HOMEOBOX GENES IN THE DEVELOPMENT AND REGENERATION OF THE CEPHALOCHORDATE *BRANCHIOSTOMA LANCEOLATUM* AND THE POLYCHAETE ANNELID *SPIROBRANCHUS LAMARCKI*

Tom Barton-Owen

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

2019

Homeobox genes in the development and regeneration of the cephalochordate *Branchiostoma lanceolatum* and the polychaete annelid *Spirobranchus lamarcki*

Tom Barton-Owen

University of St Andrews

This thesis is submitted in partial fulfilment for the degree of
Doctor of Philosophy (PhD)
at the University of St Andrews

September 2018

Candidate's declaration

I, Tom Barton-Owen, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 74,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree.

I was admitted as a research student at the University of St Andrews in September 2013.

I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date Signature of candidate

Supervisor's declaration

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

Date Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are 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. We also understand, unless exempt by an award of an embargo as requested below, 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 this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Tom Barton-Owen, have obtained, or am in the process of obtaining, third-party copyright permissions that are required or have requested the appropriate embargo below.

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

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date Signature of candidate

Date Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Tom Barton-Owen, hereby certify that no requirements to deposit original research data or digital outputs apply to this thesis and that, where appropriate, secondary data used have been referenced in the full text of my thesis.

Date Signature of candidate

Table of Contents

Table of Figures

Table of Tables

Acknowledgements

"Did you do that all by yourself?"

– Ildiko, to Simon (about an inflated balloon)

I did not. This PhD would never have been possible without the support I received from my excellent supervisors, Ildiko Somorjai and David Ferrier. I would like to express my immense gratitude for their patience, support, and guidance. Both steered me in the right direction, jogged me along, and gave me the benefit of their advice and incredible scientific knowledge more times than I can count.

Numerous others have been invaluable in the progress of this PhD. Simon Dailey and Réka Szabó produced the transcriptomes on which a large portion of this work is based, and helped me a lot with bioinformatics. Lab members including Simon, Réka, and Myles also taught me molecular techniques. I also thank past and present members of the Ferrier and Somorjai labs for their scientific input, friendship, conversation, and prolonged tea/coffee breaks; Simon, Blincko, Jack, Réka, Morag, Myles, Ashley, & Clara.

This PhD was made possible by a School of Biology PhD Apprenticeship, for which I am extremely grateful.

Thank you to the numerous excellent friends I have in St Andrews for the fun times we've had; Dana, Simon, Scott & the Shire, Masha, and the Friday night pub group.

Finally, thank you very much to my parents, who have given me invaluable support, and to my brother Giles helped me with advice on Python and a steady stream of amusing chat.

FOR POSTERITY: Simon did indeed blow that balloon up all by himself (Dailey, pers. comm.).

Abstract

The development of complex animal morphology requires the extremely sophisticated spatiotemporal coordination of cell behaviour and communication. Homeobox genes encode transcription factors that are deployed in developmental processes to control the expression of other genes in particular locations and contexts. Many homeobox genes are highly conserved and act in similar roles between distantly-related animals that derive from the roles of their ancestral orthologues. The way that these genes have differentially evolved between taxa, and the effect that these changes have on the development and morphology of animals, is critical to our understanding of metazoan evolution. One particular developmental context, the regeneration of missing tissue, offers a unique perspective on evolutionary developmental biology because of its relationship to ontogenic development and its surprising diversity of retention and process between animal taxa.

I examined the homeobox gene content of transcriptomes taken from the mature and regenerating tissue of the post-anal tail of *Branchiostoma lanceolatum*, a well-studied cephalochordate with a highly conserved genome, and the evolutionarily novel operculum of *Spirobranchus lamarcki,* a sedentarian annelid. In *S. lamarcki* regeneration, a diverse variety of homeobox genes is expressed, and the regenerative expression response is substantial. The discovery of several difficult-to-classify homeobox genes lead to the substantial expansion and improvement of the classification of a variety of homeobox genes undergoing unusual rapid and expansive evolution in the Spiralia, including dozens of TALE and PRD class genes, a new orthology group, and a strange *S. lamarcki* Hox gene.

In *B. lanceolatum*, a similar diversity of expressed genes is observed but a milder regenerative response. One transcriptomic sequence in particular, identified as *Pax3/7*, led to the discovery that this well-studied gene has a previously unnoticed duplication in cephalochordates. This discovery has implications for ongoing study of vertebrate and cephalochordate neural plate border evolution.

Publications arising from this thesis

The following publications resulted from work undertaken during the preparation of this thesis;

Derived from work not presented herein:

Plöschner, Martin, Věra Kollárová, Zbyněk Dostál, Jonathan Nylk, Thomas B. Barton-Owen, David E. K. Ferrier, Radim Chmelík, Kishan Dholakia, and Tomáš Čižmár. 2015. 'Multimode Fibre: Light-Sheet Microscopy at the Tip of a Needle'. *Scientific Reports* 5 (December).

Derived from data collected and presented in ***Chapter 3****:*

Barton-Owen, Thomas B., Réka Szabó, Ildiko M. L. Somorjai, and David E. K. Ferrier. 2018. 'A Revised Spiralian Homeobox Gene Classification Incorporating New Polychaete Transcriptomes Reveals a Diverse TALE Class and a Divergent Hox Gene'. *Genome Biology and Evolution* 10 (9): 2151–67.

Derived from data collected and presented in ***Chapter 5****:*

Barton-Owen, Thomas B., David E. K. Ferrier, and Ildikó M. L. Somorjai. 2018. 'Pax3/7 Duplicated and Diverged Independently in Amphioxus, the Basal Chordate Lineage'. *Scientific Reports* 8 (1): 9414.

Data availability

The data analysed in Chapter 3 is deposited under BioProject PRJNA433343. This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GGGS00000000. The version described in this thesis is the first version, GGGS01000000.

1. General introduction

1.1. Regeneration

Regeneration is the post-ontogenic replacement of damaged or severed tissues or structures by an animal. The ability to regenerate is common amongst the Metazoa, but displays a remarkable diversity at large and small taxonomic scales in the degree of regenerative capacity and the mechanism by which it is achieved (Brockes and Kumar 2008; Tiozzo and Copley 2015). Some animals, like some annelid and planarian worms, are able to regenerate complete adults from a small fraction of their bodies, while others, like adult humans, are almost entirely incapable of regeneration beyond homeostasis of their organs. Some animals maintain a population of pluripotent stem cells that is induced to form new tissue while others must remodel their existing tissue. Regenerative ability is also clearly frequently lost or diminished in taxa. These dissimilarities between animals lead to a multitude of evolutionary questions; for example, how did the capacity for regeneration evolve, and how many times? Which common ancestors of extant taxa could regenerate, how did they do it, and how extensively? Why do some animals lose the capacity for regeneration? Why is there such a diversity in the regenerative mechanisms of distantly- or even closely-related animals?

The post-ontogenic regenerative development of a tissue or structure is analogous if not necessarily homologous to the process that produced the original. This equivalence poses a further question; how does the regeneration of a structure relate to its original development? These questions fall within the purview of evolutionary developmental biology, which aims to understand how the evolution of morphological features is influenced by the molecular mechanisms that produce them and how those mechanisms themselves evolve.

1.1.1. Regeneration and ontogenesis

It is expected from a theoretical point of view that regeneration would, in some senses, be a recapitulation of ontogenic developmental processes. It would be bizarre for animals to have one set of developmental processes to produce the structure the first time, and a second unrelated set of processes to produce subsequent versions of the same structure. It is expected that the identities of the cells within the replacement structures are the same insofar as the replacement tissue is a faithful facsimile, and that the mechanisms for placing those cellular identities in the correct patterns to produce the structure are at least very similar.

However, regeneration cannot be a straightforward recapitulation of the complete ontogenic programme in four respects. Firstly, regeneration must only occur when a structure is damaged or severed, and so is likely to have an initiating signal produced by those events that is different to those involved in development. Secondly, there are substantial histological differences between the embryonic tissues in which the original ontogenesis occurs, and the tissues of a mature, injured animal; specifically, the general preponderance of undifferentiated pluri-/multipotent stem cells in the former, and their contrasting paucity in most examples of the latter. Thirdly, ontogenesis always occurs to produce a complete structure, whereas the required replacement tissues are likely to be only a portion of the original tissues. Finally, the completed ontogenic patterning programme typically produces a juvenile structure smaller than the regenerating one, so the process may require scaling up (although regeneration typically produces replacements smaller than the original tissue). Therefore, we expect to find both similarity of developmental control gene deployment and differences to the molecular mechanisms of ontogenesis caused by the different requirements of the two contexts.

1.1.2. Typologies, mechanisms, & principles of regeneration

Regeneration can occur at different levels within the body structures of animals, and is accordingly divided into five categories (Bely and Nyberg 2010). Three of these categories (cell, tissue, and internal organ regeneration) are relatively infrequently considered in regenerative studies because they are common parts of homeostatic body maintenance, the capacities for which are present in humans. The remaining two categories, regeneration of structures like lateral appendages and tails, and the regeneration of the whole adult body (depicted in Figure 1.1), are capacities absent from reptiles, birds, and mammals and all vertebrates (respectively) and are the subjects of intensive research.

1.1.2.1. Whole body regeneration

Whole body regeneration (WBR) is defined as the ability to regenerate any part of the adult body, albeit not necessarily at the same time. This ability is distributed broadly amongst the Metazoa; the only phyla possessing only animals with a confirmed inability for WBR are those belonging to the Ecdysozoa and the Mollusca, Rotifera, Gastrotricha, and Chaetognatha (Bely 2010; Bely and Nyberg 2010).

Planarians are small flatworms belonging to the phylum Platyhelminthes, in the Spiralia (see section 1.3.2.1). They possess a very simple body plan, typified by an unsegmented, acoelomate, dorsoventrally flattened body with an unusual blind-ended gut (*i.e.* lacking an anus). The majority of their body is filled with solid mesenchymal tissue. Planarians exhibit a remarkable morphological plasticity, including the ability to 'de-grow' in response to starvation. However, even more remarkable is their extraordinary capacity for regeneration, which includes their ability to regrow their entire body from a small piece of excised tissue as well as extensive anterior and posterior regeneration. Planarians, particularly *Schmidtea mediterranea* and *Dugesia japonica,* are important models of WBR. Their regenerative capacity is attributed to a population of stem cells called neoblasts, which are maintained throughout the adult mesenchyme, at least some of which are pluripotent (*i.e.* capable of producing any cell type in the adult body, *sensu* Frank, Plickert, and Müller 2009; Mitalipov and Wolf 2009; & Rink 2013). Planarian regeneration is comparatively well-understood (reviewed by Agata *et al.*, 2003; Reddien and Alvarado 2004; Oviedo *et al.*, 2008; Saló *et al.*, 2009; Baguña 2012; Rink 2013; Gehrke and Srivastava 2016, and introduced in more detail below). Other Platyhelminthes and the morphologically-similar but distantly related Acoela (belonging to the Xenacoelomorpha, see section

1.3.2) are used as outgroups and points of comparison to planarian regeneration and neoblast evolution (Gehrke and Srivastava 2016).

Polychaete annelids are another group of important models of regeneration including WBR, though many polychaetes have lost their ancestral ability to regenerate anterior segments (see Bely 2006, 2010; Zattara and Bely 2016), meaning they are no longer capable of whole body regeneration. Annelids also possess a population of cells called neoblasts, but they are unrelated to planarian neoblasts and not as well understood (Myohara 2012; Bely 2014). Annelid regeneration is introduced in more detail in section 1.3.2.3.

The Cnidaria include several established non-bilaterian models of whole body regeneration in the *Hydra* sp., *Hydractinia* sp., and *Nematostella* sp. (reviewed by Bode 2003; Holstein, Hobmayer, and Technau 2003; Lai and Aboobaker 2018). *Hydra* sp. are capable of the remarkable feat of regenerating an intact animal from a collection of disassociated and then reaggregated cells. Cnidaria are an example of an animal that regenerates principally by remodelling its existing tissues (*i.e.* morphallaxis, see section 1.1.2.3).

Tunicates are of interest as models of WBR because they are the only chordates with the capacity. Various *Botryllus* and *Botrylloides* spp. capable of WBR have been studied in this regard (Rinkevich, Shlemberg, and Fishelson 1995; Rinkevich *et al.*, 2007; Voskoboynik *et al.*, 2007; Brown *et al.*, 2009).

1.1.2.2. Structure regeneration

The regeneration of structures is more widespread in the Metazoa than WBR, at least in part because animals seem to be prone to apparently lose the ability to regenerate their anterior while retaining their posterior regenerative ability. Structure regeneration almost always occurs along one of two main body axes; the anteroposterior (AP) axis, as is the case with posterior regeneration, or a proximodistal (PD) axis, as is the case with appendage regeneration. Because of the homology of the AP axis among Bilateria, but the probably-independent evolution of appendages (see section 1.2.3) it may be preferable to categorise regeneration into AP vs appendage regeneration.

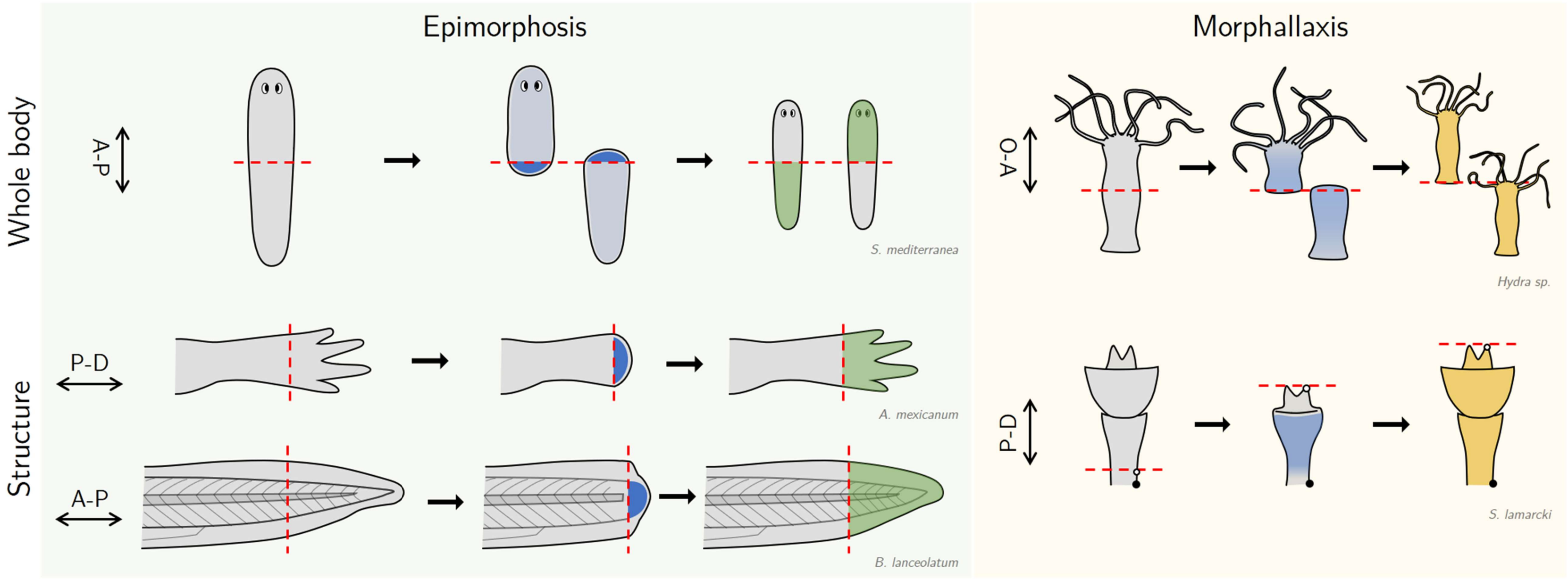


Figure 1.1. Illustration of regenerative types & mechanisms. Planes of amputation are marked with dashed red lines. Regions of cell proliferation are marked in blue throughout; in solid blue where the cell proliferation is concentrated and in pale blue where it is diffuse. In epimorphic regeneration, replacement tissue is marked in green. In morphallactic regeneration, tissue derived from remodelled original tissue and decentralised cell proliferation is marked in yellow. In serpulid operculum regeneration, black and white dots are hypothetical landmarks in the original tissue. A-P = anteroposterior axis. PD = proximodistal axis. O-A = oral-aboral axis. *S. mediterranea* = *Schmidtea mediterranea*; *A. mexicanum* = *Ambystoma mexicanum*; *B. lanceolatum* = *Branchiostoma lanceolatum; S. lamarcki* = *Spirobranchus lamarcki.* Operculum illustration adapted from Szabó and Ferrier (2014).

The most well-studied models of structure regeneration are the urodele amphibians, specifically the salamanders *Notophthalmus viridescens* (a newt) and *Ambystoma mexicanum* (the axolotl). The urodeles are the only tetrapods capable of regenerating their limbs and tails in adulthood, although that capacity is also found in the tadpoles of frogs, including the model species *Xenopus laevis*, and, in a limited sense, even the foetuses and newborn infants of mice and humans. The mechanisms of urodele limb regeneration are introduced further below. Other chordate models of structure regeneration include fin regeneration in the zebrafish *Danio rerio* and tail regeneration in the cephalochordate *Branchiostoma lanceolatum*, one of the two model regenerative systems used in this study (section 1.4.1).

Annelid models of structure regeneration are generally those which have lost the ability to regenerate anteriorly (covered in detail in section 1.3.2.3). There are also models of annelid appendage regeneration; specifically, the segmental parapodia of *Platynereis dumerilii,* and the unpaired novel head appendage of *Spirobranchus lamarcki,* the other model of regeneration used in this study (section 1.3.1). There are also arthropod regenerative model organisms, including the red flour beetle *Tribolium castaneum* (Shah, Namigai, and Suzuki 2011; Lee *et al.*, 2013)*,* the cricket *Gryllus bimaculatus* (*e.g.* Bando *et al.*, 2013), and the amphipod crustacean *Parhyale haiwaiensis* (Konstantinides and Averof 2014; Kao *et al.*, 2016; Alwes, Enjolras, and Averof 2016).

1.1.2.3. Morphallaxis & epimorphosis

The classification of regeneration into whole-body *versus* structure does not describe the mechanisms of regeneration particularly usefully. One system used for this purpose are the terms ‘morphallaxis’ and ‘epimorphosis,’ which were originally a bipartite classification proposed by T. H. Morgan (1901). Morphallactic regeneration encompasses those examples in which the replacement tissue was produced by the remodelling of mature tissue from within the animal, whereas epimorphic regeneration was that in which the replacement tissues were produced from new cells that were produced in a blastema (depicted in Figure 1.1).

Although this distinction has proved to be enduring, it has unsurprisingly not survived for more than a century of scholarship without its meaning evolving (Agata, Saito, and Nakajima 2007). As classic and new models of regeneration were developed and studied with increasingly sophisticated molecular tools, it has become apparent that few systems fall strictly within one category or the other. For example, *Hydra* sp. is considered a morphallactic system because it regenerates via the remodelling of tissues proximal to the cut site (Figure 1.1, top right). However, cell proliferation is involved in an altered, regeneration-specific pattern reminiscent of a blastema (Holstein, Hobmayer, and David 1991; Holstein, Hobmayer, and Technau 2003), even if regeneration can proceed without cell proliferation (Holstein, Hobmayer, and Technau 2003). Similarly, examples of planarian (Figure 1.3, top left) and annelid posterior regeneration are considered epimorphic because of their formation of a regenerative blastema (see below), but the expression profile of genes that specify axial identity in the original tissue changes rapidly in response to tissue loss, as if they are being genetically if not physiologically remodelled. To describe the hybridity of these systems, ‘morphallactic’ and ‘epimorphic’ have become adjectives applied to specific aspects, mechanisms, and observations within the regenerative programme of animals rather than to only the programmes themselves. However, regenerative programmes can still be categorised as broadly epimorphic or morphallactic by the presence of a blastema.

The blastema

A blastema (solid blue regions, Figure 1.1; between dashed red and blue lines, Figure 1.2) is a region underlying the wound epithelium in which are collected populations of undifferentiated, proliferating cells in a structureless mass. In various systems the cells comprising the blastema can be derived from stem cell populations like planarian and annelid neoblasts, or vertebrate satellite cells (see below), or from terminally-differentiated myofiber or fibroblast cells from the original tissue that are induced to dedifferentiate and re-enter the cell cycle in response to the wounding (see below). Although these cells are undifferentiated, they are rarely pluripotent, although a subset of planarian neoblasts are (Tiras and Aslanidi 2016; Zeng *et al.*, 2018). In vertebrates, blastema cells are multipotent but lineage-restricted (Zielins *et al.*, 2016). Having collected the necessary population of progenitor cells, the blastema grows, is polarised, and differentiates and patterns into the replacement tissue. Blastemas are widely distributed amongst the Metazoa, including in the Xenacoelomorpha (*e.g.* Srivastava *et al.*, 2014), annelids, planarians, phoronids (Emig 1973), echinoderms (Dupont and Thorndyke 2007), hemichordates (Rychel and Swalla 2009), cephalochordates (Somorjai *et al.*, 2012), amphibians, and mammals (Seifert *et al.*, 2015).

