574.2, NP_002859.1, NP_004153.2	v1g158216	TRIADDRAFT_22550	scaffold_3
NP_859062.1, NP_004279.3	v1g242338	TRIADDRAFT_54453	scaffold_3
NP_056480.1	v1g183365	TRIADDRAFT_54056	scaffold_3
NP_036417.1, NP_653234.2	v1g91046	TRIADDRAFT_30951	scaffold_15
NP_689557.1	v1g236359	TRIADDRAFT_63807	scaffold_3

Human	Nematostella	Trichoplax	Trichoplax scaffold
NP_000465.1	v1g234699	TRIADDRAFT_9370	scaffold_3
NP_000113.1	v1g96883	TRIADDRAFT_22911	scaffold_3
NP_061854.1	v1g163820	TRIADDRAFT_50999	scaffold_21
NP_003343.1, NP_008867.2, NP_008868.3	v1g91283	TRIADDRAFT_33331	scaffold_33
NP_000918.2, NP_000434.1, NP_003733.2	v1g237874, v1g82183	TRIADDRAFT_54423, TRIADDRAFT_54424	scaffold_3
NP_002197.2, NP_001138468.1, NP_001073286.1	v1g238305	TRIADDRAFT_54908	scaffold_3
NP_055369.1	v1g158777	TRIADDRAFT_23232	scaffold_3
NP_004516	v1g40010	TRIADDRAFT_27379, TRIADDRAFT_19424, TRIADDRAFT_62104	scaffold_7, scaffold_1, scaffold_35
NP_001002031.1, NP_005166.1, NP_001680.1	v1g204676	TRIADDRAFT_49770, TRIADDRAFT_37048	scaffold_1
NP_055475.2	v1g96987	TRIADDRAFT_53785	scaffold_2
NP_001070666.1, NP_775952.4	v1g99635	TRIADDRAFT_22278	scaffold_3
NP_001171867.1, NP_060894.2	v1g61432	TRIADDRAFT_55017	scaffold_3
NP_001649.1, NP_001650.1	v1g242285	TRIADDRAFT_63238	scaffold_3
NP_001186913.1	v1g226219	N/A	N/A
NP_689597.1	v1g164107	TRIADDRAFT_50116	scaffold_3
NP_002126.2, NP_006177.1	v1g152310, v1g101676	TRIADDRAFT_49897, TRIADDRAFT_21656	scaffold_2
NP_060164.3	v1g211378, v1g211665, v1g116714, v1g241935, v1g204788, v1g204789, v1g97363	TRIADDRAFT_1695	scaffold_16
NP_001153218.1, NP_004318.3, NP_060930.3, NP_060757.4	v1g239294, v1g80747	TRIADDRAFT_58026, TRIADDRAFT_55060	scaffold_7, scaffold_3
NP_006652.1	v1g94726, v1g183341, v1g110219	TRIADDRAFT_21752	scaffold_2

Human	Nematostella	Trichoplax	Trichoplax scaffold
NP_065954.1, NP_065986.2, NP_114113.1	v1g238058	TRIADDRAFT_60404	scaffold_15
NP_112185.1	v1g203674, v1g211579, v1g247750	TRIADDRAFT_54508	scaffold_3
NP_003629.1, NP_002201.1	v1g196726	TRIADDRAFT_54908	scaffold_3
NP_003208.2	v1g240213	TRIADDRAFT_63745	scaffold_3
NP_001073331.1, NP_001034933.1	v1g81002	TRIADDRAFT_35747	scaffold_3
NP_003143.2, NP_036580.2	v1g241935	TRIADDRAFT_53918	scaffold_2
NP_055117.1	v1g184753	TRIADDRAFT_63141	scaffold_16
NP_001073998.2, NP_003378.3	v1g197314	TRIADDRAFT_61805	scaffold_28
NP_116264.2	v1g240237	TRIADDRAFT_50916	scaffold_16
NP_001026886.1, NP_076973.1	v1g158987	TRIADDRAFT_63747	scaffold_3
NP_006328.2	v1g99527	TRIADDRAFT_30763	scaffold_15
NP_115970.2, NP_776183.1	v1g232597	TRIADDRAFT_54454	scaffold_3
NP_001408.2	v1g238031	TRIADDRAFT_53556	scaffold_2
NP_003876.1, NP_003927.1	v1g80476	TRIADDRAFT_59487	scaffold_11
NP_004498.1	v1g158364	TRIADDRAFT_53881	scaffold_2
NP_000828.1, NP_036438.2, NP_071435.2	v1g238030	TRIADDRAFT_31200, TRIADDRAFT_60535	scaffold_16
NP_853514.1, NP_000288.1, NP_057196.2, NP_001009944.2	v1g198568, v1g196807	TRIADDRAFT_53596	scaffold_2
NP_944490.1, NP_003121.1	v1g81972	TRIADDRAFT_22435	scaffold_3
NP_061720.2	v1g161999	TRIADDRAFT_60424	scaffold_15
NP_002172.2, NP_066382.1, NP_000184.1	v1g241466, v1g87421, v1g87496, v1g140260, v1g95413	N/A	N/A
NP_001012241.1	v1g240276	TRIADDRAFT_63804	scaffold_3

Human	Nematostella	Trichoplax	Trichoplax scaffold
NP_003031.3, NP_005061.2, NP_003750.1, NP_000333.1	v1g91031	TRIADDRAFT_54168, TRIADDRAFT_22844	scaffold_3
NP_060579.3	v1g238029	TRIADDRAFT_54458	scaffold_3
XP_001714944.3	v1g203573	TRIADDRAFT_60645	scaffold_16
NP_115766.3	v1g181320	TRIADDRAFT_6541	scaffold_16
NP_954699.1, NP_001137381.1	v1g158372	TRIADDRAFT_21066	scaffold_2
NP_002401.1	v1g158262	N/A	N/A
NP_002938.1	v1g236362	N/A	N/A
NP_001245.1	v1g80906	TRIADDRAFT_22362	scaffold_3
NP_002602.2, NP_001135858.1, NP_002601.1, NP_002603.1	v1g232588	TRIADDRAFT_20860	scaffold_2
NP_079485.1, NP_056445.3	v1g238416	TRIADDRAFT_31210	scaffold_16
NP_071903.2, NP_060583.2	v1g197260	TRIADDRAFT_21081	scaffold_2
NP_001164275.1	v1g99498	TRIADDRAFT_54234	scaffold_3
NP_001369.1, NP_004402.1	v1g158934	TRIADDRAFT_20566	scaffold_2
NP_036355.2	v1g177548	TRIADDRAFT_20417	scaffold_2
NP_001180242.1	v1g82280	TRIADDRAFT_23229	scaffold_3
NP_872327.2	v1g36807	TRIADDRAFT_55075	scaffold_3
NP_001161688.1	v1g210418, v1g196789	N/A	N/A
NP_997254.3	v1g82295	TRIADDRAFT_55046	scaffold_3
NP_055855.2	v1g201207	TRIADDRAFT_54493, TRIADDRAFT_54491	scaffold_3
NP_006832.1, NP_109599.3, NP_061849.2, NP_722518.2	v1g204678	TRIADDRAFT_53014	scaffold_2
NP_004873.3	v1g117089	TRIADDRAFT_54065	scaffold_3
NP_055132.2	v1g238335	TRIADDRAFT_20625	scaffold_2
NP_006795.3	v1g241470	TRIADDRAFT_53760	scaffold_2
NP_570857.2	v1g82385	TRIADDRAFT_55048	scaffold_3
NP_075567.2	v1g158438	TRIADDRAFT_60632	scaffold_16
NP_660298.2	v1g205666	TRIADDRAFT_54455	scaffold_3