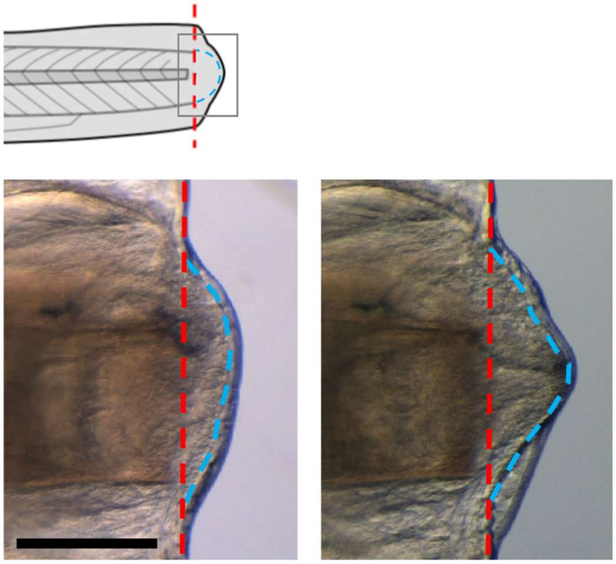


Figure 1.2. The regenerative blastema of *B. lanceolatum*. Above: Illustrative reference of the approximate position of the photographs of the amputated post-anal tail. Below: photographs of the regenerative blastema ten days (left) and fourteen days (right) after regeneration. The blastema is between the red dashed line (the plane of amputation) and the blue dashed line, which indicates the blastemal-epidermal boundary. Scale bar = 250 µm. Images reproduced with the permission of S. Blincko.

1.1.2.4. Intercalation & progressive specification

In addition to their work in reconsidering the definitions of epimorphosis and morphallaxis, Agata, Saito and Nakajima (2007) sought to reframe thought in regenerative biology by suggesting a new principle that they suggested could provide a theoretical basis for all regenerative processes; that all regeneration proceeds by a process of ‘distalization’ and ‘intercalation’. In the first process, distalization, the identity of the tissue at or near the plane of amputation is respecified to the most distant identity of the original tissue (illustrated in Figure 1.3, above). This process results in an incongruous juxtaposition of cells with disparate axial identities that are never adjacent in intact tissues (pink line & arrow, Figure 1.4). They postulated that this juxtaposition would induce the intercalation of medial identities until there was no more incongruity of identity, at which point axial patterning would be complete. This theory is cogent and was consistent with the previous and subsequent evidence until Roensch *et al.* (2013) found that axial identity genes in regenerating salamander limbs are expressed in a proximal-to-distal order rather than a distal-to-intermediate order (*i.e.* they are progressively specified, Figure 1.3, below). A reconciliation with previous studies showing early distal cell identity specification in salamander limbs (*e.g.* Echeverri and Tanaka 2005) and a reconsideration of distalization/intercalation as a unifying principle has yet to be attempted.

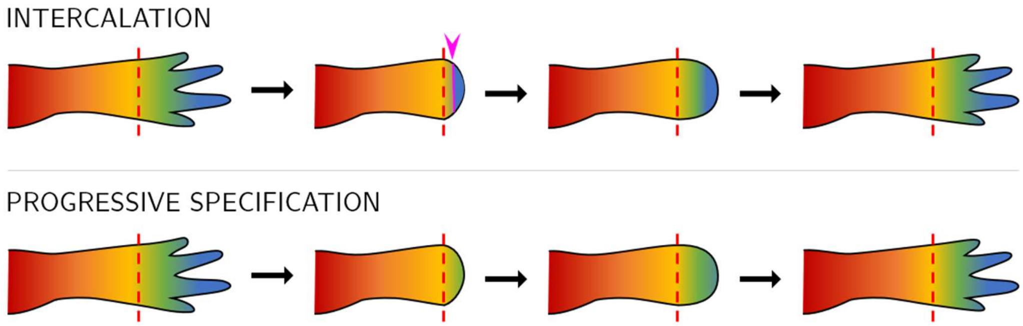


Figure 1.3. Illustration of hypothetical intercalation and progressive specification in the urodele limb. The proximodistal identity of the tissue is represented by a rainbow gradient where red indicates most proximal identity and blue, most distal. Pink arrow and line = incongruous juxtaposition of proximal and distal identities.

In order to understand the evolution of developmental processes like ontogenesis and regeneration, a widespread and fruitful approach which is integral to the field of evolutionary developmental biology is to study the mechanisms that orchestrate them. These mechanisms consist of complex networks of genetic interactions controlled by specific types of gene. Among these control genes, and the specific focus of this study, are homeobox genes, which are introduced in the next section.

1.2. Homeobox genes

Homeobox genes are a superfamily of genes that encode a homeodomain, a DNA-binding domain usually 60 amino acids in length. The homeodomain is extremely highly evolutionary conserved; homeodomains have been identified in almost all eukaryotic life (Derelle *et al.*, 2007; Mendoza *et al.*, 2013) and within animals and plants, orthology groups dating back hundreds of millions of years can usually be detected based solely on the sequence of the homeodomain. Homeodomain-containing proteins usually act as transcription factors (TFs), capable of inducing or suppressing the expression of other genes based on their ability to bind specific nearby DNA sequences and influence the transcription apparatus. This ability is utilised by cells to produce complex behaviours by the context-dependent or spatiotemporally-specific activation of genetic toolkits. To this end, homeodomain TFs, members of other TF superfamilies, and proteins involved in processes like signal transduction are assembled into intricate gene regulatory networks (GRNs) responsible for orchestrating these cellular repertoires (Davidson and Erwin 2006; Davidson and Levine 2008).

One of the most evolutionarily significant of the abilities afforded by TFs is complex embryonic development, a trait coincidental with TF repertoire complexity in embryophytes and metazoans (Mendoza *et al.*, 2013). Homeobox genes are deployed in many roles that are fundamental to the ability to develop, including specifying cellular positional identity along the various body axes and controlling the proliferation and terminal differentiation of cells into their final type. They are also implicated in the evolution of important innovations in the evolution of development, including segmentation (Chipman 2010), the head (Monsoro-Burq 2015; Kuratani, Kusakabe, and Hirasawa 2018), and paired appendages (Gehrke and Shubin 2016; Panganiban *et al.*, 1997). In both their most ancient and more recently acquired roles, the involvement of particular homeobox gene families in the GRNs responsible for making particular structures has been deeply conserved in metazoan evolution, although the precise configuration of GRNs is much more flexible. This deep conservation has made it possible to detect evidence for the homologous or independent origins of processes, features, and cell types found throughout the tree of life, functioning as an extremely powerful tool for shedding light on the evolution of developmental processes.

1.2.1. The homeodomain & transcription factor activity

The homeodomain is a globular domain, usually 60 amino acids in length, capable of DNA-binding (depicted in Figure 1.4) and protein-protein interaction. It comprises three alpha helices, the latter two of which form a helix-turn-helix motif. The homeodomain interacts with DNA *via* numerous contacts between the third helix and the major groove of the DNA, while action of the unstructured N-terminal tail in the minor groove stabilises the interaction (asterisk in Figure 1.4; Bürglin and Affolter 2016; Gehring, Affolter, and Bürglin 1994). Homeodomains target with high affinity a short (4-6 nucleotide) AT-rich motif usually centred on the nucleotide phrase 'TAAT'. The sequence of the homeodomain does affect binding specificity, but not by a particularly large degree (Berger *et al.*, 2008; Jolma *et al.*, 2013; Bobola and Merabet 2017). The degree to which homeodomains themselves are capable of high-affinity target sequence-specific binding is not considered sufficient to explain the apparent specificity of homeodomain protein activity. Instead, cooperative and synergistic binding with other DNA-binding domains in the same proteins, cooperation with other transcription factors, interaction with cofactors including the formation of hetero- and homo- dimers and oligomers, the influence of chromatin landscapes, and arrays of low-affinity binding sites (Crocker *et al.*, 2015) have all been found to be necessary to achieve the specific regulation of target genes (reviewed by Bürglin and Affolter 2016; Bobola and Merabet 2017).

1.2.2. Homeobox gene evolution

The homeobox genes of modern animals are divided into orthology groups called families, which by convention are defined as the modern orthologues and paralogues of a single homeobox gene in the common ancestor of the Bilateria (Holland, Booth, and Bruford 2007, for taxonomy, see section 1.3.2). There are also homology groups of homeobox genes that are more taxonomically-restricted and therefore do not meet this definition, like the Hox9-Hox14 genes of vertebrates (see section 1.4.4). In some cases these are within the scope of homeobox families (*i.e.* the vertebrate Posterior Hox genes all belong to the Hox9-14/AbdB family), but in others (*e.g.* broad Hox/ParaHox homology) exist outwith or between families. The families (and non-family homology groups) are grouped into 11 classes (ANTP, CERS, CUT, HNF, LIM, POU, PRD, PROS, SINE, TALE, and ZF), membership of which is on the basis of broader detectible homology; in many cases, this is on the basis of possessing other domains and domain structures; in some cases these domains are found in non-homeobox genes (*e.g.* the PRD domain, zinc finger) and others, they are found only in homeobox genes of that class (*e.g.* the CUT, POU, and PROSPERO domains). Some families possess atypical homeodomain sequences (*e.g.* TALE, PROS). Some classes contain only a single family (CERS & PROS).

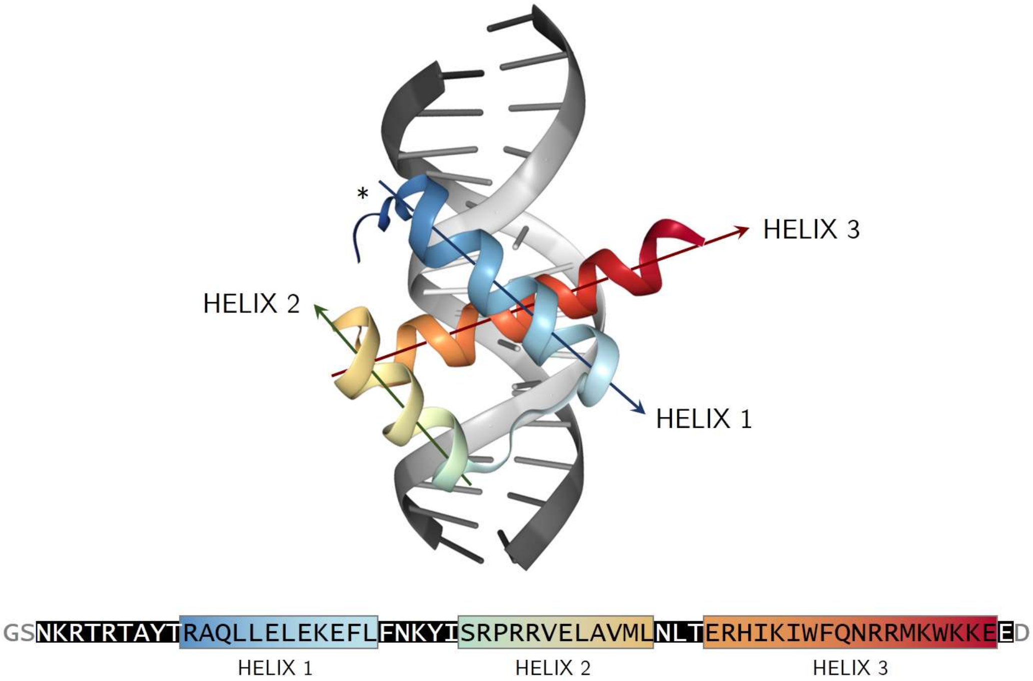

*
HELIX 3
HELIX 2
HELIX 1
GSNKRTRTAYTRAQLLELEKEFLFNKYISRPRRVELAVMLNLTERHIKIWFQNRRMKWKKEED
HELIX 1
HELIX 2
HELIX 3

Figure 1.4. The homeodomain in complex with DNA. Above: The isolated Pdx1 homeodomain (a Pdx/Xlox family member) in complex with DNA. The asterisk indicates the site of the stabilising minor groove interaction of the unstructured N-terminal tail. The crystal structure was solved by Longo, Guanga and Rose (2007), and adapted from a visualisation produced using the NGL Viewer (Rose *et al.*, 2018) from PDB record 2H1K. Below: the sequence of the visualised structure, with corresponding coloration of the helices. The underlying black box indicates the 60 amino acids of the homeodomain. Grey letters outside the black box are not part of the homeodomain.

The earliest orthological distinction that can be made in homeobox evolution is between typical 60 amino acid homeodomains and homeodomains with a three amino acid extension in the loop between the first and second alpha helix (Bertolino *et al.*, 1995; Bharathan *et al.*, 1997; Burglin 1997), called a TALE (Three Amino-acid Loop Extension) class homeodomain (Bertolino *et al.*, 1995) or a homeobox KN domain. This variant is found in all major eukaryotic lineages, and has been ascribed to the eukaryote ancestor (Derelle *et al.*, 2007; Mendoza *et al.*, 2013); as such, analyses of the early evolution of homeobox genes often consider TALE to be a major superfamily rather than just a class of homeobox gene. No other orthology grouping has been found that is not restricted to within Holozoa or plants (Figure 1.5).

Among the Holozoa, it is possible to detect orthology groups which represent the foundations of some other major classes of (non-TALE) homeobox genes, probably including CERS, LIM, POU, and PRD (Sebé-Pedrós *et al.*, 2011). The holozoan ancestor's complement of possibly less than ten homeobox genes (Sebé-Pedrós *et al.*, 2011) expanded radically to 17-20 in the eumetozoan/demosponge ancestor, 61-62 in the cnidarian/bilaterian ancestor, and 82 in the nephrozoan ancestor (Figure 1.5) (Larroux *et al.*, 2008). In the ancestral metazoan, the gene proto-classes (including those above and others unique to the Metazoa, including ANTP, SINE and HNF) were probably organised into a giga-cluster. This fragmented in the lineage predating the ancestor of the Bilateria, so that the bilaterian ancestor probably had several ANTP-class clusters; a SuperHox cluster, containing the Hox and Hox-linked genes, a related ParaHox cluster (both deriving from an ancestral ProtoHox cluster), a possible NK cluster, containing the Nk (except Nk2.1 and Nk2.2) and the NK-linked (NKL) genes, and an Nk2 cluster; as well as a SINE cluster and a possible PRD cluster (Ferrier 2016).

The majority of this gene gain probably happened in the form of tandem or small-scale duplication. This mode of gene duplication would lead to the formation of clusters, which might then be retained because of evolved constraints in the form of co-regulation of the clustered genes (as evident in the Hox and ParaHox clusters of many modern animals). It has been hypothesised that the Hox, ParaHox and NK clusters were responsible for anteroposterior patterning of neural, gut, and mesodermal tissues respectively in the bilaterian ancestor (Holland 2013). However, it is also possible that the various fragmentary clusters found in the genomes of extant animals, from which these presumptive ancestral clusters are deduced, actually represent secondary clustering produced by genomic rearrangement and subsequently retained for the same adaptive reasons (Ferrier 2016).

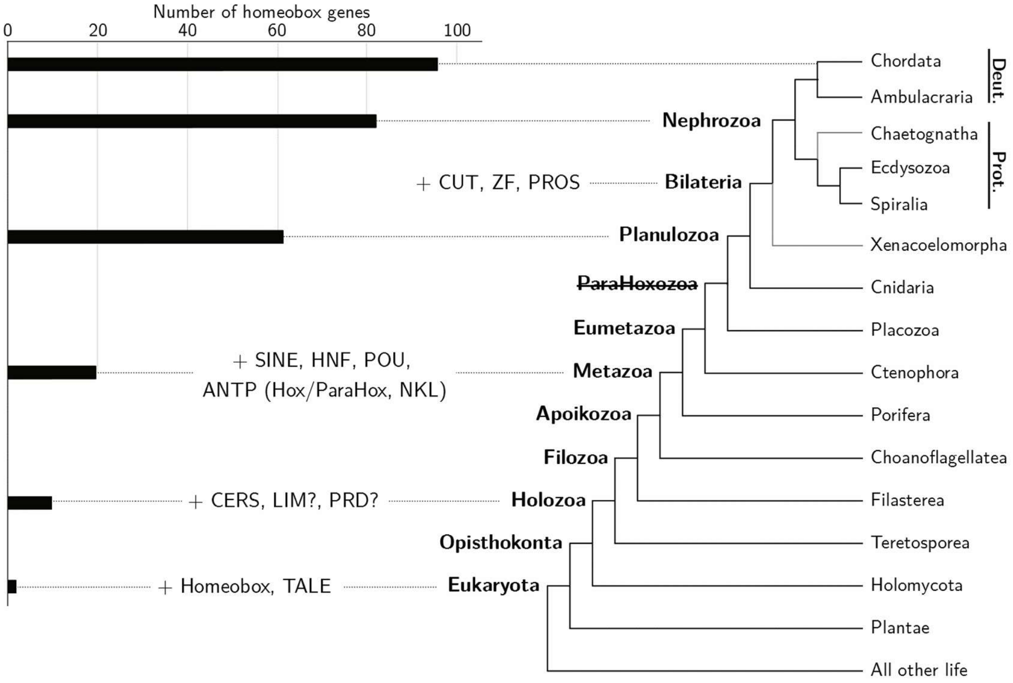


Figure 1.5. Homeobox gene evolution in Eukaryota. Estimated homeobox gene number of the most recent common ancestor of selected clades on the right is indicated with a bar chart on the left. Homeobox classes and other significant orthology groups are indicated when their first members appear in extant descendants. Cladogram topology & clade names are based on information in Ryan *et al.* (2010), Shen *et al.* (2015), Cannon *et al.* (2016), Rouse *et al.* (2016), Feuda *et al.* (2017), Hehenberger *et al.* (2017), Simion *et al.* (2017), & Paps (2018). ParaHoxozoa is struck out because of the discovery of ParaHox genes outwith this clade. Bilaterian phylogeny and terminology is discussed in section 1.3.2. Information concerning the evolution of homeobox classes is derived from the following sources: CERS (Sebé-Pedrós *et al.*, 2011); ANTP (& Hox/ParaHox), PRD, & HNF (Ferrier 2016); LIM (Sebé-Pedrós *et al.*, 2011; Ferrier 2016); SINE (Ferrier 2016; Paps and Holland 2018); POU & NKL (Paps and Holland 2018); CUT, ZF, & PROS (Larroux *et al.*, 2008; Brauchle *et al.*, 2018 [preprint]). Deut. = Deuterostomia; Prot. = Protostomia.

The deep evolutionary history of the homeobox superfamily, and the dynamics with which homeobox genes continue to evolve, is most likely intimately linked to their association with developmental processes. Homeobox genes are underrepresented amongst genes deriving from small-scale duplications (SSDs) and overrepresented amongst those deriving from whole genome duplications (WGDs) (Blomme *et al.*, 2006; Hakes *et al.*, 2007; Huminiecki and Heldin 2010; Makino, Hokamp, and McLysaght 2009; Leite *et al.*, 2018), due to a complex plurality of mechanisms relating to the dosage balance between the duplicate product and the rest of the genome, which is preserved by loss in SSDs but retention in WGDs (Conant, Birchler, and Pires 2014).

Support for this notion can also be seen in homeobox evolution in animals in which radical reconfiguration from ancestral developmental programmes has occurred, like the nematode *Caenorhabditis elegans* and the tunicate *Ciona intestinalis,* both of which have a simplified morphology which they produce with an invariant cell lineage (*i.e.* one in which cell fate is largely determined by its lineage, not by intercellular signalling) (Holland and Gibson-Brown 2003; Sulston and Horvitz 1977; Sulston *et al.*, 1983). These animals have undergone concomitant rapid mitochondrial genome evolution and nuclear genome reduction, including unusual homeobox evolution; specifically, loss of some homeobox families (Hench *et al.*, 2015; S. Wada *et al.*, 2003) gain and rapid evolution of others (Hench *et al.*, 2015), fragmentation of Hox clustering (Hench *et al.*, 2015; Ikuta *et al.*, 2004; Spagnuolo *et al.*, 2003), and redeployment of homeobox genes from more typical global temporal patterns into later development (Hench *et al.*, 2015; Schep and Adryan 2013).