Human	Nematostella	Trichoplax	Trichoplax scaffold
NP_001072.2	v1g31243	TRIADDRAFT_53561	scaffold_2
NP_004222.2	v1g226158	TRIADDRAFT_37267	scaffold_2
NP_003092.4	v1g80875	TRIADDRAFT_22890	scaffold_3
NP_219487.3	v1g205582	TRIADDRAFT_31335	scaffold_17
NP_775491.1, NP_003355.1	v1g241497	TRIADDRAFT_31186	scaffold_16
NP_059129.3, NP_001165113.1	v1g164057	TRIADDRAFT_22946	scaffold_3
NP_001092303.1	v1g236583	TRIADDRAFT_60605	scaffold_16
NP_055177.2	v1g158887	TRIADDRAFT_30551	scaffold_14
NP_005972.1, NP_055214.1	v1g161973	TRIADDRAFT_63235, TRIADDRAFT_21448	scaffold_2
NP_003165.2	v1g181209	TRIADDRAFT_54138	scaffold_3
NP_060546.2	v1g96973	TRIADDRAFT_63243	scaffold_3
NP_620158.3	v1g91203	TRIADDRAFT_49953, TRIADDRAFT_54084	scaffold_3
NP_114109.1	v1g238326	N/A	N/A
NP_110436.1	v1g248604, v1g203676	N/A	N/A
NP_001036111.1, NP_055864.2	v1g203582	TRIADDRAFT_54221	scaffold_3
NP_004068.2	v1g82156	TRIADDRAFT_54073	scaffold_3
NP_036575.1	v1g184716	TRIADDRAFT_50106	scaffold_3
NP_003070.3	v1g177453	TRIADDRAFT_53566	scaffold_2
NP_036454.1	v1g226211	TRIADDRAFT_50007	scaffold_3
NP_000879.2	v1g143492, v1g91193	TRIADDRAFT_54882	scaffold_3
NP_001034934.1	v1g90973	TRIADDRAFT_31069	scaffold_16
NP_065875.3, NP_001186346.1	v1g82036	TRIADDRAFT_53574	scaffold_2
NP_057018.1	v1g238448	TRIADDRAFT_30778	scaffold_15
NP_060726.3	v1g82024	TRIADDRAFT_55049	scaffold_3
NP_005722.1	v1g241492	TRIADDRAFT_50117	scaffold_3
NP_001124435.1, NP_001136117.1	v1g238352	TRIADDRAFT_57241, TRIADDRAFT_53906	scaffold_6, scaffold_2
NP_001035938.1	v1g241917	N/A	N/A
NP_005928.2	v1g96791	TRIADDRAFT_55039	scaffold_3
NP_001121616.1,	v1g82384	TRIADDRAFT_30981	scaffold_15

Human	**Nematostella**	**Trichoplax**	**Trichoplax scaffold**
NP_114152.3, NP_001032208.1, NP_055585.1	v1g184029	TRIADDRAFT_21576, TRIADDRAFT_57442	scaffold_2, scaffold_6
NP_000456.2	v1g39202	TRIADDRAFT_55679	scaffold_4
NP_689732.2	v1g240211	TRIADDRAFT_60596	scaffold_16
NP_001167596.1	v1g164036	TRIADDRAFT_20654	scaffold_2
NP_113609.1	v1g99563	TRIADDRAFT_30804	scaffold_15
NP_055070.1	v1g97340, v1g156534	TRIADDRAFT_33595	scaffold_36
NP_002148.1	v1g178050	TRIADDRAFT_37293	scaffold_3
NP_000989.1	v1g226223	TRIADDRAFT_35714	scaffold_3
NP_694453.2	v1g163860	TRIADDRAFT_1749	scaffold_3
NP_001078916.1	v1g81851	TRIADDRAFT_55009	scaffold_3
NP_056345.3, NP_114105.1	v1g80798	TRIADDRAFT_23196	scaffold_3
NP_001034782.1	v1g99611	TRIADDRAFT_30960	scaffold_15
NP_001034813.2	v1g163904	TRIADDRAFT_53065	scaffold_2
NP_004385.1	v1g82268	N/A	N/A
NP_005250.1, NP_005802.1	v1g196852	N/A	N/A
NP_065207.2, NP_443149.2	v1g91041	TRIADDRAFT_22976, TRIADDRAFT_22201	scaffold_3
NP_060218.1	v1g162067	TRIADDRAFT_54223	scaffold_3
NP_037422.2	v1g197451	TRIADDRAFT_54965	scaffold_3
NP_065138.2, NP_116201.7	v1g238059	TRIADDRAFT_5497	scaffold_2
NP_653309.3	v1g97653	TRIADDRAFT_22532	scaffold_3
NP_009172.2	v1g178046	TRIADDRAFT_37300	scaffold_3
NP_001906.3	v1g234636	TRIADDRAFT_23237, TRIADDRAFT_55036	scaffold_3
NP_859525.1	v1g99492	TRIADDRAFT_31116	scaffold_16
NP_079543.1	v1g158880	TRIADDRAFT_20746	scaffold_2
NP_001478.2	v1g164044	TRIADDRAFT_54870	scaffold_3
EAW58000.1	v1g241881	TRIADDRAFT_61431	scaffold_24
NP_006391.1	v1g242304	TRIADDRAFT_54892	scaffold_3
NP_005776.1	v1g158326	N/A	N/A
NP_002778.1	v1g165493	TRIADDRAFT_38289	scaffold_16
NP_002038.2	v1g162025	TRIADDRAFT_33686	scaffold_37

Human	Nematostella	Trichoplax	Trichoplax scaffold
NP_001171715.1	v1g240248	TRIADDRAFT_55106	scaffold_3
NP_001813.1, NP_001020372.2	v1g101727	TRIADDRAFT_53574	scaffold_2
NP_009141.2, NP_004473.2	v1g31464	TRIADDRAFT_22839, TRIADDRAFT_22201, TRIADDRAFT_22976	scaffold_3
NP_036203.1	v1g81873	N/A	N/A
NP_004809.2	v1g228991	TRIADDRAFT_50105	scaffold_3
NP_005680.1	v1g99533	TRIADDRAFT_20003	scaffold_1
NP_060941.2	v1g99499	TRIADDRAFT_60469	scaffold_15
NP_079423.1	v1g183953	TRIADDRAFT_22596	scaffold_3
NP_057700.3	v1g101462	TRIADDRAFT_51403	scaffold_1
NP_071682.1	v1g228048	TRIADDRAFT_2519	scaffold_3
NP_002078.1	v1g32586	TRIADDRAFT_63706	scaffold_3
NP_219481.1	v1g201177	TRIADDRAFT_22994	scaffold_3
NP_001183956.1	v1g117154	TRIADDRAFT_54405	scaffold_3
NP_065726.1	v1g224186, v1g236726	N/A	N/A
NP_001340.2	v1g205625	TRIADDRAFT_56497	scaffold_5

B.7 R command for Binomial Exact Test

The binomial extact test is executed as following:

```
R>binom.test(x, n, p=? )
```

The parameters are as following: x being the number of successes, n being the number of trials and p = ? hypothesized probability.

Appendix C

C.1 Bilaterian-Cnidarian-Placozoan (BCP) Hox PAL extended to poriferan *A. queenslandica*

Human	Amphimedon	ContigAmphi
NP_005010.2, NP_000915.1	Aqu1.217459, Aqu1.217460	13307
NP_002872.1, NP_005393.2	Aqu1.222153	13436
NP_002482.1	Aqu1.215502	13219
NP_056085.1, NP_899200.1	Aqu1.218327	13337
NP_057287.2, NP_059127.2, NP_001193638.1	Aqu1.228444	13514
NP_001120793.1, NP_009207.2, NP_006798.1	Aqu1.217708	13315
NP_001025167.2	Aqu1.222173	13436
NP_036360.3, NP_001026849.1	N/A	N/A
NP_004932.1, NP_003578.2	Aqu1.228495	13514
NP_000979.1	Aqu1.222164, Aqu1.222165, Aqu1.222166	13436
NP_803190.2, NP_113622.1	Aqu1.217015	13289
NP_689953.1	Aqu1.217736	13315
NP_036423.4	Aqu1.224063	13470
NP_060559.2	Aqu1.225679	13490
NP_060229.3, NP_060292.3	N/A	N/A
NP_001020.2	Aqu1.216592	13271
NP_065811.1, NP_055867.3	Aqu1.217545	13310
NP_001193998.1, NP_054757.1	Aqu1.220649, Aqu1.220650	13403

Human	Amphimedon	ContigAmphi
NP_115729.1	Aqu1.217446	13307
NP_079457.2	Aqu1.218251	13335
NP_954587.2	N/A	N/A
NP_659478.1, NP_060844.2, NP_076961.1	Aqu1.228417	13514
NP_079029.3	Aqu1.211849,Aqu1.211850	12926
NP_006384.1, NP_006166.3, NP_001157979.1, NP_006139.1	Aqu1.224095, Aqu1.224096, Aqu1.224097, Aqu1.224098	13470
NP_055126.1	Aqu1.211853	12926
NP_001005209.1	Aqu1.226073	13495
NP_001017957.1	Aqu1.223055	13453
NP_005585.1, NP_954984.1	Aqu1.224879	13482
NP_001028217.1	Aqu1.224061	13470
NP_055078.1	Aqu1.205815	10364
NP_003911.2	Aqu1.217007	13289
NP_659447.1	Aqu1.218934	13416
NP_055400.1	Aqu1.222121, Aqu1.222122	13436
NP_612405.2, NP_001096032.1, NP_065865.1	N/A	N/A
NP_001030022.1, NP_835227.1	N/A	N/A
NP_079178.2	Aqu1.223300	13457
NP_001120863.1	N/A	N/A
NP_066024.1, NP_997221.2, NP_002290.2, NP_004786.2	N/A	N/A
NP_005799.2, NP_872580.1	Aqu1.227965	13511
NP_524146.1, NP_524147.2, NP_002467.1	Aqu1.209650, Aqu1.209651	12507
NP_036232.2	Aqu1.206724	11132
NP_037473.3	Aqu1.205870	10409
NP_054859.2	Aqu1.227777	13510

Human	Amphimedon	ContigAmphi
NP_002786.2	Aqu1.228381	13514
NP_057399.1	N/A	N/A
NP_001106178.1	Aqu1.228427	13514
NP_036365.1, NP_001165906.1	Aqu1.205777	10335
NP_848930.1, NP_036229.1	Aqu1.230054	13521
NP_009228.2, NP_689484.3, NP_079054.3, NP_940863.3	N/A	N/A
NP_064527.1, NP_002480.1	N/A	N/A
NP_001097.2, NP_001607.1	Aqu1.224888	13482
NP_001194.1, NP_004320.2, NP_001096.1	Aqu1.227949	13511
NP_060599.1	Aqu1.228048, Aqu1.228049	13511
NP_001231.2, NP_001232.1	Aqu1.215681	13228
NP_036565.2	Aqu1.222096	13436

NP_689609.2,
NP_859076.3,
NP_060866.2,
NP_683759.1
	Aqu1.229551,
	Aqu1.217427,
	Aqu1.222125
		13520;
		13306;
		13436

NP_003133.1, NP_742067.3	Aqu1.215049	13197
NP_005792.1	Aqu1.219691	13287
NP_001001550.1, NP_004481.2, NP_005301.2	Aqu1.222595	13446
NP_003066.2, NP_003065.3	Aqu1.224867	13482
NP_001247.3, NP_001107563.1	Aqu1.218098	13329
NP_004513.1, NP_004975.2, NP_004512.1	Aqu1.228376	13514
NP_982288.1, NP_671723.1, NP_071358.1, NP_078828.2	Aqu1.229231	13519
NP_065970.2, NP_667340.2	N/A	N/A

Human	Amphimedon	ContigAmphi
NP_840101.1, NP_003059.1, NP_005976.2	N/A	N/A
NP_075559.2, NP_001093894.1	N/A	N/A
NP_060621.3	N/A	N/A
NP_079265.2	N/A	N/A
NP_037409.2	N/A	N/A
NP_061167.1, NP_006046.1	Aqu1.222155	13436
NP_079095.3	Aqu1.227282	13507
NP_002586.2, NP_148978.2, NP_002587.2	Aqu1.228858	13516
NP_003496.1, NP_003498.1, NP_001457.1, NP_003459.2, NP_114072.1	Aqu1.228355, Aqu1.228356	13513
NP_066564.2, NP_003875.3	Aqu1.227534	13508
NP_057551.1, NP_644809.1, NP_054901.1	Aqu1.222118	13436
NP_001019839.1	N/A	N/A
NP_001926.2, NP_004451.2, NP_001171507.1, NP_001927.3	Aqu1.222672	13447
NP_002889.1, NP_002888.1, NP_001171182.1	N/A	N/A
NP_002147.2	Aqu1.205528	10141
NP_835455.1	Aqu1.217476	13308
NP_004574.2, NP_002859.1, NP_004153.2	Aqu1.218101	13329
NP_859062.1, NP_004279.3	Aqu1.229783	13521
NP_056480.1	Aqu1.222113	13436
NP_036417.1, NP_653234.2	N/A	N/A
NP_689557.1	Aqu1.229597	13520
NP_000465.1	N/A	N/A

Human	Amphimedon	ContigAmphi
NP_000113.1	Aqu1.200730	2132
NP_061854.1	N/A	N/A
NP_003343.1, NP_008867.2, NP_008868.3	Aqu1.217315, Aqu1.217316	13302
NP_000918.2, NP_000434.1, NP_003733.2	Aqu1.200362	1119
NP_002197.2, NP_001138468.1, NP_001073286.1	Aqu1.229546	13520
NP_055369.1	Aqu1.228419	13514
NP_004516	N/A	N/A
NP_001002031.1, NP_005166.1, NP_001680.1	Aqu1.222001	13434
NP_055475.2	Aqu1.228520	13514
NP_001070666.1, NP_775952.4	Aqu1.225388	13487
NP_001171867.1, NP_060894.2	Aqu1.227119	13506
NP_001649.1, NP_001650.1	Aqu1.230054	13521
NP_001186913.1	Aqu1.223931	13468
NP_689597.1	Aqu1.202319	5608
NP_002126.2, NP_006177.1	N/A	N/A
NP_060164.3	N/A	N/A
NP_001153218.1, NP_004318.3, NP_060930.3, NP_060757.4	Aqu1.228380	13514
NP_006652.1	Aqu1.222756	13448
NP_065954.1, NP_065986.2, NP_114113.1	N/A	N/A
NP_112185.1	N/A	N/A
NP_003629.1, NP_002201.1	N/A	N/A
NP_003208.2	Aqu1.225824	13491
NP_001073331.1, NP_001034933.1	Aqu1.221642	13427