Homeobox gene evolution is typically highly conservative. This is particularly true of bilaterian families, which are constrained by their mostly indispensable roles in development. However, so far there have been very few (if any) genomes to have undergone a rigorous homeobox gene survey that lack previously-undescribed homeobox genes. Although this is doubtless to some extent an artefact of the numerical disparity between global species richness and rigorous homeobox gene surveys, it is illustrative of the fact that the birth of new, taxonomically-restricted homeobox genes is much more common than the loss of established gene families, and examples can be found at every level of taxonomy, from the mammal-specific genes (Leidenroth and Hewitt 2010) to genes that appear in *C. elegans* but not in other members of the *Caenorhabditis* genus (Hench *et al.*, 2015). These genes often evolve, like the earliest homeobox genes, via tandem or small-scale duplication followed by asymmetric divergence. As divergence proceeds, signals of detectable orthology (*e.g.* protein sequence, intron position, and synteny) are progressively lost, until the gene appears to be an orthologue- and paralogue-less orphan. Usually these are still classifiable to within a homeobox class, but sometimes diverge so extremely that they fall entirely outside known classes.

1.2.3. Homeobox genes in axial specification

Homeobox genes are used to mediate and integrate many types of upstream signals and regulatory mechanisms, and innumerable target genes and networks, which together are used to orchestrate the incredibly complex and precise spatiotemporal pattern of gene expression required to develop a single zygotic cell into a mature animal. A complete review of these roles is beyond the scope of this thesis, and they are mentioned below where they are relevant to the study of evolution or regeneration. However, the roles of homeobox genes in patterning body axes – particularly the AP and PD axes – are generally relevant to the study of their roles in regeneration and are reviewed briefly below.

1.2.3.1. Anteroposterior patterning

One of the inaugural, most famous, and most significant findings of evolutionary developmental biology is the deep homology underlying AP axis patterning between all bilaterian life. The signal of homology is found in Hox genes, the first group of homeobox genes to be discovered, and one with remarkable properties.

Hox genes were first discovered in *Drosophila melanogaster,* an extensively-studied dipterid fly developmental model. Mutations to these genes were found to produce ‘homeotic’ phenotypes, that is, in which the complete identity of one body part was switched to another (Garcia-Bellido and Lewis 1976; Lewis 1978). These genes were discovered to contain the homeobox, named after homeotic mutations (Gehring 1985) and to be arranged in two clusters (Kaufman, Lewis, and Wakimoto 1980; Sánchez-Herrero *et al.*, 1985). Orthologues of these genes were soons found in less derived insects (Beeman 1987) and vertebrates (reviewed by Akam 1989), and eventually in a wide variety of other animals (*e.g.* annelids [Snow and Buss 1994]; cephalochordates [Garcia-Fernàndez and Holland 1994]; basal ecdysozoans [Grenier *et al.*, 1997]; planarians [Orii *et al.*, 1999]; cnidarians

[Finnerty and Martindale 1999]; and Xenacoelomorpha [Cook *et al.*, 2004; Fritzsch *et al.*, 2008]).

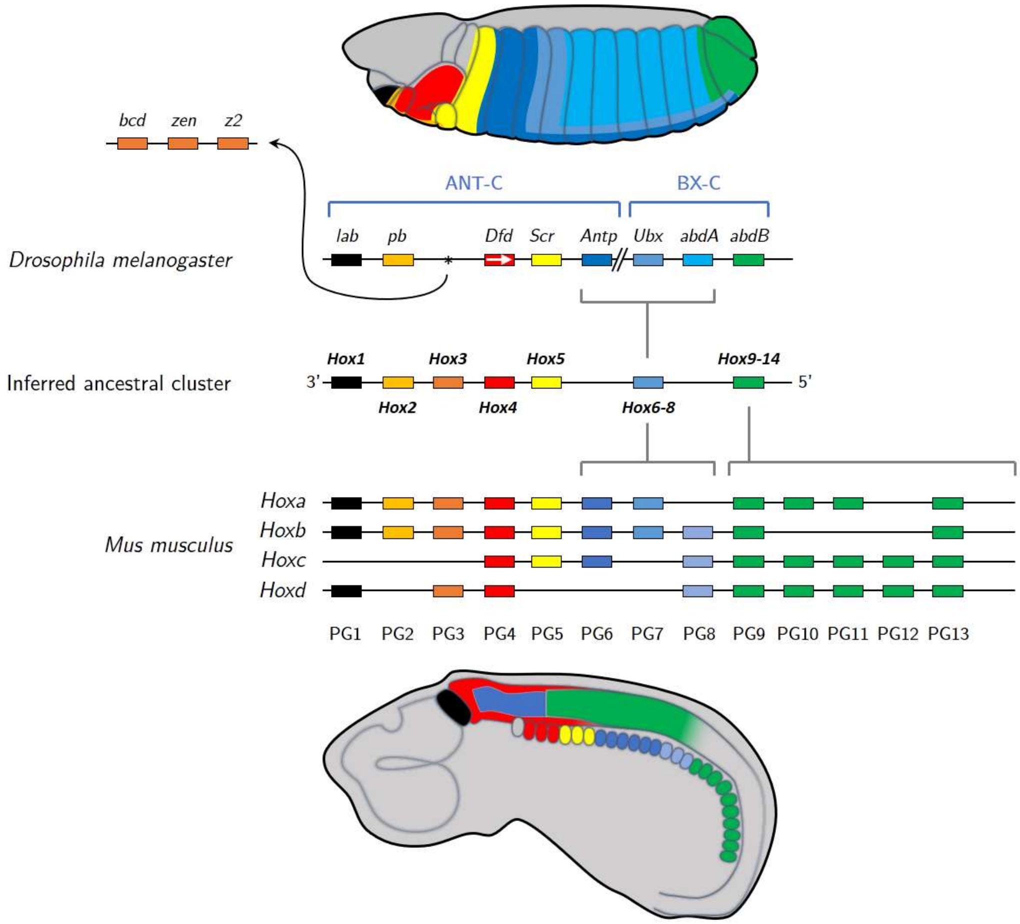


Figure 1.6. The Nephrozoan Hox cluster, showing the clustering and spatial colinearity of expression in two modern model organisms and the inferred ancestral cluster. All genes are orientated right-to-left unless otherwise indicated (white arrow). Adapted from Mallo & Alonso (2013) and Carroll (1995), and modified per Stauber *et al.* (1999), Zeltser *et al.* (1996), and Balavoine *et al.* (2002) to correct the omission of mouse HoxB13, an ancestral *Hox3* and the *D. melanogaster Hox3* paralogues, and per Gaunt (2015) to show gene and cluster orientation. Top and bottom: illustrative drawings of *D. melanogaster* larva and *M. musculus* embryo respectively, showing the regions of corresponding gene expression. ANT-C = Antennapedia complex; BX-C = Bithorax complex; PG = paralogy group.

From these data, it became apparent that in nephrozoan animals (*i.e.* Protostomia + Deuterostomia; see section 1.3.2), Hox genes fall into seven gene families (Figure 1.6); Hox1/lab, Hox2/pb, Hox3/zen, (*i.e.* Anterior Hox genes) Hox4/Dfd, Hox5/Scr, Hox6-8/AbdA (*i.e.* Medial Hox genes) and Hox9-14/AbdB (*i.e.* Posterior Hox genes). These were ancestrally arranged in a single cluster, but these have frequently become secondarily disorganised, split, atomised (reviewed by Duboule 2007) or, in vertebrates, been duplicated fourfold followed by subsequent paralogue loss (see section 1.4.2). The evolution of Hox gene clusters has been extensively studied (reviewed by Balavoine, de Rosa, and Adoutte 2002; Monteiro and Ferrier 2006; Duboule 2007; Ferrier 2010; Lanfear 2010; Moreno and Martínez 2010; Ikuta 2011; Mallo and Alonso 2013; Pascual-Anaya *et al.*, 2013; Gaunt 2015; Barucca, Canapa, and Biscotti 2016; Ferrier 2016; Hrycaj and Wellik 2016; Thomas-Chollier and Martinez 2016 and others).

Hox genes are expressed along the anteroposterior axis of most bilaterian animals in a position that corresponds to their place in the cluster (Figure 1.6); genes at the 3' end of the cluster are expressed in the anterior, the central genes in the centre of the developing body, and the genes at the 5' end in the posterior. This phenomenon is referred to as spatial colinearity. In addition, but less commonly found, the Hox cluster also exhibits temporal colinearity, in which genes are expressed in a 3'-to-5', anterior-to-posterior order. The taxonomic distribution of these phenomena suggests that both may have been present in the ancestor of Nephrozoa, but their precise nature, their mechanistic basis relative to Hox clustering and to one another, their adaptive and functional significance, and the forces under which they evolved and were subsequently lost, are still incompletely understood.

1.2.3.2. Appendage development & proximodistal patterning

Various taxa of bilaterian animals possess symmetric paired appendages, including vertebrates, arthropods, and polychaete worms, all of which are segmented. The initiation of the body-wall outgrowths from which these are produced, and the proximodistal polarization/axis establishment of these appendages are controlled by a set of common genes, including the homeobox gene families Meis (a.k.a. homothorax), Pbx (a.k.a. extradenticle) and Dlx (a.k.a. Distal-less) (Grimmel, Dorresteijn, and Fröbius 2016) in all three groups. The paired appendages of vertebrates and arthropods are usually considered to be non-homologous (Panganiban *et al.*, 1997; Gehrke and Shubin 2016; Tabin, Carroll, and Panganiban 1999; Pueyo and Couso 2005; Winchell, Valencia, and Jacobs 2010; Winchell and

Jacobs 2013; Grimmel, Dorresteijn, and Fröbius 2016), meaning that these genes were convergently co-opted (probably as part of a GRN like the AP head axis patterning network, Lemons *et al.*, 2010). Relationships between structures with homology of control genes, but which do not share homologous derivation from an ancestral structure, have been referred to as 'deep homology' (Shubin, Tabin, and Carroll 2009) or 'homocracy' (Nielsen and Martinez 2003).

Although there are similarities in the initiation of body wall outgrowths and outgrowth polarization, the proximodistal patterning of arthropods and vertebrates is dissimilar (Pueyo and Couso 2005), with the former continuing to use Meis/Pbx orthologues in the proximal tissue and Dlx orthologues in the distal tissues (reviewed by Tweedt 2017) and the latter using a group of signalling pathways (Delgado and Torres 2017) that eventually activate nested posterior Hox genes (Zakany and Duboule 2007; Mariani 2010).

1.2.4. Homeobox genes in regeneration

Regeneration represents an interesting puzzle in evolutionary developmental biology because of its fascinating relationship to the ontogenesis of the same structures it recreates (see section 1.1). Because of their anciently conserved roles relating to developmental specification of cell fate and axial identity, homeobox genes can be a useful tool to understand this relationship. The extent to which homeobox genes recapitulate their ontogenic roles in regeneration can be seen as a proxy for the extent to which regeneration is a general recapitulation of ontogenesis. However, regeneration cannot be a direct recapitulation of ontogenesis because of the necessity of producing new cells in a terminally-differentiated environment, as well as in the initiation of regeneration in the event of injury (see section 1.1.1). Homeobox genes also perform regeneration-specific roles like in wound healing and stem-like cell maintenance and recruitment.

1.2.4.1. Homeobox genes in adult tissue

A wide variety of homeobox genes are constitutively expressed in adult tissue, though this type of expression is infrequently surveyed systematically. A recent study in human tissue transcriptomes found a wide diversity of both broad (predominantly in the

HNF, TALE, ZF and CERS classes) and tissue-specific (predominantly in the ANTP, PRD, LIM and POU classes) expression of homeobox genes (Dunwell and Holland 2016). In many cases, the genes ares being deployed to regulate the minutiae of cellular homeotic function (*e.g.* Charest-Marcotte *et al.*, 2010). However, in many animals, including mammals (Donoghue *et al.*, 1992; Grieshammer, Sassoon, and Rosenthal 1992; Chang *et al.*, 2002; Rinn *et al.*, 2006; Ackema and Charité 2008; referenced & reviewed by Wang, Helms, and Chang 2009), annelids (Bakalenko *et al.*, 2013), and planaria (Reddien 2011; Currie *et al.*, 2016), Hox genes are expressed in adult tissues in various configurations, where they function to specify the axial identity of adult tissue long after their roles in developmental axial specification (see section 1.2.3).

The specifics of this constitutive expression vary substantially between clades. In nereid polychaetes, Hox and ParaHox genes are expressed in gene-specific and mostly colinear nested gradients along the AP axis, largely within the ventral nerve cord and ectoderm (Hox) and digestive tract (ParaHox) (Bakalenko *et al.*, 2013; M. A. Kulakova, Cook, and Andreeva 2008). *Alitta virens* undergoes constant, almost life-long post-larval growth via the posterior addition of segments, and these expression patterns shift to maintain AP proportionality as the animal elongates. There are substantial differences between the deployment of Hox genes to pattern larval and post-larval segments (Bakalenko *et al.*, 2013). The planarian *Schmidtea mediterranea* expresses its Hox genes in both AP-axial and radial regions, but their relationship to developmental expression is not yet known (Currie *et al.*, 2016).

Adult mammal fibroblasts and muscle cells have an extremely stable record of their identity in the form of the gene-specific regions of Hox expression. The differential and combinatorial deployment of Hox genes from their four Hox clusters, (derived from the 2R-WGD, see section 1.4.2), is thought to be sufficient to give positional information in the anteroposterior, dorsoventral, and proximodistal body axes, but oddly is substantially simplified from the Hox gene code in developmental axial specification (Wang, Helms, and Chang 2009). The mature Hox code appears to be locked into cell lineages epigenetically, and is resistant to change both via transplantation and *ex vivo* culture (Rinn *et al.*, 2006).

Unfortunately, far less is known about constitutive adult homeobox expression in urodele amphibians, the only tetrapods capable of adult limb and tail regeneration. In axolotl, a LIM-class homeobox gene, *Lmx1b*, was found expressed at low levels in the dorsal skin of the mature arm, in a region apparently analogous to its expression in limb development. Unlike the cell-lineage stability of mammal Hox genes, expression of this gene was reactive to ventralizing retinoic acid treatment (Satoh and Makanae 2014). The central nervous system of the tail of the newt *Pleurodeles waltl* was found to express *Hoxa9, HoxC12* and *HoxC13* approximately colinearly in adulthood, and upregulate these genes strongly during its regeneration (Nicolas *et al.*, 2003).

1.2.4.2. Homeobox genes in wound healing

Wound healing is a complicated process in which the wound must be obstructed to prevent the loss of blood, cleared of damaged tissue and potentially exogenous matter, and then healed over to re-establish epidermal integrity and prevent infection. In vertebrates, several homeobox genes including members of the Hox3 (Uyeno *et al.*, 2001; Mace *et al.*, 2005, 2009), Hox8 (Jain *et al.*, 2008), Hox13 (Stelnicki *et al.*, 1998), Msx (Carlson, Bryant, and Gardiner 1998; Yeh *et al.*, 2009), and Prrx (Stelnicki *et al.*, 1998; White *et al.*, 2003) families/paralogy groups have been implicated in scarless wound-healing contexts (reviewed by Kuri, Belek, and Boudreau 2011; & Kachgal, Mace, and Boudreau 2012), including in epidermal migration. The differences between wound healing with regenerative capacity and without (leading to scarring) has been associated with the expression of several of these genes (particularly Msx, reviewed by Yokoyama 2008) and other immunological processes (Godwin and Rosenthal 2014).

1.2.4.3. Homeobox genes in stem-like cells & blastemas

Pax3/7

In most vertebrates (Le Grand and Rudnicki 2007) and probably crustaceans (Konstantinides and Averof 2014) and cephalochordates (Somorjai *et al.*, 2012), a population of proliferative undifferentiated cells, which derive from muscle progenitor cells, persists into the adult. In the event of tissue damage, these cells leave their quiescent state and start to proliferate, migrate to the wound site, and produce a substantial portion of the cells that comprise the blastema. These cells are initially specified and stably maintained in their satellite state by the expression of members of the *Pax3/7* family, a PRD class homeobox gene. These cells can be genetically ablated in vertebrates by deleting *Pax7,* a *Pax3/7* paralogue (Murphy *et al.*, 2011), and doing so is deleterious to the regenerative capacity of mice (Murphy *et al.*, 2011; Sambasivan *et al.*, 2011; Frederic Relaix and Zammit 2012). Once in the blastema, satellite cells are multipotent (Asakura, Rudnicki, and Komaki 2001) but principally contribute to the regeneration of skeletal muscle. However, *Pax3/7*$^{+}$ satellite cells are not the principle component of all vertebrate regeneration systems (Sandoval-Guzmán *et al.*, 2014). Satellite cells and *Pax3/7* are covered in greater detail in Chapter 5.

Msx

Another strategy for producing undifferentiated, proliferative cells to populate the blastema is by the dedifferentiation of mature multinucleate myotubes proximal to the wound site into mononucleate cells (reviewed by Frasch 2016; Wang and Simon 2016). In both mice (Odelberg, Kollhoff, and Keating 2000) and newts (Kumar *et al.*, 2004), this can be induced by *Msx*, an NK-linked ANTP-class homeobox gene, expression of which also leads to proliferation (Odelberg, Kollhoff, and Keating 2000). Dedifferentiated muscle cells are a major contributor to the newt blastema (Echeverri, Clarke, and Tanaka 2001), but their progeny seem to contribute only to the replacement muscle tissue and not other cell types (Sandoval-Guzmán *et al.*, 2014). The roles of *Msx* in initiating muscle dedifferentiation in regeneration is thought to be related to its developmental deployment to inhibit differentiation of muscle progenitor cells (and various other progenitor cells, including bone and neural crest) (Hu *et al.*, 2001; Kuwajima *et al.*, 2004; Lee, Habas, and Abate-Shen 2004; Brunelli and Cossu 2005; Ryoo, Lee, and Kim 2006; Han *et al.*, 2007; Bhatt, Diaz, and Trainor 2013).

Despite the ability to induce muscle dedifferentiation via ectopic *Msx* expression in mice, the use of myotube dedifferentiation in regeneration has only been observed in newts. However, *Msx* also has important and vital roles in regeneration in zebrafish

(Akimenko *et al.*, 1995; Thummel *et al.*, 2006), frogs (Beck, Christen, and Slack 2003; Beck *et al.*, 2006), and foetal and newborn mice (Reginelli *et al.*, 1995; Han *et al.*, 2003). This expression is associated with its regulation by and mediation of Bone Morphogenetic Protein (BMP) and Fibroblast Growth Factor (FGF) signalling (Beck, Christen, and Slack 2003; Beck *et al.*, 2006; Han *et al.*, 2003; Yokoyama 2008). In this role, *Msx* expression appears in the wound epidermis (including in non-regenerative wounds) and in the blastema (Carlson, Bryant, and Gardiner 1998; Koshiba *et al.*, 1998; Endo, Tamura, and Ide 2000; Yokoyama 2008) where it may regulate the growth of the blastema (Park, Ju, and Kim 2009). This broader role has been related to its developmental expression in limb buds (Lallemand *et al.*, 2005), in the Apical Ectodermal Ridge (where present) and underlying mesoderm (Carlson, Bryant, and Gardiner 1998; Koshiba *et al.*, 1998; Yokoyama 2008). *Msx* has also been described in the regenerative blastema of a cephalochordate (see section 1.4.1 and Somorjai *et al.*, 2012), regulating planarian neoblasts in the cephalic blastema in concert with BMP (Mannini *et al.*, 2008), and even in the transdifferentiation of cnidarian muscle (Galle, Yanze, and Seipel 2005), indicating that these roles may predate the vertebrate lineage.

Other homeobox genes

Prrx-1, a member of the *Prrx* PRD-class homeobox gene family, is expressed early in multipotent cells in the amphibian blastema, but expression ceases in fully differentiated cells (Suzuki *et al.*, 2005, 2007; Satoh *et al.*, 2011; Yokoyama 2008; Lehrberg and Gardiner 2015). It is not clear what role *Prrx-1* is performing in this context, but it may relate to its roles in patterning developing limb buds (Kuratani *et al.*, 1994; Martin and Olson 2000).

Another gene of interest is *Oct4*, a paralogue of the POU class V homeobox genes, which are restricted to vertebrates (Onichtchouk 2016). This gene acts, as part of a core regulatory circuitry including *Sox2, Klf4, c-myc* and *Nanog*, as the ‘gatekeeper of pluripotency’ in embryonic stem cells, preventing them from differentiating away from a pluripotent state (Pesce and Schöler 2001; Boyer *et al.*, 2005; Tantin 2013; Onichtchouk and Driever 2016). Mature cells can be induced *in vitro* to form pluripotent stem cells by activating various *Oct4*-inclusive subsets of the regulatory circuit (Takahashi and Yamanaka 2006; Takahashi *et al.*, 2007; Hanna *et al.*, 2008; Huangfu *et al.*, 2008; Kim *et al.*, 2009; Adachi and Schöler 2012). Although the Pou5 genes are restricted to vertebrates, roles of POU homeobox genes in stem cell identity may be deeply conserved (Gold, Gates, and Jacobs 2014), as a putative *Pou4* orthologue (Gold, Gates, and Jacobs 2014) and homologues of many up- and downstream members of the vertebrate pluripotency regulatory network were found to be involved in planarian stem cells (Önal *et al.*, 2012) and a putative *Pou3* orthologue (Gold, Gates, and Jacobs 2014) was found to be capable of inducing stem-ness in mature cnidarian cells (Millane *et al.*, 2011). However, *Oct4*/other POU genes are not strongly associated in the literature with pluripotency-related expression in vertebrate regeneration, even when other core pluripotency network genes were found in newts and echinoderms (Maki *et al.*, 2009; Mashanov, Zueva, and García-Arrarás 2014). *Oct4* was found to be necessary but not upregulated in zebrafish fin regeneration (Christen *et al.*, 2010) and unnecessary for the self-renewal of adult somatic stem cells in mice (Lengner *et al.*, 2007). It is thought that *Oct4*'s limited involvement in vertebrate regeneration is because blastemal tissues do not achieve a state of pluripotency (Christen *et al.*, 2010), but it is of interest to regenerative systems where pluripotent cells might be involved.

Several other homebox genes have been found to be involved in stem cells in vertebrates; many of these are vertebrate-specific orthology groups, including *Nanog*/*ENK* (Chambers *et al.*, 2003; Mitsui *et al.*, 2003; Wang *et al.*, 2003; Booth and Holland 2004), *Hesx* (Webb *et al.*, 1993) and *Rhox* genes (Song *et al.*, 2016). Hox genes also control bone-marrow-derived stem/progenitor cells in mammal cutaneous healing and homeostatic organ repair/regeneration (reviewed by Mahdipour and Mace 2011; Seifert *et al.*, 2015; Wells and Watt 2018). *HoxA9* and *HoxA13* are expressed in axolotl regeneration (Gardiner *et al.*, 1995; Gardiner and Bryant 1996) and *HoxC10* in *Xenopus* (Christen *et al.*, 2003) before the appearance of the blastema, indicating potential roles in dedifferentiation or cell recruitment. *HoxC10* expression has been observed in the blastema-building stage of axolotl forelimb regeneration, even though this gene is not involved in forelimb ontogenesis

(Carlson *et al.*, 2001; Bryant, Endo, and Gardiner 2002). *HoxC13* paralogues have been found to affect blastema size in zebrafish (Thummel *et al.*, 2007).

Various homeobox genes have been identified as distinctive of planarian neoblasts, including two Posterior Hox genes, *Pbx*, *Nk2.1*, *Meis*, *zeb*-1, and other genes from the ZF, CUT, and LIM classes (Önal *et al.*, 2012; Abnave *et al.*, 2017), of which *zeb-1* has been shown to control neoblast migration (Abnave *et al.*, 2017)

1.2.4.4. Homeobox genes in axial identity and patterning in regeneration

The deployment of homeobox genes to govern the axial patterning of regenerating tissue is an area of extremely active research. The most commonly identified genes in these roles are the Hox genes and their cofactors, TALE-class genes.

Annelids & planarians

Nereid annelids and planarians respond similarly to posterior amputation with regards to their expression of Hox genes. In the errantian polychaete *Alitta virens*, *Lox5, Lox2* and *Post2* (two Medial Hox genes and a Posterior one; see Chapter 3) and the ParaHox gene *Cdx* respond within four hours of amputation, during the wound healing process but long before any visible regeneration is underway (Novikova *et al.*, 2013; Kulakova, Cook, and Andreeva 2008). *Lox2* and *Post2*, which are in the adult constitutively expressed in the posterior and in the experimental conditions had their entire domain of expression excised, are re-expressed *de novo* in the neural tissue of the segments made posterior-most by amputation. The expression domains of *Hox2, Hox3* and *Hox5* respond within 10 hours, the former two reestablishing their expression domains at the extreme posterior, followed by *Hox7* (18 hpa) and *Hox1, Hox4* and *Lox4* changing only after the appearance of new structures (Novikova *et al.*, 2013). Differences reported in the regeneration of the related errantian *P. dumerilii* (Pfeifer, Dorresteijn, and Fröbius 2012) were attributed to methodological differences by Novikova *et al.* (2016) but illustrate that much is left to be discovered about annelid regenerative homeobox deployment.

In contrast, the regeneration of *Capitella teleta,* a member of a different major annelid clade (the Sedentaria – see section 1.3.2.2 and Figure 1.10), is very different, being typified instead by an almost completely static Hox code, in which three Hox genes shift their anterior expression boundary by 1-2 segments after amputation. However, similarities of *Hox3* and *Post2* expression in the blastema was observed (de Jong and Seaver 2016). Direct comparison between nereids and capitellids is rendered difficult by the contrast between the general homonomy of segments in the former and the morphological distinction between thoracic and abdominal segments in the latter.

The planarian *Dugesia japonica* has been observed to have a broadly similar regenerative response in Hox regulation to *A. virens*. *Lox5* (Orii *et al.*, 1999) and *Abd-B* (a Posterior Hox gene) (Nogi and Watanabe 2001) are expressed in an anteroposterior gradient with strong expression at the posterior. When the posterior is removed, the gradient shifts rapidly such that the new posterior is expressing *Lox5* and *Abd-B* strongly and the proportionality of the gradient is restored. Similarly, in the excised posterior, *Lox5* and *Abd-B* expression are abolished in the new anterior tissues so that the gradient is restored. These genes are probably responding to shifts in, and mediating, Wnt/β-catenin signalling, the defining signal of anteroposterior identity in *S. mediterranea* (Gurley, Rink, and Alvarado 2008; Gurley *et al.*, 2010). Data are still lacking on the expression of Hox genes in *S. mediterranea* regeneration.

Vertebrates

Urodele regenerating limbs are repatterned by *Meis1*, *Meis2, HoxA9, HoxA11,* and *HoxA13* (Gardiner *et al.*, 1995; Mercader, Tanaka, and Torres 2005; Mercader *et al.*, 2009; McCusker and Gardiner 2013; Nacu *et al.*, 2013; Roensch *et al.*, 2013; Roselló-Díez *et al.*, 2014, reviewed by Stocum 2017). Specifically, *HoxA9*, *HoxA11* and *HoxA13* are expressed in a nested proximal-to-distal spatiotemporal sequence, determining the positional identity of the blastema cells (Roensch *et al.*, 2013), so that *HoxA9*$^{+}$/*HoxA11*$^{-}$/*HoxA13*$^{-}$ expression specifies the upper arm, *HoxA9*$^{+}$/*HoxA11*$^{+}$/*HoxA13*$^{-}$ expression specifies the lower arm, and *HoxA9*$^{+}$/*HoxA11*$^{+}$/*HoxA13*$^{+}$ expression specifies the foot. This proximal-to-distal colinear pattern emerges from an earlier blastemal pattern which ‘violates’ colinearity (Gardiner *et al.*, 1995), but which is thought to be unrelated to patterning. *HoxA11* and *HoxA13* expression has also been observed in the ‘patternless’ regeneration of juvenile individuals of the frog *Xenopus laevis*, which form a spiked blastema but are not capable of patterning it into a replacement limb (reviewed by Suzuki *et al.*, 2006).

Meis1 and *Meis2* also contribute to the proximal identity of the limbs (Mercader, Tanaka, and Torres 2005), under the control of retinoic acid gradients (Mercader *et al.*, 2009), and controlling the *HoxA13* expression domain (Roselló-Díez *et al.*, 2014). These findings are significant because of their substantial similarity to expression patterns observed during the developmental patterning of the limb bud. However, differences have been observed; for example, the important limb developmental Hox and Meis regulator and cofactor *Pbx* (reviewed by Capellini, Zappavigna, and Selleri 2011) does not seem to play an important role in regeneration (Mercader *et al.*, 2009).

Understanding the roles of homeobox genes in regenerative processes requires studying their deployment in specific organisms. The choice of these organisms is of paramount importance because only relatively few can be selected to aid our understanding of the evolution of the breadth of animal diversity. In this study, I have employed two animals from distantly related phyla and with dissimilar modes of regeneration. The next two sections introduce these organisms, their life history, morphology, evolution, regenerative process, and genetic resources.

1.3. *Spirobranchus lamarcki*

Spirobranchus (formerly *Pomatoceros*) *lamarcki* is a polychaete annelid belonging to the Serpulidae, a group of sessile tube-building worms. After its free-swimming trochophore stage, *S. lamarcki* settles, metamorphoses, and builds calcareous tubes on hard substrates in the intertidal areas of European shores. From the mouth of these tubes it extends its branchial crown, an array of tentacles used for filter-feeding and gas exchange. When the animal perceives a threat from predation or other mechanical damage, it rapidly withdraws its the branchial crown into the tube, plugging the mouth with a specialised head appendage; the operculum (illustrated in Figure 1.7).

The operculum comprises a distal cup-like structure terminating in a concave calcareous plate (bearing 0-3 central spines) which fits in the circular tube mouth; and a muscular filament or peduncle by which the cup is attached to the anterior dorsal thorax. The cup and peduncle are separated by a groove formed by an inwardly intruding flange of cuticle and epidermis. The peduncle is approximately triangular in cross-section, contains nervous tissue, a blood vessel (Bubel 1983b), and muscle (Bubel 1983a) and is marked by alternate bands of black and white pigmentation. The proximal edge of the proximal-most band of black pigmentation is the site of the Easy Break Point (EBP), a plane through the peduncle at which the animal can autotomise the appendage in the event of distal injury (Bubel and Thorp 1985).

1.3.1. *S. lamarcki* regeneration

Following autotomy or dissection at the EBP, *S. lamarcki* can regrow its operculum over the course of approximately one to two weeks (illustrated in Figure 1.7, detailed in Bubel and Thorp 1985a, 1985b; Szabó and Ferrier 2014). The process of regeneration starts with a rapid contraction of the cut surface to minimise blood loss, followed by healing. Within the first 24 hours, the apices of the triangular amputation surface begin to protrude (eventually becoming the spines of the regenerating opercular plate) and the stump elongates. The medial region of the elongating stump swells by 1 day post amputation (dpa) and continues to grow, developing an identifiable rim and cup shape by 3 dpa. At this point, calcification starts to be visible at the base of the spines and continues across the plate as it becomes more distinct in shape. The groove separating the cup and peduncle appears at the onset of calcification and the ‘wings’ that project from the lateral sides of the peduncle start to appear later (5 dpa). From this point on, pigmentation appears and growth is limited to the maturation of existing structures.

The initial elongation of the peduncle stump involves little cell proliferation (Szabó and Ferrier 2014). A rapid increase in non-regionalised cell proliferation in the presumptive peduncle and cup regions is seen in the morphogenesis of the cup and rim regions, though proliferation is not found in the spine, plate and rim. In late regeneration, proliferation seems to become restricted to the wings and the edges of the peduncle from which they protrude. Early proliferation is restricted to the epidermis but later appears in the interior structures (excluding cup region connective tissue). Proliferation is found to occur just proximally to the amputation site but not elsewhere in the animal.

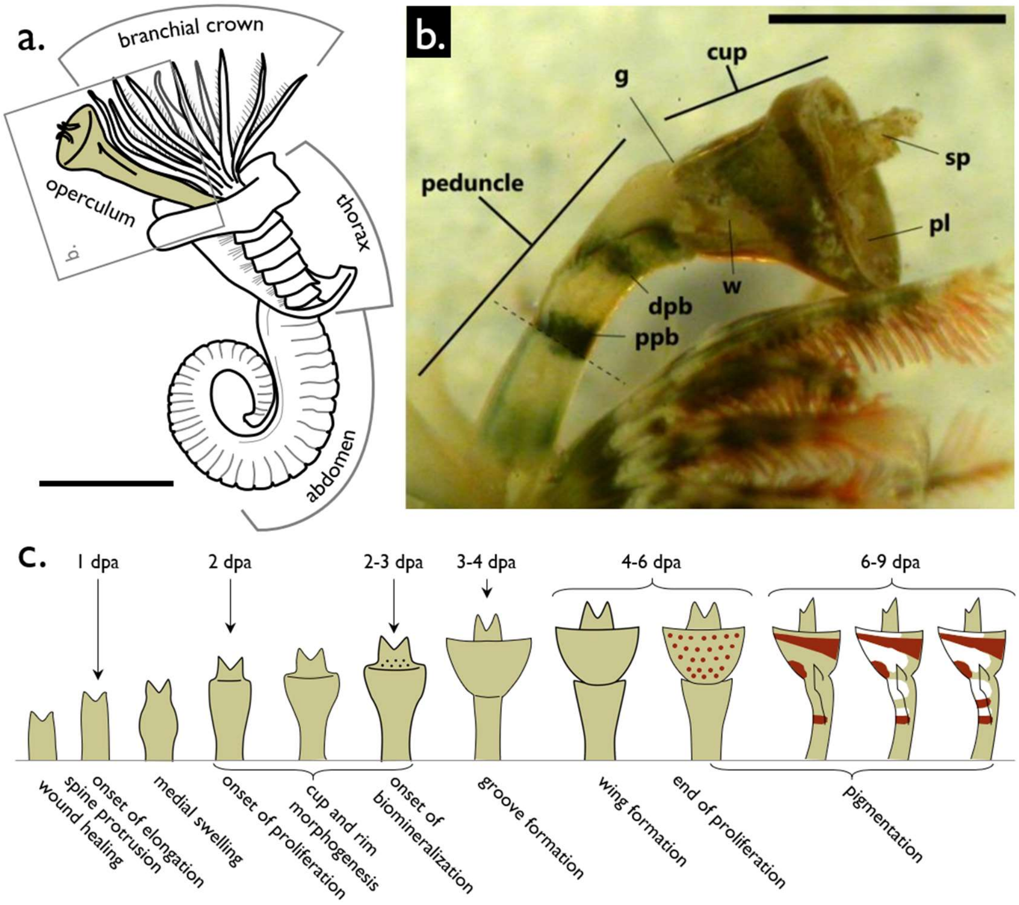


Figure 1.7. *S. lamarcki* morphology and regeneration. **(a)** Illustrative line drawing of *S. lamarcki* adult gross morphology. The approximate location and orientation of the view presented in (b) is indicated with a grey box. Scale bar approximately 2-4 mm. **(b)** The morphology of the *S. lamarcki* operculum. g = groove; sp = spine(s); pl = plate; w = wing; dpb = distal pigment band; ppb = proximal pigment band. The dashed line indicates the Easy Break Point. Scale bar approximately 1 mm. **(c)** Illustrative line drawings of the process of *S. lamarcki* operculum regeneration, including indications of the approximate timescale (above) and morphological and cellular processes (below). The right-most three drawings are lateral views; the rest are dorsal/ventral views. **(b)** adapted from Szabó (2015) and **(c)** adapted from Szabó & Ferrier (2014).

S. lamarcki operculum regeneration does not employ a blastema. Substantial (and some functionally important) regions of the replacement tissue (*i.e.* the spines, plate, rim, and connective tissue of the cup region) seem to derive from the remodelling of previously proximal tissue with no apparent contribution from new cells (Szabó and Ferrier 2014), and cellular proliferation is found in most developing replacement tissue other than those deriving from the amputation surface. This regenerative process can therefore be considered classically morphallactic (see section 1.1.2.3).

The process of *S. lamarcki* regenerative biomineralization is also a topic of interest (see sections 1.3.2 and 1.3.4). Growth of the plate proceeds by the emergence, growth and eventual merging of smooth mineral tiles, which become predominantly aragonitic as regeneration proceeds (Szabó 2015).

S. lamarcki is also capable of regenerating the individual radioles of its branchial crown within a similar timescale (Miles & Ferrier, unpublished). There is some indication that these structures may utilise a different regenerative process, possibly including a blastema-like structure (Ferrier, pers. comm.). In contrast, *S. lamarcki* possesses a poor capacity for posterior regeneration of abdominal segments, which proceeds very slowly, and no capacity for anterior regeneration (Miles & Ferrier, unpublished).

A previous member of the Ferrier laboratory, Dr R. Szabó, produced a transcriptome derived from the tissues of the mature, early regenerating (2 dpa) and late regenerating (6 dpa) operculum using the Illumina HiSeq2000 platform (Szabó 2015; Szabó and Ferrier 2015). This transcriptome is the subject of a detailed homeobox survey in the present study.

1.3.2. Annelid evolution, regeneration and genomics

In order to study *S. lamarcki* from an evolutionary developmental viewpoint, it is necessary to contextualise it within annelid evolution and the Annelida within broader evolutionary history. For this purpose, an overview of these histories is given below.

The Bilateria are a taxonomic group distinguished by the synapomorphy of bilateral symmetry, although some groups (*e.g.* the echinoderms) have secondarily partially lost this trait. Bilaterians were previously neatly divided into two groups. The Deuterostomia, united by their developmental derivation of the anus from the blastopore and subsequent mouth development from non-blastopore cells (hence, *Deutero-stomia*, ‘second mouth'), comprises the Ambulacraria (Hemichordata + Echinodermata) and the Chordata (Tunicata + Cephalochordata + Vertebrata). The second group, the Protostomia, are united by the ancestral trait of producing the mouth from the blastopore (*Proto-stomia*, 'first mouth'), although some clades have deuterostomic or amphistomic development (Harzsch, Müller, and Perez 2015). This group contains two major groups; the Ecdysozoa (animals that grow by ecdysis), consisting of the Arthropoda, Tardigrada, Onychophora, and several groups of worm-like animals including the Nematoda; and the Spiralia (animals with spiralian cleavage) or Lophotrochozoa (see section 1.3.2.1).

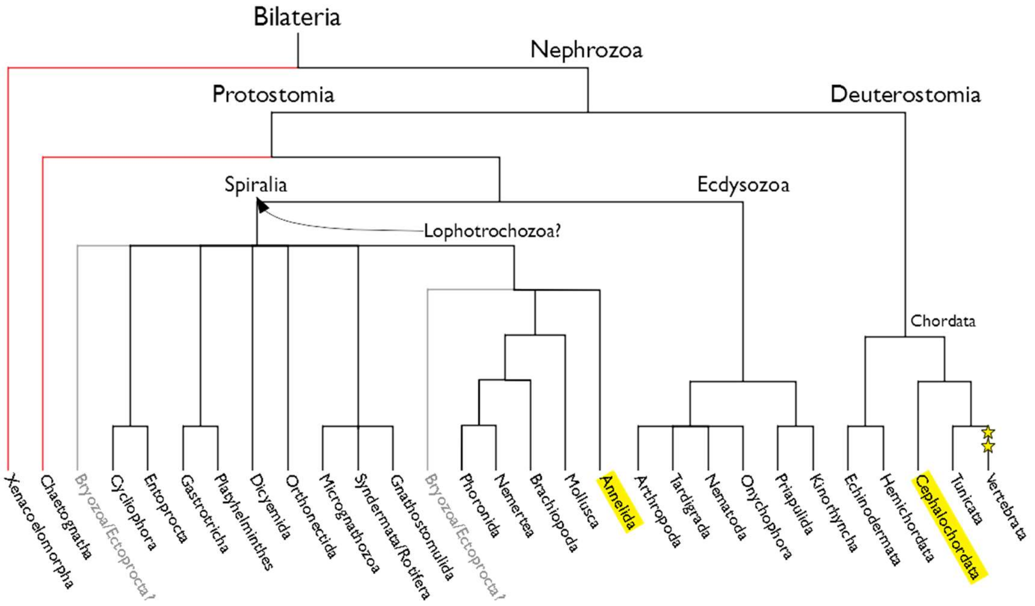


Figure 1.8. Cladogram of the taxonomic relationships of the Bilateria. Difficult-to-classify taxa are marked with red lines. Bryozoa, which are difficult to place and on which the nomenclature of the Spiralia/Lophotrochozoa depend, are marked with grey lines and text. The phyla/subphyla to which the principle model animals studied herein belong are highlighted yellow. The two whole genome duplications in the vertebrates are marked with yellow stars; paleopolyploidy events are not marked in other groups. The placement of the Xenacoelomorpha is based on Cannon *et al.* (2016) and Rouse *et al.* (2016). The placement of the Chaetognatha is based on Shen *et al.* (2015). The internal topology of the Spiralia is based on Luo *et al.* (2018); of Ecdysozoa, on Giribet & Edgecombe (2017); and of Deuterostomes, on Putnam *et al.* (2008).

The neat topologies described above are complicated by the presence of several cryptic, difficult-to-classify taxa. Recent evidence suggests that the cryptic clade of Xenacoelomorpha, consisting of morphologically and phylogenetically difficult-to-classify marine worms, which was previously placed in the Deuterostomia, actually belongs as the sister group to the Protostomia + Deuterostomia (*i.e.* the Nephrozoa) (Cannon *et al.*, 2016; Rouse *et al.*, 2016). The Chaetognatha also defy easy classification; they exhibit deuterostomic development but have been consistently placed among the Protostomia (Helfenbein *et al.*, 2004; Helmkampf, Bruchhaus, and Hausdorf 2008; Papillon *et al.*, 2004; Shen *et al.*, 2015) albeit inconsistently placed within it. The literature is currently divided between considering them members of the Ecdysozoa (Perez, Müller, and Harzsch 2014), members of the Gnathifera (Rotifera + Chaetognatha), in the Spiralia (Fröbius and Funch 2017), and sisters to all other protostomes (*e.g.* Marlétaz *et al.*, 2006; Shen *et al.*, 2015). A summary of the phylogeny of the Bilateria is presented in Figure 1.8.