Human	Amphimedon	ContigAmphi
NP_003143.2, NP_036580.2	Aqu1.224126	13470
NP_055117.1	Aqu1.204252	8738
NP_001073998.2, NP_003378.3	Aqu1.203711	7988
NP_116264.2	Aqu1.217714, Aqu1.217715, Aqu1.217716	13315
NP_001026886.1, NP_076973.1	Aqu1.228852	13516
NP_006328.2	Aqu1.218117	13329
NP_115970.2, NP_776183.1	N/A	N/A
NP_001408.2	Aqu1.211855	12926
NP_003876.1, NP_003927.1	Aqu1.229543	13520
NP_004498.1	N/A	N/A
NP_000828.1, NP_036438.2, NP_071435.2	N/A	N/A
NP_853514.1, NP_000288.1, NP_057196.2, NP_001009944.2	N/A	N/A
NP_944490.1, NP_003121.1	N/A	N/A
NP_061720.2	Aqu1.212008	12947
NP_002172.2, NP_066382.1, NP_000184.1	Aqu1.217859	13319
NP_001012241.1	N/A	N/A
NP_003031.3, NP_005061.2, NP_003750.1, NP_000333.1	Aqu1.218116	13329
NP_060579.3	N/A	N/A
XP_001714944.3	N/A	N/A
NP_115766.3	Aqu1.224871	13482
NP_954699.1, NP_001137381.1	Aqu1.206501	10968
NP_002401.1	N/A	N/A
NP_002938.1	N/A	N/A
NP_001245.1	Aqu1.218100	13329

Human	Amphimedon	ContigAmphi
NP_002602.2, NP_001135858.1, NP_002601.1, NP_002603.1	N/A	N/A
NP_079485.1, NP_056445.3	Aqu1.218590	13344
NP_071903.2, NP_060583.2	Aqu1.204976	9624
NP_001164275.1	Aqu1.218108	13329
NP_001369.1, NP_004402.1	Aqu1.211857	12926
NP_036355.2	Aqu1.224923	13482
NP_001180242.1	Aqu1.208035	11908
NP_872327.2	Aqu1.228462	13514
NP_001161688.1	N/A	N/A
NP_997254.3	N/A	N/A
NP_055855.2	Aqu1.225405	13487
NP_006832.1, NP_109599.3, NP_061849.2, NP_722518.2	Aqu1.216196	13252
NP_004873.3	Aqu1.217558	13310
NP_055132.2	Aqu1.211945, Aqu1.211946	12937
NP_006795.3	Aqu1.220677	13404
NP_570857.2	Aqu1.211591, Aqu1.211590	12892
NP_075567.2	Aqu1.209646	12507
NP_660298.2	Aqu1.221080	13414
NP_001072.2	N/A	N/A
NP_004222.2	Aqu1.214464	13164
NP_003092.4	Aqu1.222138, Aqu1.222139, Aqu1.222140	13436
NP_219487.3	N/A	N/A
NP_775491.1, NP_003355.1	Aqu1.213750, Aqu1.213751	13108
NP_059129.3, NP_001165113.1	Aqu1.217454	13307
NP_001092303.1	Aqu1.217734, Aqu1.217735	13315

Human	Amphimedon	ContigAmphi
NP_055177.2	Aqu1.224903	13482
NP_005972.1, NP_055214.1	N/A	N/A
NP_003165.2	Aqu1.212733	13026
NP_060546.2	Aqu1.218111	13329
NP_620158.3	Aqu1.222109	13436
NP_114109.1	Aqu1.224887	13482
NP_110436.1	Aqu1.210503	12702
NP_001036111.1, NP_055864.2	Aqu1.227126	13506
NP_004068.2	Aqu1.222101	13436
NP_036575.1	Aqu1.223182	13456
NP_003070.3	Aqu1.224100	13470
NP_036454.1	Aqu1.228391	13514
NP_000879.2	Aqu1.219978	13384
NP_001034934.1	Aqu1.217705	13315
NP_065875.3, NP_001186346.1	N/A	N/A
NP_057018.1	Aqu1.225423	13487
NP_060726.3	Aqu1.228466	13514
NP_005722.1	Aqu1.228370	13514
NP_001124435.1, NP_001136117.1	Aqu1.202016	4992
NP_001035938.1	N/A	N/A
NP_005928.2	Aqu1.229594	13520
NP_001121616.1,	Aqu1.217122	13294
NP_114152.3, NP_001032208.1, NP_055585.1	Aqu1.209644	12507
NP_000456.2	N/A	N/A
NP_689732.2	Aqu1.224068	13470
NP_001167596.1	Aqu1.218939	13353
NP_113609.1	Aqu1.217443	13307
NP_055070.1	Aqu1.224092	13470
NP_002148.1	Aqu1.222102	13436

Human	Amphimedon	ContigAmphi
NP_000989.1	Aqu1.217542	13310
NP_694453.2	Aqu1.229595	13520
NP_001078916.1	Aqu1.222151	13436
NP_056345.3, NP_114105.1	Aqu1.218112	13329
NP_001034782.1	Aqu1.221184	13416
NP_001034813.2	Aqu1.227512	13508
NP_004385.1	N/A	N/A
NP_005250.1, NP_005802.1	N/A	N/A
NP_065207.2, NP_443149.2	N/A	N/A
NP_060218.1	N/A	N/A
NP_037422.2	Aqu1.217544	13310
NP_065138.2, NP_116201.7	Aqu1.227565	13508
NP_653309.3	Aqu1.227515	13508
NP_009172.2	Aqu1.228464	13514
NP_001906.3	N/A	N/A
NP_859525.1	Aqu1.217018	13289
NP_079543.1	Aqu1.213661	13103
NP_001478.2	Aqu1.224166	13471
EAW58000.1	N/A	N/A
NP_006391.1	N/A	N/A
NP_005776.1	Aqu1.222033	13435
NP_002778.1	N/A	N/A
NP_002038.2	Aqu1.222129	13436
NP_001171715.1	N/A	N/A
NP_001813.1, NP_001020372.2	N/A	N/A
NP_009141.2, NP_004473.2	N/A	N/A
NP_036203.1	N/A	N/A
NP_004809.2	Aqu1.222141, Aqu1.222142	13436
NP_005680.1	N/A	N/A

Human	Amphimedon	ContigAmphi
NP_060941.2	Aqu1.222313	13439
NP_079423.1	Aqu1.222110, Aqu1.222111	13436
NP_057700.3	Aqu1.229487	13520
NP_071682.1	N/A	N/A
NP_002078.1	N/A	N/A
NP_219481.1	N/A	N/A
NP_001183956.1	Aqu1.217125	13294
NP_065726.1	N/A	N/A
NP_001340.2	Aqu1.207434	11607

C.2 l-ParaHox PAL gene list

Human	Chromosomal segment	Trichoplax adhaerens
ENSG00000032742	13.1	TRIADDRAFT_56242
ENSG00000102710	13.1	TRIADDRAFT_56408
ENSG00000102743	13.1	TRIADDRAFT_5000
ENSG00000120688	13.1	TRIADDRAFT_56241
ENSG00000120694	13.1	TRIADDRAFT_25019
ENSG00000120697	13.1	TRIADDRAFT_25674
ENSG00000120699	13.1	TRIADDRAFT_25085
ENSG00000122042	13.1	TRIADDRAFT_25794
ENSG00000132953	13.1	TRIADDRAFT_56536
ENSG00000132963	13.1	TRIADDRAFT_56262
ENSG00000133101	13.1	TRIADDRAFT_24944
ENSG00000133105	13.1	TRIADDRAFT_12560
ENSG00000133105	13.1	TRIADDRAFT_56600
ENSG00000133119	13.1	TRIADDRAFT_26016
ENSG00000139505	13.1	TRIADDRAFT_56124
ENSG00000150456	13.1	TRIADDRAFT_25578
ENSG00000165487	13.1	TRIADDRAFT_56755
ENSG00000172915	13.1	TRIADDRAFT_25672