1.3.2.1. The Spiralia & the Lophotrochozoa

The internal phylogeny of the Spiralia is not well-resolved. A sub-clade containing at least the Annelida, Mollusca, Brachiopoda, Nemertea, and Phoronida and excluding most other Spiralia is usually recovered in recent analyses (Cannon *et al.*, 2016; Kocot *et al.*, 2017; Laumer *et al.*, 2015; Luo *et al.*, 2018; Struck *et al.*, 2014) (see Figure 1.8) but the deeper relationships between the other clades have not been resolved so consistently. One particularly prominent issue is the location of the Bryozoa (synonymous with Ectoprocta), which is sometimes reconstructed as a basal or internal member of this sub-clade (Helmkampf, Bruchhaus, and Hausdorf 2008; Laumer *et al.*, 2015; Nesnidal *et al.*, 2010; Struck *et al.*, 2014), and sometimes not (Cannon *et al.*, 2016; Dunn *et al.*, 2008; Hausdorf *et al.*, 2010; Kocot *et al.*, 2017; Paps, Baguñà, & Riutort, 2009a, 2009b). This clade somesometimes also contains the Entoprocta (Dunn *et al.*, 2008; Struck *et al.*, 2014).

The location of the Bryozoa is significant because it is critical to an issue of nomenclature that is contentious in the present literature; specifically, whether the group of non-ecdysozoan, non-chaetognath protostomes should be referred to as Spiralia or Lophotrochozoa (Figure 1.9). Halanych *et al.* (1995) coined the latter term to describe the group of Bryozoa, Articulates + Inarticulates (*i.e.* Brachiopoda), Phoronida, Mollusca, and Annelida, without including in their phylogeny any other protostomes than arthropods (expanded to Ecdysozoa in Figure 1.9). If the other spiralian phyla belong in position A in Figure 1.9, the Lophotrochozoa is synonymous with the Spiralia; if they instead belong in position B, then Lophotrochozoa are a sub-clade of the Spiralia. Many researchers refer to the entire group as Lophotrochozoa, presumably on the basis that they accept the former scenario. Others refer to the group as Spiralia and to Lophotrochozoa as a sub-clade, a usage referred to by proponents of the Spiralia to as *sensu stricto* as opposed to the *sensu lato* of the former usage. However, if there indeed is no minimal clade including Mollusca, Annelida, Brachiopoda, Phoronida, and Bryozoa that does not also include all other spiralian animals, this usage is not any more *sensu lato* than that originally proposed. In this case, the Spiralia, which was named first (Schleip, 1929; cited in Costello & Henley, 1976), should presumably take precedence unless it is demonstrated that the Lophotrochozoa is a more taxonomically apt name. In this study, I have elected to refer to the entire group as Spiralia, and to a sub-clade containing at least Mollusca, Annelida, Brachiopoda, Phoronida, Nemertea, and Bryozoa as the Lophotrochozoa.

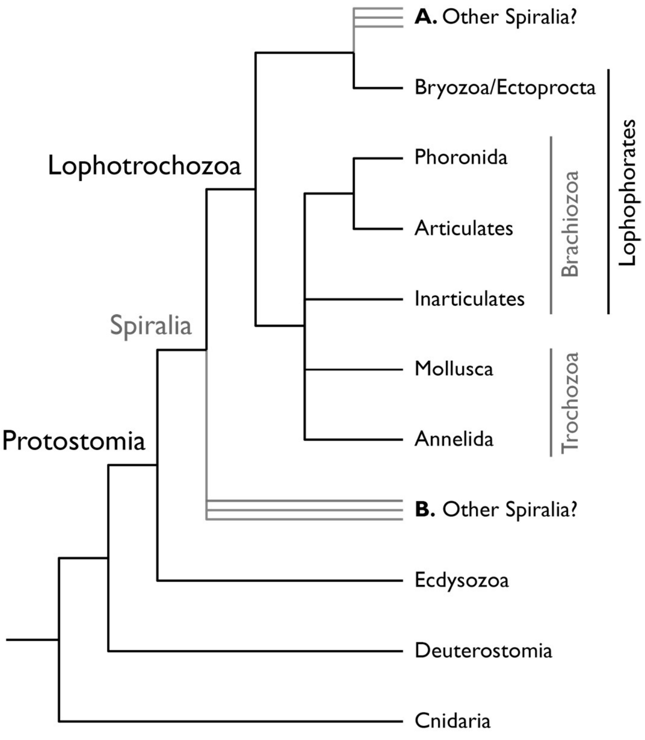


Figure 1.9. Cladogram illustrating the difficulty with the nomenclature of Spiralia/Lophotrochozoa, reproduced and adapted from Halanych *et al.*,, (1995) (in black). Labels not included in Figure 2 of Halanych *et al.*, (1995) are marked in grey. If the other spiralian groups (branches in polytomy) belong in position A, Lophotrochozoa is synonymous with Spiralia; if they belong in position B, the Lophotrochozoa is a subset of Spiralia.

1.3.2.2. The Annelida

The Annelida are a group of segmented worms, which typically possess a long, vermiform body comprised of similar (homonomous) segments terminated at the anterior and posterior by asegmental caps of tissue. Blood vessels, a gut, and a ventral nerve cord run the length of the body, and the latter is connected to an anterior dorsal brain. The trunk segments typically have lateral appendages; the chaetae (bundles of bristles), used in locomotion and sensation, or parapodia, used in locomotion.

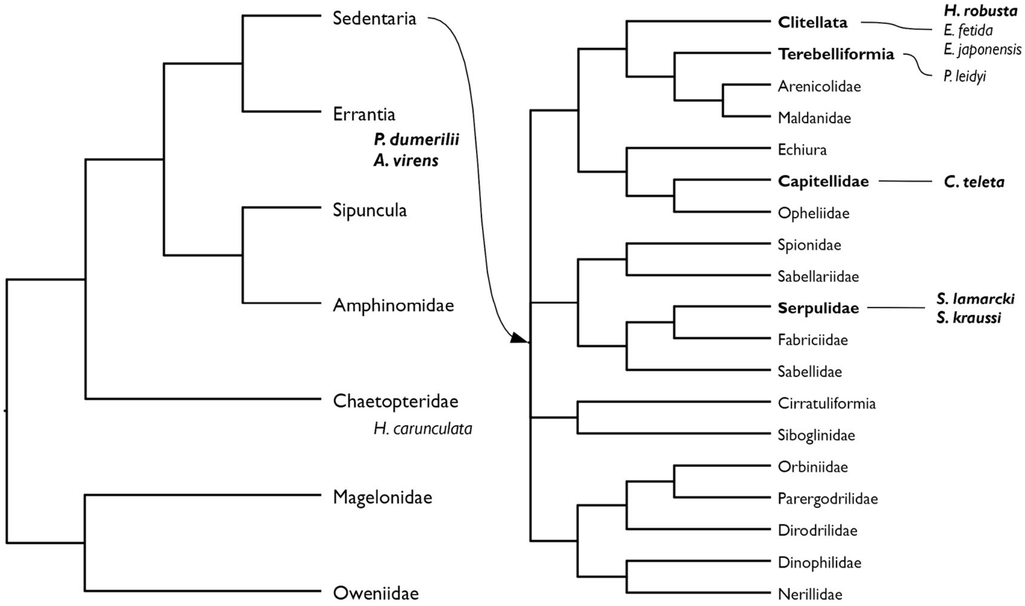


Figure 1.10. Cladogram of the current state of annelid phylogeny, adapted from Weigert & Bleidorn (2016). The position of species mentioned in this study are indicated; species from which sequence material has been used are emboldened. The cladogram of the Errantia has been omitted because of the low taxon sampling in this family. Abbreviations: *P. dumerilii* = *Platynereis dumerilii*; *A. virens* = *Alitta virens*, *H. robusta* = *Helobdella robusta*; *E. fetida* = *Eisenia fetida; E. japonensis* = *Enchytraeus japonensis; P. leidyi* = *Pristina leidyi*; *C. teleta* = *Capitella teleta*; *S. lamarcki* and *kraussi* = *Spirobranchus lamarcki* and *kraussi*.

Most of the >18,000 described species of annelid belong to one of two major clades, the Errantia, being typified by an errant lifestyle, and the Sedentaria, typically sedentary worms (Andrade *et al.*, 2015; Struck *et al.*, 2011, 2015; Weigert *et al.*, 2014) (summarised in Figure 1.10). Previous phylogenies based on morphology and trait studies had posited the existence of monophyletic clades of Polychaeta, annelids with many chaetae, and Oligochaeta, with few, but the latter (now understood to be a paraphyletic subset of the Clitellata — Timm and Martin 2015) were found in phylogenetic analysis to belong to the Sedentaria (Struck *et al.*, 2011, 2007), making the Polychaeta also paraphyletic. Phylogenetic studies have also produced other annelid evolutionary surprises, including that the Sipuncula and Echiura (previously considered to be different phyla) are derived annelids which have lost segmentation (Struck *et al.*, 2007).

Amongst the Sedentaria are a number of families of sessile tube-building 'fan worms' that use a branchial crown of tentacles to filter-feed. The original sub-order that united these animals, Sabellida, is now known to be paraphyletic (including, for example, Oweniidae), but a clade of the same name containing only Serpulidae, Fabriciidae, and Sabellidae (supported by molecular phylogenies) is now used (Bok *et al.*, 2017). The operculum is found only in the Serpulidae, and is considered to be an evolutionarily novel modification of a radiole (*i.e.* tentacle in the branchial crown) (ten Hove and Kupriyanova 2009). Opercula are morphologically highly diverse (ten Hove and Kupriyanova 2009; Wong *et al.*, 2014), bearing elaborate horns, pinnules, and chitinous or calcareous reinforcement, and can be several in number, non-functional/rudimentary, or absent (ten Hove and Kupriyanova 2009).

1.3.2.3. Regeneration in other annelids

Unlike *S. lamarcki*, most annelids are competent anterior and posterior regenerators; some are capable of regenerating a complete individual from a single intact medial segment (Bely 2006). The capacity for AP axial regeneration was very probably present in the annelid ancestor (Zattara and Bely 2016). In typical annelid axial regeneration (Bely 2014; Özpolat and Bely 2016), the wound is plugged rapidly by muscular contraction before epithelial integrity is re-established. Cell migration, in most clades involving only phagocytotic cells that aid in wound plugging, plays an important role in early wound healing. Underlying the wound site, in concert with remodelling of associated connective tissues, mature myotubes dedifferentiate to a myoblast-like form and become proliferative, along with dedifferentiated cells from the ectoderm and endoderm.

These cells contribute to the growth of a blastema, which often requires the severed ventral nerve cord and nervous input to form and are highly polar. Contributions of blastemal cells to the replacement tissues seems to be entirely restricted along germ lines; blastemal cells from dedifferentiated epidermis produce the replacement epidermis and foregut, and replacement nerve cells are produced from internalised dedifferentiated blastemal cells and by re-innervation from proximal nervous tissue; replacement muscle tissue is produced from dedifferentiated myoblast-like cells, blood vessels, or neoblasts (see below); and replacement endodermal cells are derived from old endodermal cells.

The blastema begins to pattern, as muscle fibres grow into the blastema, re-innervation matures, and, in posterior regeneration, the segment addition zone starts generating new segments which are added to the posterior growth zone in a process that closely resembles accelerated normal post-embryonic segment addition. Some species also undergo the morphallactic remodelling of existing tissues.

However, a great deal of variety exists within annelid regenerative mechanisms. In the regeneration of most oligochaetes, for example, dissepimentary cells referred to as neoblasts (but functionally and morphologically distinct from planarian regenerative neoblasts) almost certainly migrate to the wound site (Zattara, Turlington, and Bely 2016) where they participate in regeneration, possibly contributing stem-like proliferative cells to the blastema (Bely 2014; Myohara 2012; Özpolat and Bely 2016). The issue of the regenerative role of neoblasts is still not settled, though it appears that they are more important in asexual reproduction than regeneration (Myohara 2012). Diversity of annelid regeneration can be more extreme; for example, two quite closely related species of sabellids were found to employ different mechanisms for anterior regeneration, one undergoing morphallactic tissue remodelling as well as epimorphic growth and the other using only epimorphic processes (Licciano *et al.*, 2012).

A huge variety of annelid species have been the subject of regenerative experiments. Among the best developed are the polychaetes *Capitella teleta*, *P. dumerilii*, (for both of which genomes are available, see below), and *Alitta* (formerly *Nereis*) *virens*, *Pristina leidyi,* and the oligochaete *Enchytraeus japonensis.* All these models perform AP axis regeneration, which can involve the regeneration of the common annelid segmental appendages, chaetae or parapodia. However, *S. lamarcki* is currently the only annelid model of non-segmental appendage regeneration.

1.3.2.4. Annelid genomes

The genomes of four annelid species are currently publicly available; specifically, *C. teleta* (Simakov *et al.*, 2013), the freshwater leech *Helobdella robusta* (Simakov *et al.*, 2013), the earthworm *Eisenia fetida* (Zwarycz *et al.*, 2016), and *S. lamarcki* (Kenny *et al.*, 2015). A fifth, *P. dumerilii*, is complete but not published or publicly available (see Zantke *et al.*, 2014; Chou *et al.*, 2018). I was kindly granted access by the Arendt lab for this study.

The genomes of *P. dumerilii* and *C. teleta* are more alike those of invertebrate deuterostome models like amphioxus – and by extension, the bilaterian/nephrozoan ancestor – than those of other groups (Ferrier 2012), including major ecdysozoan model species. These similarities include in gene structure and orthology (Hui *et al.*, 2009; Raible *et al.*, 2005) as well as in macrosynteny and in a principle component analysis (Hui *et al.*, 2012; Simakov *et al.*, 2013). Although no whole genome comparison has yet been performed, *S. lamarcki* seems to possess a similar high degree of conservation (Takahashi *et al.*, 2009), although another *Spirobranchus* species was found to have a highly derived mitogenome (Seixas, Russo, and Paiva 2017).

In contrast to the polychaetes, the genomes of the Clitellata are highly dynamic. Although a principle component analysis placed the genome of *H. robusta* close to *C. teleta* and *B. floridae,* significant conservation of macrosynteny with other species was not found; instead, *H. robusta* has fewer ancestral bilaterian genes, and many more novel introns (Simakov *et al.*, 2013). Extreme homeobox gene expansions were found between the *Capitella*/Clitellata ancestor and the *Eisenia*/*Helobdella* ancestor (Zwarycz *et al.*, 2016).

1.3.2.5. Annelid transcriptomes

A diverse selection of annelid species have been the subject of transcriptome sequencing for a variety of ecological and evolutionary applications (Altincicek and Vilcinskas 2007; Holder *et al.*, 2013; Kenny and Shimeld 2012; Kvist *et al.*, 2013; Mehr *et al.*, 2015; Neave *et al.*, 2012; Nyberg *et al.*, 2012; Owen *et al.*, 2008; Riesgo *et al.*, 2012; Zakas *et al.*, 2012). Among these, a transcriptome of *P. leidyi* regeneration is the only other annelid regeneration transcriptome, in which were identified the homeobox genes *Otx, Hox-Z, Msx* and *Dlx* (Nyberg *et al.*, 2012).

1.3.3. *S. lamarcki* homeobox genes

The classification of *S. lamarcki* homeobox genes began with the discovery of two *Dlx* paralogues, *Dlxa* and *Dlxb,* which were found (contrary to expectations – see section 1.2.3.2) not to be involved in appendage development (McDougall 2008; McDougall *et al.*, 2011), and identification of the *S. lamarcki* ParaHox genes (present in a single copy each), which were found to be clustered and possibly expressed in the developing operculum (excluding *Xlox*) (Hui 2008). A developmental transcriptome from 24-72 hours post-fertilization *S. lamarcki* trochophores allowed the positive identification of 37 homeobox gene sequences (including *Dlxa* and *Cdx*) and the putative identification of 14 (including *Dlxb*) (Kenny and Shimeld 2012). These included a *Pax beta* gene like that of *H. robusta* (Schmerer, Savage, and Shankland 2009) and several difficult-to-classify or divergent genes, including *Abox-like*, *Hmbox-like, Nk3-like, Paired-like* and a TALE class gene identified as *Mkx2.* No comprehensive homeobox survey of the available genome (Kenny *et al.*, 2015) has yet been performed.

1.3.4. *S. lamarcki* as a model of regeneration

S. lamarcki has several features that distinguish it as a valuable model for studying development and regeneration. The system itself is tractable, being easy to collect on loose rocks from shores in Europe and keep in a low-maintenance aquarium. The animals are relatively easily extracted from their calcareous tubes, which often induces gravid animals to spawn, making developmental work easy. Consistent and rapid experimental induction of regeneration is also undemanding and occurs over a convenient timescale. The operculum itself is a fascinating subject for study, as an evolutionary novel non-segmental head appendage unique to the Serpulidae (see section 1.3.2.2). It is also histologically diverse, containing muscle, epidermis, cuticle, a blood vessel, radiolar eyes (Bok *et al.*, 2017), and, significantly, the mineralised plate, a rare example of annelid biomineralization. With the publication of the *S. lamarcki* genome (Kenny *et al.*, 2015), this system is well founded for ongoing study.

Wide taxon sampling is vital to gain a picture of the diversity of developmental and regenerative mechanisms broad enough to answer the questions posed in evolutionary developmental biology. *S. lamarcki* is well-placed for these comparisons; as a member of the Spiralia, a long-neglected clade which is now receiving greater attention; as a member of the Annelida, which are important and well-established models of epimorphic axial regeneration, and even within the Serpulidae, where interesting comparisons could illuminate developmental genetic mechanisms behind the diversity of operculum morphology.

1.4. *Branchiostoma lanceolatum*

Branchiostoma lanceolatum is an amphioxus (or lancelet) which lives in sandy seabeds in depths of 0.5 to 40 metres in coastal European waters. After its planktonic embryonic and larval stage, metamorphosis occurs and adults (c. 5-50 mm long) habitually sit buried in sand with only their head protruding, filter-feeding on microalgae from the water column. The adults produce gametes in transient lateral gonads and broadcast them at yearly intervals (Bertrand and Escriva 2011). This lifestyle is true of all lancelets, with some variation (symmetry of gonad development, spawning interval) except for *Asymmetron inferum,* which is specialised for whale-fall environments and has been found living in these toxic sulphide-rich conditions in excess of 200 m desep (Nishikawa 2004; Kon *et al.*, 2007).

Amphioxus have a relatively simple body plan (depicted in Figure 1.11a), thought to be the most conserved of extant chordates from the chordate ancestor (Bertrand and Escriva 2011). The notochord, a rod of stiff tissue against which the muscles act, runs down the midline. Directly dorsal to this is the hollow nerve cord, which terminates with a brain (also called a cerebral vesicle or intercalated region) and a frontal eye at the anterior end (Wicht and Lacalli 2005) and extends to the posterior limit of the notochord. Laterally surrounding this is segmented muscle, which runs the anteroposterior length of the notochord. On the anterior ventrum, the animal possesses an oral cavity enclosed by a basket of cirri. This leads into a large pharynx lined with gills and then into the digestive tract, which terminates at the anus near the posterior. These elements (post-anal tail, pharynx, segmented muscle, dorsal nerve cord, and notochord) are considered synapomorphies of the Chordata although debate exists about whether some structures have homologues in Ambulacraria (Rychel *et al.*, 2006).

1.4.1. *B. lanceolatum* regeneration

The regenerative capacity of amphioxus was first observed in *A. lucayanum* in 1893 by E. A. Andrews, who described in a paragraph its strong capacity for regrowth of the post-anal tail after amputation. Early studies of *B. lanceolatum* reported either the lack (Bert 1867; Nusbaum 1905; referenced in Somorjai 2017) or paucity (Biberhofer 1906; Probst 1930; referenced in Somorjai 2017) of regenerative capacity; in contrast, more recent studies of *Branchiostoma* species indicated that the genus, including *B. lanceolatum*, are capable of regeneration (Bone 1992; Pegeta 1992; Silva, Mendes, and Mariano 1998; Q. Zhang *et al.*, 2009; reviewed by Somorjai 2017), including of the oral cirri (Kaneto and Wada 2011). However, a systematic investigation of posterior regeneration using molecular tools was not attempted until recently (Somorjai *et al.*, 2012; Somorjai, Escrivà, and Garcia-Fernàndez 2012).

When the post-anal tail of a young adult amphioxus is severed, wound healing occurs between two and seven dpa (depicted in Figure 1.12A and B, day 0 to 6, stage 0 to 1a). Swelling underlying the wound epithelium, comprised of the accumulation of connective tissues and mesenchymal cells, is visible after ten days, accompanied by fragmentation of muscle fibres just proximal to the amputation plane (Figure 1.12, stage 1b). This swelling is identifiable as a blastema bud by 14 dpa at the earliest (Figure 1.12, stage 2). During this time, abundant *Pax3/7*-expressing cells have been observed in the blastema, and *Pax3/7* expression and Pax3/7 protein detection remains strong in the boundary between the mature/fragmenting muscle and the blastema, and in the elongating neural tube, until stage 3 (Figure 1.12D; Figure S6 in Somorjai *et al.*, 2012). These cells probably derive from a population of *Pax3/7*-expressing cells that are present (albeit rare) underlying the basal lamina of muscle in uninjured adult amphioxus, which are thought to be homologous to vertebrate satellite cells (Somorjai *et al.*, 2012). These cells are abundant in small/young adults and decline in larger/older adults; this decline was found to be significantly correlated with decline in the extent of regeneration (Somorjai *et al.*, 2012). *Msx* expression is also evident in the blastemal mesoderm and wound epithelium at stage 2 (Figure 1.12C, and Somorjai *et al.*, 2012).

The blastema continues to develop and differentiate over the course of approximately seven days (until 21 dpa; Figure 1.12, stage 3). During this time, a neural ampulla proceeds from the severed dorsal nerve cord, and the notochord at the amputation site becomes filled with apparently undifferentiated, stellate cells which extend into the blastema while dedifferentiation proceeds proximally. In the weeks following, regeneration proceeds with ongoing notochord proximal degradation and posterior extension; growth, posterior extension and lateral flattening of the blastema; re-formation of the fragmented muscle; and anteroposterior patterning of the regenerating neural tube (Figure 1.12, stage 4). By week six, the notochord and musculature have matured to the extent that they are plainly visible, including the presence of mature muscle fibres beyond the plane of amputation. Fifteen weeks after amputation, the replacement tail is difficult to distinguish from the original in form but is smaller. Amphioxus were found to lose regenerative capacity as they age and grow (Somorjai *et al.*, 2012) but are capable of regenerating multiple times as juveniles (Somorjai, Escrivà, and Garcia-Fernàndez 2012).

As well as *Pax3/7* and *Msx,* the expression of other conserved vertebrate markers and pathways of regeneration has been observed. The transcription factor *SoxB2* was found to be expressed in the regenerating neural ampulla (stage 3, Somorjai *et al.*, 2012), orthologues of which are associated with neurogenic differentiation in vertebrates (Sandberg, Källström, and Muhr 2005; Whittington *et al.*, 2015) and nematodes (Vidal *et al.*, 2015). The Bone Morphogenetic Protein (BMP) antagonist *chordin* was found to be expressed in the undifferentiated notochord cells (stage 3, Somorjai *et al.*, 2012); significantly, *chordin* is involved in embryonic notochord ontogenesis in amphioxus (Somorjai *et al.*, 2008; Yu *et al.*, 2007). The signalling molecule *Wnt5* was found to be expressed in a similar spatiotemporal pattern to *Msx* in the blastema, (but unlike *Msx,* also in the notochord) (stage 2, Somorjai, Escrivà, and Garcia-Fernàndez 2012), which is involved in regenerative control in vertebrates (Ghosh *et al.*, 2008; Knapp *et al.*, 2013; Sugiura *et al.*, 2009) and planaria (Gurley *et al.*, 2010). β-catenin, the nuclear actuator of canonical Wnt signalling

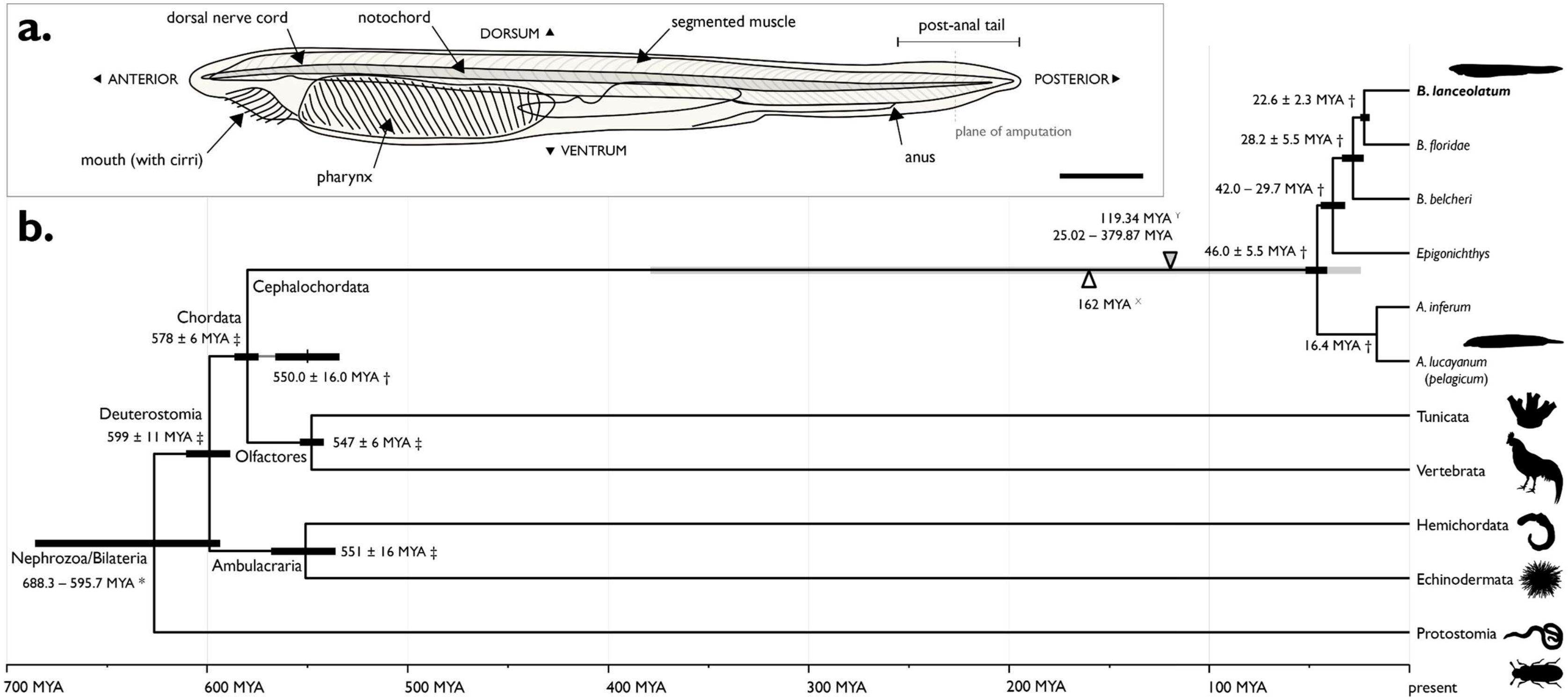


Figure 1.11.Cephalochordate morphology and evolution. **(a)** Illustrative line drawing of *Branchiostoma lanceolatum*, showing the major morphological features. The dashed line indicates the plane of amputation when inducing regeneration (see section 1.4.4). Scale bar = ~2-10 mm. **(b)** Phylogeny showing the timing of deuterostome divergence events with a focus on amphioxus model species. Timings marked with a dagger (†) are from Igawa *et al.* (2017). Timings marked with a double dagger (‡) are from Delsuc *et al.* (2018). Timings marked with an asterisk (*) are from dos Reis *et al.* (2015). Previous estimates of the divergence of extant cephalochordates from Yue *et al.* (2014) are marked with a superscript Y (^{Y}) and a grey arrowhead and bar; and from Nohara *et al.* (2005) with a superscript X (^{X}) and a white arrowhead. MYA = million years ago; *B.* = *Branchiostoma*; *A.* = *Asymmetron*.

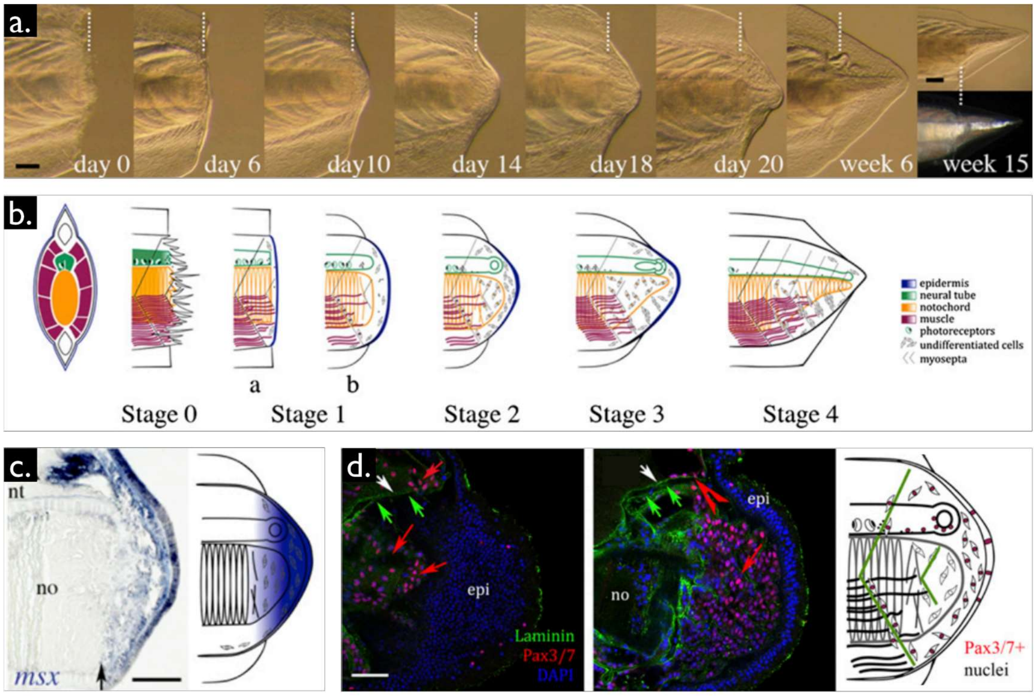


Figure 1.12. Regeneration in *B. lanceolatum*, reproduced & adapted from Somorjai *et al.*, (2012). **(a)** Images of post-anal tail regeneration of a small adult from amputation (day 0) to morphological maturity (week 15), using polarized light. Dashed white line indicates the plane of amputation. **(b)** Schematic summary of the process of regeneration, with approximately cotemporal layout to (a). **(c)** Whole mount *in situ* hybridization image (left) and schematic summary (right) of the expression of *Msx* in the stage 2 blastemal mesenchyme and wound epithelium. **(d)** Confocal images (left & centre) and schematic summary (right) showing Pax3/7 protein (red) in degrading muscle fibres (left, red arrows) and blastemal mesoderm (centre, red arrows) In the schematic, green lines indicate myoseptal boundaries and red dots indicate Pax3/7+ nuclei. Abbreviations: nt = neural tube; no = notochord; epi = epidermis. Scale bar throughout = 100 µm.

and a component of adherens junctions (Bienz 2005), was found to be present in the cell membranes (but not nuclei) of the nascent notochord region of the blastema (stage 2, Somorjai, Escrivà, and Garcia-Fernàndez 2012). Wnt/β-catenin signalling has well-established roles in controlling regeneration in vertebrates (Kawakami *et al.*, 2006; Yokoyama *et al.*, 2007; Song *et al.*, 2015; Wehner and Weidinger 2015; Strand 2016) and planarians (Gurley, Rink, and Alvarado 2008) (reviewed by Somorjai 2017).

1.4.2. Deuterostome & cephalochordate evolution

The Deuterostomia are a group of animals with the synapomorphic characteristic of the secondary developmental ontogenesis of their mouth, while the anus derives from the blastopore (see section 1.3.2 for the context in bilaterian evolutionary history). The Deuterostomia comprise the Ambulacraria (Echinodermata + Hemichordata) and the Chordata. Cephalochordata, the group comprised of all extant amphioxus, are the basal-most branching of the chordates, splitting from the Olfactores (Tunicata + Vertebrata) about 534-583 million years ago (Figure 1.11b, Delsuc *et al.*, 2018; Igawa *et al.*, 2017). Cephalochordates had traditionally been thought to be the closest relative of vertebrates on the basis of their closer morphology (classed as the Euchordata), but molecular data has since re-placed them as the basal-most chordate group (Delsuc *et al.*, 2006)

B. lanceolatum is one of five commonly-studied amphioxus species. *B. floridae* is (marginally) the most extensively studied and can be found on the coasts of Florida and the Gulf of Mexico, and *B. belcheri* and *B. japonicum* are found on Asian coasts. *A. lucayanum* (tentatively renamed *A. pelagicum* by Igawa *et al.* [2017]) is a more recently studied amphioxus model, usefully placed as the most basally-branching of known extant amphioxus genera, although complicated by the presence of cryptic species complexes (Igawa *et al.*, 2017). Estimates had placed the divergence of all known extant amphioxus species at ~120 MYA (Nohara *et al.*, 2005) or 162 MYA (Yue *et al.*, 2014), though a more recent estimate indicates that this event was much more recent, at 46.0 ± 5.5 MYA (Igawa *et al.*, 2017). Recent data on the timing of deuterostome evolution are summarised in Figure 1.11b.

1.4.3. Cephalochordate genomics

The Cephalochordata has one of the most dense coverages of genomic sequencing of any major clade in the tree of life; of the approximately 30 known species (Poss and Boschung 1996), four genomes are currently available: *B. floridae* (Holland *et al.*, 2008; Putnam *et al.*, 2008), *B. belcheri* (Huang *et al.*, 2014), *B. lanceolatum* (Marletaz *et al.*, in press), and *A. lucayanum* (or *A. pelagicum*) (Yue *et al.*, 2016). Various analyses of the content and organisation of the genomes have concluded that amphioxus genomes are highly conserved from the chordate ancestor in terms of gene gain and loss, intron position and number, non-coding elements, and other regulatory mechanisms (Holland *et al.*, 2008; Putnam *et al.*, 2008; Louis, Roest Crollius, and Robinson-Rechavi 2012; Paps, Holland,

and Shimeld 2012; Somorjai *et al.*, 2018). Amphioxus genomes are observed to be extremely polymorphic (Holland *et al.*, 2008; Putnam *et al.*, 2008; Huang *et al.*, 2014).

The conserved states of amphioxus genomes stand in contrast to the other extant chordate sub-phyla, both of which have undergone major events in their genomic evolutionary history that have caused their substantial divergence from their reconstructed chordate ancestor. The ancestor of vertebrates underwent two rounds (referred to as 2R) of WGD (Dehal and Boore 2005; Putnam *et al.*, 2008), meaning they typically have 1-4 ohnologues (*i.e.* paralogues originating from WGD) for every amphioxus gene. The massive increase in genomic material in the vertebrate ancestor – and the stoichiometric constraints that seem to have made the loss of particular types of gene, (including developmental control genes; (Brunet *et al.*, 2006; J. C. Davis and Petrov 2004; Roux and Robinson-Rechavi 2008) deleterious (Birchler and Veitia 2012; Makino and McLysaght 2010; Xie *et al.*, 2016) – lead to a substantial increase in the complexity of the vertebrate genome.

Tunicates are yet more highly derived with respect to both body plan and genomic architecture, having undergone a radical simplification and remodelling of both, including a departure from ancestral chordate gene repertoire, synteny, clustering, and intron size/placement (Dehal *et al.*, 2002; Denoeud *et al.*, 2010; Louis, Roest Crollius, and Robinson-Rechavi 2012). Consequently, amphioxus genomes are accepted as the closest extant genome to that of the chordate ancestor, albeit with the important caveats that it should neither be used as a proxy for it, nor be considered 'ancestral' (Louis, Roest Crollius, and Robinson-Rechavi 2012).

1.4.4. Cephalochordate homeobox genes

Due to their positioning as the sister group to Olfactores (and because of their previous placement as the sister group to vertebrates), the homeobox gene complement and organisation in the cephalochordate genome have been the subject of intense study for more than 25 years. The general picture of amphioxus homeobox evolution is a conservative one (Paps, Holland, and Shimeld 2012; Takatori *et al.*, 2008). They possess 133 genes to the 235 found in humans and the 96 inferred in the ancestral chordate (Takatori *et al.*, 2008). This disparity is mostly attributable to the 2R paleopolyploidy events in early vertebrate evolution, but also to small-scale gains in both vertebrates and amphioxus and losses in Olfactores and vertebrates. In contrast, no homeobox gene family thought to be present in the chordate ancestor has been lost by cephalochordates, a pattern of conservation that is also true of other TF families (Paps, Holland, and Shimeld 2012), receptor tyrosine kinases (D'Aniello *et al.*, 2008), and the Wnt signalling pathway (Somorjai *et al.*, 2018).

Table 1.1. Summary of the homeobox gene family complement changes in the vertebrate and amphioxus lineages. Family names as used on HomeoDB (Zhong and Holland 2011). Families with duplications not described before the present study are marked with an asterisk.

Homeobox class	Vertebrate gains	Vertebrate losses	Cephalochordate gains	Cephalochordate duplications
ANTP	-	Abox Bari Msxlx Nedx Nk7 Ro	Ankx Hx Lcx	Hox9-13(15) Evx Mnx Emx Nedx Nk1 Ventx
PRD	Hesx Mix	Repo	Aprd1-6	Pax3/7* Uncx
LIM	-	-	-	Lhx2/9
POU	(Pou5)	-	-	Pou3
HNF	-	-	Ahnf	Hmbox
TALE	-	-	Atale	Irx
CUT	Satb	Compass	Acut	-
ZF	Adnp	-	Azfh	-
Other	-	-	Ahbx Muxa Muxb	-

Although the cephalochordate lineage did not undergo WGD events, it has nonetheless increased its homeobox complement, both in terms of new homeobox genes without detectible direct orthology (listed as gains in Table 1.1) and genes identifiable as paralogues of existing families (duplications in Table 1.1), although, because all homeodomain-containing genes (presumably) share homology, both types arose via the same mechanisms and differ in the extent of sequence divergence. Expression data for amphioxus-specific homeobox genes are almost completely absent; *Lcx* (first mentioned in Luke *et al.*,

2003) is purported to be expressed in the developing gut, hence its name *Lunchbox* (G.N. Luke pers. comm.; referenced on HomeoDB) but no published data exist on *Ankx*, *Hx*, *Aprd1-6*, *Ahnf*, *Atale*, *Acut*, *Azfh*, *Ahbx*, *Muxa* or *Muxb* expression. The cephalochordate-specific homeobox genes and duplications are all thought to have been produced by small-scale DNA-mediated duplications, although 176 retrogenes have been found in amphioxus (Casola and Betrán 2017; Chen *et al.*, 2011). So far, no differences between the homeobox gene complements of the cephalochordate species have been recorded, although the possibility is suggested by the different chromosome counts of *B. floridae* (38) and *A. lucayanum* (34) (Holland, Holland, and Heimberg 2015). Most available evidence suggests homeobox genes (Somorjai *et al.*, 2008) and TFs and signalling genes in general (Aldea *et al.*, 2015; Yong *et al.*, 2017) are deployed in conserved domains and timings, although differences in Hox gene expression have been observed (Pascual-Anaya *et al.*, 2012).

1.4.5. *B. lanceolatum* as a model of regeneration

Cephalochordates are a well-placed and significant model in evolutionary developmental biology. Their phylogenetic positioning as the earliest-branching chordate is ideally placed to shed light on not only the ancestor of all chordates, but beyond that to the deuterostome, olfactorean, and bilaterian/nephrozoan ancestors. Research on amphioxus development and genetics has had a significant impact on our understanding of these nodes.

The attention of the amphioxus community is split between the four *Branchiostoma* species and *A. lucayanum*/*A. pelagicum.* Despite this, no species is particularly privileged in available resources over the others; even though early scholarship was mostly focussed on *B. floridae*, *B. lanceolatum* now has genomic and laboratory resources that are approximately equal to its sister species, and in some cases greatly exceed it (*e.g.* Marletaz *et al.*, in press), particularly in regard to regeneration (Somorjai *et al.*, 2012; Somorjai, Escrivà, and Garcia-Fernàndez 2012; Dailey 2017). However, it does suffer some biological disadvantages relative to other amphioxus in its slower rate of development, which make it more difficult to raise *B. lanceolatum* larvae through metamorphosis.