Human	Chromosomal segment	Trichoplax adhaerens
ENSG00000172915	13.1	TRIADDRAFT_25032
ENSG00000244754	13.1	TRIADDRAFT_56283
ENSG00000010671	13.1	TRIADDRAFT_24853
ENSG00000067177	13.1	TRIADDRAFT_25466
ENSG00000080572	X.6	TRIADDRAFT_63939
ENSG00000085224	X.6	TRIADDRAFT_25002
ENSG00000089682	X.6	TRIADDRAFT_8568
ENSG00000101811	X.6	TRIADDRAFT_56724
ENSG00000102144	X.6	TRIADDRAFT_63295
ENSG00000102383	X.6	TRIADDRAFT_25400
ENSG00000123570	X.6	TRIADDRAFT_56572
ENSG00000126953	X.6	TRIADDRAFT_26075
ENSG00000131269	X.6	TRIADDRAFT_56527
ENSG00000147099	X.6	TRIADDRAFT_25928
ENSG00000147162	X.6	TRIADDRAFT_56833
ENSG00000147174	X.6	TRIADDRAFT_56122
ENSG00000147224	X.6	TRIADDRAFT_25357
ENSG00000165240	X.6	TRIADDRAFT_56323
ENSG00000188419	X.6	TRIADDRAFT_63951
ENSG00000198034	X.6	TRIADDRAFT_37748
ENSG00000038274	5.4	TRIADDRAFT_25402
ENSG00000038274	5.4	TRIADDRAFT_25365
ENSG00000081791	5.4	TRIADDRAFT_56780
ENSG00000091010	5.4	TRIADDRAFT_25765
ENSG00000113643	5.4	TRIADDRAFT_56467
ENSG00000123643	5.4	TRIADDRAFT_36005
ENSG00000123643	5.4	TRIADDRAFT_56468
ENSG00000123643	5.4	TRIADDRAFT_56216
ENSG00000131507	5.4	TRIADDRAFT_56783
ENSG00000155506	5.4	TRIADDRAFT_56648
ENSG00000155506	5.4	TRIADDRAFT_56647
ENSG00000155508	5.4	TRIADDRAFT_26102

Human	Chromosomal segment	Trichoplax adhaerens
ENSG00000164576	5.4	TRIADDRAFT_25386
ENSG00000014824	4.2	TRIADDRAFT_25311
ENSG00000065882	4.2	TRIADDRAFT_13887
ENSG00000075539	4.2	TRIADDRAFT_25535
ENSG00000075539	4.2	TRIADDRAFT_56185
ENSG00000078140	4.2	TRIADDRAFT_24883
ENSG00000090989	4.2	TRIADDRAFT_25724
ENSG00000109189	4.2	TRIADDRAFT_50293
ENSG00000109680	4.2	TRIADDRAFT_57017
ENSG00000121892	4.2	TRIADDRAFT_56191
ENSG00000124406	4.2	TRIADDRAFT_25047
ENSG00000151806	4.2	TRIADDRAFT_56304
ENSG00000169299	4.2	TRIADDRAFT_26086
ENSG00000183783	4.2	TRIADDRAFT_25991
ENSG00000215203	4.2	TRIADDRAFT_56655

C.3 l-ParaHox PAL extended to poriferan *A. queenslandica* gene list

Here is merged the information from scaffold 38 from *T. adhaerens.*

Human	**Amphimedon queenslandica**	**Amphimedon contig**
ENSG00000032742	Aqu1.213626	Contig13101
ENSG00000102710	Aqu1.217641	Contig13313
ENSG00000102743	N/A	N/A
ENSG00000120688	N/A	N/A
ENSG00000120694	N/A	N/A
ENSG00000120697	Aqu1.222773	Contig13448
ENSG00000120699	N/A	N/A
ENSG00000122042	N/A	N/A
ENSG00000132953	Aqu1.220047	Contig13386
ENSG00000132963	Aqu1.210844	Contig12764
ENSG00000133101	Aqu1.222748	Contig13448
ENSG00000133105	N/A	N/A
ENSG00000133105	N/A	N/A

Human	**Amphimedon queenslandica**	**Amphimedon contig**
ENSG00000133119	Aqu1.209201	Contig12374
ENSG00000139505	Aqu1.220336	Contig13395
ENSG00000150456	Aqu1.221740	Contig13429
ENSG00000165487	Aqu1.215408	Contig13214
ENSG00000172915	N/A	N/A
ENSG00000172915	N/A	N/A
ENSG00000244754	N/A	N/A
ENSG00000010671	N/A	N/A
ENSG00000067177	Aqu1.225488	Contig13489
ENSG00000080572	Aqu1.222724	Contig13448
ENSG00000085224	Aqu1.227733	Contig13509
ENSG00000089682	N/A	N/A
ENSG00000101811	Aqu1.223249	Contig13456
ENSG00000102144	Aqu1.228147	Contig13512
ENSG00000102383	N/A	N/A
ENSG00000123570	Aqu1.216692	Contig13276
ENSG00000126953	Aqu1.221114	Contig13414
ENSG00000131269	Aqu1.209887	Contig12565
ENSG00000147099	Aqu1.225542	Contig13489
ENSG00000147162	Aqu1.219980	Contig13384
ENSG00000147174	Aqu1.227337	Contig13507
ENSG00000147224	Aqu1.227513	Contig13508
ENSG00000165240	Aqu1.227660	Contig13509
ENSG00000188419	Aqu1.228123	Contig13512
ENSG00000198034	Aqu1.225522	Contig13489
ENSG00000038274	Aqu1.216533	Contig13268
ENSG00000038274	same	same
ENSG00000081791	N/A	N/A
ENSG00000091010	N/A	N/A
ENSG00000113643	Aqu1.228111	Contig13512
ENSG00000123643	Aqu1.228679	Contig13515
ENSG00000123643	same	same
ENSG00000123643	same	same

Human	**Amphimedon queenslandica**	**Amphimedon contig**
ENSG00000131507	Aqu1.209203	Contig12374
ENSG00000155506	Aqu1.214369	Contig13157
ENSG00000155506	same	same
ENSG00000155508	Aqu1.227676	Contig13509
ENSG00000164576	Aqu1.227715	Contig13509
ENSG00000014824	Aqu1.204139	Contig8594
ENSG00000065882	Aqu1.216617, Aqu1.216618	Contig13273
ENSG00000075539	N/A	N/A
ENSG00000075539	N/A	N/A
ENSG00000078140	Aqu1.229828	Contig13521
ENSG00000090989	Aqu1.225032	Contig13484
ENSG00000109189	Aqu1.222734	Contig13448
ENSG00000109680	Aqu1.220963	Contig13411
ENSG00000121892	Aqu1.212912	Contig13047
ENSG00000124406	Aqu1.214243	Contig13148
ENSG00000151806	Aqu1.227159	Contig13506
ENSG00000169299	Aqu1.204505	Contig9067
ENSG00000183783	N/A	N/A
ENSG00000215203	Aqu1.210885	Contig12771

C.4 Python code of the Monte-Carlo simulation

The source for this code is in CD Appendix C>C4>simulation_code.py, add.txt and README.txt

C.5 Python codes for retrieving orthologues from scaffold 13506 of *Amphimedon queenslandica*, *Capitella teleta* and *Lottia gigantea* *genomes*

The source for this code is in CD Appendix C>C5>CteHbxLoc.py, FilterBlastAmphi.py, FilterBlastAMphi2.py, FilterBlastCTE.py, FilterBlastCTE2.py, FilterBlastLGI.py, FilterBlastLGI2.py, LgiHboxLoc.py, Prots13506.py

C.6 BCP Hox PAL extended to the choanoflagelate *M.* brevicollis gene list

Human	Monosiga brevicollis	Monosiga scaffolds
NP_005010.2, NP_000915.1	Monbr1_34427	scaffold_37
NP_002872.1, NP_005393.2	Monbr1_35711	scaffold_2
NP_002482.1	N/A	N/A
NP_056085.1, NP_899200.1	Monbr1_37136	scaffold_10
NP_057287.2, NP_059127.2, NP_001193638.1	Monbr1_9993	scaffold_18
NP_001120793.1, NP_009207.2, NP_006798.1	N/A	N/A
NP_001025167.2	Monbr1_29913	scaffold_40
NP_036360.3, NP_001026849.1	Monbr1_25959	scaffold_12
NP_004932.1, NP_003578.2	Monbr1_19544,Monbr1_17513	scaffold_3
NP_000979.1	N/A	N/A
NP_803190.2, NP_113622.1	N/A	N/A
NP_689953.1	Monbr1_33474	scaffold_20
NP_036423.4	Monbr1_33070	scaffold_15
NP_060559.2	N/A	N/A
NP_060229.3, NP_060292.3	N/A	N/A
NP_001020.2	Monbr1_3665	scaffold_6
NP_065811.1, NP_055867.3	Monbr1_15180	scaffold_4
NP_001193998.1, NP_054757.1	Monbr1_16574	scaffold_7
NP_115729.1	Monbr1_15611	scaffold_5
NP_079457.2	N/A	N/A
NP_954587.2	N/A	N/A
NP_659478.1, NP_060844.2, NP_076961.1	N/A	N/A
NP_079029.3	Monbr1_37859	scaffold_17

Human	Monosiga brevicollis	Monosiga scaffolds
NP_006384.1, NP_006166.3, NP_001157979.1, NP_006139.1	Monbr1_34066	scaffold_29
NP_055126.1	N/A	N/A
NP_001005209.1	N/A	N/A
NP_001017957.1	Monbr1_26903	scaffold_16
NP_005585.1, NP_954984.1	Monbr1_34333	scaffold_34
NP_001028217.1	Monbr1_7690	scaffold_8
NP_055078.1	Monbr1_7044	scaffold_6
NP_003911.2	Monbr1_28544	scaffold_27
NP_659447.1	Monbr1_37690	scaffold_15
NP_055400.1	Monbr1_33674	scaffold_22
NP_612405.2, NP_001096032.1, NP_065865.1	N/A	N/A
NP_001030022.1, NP_835227.1	N/A	N/A
NP_079178.2	N/A	N/A
NP_001120863.1	N/A	N/A
NP_066024.1, NP_997221.2, NP_002290.2, NP_004786.2	N/A	N/A
NP_005799.2, NP_872580.1	Monbr1_34461	scaffold_37
NP_524146.1, NP_524147.2, NP_002467.1	Monbr1_39222	scaffold_43
NP_036232.2	Monbr1_18914	scaffold_8
NP_037473.3	Monbr1_38170	scaffold_21
NP_054859.2	Monbr1_9161	scaffold_14
NP_002786.2	Monbr1_32253	scaffold_9
NP_057399.1	Monbr1_13000	scaffold_52
NP_001106178.1	Monbr1_31879	scaffold_6
NP_036365.1, NP_001165906.1	Monbr1_24227	scaffold_6
NP_848930.1, NP_036229.1	Monbr1_5315	scaffold_2

Human	Monosiga brevicollis	Monosiga scaffolds
NP_009228.2, NP_689484.3, NP_079054.3, NP_940863.3	Monbr1_9070	scaffold_13
NP_064527.1, NP_002480.1	N/A	N/A
NP_001097.2, NP_001607.1	Monbr1_27170	scaffold_17
NP_001194.1, NP_004320.2, NP_001096.1	Monbr1_27170	scaffold_17
NP_060599.1	Monbr1_13418	scaffold_2
NP_001231.2, NP_001232.1	Monbr1_33137	scaffold_16
NP_036565.2	Monbr1_37521	scaffold_14
NP_689609.2, NP_859076.3, NP_060866.2, NP_683759.1	Monbr1_23351,Monbr1_31729	scaffold_4,scaffold_5
NP_003133.1, NP_742067.3	Monbr1_34421	scaffold_36
NP_005792.1	Monbr1_3654	scaffold_2
NP_001001550.1, NP_004481.2, NP_005301.2	N/A	N/A
NP_003066.2, NP_003065.3	Monbr1_32596	scaffold_11
NP_001247.3, NP_001107563.1	Monbr1_26880	scaffold_16
NP_004513.1, NP_004975.2, NP_004512.1	Monbr1_21638	scaffold_21
NP_982288.1, NP_671723.1, NP_071358.1, NP_078828.2	Monbr1_20758	scaffold_8
NP_065970.2, NP_667340.2	Monbr1_16139	scaffold_6
NP_840101.1, NP_003059.1, NP_005976.2	Monbr1_34432	scaffold_37
NP_075559.2, NP_001093894.1	N/A	N/A
NP_060621.3	N/A	N/A
NP_079265.2	Monbr1_32192	scaffold_8
NP_037409.2	Monbr1_25079	scaffold_9

Human	Monosiga brevicollis	Monosiga scaffolds
NP_061167.1, NP_006046.1	N/A	N/A
NP_079095.3	Monbr1_15103	scaffold_4
NP_002586.2, NP_148978.2, NP_002587.2	Monbr1_32324	scaffold_9
NP_003496.1, NP_003498.1, NP_001457.1, NP_003459.2, NP_114072.1	N/A	N/A
NP_066564.2, NP_003875.3	Monbr1_14245	scaffold_3
NP_057551.1, NP_644809.1, NP_054901.1	Monbr1_32962	scaffold_15
NP_001019839.1	N/A	N/A
NP_001926.2, NP_004451.2, NP_001171507.1, NP_001927.3	Monbr1_34096	scaffold_30
NP_002889.1, NP_002888.1, NP_001171182.1	Monbr1_22121	scaffold_2
NP_002147.2	Monbr1_37718	scaffold_16
NP_835455.1	N/A	N/A
NP_004574.2, NP_002859.1, NP_004153.2	Monbr1_34712	scaffold_47
NP_859062.1, NP_004279.3	N/A	N/A
NP_056480.1	Monbr1_32791	scaffold_13
NP_036417.1, NP_653234.2	Monbr1_29429	scaffold_34
NP_689557.1	N/A	N/A
NP_000465.1	N/A	N/A
NP_000113.1	Monbr1_32554	scaffold_11
NP_061854.1	Monbr1_21911	scaffold_2
NP_003343.1, NP_008867.2, NP_008868.3	Monbr1_28202	scaffold_24
NP_000918.2, NP_000434.1, NP_003733.2	Monbr1_19578	scaffold_3

Human	Monosiga brevicollis	Monosiga scaffolds
NP_002197.2, NP_001138468.1, NP_001073286.1	N/A	N/A
NP_055369.1	Monbr1_25282	scaffold_9
NP_004516	N/A	N/A
NP_001002031.1, NP_005166.1, NP_001680.1	N/A	N/A
NP_055475.2	N/A	N/A
NP_001070666.1, NP_775952.4	Monbr1_28955	scaffold_30
NP_001171867.1, NP_060894.2	Monbr1_17480	scaffold_3
NP_001649.1, NP_001650.1	Monbr1_35269	scaffold_26
NP_001186913.1	Monbr1_37042	scaffold_10
NP_689597.1	Monbr1_32801	scaffold_13
NP_002126.2, NP_006177.1	N/A	N/A
NP_060164.3	Monbr1_29375,Monbr1_38850	scaffold_34,scaffold_32
NP_001153218.1, NP_004318.3, NP_060930.3, NP_060757.4	Monbr1_25079,Monbr1_13806	scaffold_9,scaffold_2
NP_006652.1	Monbr1_35957	scaffold_3
NP_065954.1, NP_065986.2, NP_114113.1	Monbr1_24942	scaffold_8
NP_112185.1	N/A	N/A
NP_003629.1, NP_002201.1	N/A	N/A
NP_003208.2	N/A	N/A
NP_001073331.1, NP_001034933.1	N/A	N/A
NP_003143.2, NP_036580.2	N/A	N/A
NP_055117.1	Monbr1_19429	scaffold_2
NP_001073998.2, NP_003378.3	N/A	N/A
NP_116264.2	Monbr1_23622	scaffold_4
NP_001026886.1, NP_076973.1	N/A	N/A
NP_006328.2	Monbr1_28065	scaffold_23