1.5. Thesis aims and chapter overviews

Homeobox genes are a powerful evolutionary tool for the qualitative detection of mechanistic homology because of the high degree of conservation with which they evolve both as gene orthology groups and as participants in ancient functions and GRNs. Homeobox gene deployment is therefore useful to reveal deeply conserved mechanisms deployed in regenerative programmes among bilaterian life, both of those that are conserved from ancestral regenerative capacity and developmental mechanisms that are redeployed in a modified regenerative context or recapitulated directly. For this purpose, two highly dissimilar regenerative models were chosen; namely, *S. lamarcki,* a protostome capable of the morphallactic regeneration of an evolutionarily novel appendage; and *B. lanceolatum*, a deuterostome capable of the epimorphic regeneration of its post-anal tail, an ancient chordate synapomorphy. Mechanistic and regulatory homology (or even homocracy) between these two models could help illuminate the evolution of regeneration.

The aim of this thesis was initially to classify and compare the homeobox genes deployed in the two dissimilar regenerative modes employed by *S. lamarcki* and *B. lanceolatum.* To achieve this, I analysed the homeobox content of transcriptomes of regenerative tissues from each species. As a result of these analyses, several significant findings of previously undescribed homeobox genes were made that warranted further investigation.

Described in **Chapter 3**, the survey of the transcriptomes of the mature and regenerating operculum of *S. lamarcki* found several difficult-to-classify homeobox genes. To answer the questions of their orthology – normally not difficult to establish, even in spiralians – I undertook an extensive survey and molecular phylogenetic classification of several types of homeobox from a variety of spiralian genomes. The results of this survey include the description of a highly divergent *S. lamarcki Antp* gene, the expansion and revision of an existing system of classification for previously-observed radiations of homeobox genes amongst the Spiralia, and the discovery of several previously-undocumented homeobox orthology groups.

In **Chapter 4**, a survey of the homeobox genes present in transcriptomes of mature and regenerating post-anal tails of *B. lanceolatum* is performed, leading to the relatively uncomplicated classification of a diverse set of homeobox genes, including several Anterior, Medial and Posterior Hox genes. Amongst the results of the survey was a *Pax3/7* sequence that more closely resembled the previously described *B. belcheri* sequence than that of *B. lanceolatum* or *B. floridae,* leading to the discovery of a previously unknown cephalochordate *Pax3/7* duplication.

In **Chapter 5**, I describe this *Pax3/7* duplication in more depth. It is observed that *Pax3/7* tandemly duplicated in the cephalochordate ancestor and diverged substantially in coding and adjacent non-coding sequence, but that the duplicates have been retained with a very high degree of conservation since the radiation of extant cephalochordates. I also establish the paralogue-specific expression profiles in the early- and mid-development of *B. lanceolatum*, demonstrating that the paralogues have undergone differential evolution of expression.

In **Chapter 6**, I relate the conserved roles of homeobox genes in other instances of regeneration to the presence or absence of these genes in the regenerative transcriptomes of *B. lanceolatum* and *S. lamarcki.* I also discuss the value of manual curation of genetic data as illustrated by the analyses presented herein, and how this could be reconciled with the problem of ever greater data availability in genome biology.

2. Materials & Methods

2.1. Animal collection and husbandry

2.1.1. *Spirobranchus lamarcki*

Adult *Spirobranchus lamarcki* were collected on rocks at East Sands in St Andrews Bay, Scotland, and were maintained in the nearby aquarium at the Gatty Marine Laboratory in a flow-through aquarium system at ambient temperature. The animals were not fed apart from the adequate supply of food in the incoming seawater.

2.1.2. Amphioxus species

Adult European amphioxus (*B. lanceolatum*) were collected in Argelés-sur-Mer, France and Faro, Portugal. After transport to the Scottish Oceans Institute, St Andrews, Scotland, the animals were kept in a semi-closed circulated aquarium at 19°C ± ~1°C. Animals were usually kept in the sand from their collection site at an approximate depth of 30 mm, in seawater (drawn from St Andrews Bay by the Gatty Marine Laboratory pumps) at a depth of approximately 0.1-0.2 m. The water was constantly circulated through UV sterilizers and reservoirs containing biological filter media (brushes, foam, and bio-balls). The animals were kept at a density of approximately 50 animals to a small (0.21 m x 0.21 m) tank or 200 animals to a large (0.21 m x 0.39 m) tank. The aquarium was subjected to an artificial daytime of 13 hours using fluorescent lighting.

Cultures of *Isochrysis* (*Tisochrysis*) *lutea*, *Nannochloropsis oculata*, *Rhonomonas reticulate var. reticulate* (sourced from the Culture Collection of Algae and Protozoa), and *Tetraselmis* sp. (sourced from Florida Aqua Farms), were grown according to the manufacturer's instructions in Micro Algae Grow (Florida Aqua Farms), a Guillard F/2 formula, in 1 µm-filtered seawater. They were concentrated by centrifugation and re-suspended in seawater to an optical density of approximately 100-150 absorption units at a wavelength of 740 nm. These concentrates were introduced into the tanks at a rate of approximately 2 mL per small tank and 4 mL per large tank per day.

2 mL of GroTech Nutrimatine and Vitamino nutrient supplements and 10 drops

of Interpret Liquifry No. 1 lipid supplement per 50 mL algal concentrate was added to the food daily, a quantity calculated according to the manufacturer's instructions for supplement volume per aquarium volume. Tank flow-through was often stopped during feeding (for three hours twice a day or overnight) to retain the food in the tanks for longer.

Asymmetron lucayanum adults were collected in Bimini, Bahamas. The animals were maintained per the protocol described in Holland *et al.* (2015) in 25°C non-circulating water at the Scripps Institution of Oceanography in California, U.S.A. and fed a mixture of *Isochrysis* (*Tisochrysis*) *lutea*, *Isocrhysis* sp., *Pavolva lutheri* and *Thalassiosira floridana*. Water was changed daily. An automated artificial moon system was used to recreate the natural lunar cycle during their 10 hour night period.

Algal cultures were prepared in the Scottish Oceans Institute by Joseph Chapman. UK-based animal husbandry was a collective effort by the present author, Dr Simon Dailey, and Dr Somorjai.

2.2. Transcriptome collection, sequencing, and assembly

2.2.1. *Spirobranchus lamarcki*

S. lamarcki regenerating opercula transcriptome collection, sequencing and assembly is described fully in Szabó (2015). The process is briefly described below.

Animals were gently extracted from their habitation tubes using forceps. Animals were kept in Nunclon 4-well plates in 1 mL filtered seawater (changed daily) in the dark in an air-conditioned room at 15-18°C. Animals intended for RNA collection were kept in filtered seawater (FSW) re-filtered using 0.22 µm PES syringe filters.

Animals with mature opercula were selected for morphologically perfect opercula and general health. A scalpel was used to cut the opercular filament at the easy break point (EBP), identifiable as the proximal edge of the proximal pigment band. In regenerating animals, some proximal tissue was included but minimised where possible. The opercula were gently cleaned using forceps to minimise the inclusion of debris and epibionts. Animals that autotomised their operculum during tube extraction, cleaning or dissection were not used for regeneration experiments.

Samples of mature opercula (n=22) were fixed immediately in RNAlater (Ambion). Non-calcifying 2 days post-operation (dpo) (n=19) regenerating opercula and calcifying 6 dpo regenerating opercula (n=24) were washed in RNAse free water and fixed in TRIsure (Bioline). RNAlater-fixed samples were washed in RNAse free water and transferred to TRIsure for extraction. Extraction was performed approximately per the manufacturer's instructions. Sample purity, concentration and integrity were estimated using a NanoDrop and examination on 1% agarose gels. Samples were stored at -80°C.

One pooled sample of total RNA from each of the stages was submitted for Illumina HiSeq2000 sequencing at the Wellcome Trust Centre for Human Genomics, Oxford. The resulting 3 datasets were checked for quality using FastQC and filtered through adapter removal, quality checks and 3' end trimming using the NGS-QC Toolkit (Patel & Jain, 2012). The datasets were combined and assembly was performed using Trinity (Grabherr *et al.*, 2011) with a default *k*-mer size of 25. A protein prediction and CD-HIT/CD-HIT-EST (W. Li and Godzik 2006; Fu *et al.*, 2012) pipeline was performed to cluster the contigs and reduce redundancy.

The regenerative transcriptomic experimental design and execution was carried out by Dr Réka Szabó; the assembly and annotation processes were carried out by Dr Szabó and Dr Miguel Pinheiro.

2.2.2. *Branchiostoma lanceolatum*

B. lanceolatum regenerating tail transcriptome collection, sequencing and assembly are described fully in Dailey (2017). The process is briefly described below.

Animals were anaesthetised with a 2-in-50,000 emulsion of clove oil in sterile 0.22 µm FSW until unresponsive (≤5 minutes), and stored briefly in fresh sterile 0.22 µm FSW. Unresponsive animals were placed on the sterile lid of a petri dish with minimal FSW. The posterior half of the post-anal tail was dissected away using a disposable razor blade, removed from the plate with sterile forceps, washed in RNAse-free milliQ water, and stored at -80°C in RNAlater (Thermo Scientific). Care was taken to amputate mature tails in the same place between animals.

Following amputation of the mature tail, animals were placed in a fresh sterile

0.22 µm FSW and fed for 11 days, followed by a three day starvation. After this time (14 days) the regenerating tails were excised and preserved in the same way as the mature tails. Care was taken to amputate regenerating blastemas with as little proximal collar tissue as possible.

Total RNA was extracted from the TRIsure-preserved samples using the manufacturer's protocol and the non-regenerating (uncut) and regenerating RNA from 5-10 individuals each pooled for sequencing. 454 pyrosequencing was performed in the lab of Dr Nori Satoh at the Okinawa Institute for Science and Technology, Japan.

The resulting 824,736 uncut and 645,912 regeneration-specific reads were cleaned of adapter sequences and low quality regions with clean_reads v0.2.3 (COMAV Institute). The assembly was performed in Newbler 2.6, incorporating <2000 base pair (bp) long sequences from the broad developmental transcriptomic dataset generated by Oulion *et al.*, (2012). This assembly was processed in cd_hit_est to retain only a single example of clusters of sequences with ≥97.5% similarity to one another, and merged with the remaining (>2000 bp) data from Oulion *et al.* (2012). gsMapper (v2.8) was used with default settings to map the uncut and regenerating read sets against the transcriptome assembly and against the genomic reference transcript set (unpublished) and the read counts were then normalised using DESeq2.

The regeneration and RNA extraction portions of this protocol were performed entirely by Dr Ildikó M.L. Somorjai. Processing and assembly were performed by Dr Simon Dailey, with assistance from Dr Miguel Pinheiro.

2.3. Genomic and transcriptomic searches

2.3.1. Tools and databases used

A complete list of the genomic and transcriptomic databases from which information was retrieved is presented in Appendix 2.1. A list of software tools used in the course of the following analyses is presented in Table 2.1. Homeobox gene searches were performed using appropriate components of the BLAST+ software suite, using a variety of previously classified homeodomain and homeobox gene sequences (detailed below).

Table 2.1. Software tools used in the following analyses.

Tool name	Versions	Publication/Company	Usage
Local software			
4peaks	1.7.2	*Nucleobytes*	Electropherogram visualization & sequence editing
BLAST	2.4.0	Camacho *et al.*, 2009	Local sequence database creation and searching
Bowtie 2	2.2.5	Langmead and Salzberg 2012	Read mapping
Canopy	1.7.4.33 48	*Enthought*	Python scripting environment
ClustalX	2.1	Larkin *et al.*, 2007	Alignment format conversion
FigTree	1.4.3	Rambaut 2007	Phylogeny visualization
FindTar	3.11.12	Ye *et al.*, 2008	MicroRNA target prediction
Jalview	2.x	Waterhouse *et al.*, 2009	Alignment editing & visualization
MAFFT	7.245	Katoh and Standley 2013	Nucleotide & protein sequence alignment
MEGA-Proto/CC	7.0.26	Kumar *et al.*, 2012	Maximum likelihood phylogeny
MEGA7	7.0.25	Kumar, Stecher, and Tamura 2016	Model selection, draft NJ & ML trees
MiRanda	3.3a	Enright *et al.*, 2003	MicroRNA target prediction
modelgenerator	0.85	Keane *et al.*, 2006	Model selection
MrBayes	3.2.6	Ronquist and Huelsenbeck 2003	Bayesian phylogeny
PHYLIP	3.695	Felsenstein 1989	Neighbour-joining phylogeny
Python	2.7	*Python Software Foundation*	Scripting language
RNAhybrid	2.1.2	Rehmsmeier *et al.*, 2004	MicroRNA target prediction
Web services			
VISTA	-	Mayor *et al.*, 2000; Frazer *et al.*, 2004	VISTA analysis (global DNA alignment)
(AVID)	-	Bray, Dubchak, and Pachter 2003	Genome alignment (used by VISTA)
CIPRES Science Gateway	3.3	Miller, Pfeiffer, and Schwartz 2010	MrBayes web host and interface
NCBI CDD Search	-	Marchler-Bauer *et al.*, 2011, 2015	Conserved domain search
NCBI Web BLAST	-	Johnson *et al.*, 2008	BLAST searches of the NCBI database
PITA	02/2017	Kertesz *et al.*, 2007	MicroRNA target prediction
Primer3web	4.1.0	Koressaar and Remm 2007; Untergasser *et al.*, 2012	Primer design
Python scripts			
0_1_simplegetseqnamesfromfasta.py	-	(present author)	Retrieving sequence names from fasta files
0_2_filternamesforPhyLIP.py	-	(present author)	Filtering 10 character names for PHYLIP input (uniqueness & length)
0_3_replaceseqsnameinMA.py	-	(present author)	Replaces old sequence names with 10 character names for PHYLIP

microVISTA.py	-	(present author)	Recreate VISTA analysis on custom alignment
sam2fasta.py	-	(Chang Park)	Converting Bowtie output to fasta format
treecomparator.py + comparatordata.py	-	(present author)	Maps support values from equivalent nodes from input tree onto target input tree

2.3.2. Further searches in *Spirobranchus*

2.3.2.1. Deep homeodomain searches

A number of *S. lamarcki* transcriptome sequences could not be confidently placed in canonical families. Lophotrochozoans are known to possess a variety of difficult-to-classify, divergent, and non-canonical homeobox gene families in the TALE and PRD classes (*e.g.* Paps *et al.*, 2015). To contribute robustly to the classification of non-canonical TALE class genes, BLAST searches for TALE-class homeodomains were undertaken of the available genomes of *S. lamarcki*, *Capitella teleta, Platynereis dumerilii, Helobdella robusta, Lingula anatina, Lottia gigantea,* and *Patella vulgata* using as queries canonical and non-canonical homeodomain-containing sequences from the *S. lamarcki* transcriptome, from *C. teleta, Crassostrea gigas,* and *Pinctada fucata* as classified in (Paps *et al.*, 2015), and from *Spirobranchus* (formerly *Pomatoleios*) *kraussii* and *Nipponacmea fuscoviridis* SPILE (Spiralian TALE) sequences retrieved from the NCBI database (Morino, Hashimoto, and Wada 2017).

Results from these searches that could not be putatively placed in a canonical family (*Irx*, *Meis*, *Mkx*, *Pbx*, *Pknox* or *Tgif*) by sequence alignment were used as queries to retrieve more divergent homeodomains.

These recursive searches were used in the *S. lamarcki* genome and regenerative transcriptome until no new sequences were retrieved. However, due to the time-intensive nature of these searches, complete search saturation was not attempted for all the listed species. It is almost inevitable that the set of homeodomains found for some of these species is incomplete, and probable that future searches using an expanded query set and/or improved genome assemblies will identify TALE class homeodomains in *S. lamarcki* that have been missed in this analysis. These sequences were aligned against canonical

TALE homeodomain sequences from *T. castaneum* and *Drosophila melanogaster* and the alignment used to produce neighbour-joining, maximum likelihood, and Bayesian phylogenies, the support values of which were later merged onto the Bayesian topology (see section 2.4.6).

Similar search, alignment, and phylogenetic analyses were undertaken for PRD-class homeodomains, using as queries canonical homeodomain sequences from *C. gigas*, *T. castaneum* and *B. floridae* and non-canonical homeodomain sequences from *C. gigas* (PRD1, 3-9), *B. floridae* (Aprd1-6) and the unidentified *S. lamarcki* PRD-like and 'ceh-37'-like sequences. These searches were saturated much more quickly than the TALE-class queries.

2.3.2.2. Other homeodomain searches

To classify an Nk class *S. lamarcki* transcriptome sequence that could not be placed in a canonical family, putative and previously identified Nk1-7, Msx, Tlx, and Lbx sequences were collected from the genomes of *S. lamarcki*, *C. teleta, P. dumerilii, H. robusta, L. anatina, C. gigas L. gigantea, P. vulgata, Apis mellifera, D. melanogaster, T. castaneum,* and *B. floridae*, including Nk class sequences previously described in *C. gigas* (Paps *et al.*, 2015) and newly found in *L. anatina* and *L. gigantea*. These sequences were aligned used to produce neighbour-joining, maximum likelihood, and Bayesian phylogenies, the support values of which were later merged onto the Bayesian topology (see section 2.4.6).

To classify a 'Hox-like' *S. lamarcki* transcriptome sequence that could not be placed in a canonical family, putative and previously identified Hox and ParaHox sequences were collected from *S. lamarcki*, *C. teleta, A. virens, P. dumerilii, H. robusta, L. anatina, L. gigantea, Euprymna scolopes*, *Octopus bimaculoides*, *B. floridae*, *Saccoglossus kowalevskii, Strongylocentrotus purpuratus, Nematostella vectensis,* and *Sycon ciliatum.* These sequences were aligned and used to produce neighbour-joining, maximum likelihood, and Bayesian phylogenies, the support values of which were later merged onto the Bayesian topology (see section 2.4.6).

2.4. Phylogenetic analyses

2.4.1. Alignment preparation

A suitable outgroup sequence was chosen and added to the alignment. Scripts were written in Python 2.7 to automate some aspects of the data preparation, and are included in Appendix 2.2. A unique 10-character name (verified using the `0_2_filter-namesforPhyLIP.py` script) was assigned to each of the sequences in the alignment (retrieved using the `0_1_simplegetseqnamesfromfasta.py` script) and a copy of the original alignment overwritten using the `0_3_replaceseqsnameinMA.py` script. The renamed alignment was trimmed to the appropriate region in Jalview 2.x (Waterhouse *et al.*, 2009) (in most cases just the homeodomain) and sequences with too little sequence coverage removed. Alignments were exported as the `.phy` and `.nxs` file formats required by PHYLIP and MrBayes respectively using ClustalX 2.1 (Larkin *et al.*, 2007).

2.4.2. Model selection

The best-fit matrix of amino-acid evolution for each alignment was selected using ModelGenerator v0.85 (Keane *et al.*, 2006) using 4 gamma categories. Where possible the recommended matrix was used in subsequent phylogenetic analyses; where the model was not supported, the default was used instead.

2.4.3. Neighbour-joining analyses

Neighbour joining analyses were performed in PHYLIP 3.695 (Felsenstein 1989). `seqboot` was used to produce 1000 replicates. `protdist` was used to analyse these, with gamma categories if +G was recommended by ModelGenerator, with the coefficient of variation of substitution rate among positions set to the reciprocal of the square root of the alpha value produced by ModelGenerator. Neighbour-joining trees for each of the 1000 replicates were generated in `neighbor`, having specified the position of the outgroup root. Finally, a consensus tree with bootstrap support values was generated using `consense`, with the trees treated as rooted.

2.4.4. Maximum likelihood analyses

A MEGA Analysis Options (`.mao`) file was prepared in the MEGA Prototyper for a maximum likelihood analysis using 1000 bootstraps and the conditions recommended in ModelGenerator, and run using MEGA-CC (S. Kumar *et al.*, 2012).

2.4.5. Bayesian analyses

Bayesian analyses were run on the CIPRES Science Gateway (Miller, Pfeiffer, and Schwartz 2010), using MrBayes 3.2.6 on XSEDE (Ronquist and Huelsenbeck 2003). The analysis was given 168 hours to run and a generation limit of 500,000,000 (`ngen=500000000`) but was specified to stop early (`stoprule=YES`) if the convergence diagnostic fell below 0.01 (`stopval=0.01`). No analysis exhausted the time or generation limit before falling below 0.01. The model and other settings recommended by ModelGenerator were specified using the `lset rates=` and `ngammacat=`, and `prset aamodelpr=` and `statefreqpr=` options.

2.4.6. Consensus tree generation

A script was written in Python 2.7 to map the support values (bootstraps from neighbour-joining and maximum likelihood analyses and posterior probabilities from Bayesian analyses) from nodes on each tree to equivalent nodes (where they exist) on a target tree. The script and a user guide/technical document is presented in Appendix 2.2. Node equivalency is based on contents but not identity of internal nodes, so that the node marked by the green circle in Figure 2.1 A & B is equivalent, despite not being identical.