Human	Monosiga brevicollis	Monosiga scaffolds
NP_115970.2, NP_776183.1	N/A	N/A
NP_001408.2	Monbr1_5818	scaffold_3
NP_003876.1, NP_003927.1	N/A	N/A
NP_004498.1	Monbr1_7101	scaffold_6
NP_000828.1, NP_036438.2, NP_071435.2	Monbr1_25917	scaffold_12
NP_853514.1, NP_000288.1, NP_057196.2, NP_001009944.2	Monbr1_31037	scaffold_2
NP_944490.1, NP_003121.1	Monbr1_35370	scaffold_24
NP_061720.2	Monbr1_39292	scaffold_47
NP_002172.2, NP_066382.1, NP_000184.1	N/A	N/A
NP_001012241.1	N/A	N/A
NP_003031.3, NP_005061.2, NP_003750.1, NP_000333.1	N/A	N/A
NP_060579.3	N/A	N/A
XP_001714944.3	N/A	N/A
NP_115766.3	Monbr1_3397	scaffold_3
NP_954699.1, NP_001137381.1	N/A	N/A
NP_002401.1	N/A	N/A
NP_002938.1	N/A	N/A
NP_001245.1	Monbr1_28402	scaffold_26
NP_002602.2, NP_001135858.1, NP_002601.1, NP_002603.1	Monbr1_31036	scaffold_2
NP_079485.1, NP_056445.3	Monbr1_30743	scaffold_2
NP_071903.2, NP_060583.2	N/A	N/A
NP_001164275.1	N/A	N/A
NP_001369.1, NP_004402.1	Monbr1_15003	scaffold_4
NP_036355.2	Monbr1_29200	scaffold_32

Human	Monosiga brevicollis	Monosiga scaffolds
NP_001180242.1	Monbr1_27495	scaffold_19
NP_872327.2	N/A	N/A
NP_001161688.1	N/A	N/A
NP_997254.3	N/A	N/A
NP_055855.2	Monbr1_32550	scaffold_11
NP_006832.1, NP_109599.3, NP_061849.2, NP_722518.2	Monbr1_24806	scaffold_8
NP_004873.3	Monbr1_10588	scaffold_22
NP_055132.2	N/A	N/A
NP_006795.3	N/A	N/A
NP_570857.2	Monbr1_26140	scaffold_13
NP_075567.2	Monbr1_22898	scaffold_3
NP_660298.2	N/A	N/A
NP_001072.2	N/A	N/A
NP_004222.2	Monbr1_34619	scaffold_42
NP_003092.4	Monbr1_20830	scaffold_8
NP_219487.3	N/A	N/A
NP_775491.1, NP_003355.1	Monbr1_28739	scaffold_28
NP_059129.3, NP_001165113.1	Monbr1_15767	scaffold_5
NP_001092303.1	N/A	N/A
NP_055177.2	Monbr1_33402	scaffold_19
NP_005972.1, NP_055214.1	Monbr1_33910	scaffold_26
NP_003165.2	Monbr1_34314	scaffold_34
NP_060546.2	Monbr1_37775	scaffold_16
NP_620158.3	Monbr1_34414	scaffold_36
NP_114109.1	N/A	N/A
NP_110436.1	N/A	N/A
NP_001036111.1, NP_055864.2	N/A	N/A
NP_004068.2	N/A	N/A
NP_036575.1	Monbr1_37986	scaffold_19
NP_003070.3	N/A	N/A
NP_036454.1	Monbr1_14817	scaffold_3

Human	Monosiga brevicollis	Monosiga scaffolds
NP_000879.2	N/A	N/A
NP_001034934.1	Monbr1_15806	scaffold_5
NP_065875.3, NP_001186346.1	Monbr1_30343	scaffold_48
NP_057018.1	Monbr1_35337	scaffold_47
NP_060726.3	Monbr1_30764	scaffold_2
NP_005722.1	Monbr1_37027	scaffold_9
NP_001124435.1, NP_001136117.1	N/A	N/A
NP_001035938.1	N/A	N/A
NP_005928.2	Monbr1_26877	scaffold_16
NP_001121616.1,	Monbr1_26866	scaffold_17
NP_114152.3, NP_001032208.1, NP_055585.1	Monbr1_30343	scaffold_48
NP_000456.2	Monbr1_38485	scaffold_26
NP_689732.2	N/A	N/A
NP_001167596.1	Monbr1_30285	scaffold_47
NP_113609.1	Monbr1_542	scaffold_3
NP_055070.1	N/A	N/A
NP_002148.1	Monbr1_26593	scaffold_15
NP_000989.1	Monbr1_37079	scaffold_10
NP_694453.2	Monbr1_6550	scaffold_4
NP_001078916.1	N/A	N/A
NP_056345.3, NP_114105.1	N/A	N/A
NP_001034782.1	Monbr1_10965	scaffold_25
NP_001034813.2	N/A	N/A
NP_004385.1	N/A	N/A
NP_005250.1, NP_005802.1	N/A	N/A
NP_065207.2, NP_443149.2	Monbr1_27397	scaffold_19
NP_060218.1	Monbr1_31693	scaffold_5
NP_037422.2	Monbr1_33648	scaffold_22
NP_065138.2, NP_116201.7	N/A	N/A
NP_653309.3	N/A	N/A

Human	Monosiga brevicollis	Monosiga scaffolds
NP_009172.2	Monbr1_28152	scaffold_24
NP_001906.3	N/A	N/A
NP_859525.1	Monbr1_21200	scaffold_3
NP_079543.1	N/A	N/A
NP_001478.2	N/A	N/A
EAW58000.1	N/A	N/A
NP_006391.1	Monbr1_32065	scaffold_7
NP_005776.1	N/A	N/A
NP_002778.1	Monbr1_33583	scaffold_21
NP_002038.2	Monbr1_37647	scaffold_15
NP_001171715.1	Monbr1_32385	scaffold_10
NP_001813.1, NP_001020372.2	Monbr1_27594	scaffold_20
NP_009141.2, NP_004473.2	Monbr1_27397	scaffold_19
NP_036203.1	N/A	N/A
NP_004809.2	Monbr1_10827	scaffold_24
NP_005680.1	Monbr1_20835	scaffold_8
NP_060941.2	Monbr1_22194	scaffold_2
NP_079423.1	N/A	N/A
NP_057700.3	Monbr1_32671	scaffold_12
NP_071682.1	N/A	N/A
NP_002078.1	Monbr1_28514	scaffold_26
NP_219481.1	Monbr1_26159	scaffold_13
NP_001183956.1	Monbr1_30960	scaffold_2
NP_065726.1	N/A	N/A
NP_001340.2	Monbr1_38870	scaffold_33

C.7 l-ParaHox PAL extended to the choanoflagelate *M. brevicolis* gene list

Human	Monosiga brevicollis	Monosiga scaffold
ENSG00000032742	Monbr1_11191	scaffold_27
ENSG00000102710	N/A	N/A
ENSG00000102743	Monbr1_35981	scaffold_3
ENSG00000120688	N/A	N/A

Human	Monosiga brevicollis	Monosiga scaffold
ENSG00000120694	Monbr1_34504	scaffold_38
ENSG00000120697	N/A	N/A
ENSG00000120699	Monbr1_34582	scaffold_41
ENSG00000122042	N/A	N/A
ENSG00000132953	N/A	N/A
ENSG00000132963	N/A	N/A
ENSG00000133101	Monbr1_14677	scaffold_3
ENSG00000133105	N/A	N/A
ENSG00000133105	N/A	N/A
ENSG00000133119	Monbr1_38211	scaffold_21
ENSG00000139505	Monbr1_26246	scaffold_13
ENSG00000150456	N/A	N/A
ENSG00000165487	N/A	N/A
ENSG00000172915	Monbr1_8517	scaffold_11
ENSG00000172915	N/A	N/A
ENSG00000244754	Monbr1_25386	scaffold_10
ENSG00000010671	Monbr1_1610	scaffold_2
ENSG00000067177	N/A	N/A
ENSG00000080572	Monbr1_25953	scaffold_12
ENSG00000085224	Monbr1_28926	scaffold_30
ENSG00000089682	N/A	N/A
ENSG00000101811	N/A	N/A
ENSG00000102144	Monbr1_24772	scaffold_8
ENSG00000102383	Monbr1_22137	scaffold_2
ENSG00000123570	Monbr1_35292	scaffold_31
ENSG00000126953	N/A	N/A
ENSG00000131269	Monbr1_20835	scaffold_8
ENSG00000147099	Monbr1_34892	scaffold_2
ENSG00000147162	Monbr1_27585	scaffold_20
ENSG00000147174	Monbr1_23840	scaffold_5
ENSG00000147224	Monbr1_33328	scaffold_18
ENSG00000165240	Monbr1_27752	scaffold_21
ENSG00000188419	Monbr1_17747	scaffold_3
ENSG00000198034	Monbr1_33368	scaffold_19
ENSG00000038274	N/A	N/A

Human	Monosiga brevicollis	Monosiga scaffold
ENSG00000038274	N/A	N/A
ENSG00000081791	N/A	N/A
ENSG00000091010	N/A	N/A
ENSG00000113643	Monbr1_39368	scaffold_54
ENSG00000123643	Monbr1_1039	scaffold_5
ENSG00000123643	Monbr1_1039	scaffold_5
ENSG00000123643	Monbr1_33121	scaffold_16
ENSG00000131507	N/A	N/A
ENSG00000155506	Monbr1_1778	scaffold_6
ENSG00000155506	N/A	N/A
ENSG00000155508	Monbr1_1697	scaffold_2
ENSG00000164576	N/A	N/A
ENSG00000014824	Monbr1_9689	scaffold_16
ENSG00000065882	Monbr1_2868	scaffold_4
ENSG00000075539	Monbr1_22972	scaffold_3
ENSG00000075539	N/A	N/A
ENSG00000078140	Monbr1_39004	scaffold_36
ENSG00000090989	Monbr1_32259	scaffold_9
ENSG00000109189	Monbr1_29678	scaffold_37
ENSG00000109680	Monbr1_31684	scaffold_5
ENSG00000121892	Monbr1_32239	scaffold_8
ENSG00000124406	Monbr1_8524	scaffold_11
ENSG00000151806	Monbr1_9705	scaffold_16
ENSG00000169299	Monbr1_36937	scaffold_9
ENSG00000183783	Monbr1_39362	scaffold_53
ENSG00000215203	Monbr1_37035	scaffold_10

As it is mentioned before scaffold 5 and scaffold 38 of *T. adhaerens* are linked as part of the ParaHox loci.

Trichoplax adharens	Monosiga brevicollis	Monosiga scaffold
TRIADDRAFT_62201	Monbr1_13875	scaffold_2
TRIADDRAFT_51183	Monbr1_35161	scaffold_9
TRIADDRAFT_33760	Monbr1_27644	scaffold_20
TRIADDRAFT_33711	Monbr1_8385	scaffold_11
TRIADDRAFT_33763	Monbr1_32548	scaffold_11

Trichoplax adharens	Monosiga brevicollis	Monosiga scaffold
TRIADDRAFT_62217	Monbr1_34641	scaffold_43
TRIADDRAFT_33746	Monbr1_12454	scaffold_39
TRIADDRAFT_33724	Monbr1_34109	scaffold_30
TRIADDRAFT_33732	Monbr1_32910	scaffold_14
TRIADDRAFT_62226	Monbr1_18559	scaffold_6
TRIADDRAFT_62227	Monbr1_34641	scaffold_43
TRIADDRAFT_5826	Monbr1_32572	scaffold_11

Appendix D

D.1 Fasta files contaning several homeobox genes multiple alignments

The files are in CD Appendix D>D1>
ANTPbfltca_SciLsp.fa
ANTPbfltca_SciLsp.phy
NKS_CDXS_SCI_11_Feb_2013.fa
NKS_CDXS_SCI_11_Feb_2013.phy
NKS_CDXS_SCI4_10_Mar_2013_2.fa
NKS_CDXS_SCI4_10_Mar_2013_2.phy
NKS_CDXS_SCI4_10_Mar_2013.fa
NKS_CDXS_SCI4_10_Mar_2013.phy
Tca_Bfl_sponges.aln
Tca_Bfl_sponges.fa
Sycon_34059.fa
Leucosolenia_70333.fa

D.2 Fasta files of 34059 scaffold of *Sycon* and its proteins

The files are in CD Appendix D>D1>
2815_cdna_prot.fa
13732_cdna_prot.fa
22551_cdna_prot.fa
24615_cdna_prot.fa
25811_cdna_prot.fa
42087_cdna_prot.fa
42474_cdna_prot.fa
nke_SF35-2011-10-31 C21 AS sp6.TXT

D.3 Multiple alignments and phylogenies of the proteins in 34059 scaffold

The files are in CD Appendix D>D1>

Appendix E

E.1 Fasta files of different homeobox genes of *Tribolium* and *Branchiostoma*

The files are in CD Appendix E>E1>
ANTP_Bfl.fa
ANTP_Tca.fa
CERS_Bfl.fa
CERS_Tca.fa
CUT_Bfl.fa
CUT_Tca.fa
HNF_Bfl.fa
LIM_Bfl.fa
LIM_Tca.fa
POU_Bfl.fa
POU_Tca.fa
PRD_Bfl.fa
PRD_Tca.fa
PROS_Bfl.fa
PROS_Tca.fa
SINE_Bfl.fa
SINE_Tca.fa
TALE_Tca.fa
ZF_Bfl.fa
ZF_Tca.fa

E.2 Python codes for retrieving homeobox genes from *Strigamia maritima*

The files are in CD Appendix E>E2>
saturated_list_SMAR_1B.py
saturated_list_SMAR_2.py
saturated_list_SMAR_3.py
saturated_list_SMAR.py

E.3 Multiple alignments and phylogenies of *Strigamia*, *Tribolium* and *Branchiostoma*

The files are in CD Appendix E>E3>

AllHboxes_60aa_GOOD3.fa
AllHboxes_60aa_GOOD3.phy
bootstrap_tree_1000_NJ
distance_tree_NJ
distance_tree_NJ.pdf

E.4 Excel table of orthologues of *Strigamia*, *Tribolium* and *Branchiostoma*

The files are in CD Appendix E>E4>Sma_hboxes_account.xls

E.5 Multiple alignments and phylogenies of each homeobox class orthologues of *Strigamia*, *Tribolium* and *Branchiostoma*

The folders are in CD Appendix E>E5>
ANTP
CERS
CUT
HNF
LIM
POU
PRD
PROS
SINE
TALE
ZF

E.6 Statistical analyses of scaffold 48457

The file is in CD Appendix E>E6> stats_hox3_xlox.xls