The trees in Figure 2.1 would be encoded in the Newick (`.nwk` or `.tree`) format as follows:

```
A: ((a,b),(c,d))     B: (((a,b),c),d)
```

The script works by extracting the contents of each set of brackets, equivalent to each node:

```
A: (a,b) (c,d) (a,b,c,d)

B: (a,b) (a,b,c) (a,b,c,d)
```

Each node in each tree is associated with a support value. The script then performs a simple search of each node in the target tree against the nodes in the other two trees. If an equivalent node is present, it takes the associated support value, and then maps back the support values to each node in the original tree. The script is limited by its inability to detect equivalent nodes in unrooted trees.

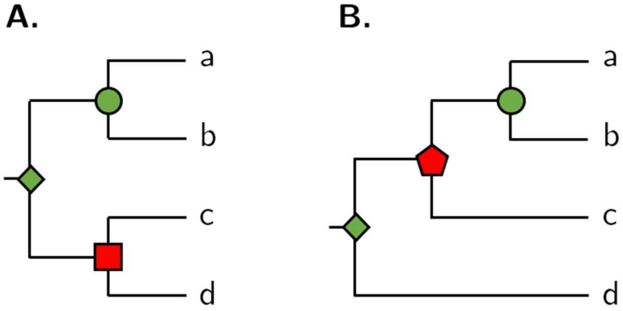


Figure 2.1. An example cladogram showing topological equivalence. The node marked by a green diamond is equivalent in both cladograms, despite having a different internal topology, because both contain sequences a, b, c, and d, and no others. The nodes marked by red shapes are not found in both trees.

The trees were formatted, inputted and run according to the instructions in the user guide (Appendix 2.2). The Bayesian trees were chosen as the target tree onto which the support values were mapped.

2.4.7. Tree visualization

Short format sequence names in the resulting trees were replaced with more informative names using a shell script to automate batches of `sed` commands (example included in Appendix 2.2). The trees were opened with FigTree 1.4.2 (Rambaut 2007). Clades in which at least one of the support values was equal or greater to 70% (*i.e.* 700 bootstraps or 0.7 posterior probability) were highlighted. In large trees, clades corresponding to established gene families were collapsed to aid visualisation. The trees were exported as Scalable Vector Graphics (.svg) files and were formatted for final presentation in LibreOffice Draw.

2.5. Amphioxus genome analyses

2.5.1. VISTA analysis

Scaffolds containing the *Pax3/7* cluster were identified in the available amphioxus genomes using BLAST searches and retrieved. In the case of the *B. lanceolatum* and *B. belcheri* genomes, these were complete uninterrupted scaffolds (Sc0000222 and scaffold5, respectively). The *B. belcheri* scaffold5 was trimmed to a 2 megabase (mb) window centred approximately on the *Pax3/7* cluster. In the *B. floridae* genome, the cluster is found on two contigs (Cont4522 and 4524) placed alongside one another in scaffold23. In the *A. lucayanum* genome, the cluster is spread over 17 small scaffolds. These scaffolds were manually curated on the basis of comparison to the *Branchiostoma* genomes. Some scaffolds were found to belong inside other scaffolds; that is, the beginning and end of the latter scaffold belong either side the former scaffold, presumably the result of an inaccurate assembly. The points at which the former scaffolds belonged in the latter scaffolds were always found to be inside regions of repeated 'N's, (*i.e.* uncalled bases). These scaffolds were split into two parts as part of the manual curation to allow the alignment software to correctly place the sequences. In places where the previous manual assembly of *A. lucayanum Pax3/7a* and *Pax3/7b* from SRA data contained more information than the genomic scaffolds (*i.e.* where the genomic scaffolds stopped or started inside an exon covered by the SRA model), the genomic scaffold was extended using the SRA model. Other manual changes included the reverse complementation of some sequences to bring them all into the same orientation, and the 'padding' of the 5' and 3' end of one particularly short scaffold with 'N's to stop the alignment software ignoring it. Details of the genome assemblies from which the scaffolds were taken is included in Appendix 2.1. Details of curative changes made to scaffolds are included in Appendix 2.4.

Annotations were prepared denoting the positions of the *Pax3/7a* and *Pax3/7b* exons on the *B. lanceolatum* Sc0000222 scaffold. These annotations are included in Appendix 2.4.

These data were uploaded to the mVISTA (Mayor *et al.*, 2000; Frazer *et al.*, 2004) web interface (http://genome.lbl.gov/vista/mvista/submit.shtml). Because the genomes were in draft form (*i.e.* the data for *B. floridae* and *A. lucayanum* are comprised of several scaffolds), the alignment was performed using AVID (Bray, Dubchak, and Pachter 2003). The results were visualised in the VISTA web viewer.

2.5.2. CNE region plot

The *B. lanceolatum Pax3/7* cluster was used as a query in a BLAST search against the *B. floridae/A. lucayanum* conserved non-coding element (CNE) database (Yue *et al.*, 2016) (details in Appendix 2.1). The 137 CNEs retrieved by the search were aligned to the genomic sequence, and the coverage visualised using spreadsheet tools.

2.5.3. 'microVISTA' analysis

In an attempt to detect conservation in the non-coding regions of *Branchiostoma Pax3/7a* and *Pax3/7b* (*i.e.* regions of the sequence conserved in both paralogues from the pre-duplication pro-orthologue), a Python script (`microVISTA.py`, Appendix 2.2) was written to replicate a VISTA analysis on a custom alignment. The script takes two aligned sequences and, in a rolling window of a size specified in the script arguments, calculates the percentage positive identity between the two alignments (a match of 'N' or '-' does not qualify as a positive identity). For each position in the alignment, the script produces an integer value of its score (percent of identities in the window starting on that position) and a binary record of whether each sequence has a gap ('-') character.

The resulting output was transferred to spreadsheet software and visualised using the graphing functions. Annotations from the VISTA analysis were used to highlight the UTR and coding regions.

An alignment of *B. belcheri Pax3/7a* against *Pax3/7b*, each with 1 kb of 5' and 3' intergenic space, was prepared. Unfortunately, an alignment (MAFFT, default settings) of the two complete loci was unable to reconstruct the gene structure, that is, several exons in the 5' did not align with one another. To impel the correct reconstruction, each region was aligned with its homologous region separately: *Pax3/7a* 5' intergenic sequence was aligned against *Pax3/7b* 5' intergenic sequence, Pax3/7a exon 1 was aligned with *Pax3/7b* exon 1, Pax3/7a intron 1 was aligned with Pax3/7b intron 1, *etc.* (MAFFT, default settings). For the purpose of this alignment, *Pax3/7a* exon '0' (actually exon 1) was counted as 5' intergenic space because of its apparent loss in *Pax3/7b*. These alignments were then concatenated in sequence to make a forced alignment of the two *Pax3/7* loci and used as an input to microVISTA.py with a window of 75.

For comparison against expected background noise, two random alignments were prepared. Each was a pair of randomly generated As, Gs, Cs and Ts (25,000 characters in length). The first was used as an input without further processing; the second was aligned using MAFFT with default settings. Each was used as an input to microVISTA.py with a window of 75.

Finally, an alignment was prepared using only the coding regions of *B. lanceolatum Pax3/7a* and *Pax3/7b*, aligned whole against one another. This alignment was used as an input for microVISTA.py with a window of 45.

2.5.4. mRNA analysis

A search for the presence of microRNA (miRNA) binding sites in the 3' UTR of *Branchiostoma Pax3/7a* and *Pax3/7b* was undertaken but not completed, using a methodology adapted from Candiani *et al.* (2011).

The sequences of the 3' UTRs of the *Pax3/7* genes of *B. lanceolatum, B. floridae,* and *Mus musculus* were collected. The original *B. lanceolatum* transcriptomic *Pax3/7b* contig contains a 1139 bp 3' UTR, and BLAST searches of this region and the 3' sequence from the *B. lanceolatum* genome against the unassembled reads from the regenerative transcriptome or against a *B. lanceolatum* developmental sequence read archive (accession number SRR2057056) did not produce evidence for extending this sequence. An alignment with the *B. floridae* genome was used as a basis to predict the extent of the *B. floridae Pax3/7b* 3' UTR.

The published *B. floridae Pax3/7*(*a*) sequence (accession number EEN66816.1) contains a 380 bp 3' UTR. An alignment with the *B. lanceolatum* genome was used as a basis for the prediction of the extent of the *B. lanceolatum Pax3/7a* 3' UTR.

The 3' UTRs for *M. musculus Pax3* and *Pax7* were taken from mRNA sequences retrieved from the NCBI database; *Pax3* transcript variant 1 (NM_008781.4) (2041 bps) and *Pax7* (NM_011039.2) (2706 bps).

A fasta format mature microRNA library for *B. lanceolatum* was produced from Zhou *et al.* (2012), Table S1, by concatenating the first, second, sixth, and seventh columns (Name, Scaffold, Mature_start_position, and Mature_end_position, respectively) into fasta sequence names for the sequences in the eighth column (Mature_sequence). A mature microRNA library for *M. musculus* was downloaded from miRBase (Kozomara and Griffiths-Jones 2011). Details of both databases are included in Appendix 2.1.

Table 2.2. Settings used in *in silico* microRNA target prediction tools, after Candiani *et al.* (2011).

Parameter	Arg	Value	Unit
miRanda			
Gap opening penalty	-go	-8	-
Gap extension penalty	-ge	-2	-
Score threshold	-sc	60	-
Energy threshold	-en	-20	kcal/mol
Scaling parameter	-scale	2	-
RNAhybrid			
ΔG	-e	-20	kcal/mol
Seed 2 to 8	-d	2,8	-
FindTar			
Loop score	-loop	15	-
ΔG	-energy	-15	kcal/mol

MicroRNA target sites were putatively accepted on the basis of prediction by any three of four *in silico* prediction tools used in this analysis; namely, miRanda (Enright *et al.*, 2003), RNAhybrid (Rehmsmeier *et al.*, 2004), FindTar (Ye *et al.*, 2008), and PITA (Kertesz *et al.*, 2007). Where possible, this analysis used parameters consistent with Candiani *et al.* (2011). A summary of the parameters and the command line arguments used is given in Table 2.2. For PITA, only hits with an energetic score (ddG) <-10 were considered to be significant.

The predictions were compared, and instances where hits were predicted within an approximately 20 bp window by at least three of the four tools were selected for manual inspection. MiRanda, RNAhybrid and FindTar produce alignments of the predicted target/miRNA binding, and these were compared for consistency. Where all three alignments predicted identical or very similar targets and binding configurations, the hit was considered a robust candidate. Where only two of the alignment-reporting tools produced a hit, or one of the three contradicted the others, the match was considered a tentative candidate. If all predicted alignments contradicted one another, the hit was rejected.

2.6. Laboratory techniques

2.6.1. Spawning and embryo husbandry

Gravid *Branchiostoma lanceolatum* adults were collected near Argelés-sur-Mer, France, and maintained in the Scottish Oceans Institute per Section 2.1.2 for at least a month before spawning was induced. Spawning was induced at or near peak gravidity following the protocol of Fuentes *et al.*, (2007). Gamete type and maturity was determined under a dissecting microscope and individuals chosen based on their sex. Chosen individuals (usually 5-10 animals of each sex) were placed in 0.22 µm FSW at 16.5°C and incubated in a water-bath at 22°C for 28-32 hours. The end of the heat shock was timed to shortly precede the end of the aquarium artificial daylight hours. Animals were transferred to disposable plastic cups containing approximately 40 mL of fresh room-temperature (18°C) FSW and placed in absolute darkness. The cups were inspected at hourly intervals under monochromatic red light, with care taken to ensure the animals were not exposed to other light sources. Animals were removed from the darkness when they had spawned or after approximately six hours.

Sperm were decanted into 50 mL Falcon tubes and rested on ice. Eggs were poured into Petri dishes, the surface of which had been abraded with kitchen scourers to prevent embryo adhesion, and which had afterwards been rinsed. The eggs were divided such that one female that had completely emptied a full set of gonads produced six plates, to which was added fresh FSW to a depth of ~5 mm. The viability of the sperm of each male was tested in scratched mini Petri dishes by applying a drop of sperm to approximately 10-20 eggs and monitoring the fertilization for abnormalities in progression and chorion shape. Sperm that took longer than a few minutes to produce visible choria or that produced wrinkled choria were excluded from subsequent fertilizations. Three drops of viable sperm were applied to each plate, which was agitated gently to mix. All gametes were either poured or handled with micropipettes, using tips from which the last ~5 mm had been cut. The plates were inspected after a few minutes to verify that fertilization was proceeding normally. The fertilization time and parent identities were recorded on the lids of each plate. Notes were also made on plates of eggs that took longer than a few minutes to produce visible choria or that produced wrinkled choria.

The plates of developing embryo were incubated in darkness at 19°C. At this temperature, after approximately 16 hours post fertilization (hpf), the embryos hatch from their choria by releasing a hatching enzyme. Continued exposure to hatching enzyme and overcrowding can be detrimental to the continuing development of the embryos, so each plate of embryos was divided into two fresh plates and the water partially replaced with fresh FSW after hatching.

Development of the embryos and larvae was monitored with reference to the staging described in Hirakow and Kajita (1991, 1994) with modifications from Zhang *et al.*, (2013). Animals were fixed at a variety of developmental stages from 4 cells to L2 (see section 2.6.2).

Asymmetron lucayanum adults, maintained per section 2.1.2, spawn of their own accord 1-2 days before the new moon (N. D. Holland, Holland, and Heimberg 2015). On these days, gravid males and females were identified, separated into Petri dishes of FSW + 100 µg/mL penicillin, and left in the dark with checks at 1.5 hourly intervals. After the spawning, the sperm and eggs were treated and fertilization executed as above. Developing embryos were incubated at 25°C but otherwise treated as above. Embryos and larvae were fixed as described in section 2.6.2.

B. lanceolatum spawning and embryonic husbandry were carried out in collaboration with Drs Dailey and Somorjai using animals collected by Dr Somorjai and F. Alier

Pous. *A. lucayanum* spawning, embryonic husbandry, and fixation were performed entirely by Dr Somorjai.

2.6.2. Material fixation

Fertilized eggs, embryos and larvae were staged according to Hirakow and Kajita (1991, 1994) with modifications from Zhang *et al.*, (2013). At the desired stage, the samples were concentrated in their plates by swirling or using a 40 µM cell filter. They were then transferred in a minimal volume of FSW to screw-top vials, to which were added as much fresh, RNAse free 4% PFA in MOPS salts (0.1M MOPS, 2mM $MgSO_4$, 1mM EGTA, & 0.5M NaCl, prepared per Appendix 2.3) as possible. The PFA was partially replaced after a few minutes to maximise final PFA concentration and the tube transferred to 4°C for an overnight (~16 hour) fix.

After the fixation step, tubes of embryos were washed 3 times in chilled RNAse-free 70% ethanol (handled with nuclease-free filter micropipette tips) and stored at -20°C until the whole mount *in situ* hybridization experiment (section 2.6.4).

2.6.3. Cloning and probe synthesis

B. lanceolatum embryos were fixed at a variety of developmental stages using TRIsure (Bioline). *A. lucayanum* embryos were fixed in RNAlater (Invitrogen) and transferred to TRIsure. RNA samples from both were extracted using the TRIsure supplier's protocol. Oligo(dT) primers were used with the Tetro cDNA Synthesis kit (Bioline) to produce cDNA libraries.

Gene-specific primers were designed using the Primer3 online interface (Koressaar and Remm 2007; Untergasser *et al.*, 2012) using template sequence from the *B. lanceolatum* transcriptome (primer sequences are presented in Table 2.3). The initial *B. lanceolatum Pax3/7b* amplicon included regions with high nucleotide sequence similarity between *Pax3/7a* and *Pax3/7b*. Therefore, a second amplicon containing only the divergent 3' region was cloned from cDNA using a second primer pair. A pre-existing *B. lanceolatum Pax3/7a* clone prepared by Somorjai *et al.*, (2008), also containing the 5' regions of strong similarity, was similarly used as a template for a *Pax3/7a* 3' sub-clone. These 3' sub-clones were necessary to achieve paralogue specificity in the *in situ* hybridisation probe. In *Asymmetron,* both probe templates were cloned directly from the cDNA.

Polymerase chain reactions (PCRs) were performed using BIOTAQ kits (Bioline) according to the manufacturer's instructions. The particulars of each reaction are recorded in Table 2.4. Apart from variables presented in Table 2.4, all reactions were prepared on ice with 2.5 µL 10x NH4 buffer and 0.25 µL 100 mM dNTP mix per 25 µL reaction. The PCR program was begun with a pre-heated lid and an initial denaturation step of 95°C for 05:00 minutes, followed by 36 cycles of 95°C for 00:30, reaction-specific temp for 00:30, and 72°C for 00:45. A final extension step was carried out at 72°C for 06:00 and the tubes suspended at 6°C until storage at 4°C or -20°C.

PCR products were inspected using agarose gel electrophoresis. Gels were prepared with 1% agarose in 1x TAE buffer with ethidium bromide at approximately 0.5 µg/ml. Samples were mixed with loading dye to 1x final concentration and loaded alongside 4 µL of Bioline Hyperladder 100bp+ (concentration as supplied). Gels were placed in standard electrophoresis tanks and subjected to 100 V for 30-50 minutes. Gels were visualised and photographed in a UVP GelDoc-IT ultraviolet (UV) gel doc. Bands matching the predicted size were excised under UV transillumination with a fresh disposable razor blade and the amplicons were extracted using the Isolate II column purification kit (Bioline).

Table 2.3. Primers used to clone *Pax3/7a* and *Pax3/7b* amplicons from *Branchiostoma lanceolatum* cDNA (*Pax3/7b* large clone and *Pax3/7a* and *Pax3/7b* probes), *Asymmetron lucayanum* cDNA (*Pax3/7a* and *Pax3/7b* probes) and the pGEM-T vector into which the amplicons were ligated (Universal M13)

Target	Forward primer	Reverse primer	Amplicon position	Amplicon length
Pax3/7a (long)	(Somorjai *et al.*, 2008)		303-1144	842
Pax3/7a (probe)	CTGGAGGAAGCAGCAGGG	GCCCAGTCCGTTCACCAA	774-1095	322
Pax3/7b (long)	GAAGACGGAGAGAAGAAACGGT	CCCGTACTGATAGGTGTCCATG	463-1275	813
Pax3/7b (probe)	CTTCAACCACCTGCTACCCA	CCCGTACTGATAGGTGTCCATG	780-1275	496
Universal M13	GTAAACGACGGCCAGT	AACAGCTATGACCATG	n/a	n/a

Table 2.4. PCR conditions used to produce various gene amplicons from *Branchiostoma lanceolatum* cDNA and probe templates from miniprepped plasmid.

Reaction	Annealing temp. (°C)	Vol (µL) per 25 µL reaction			
		$MgCl_2$	Taq	Template	Primers (each)
Pax3/7b (long)	54	1.0	0.25	0.5	0.5
Pax3/7a (sub-clone)	54	1.0	0.25	0.25	2.5
Pax3/7b (sub-clone)	54	1.0	0.25	0.25	2.5
M13 probe template	58	1.0	0.25	0.5	0.5

The amplicons were ligated into pGEM-T Easy vector (Promega) overnight at 4°C according to the manufacturer's instructions, and transformed into the XL10-Gold (Stratagene) competent *Escherichia coli* cell strain using the following heat shock protocol. Standard microbiological sterile technique was observed throughout. Cells were stored long-term in buffer at -80°C in 50 µL aliquots. Aliquots were defrosted on ice and the ligation reaction mix added and mixed in by gentle flicking. After 10 minutes, the tubes were shocked at 42°C for 45 seconds before being returned to ice for 2 minutes. 200 µL of room-temperature LB broth was added and the tube rotated at 37°C for 15 minutes. The resulting transformed bacteria were spread on agar plates containing 50 µg/mL ampicillin and coated with 4 µL of 200 mg/mL IPTG, 40 µL of 20 mg/mL X-galactose and 10 µL of 100 mg/ml ampicillin. The plates were sealed with Parafilm and incubated for 16 hours at 37°C, after which time bacteria which have been successfully transformed with a plasmid containing the amplicon were visible as white colonies.

Selected white colonies were picked with a sterile filter tip and cultured overnight with rotation at 37°C in 6 mL of LB broth containing 50 µg/mL ampicillin. These colonies' vectors were also verified as containing a single copy of the correct amplicon using colony PCR with M13 primers and gel electrophoresis. 750 µL of these cultures was thoroughly mixed with 250 µL 80% sterile glycerol in H_20 in screw-top phials by vortexing, snap-frozen in liquid nitrogen, and placed at -80°C for long-term preservation of the stock. Plasmids were extracted from the remaining culture using peqGOLD (Peqlab) or Promega plasmid miniprep kits using the manufacturer's protocol and the amplicons were sequenced using Universal M13F or M13R primers (Table 2.3) at the University of Oxford Zoology

Sequencing Service for verification. The chromatograms were inspected and corrected in 4peaks (nucleobytes) and the sequences were verified by alignment in Jalview 2.x (Waterhouse *et al.*, 2009).

Templates for antisense probe synthesis were produced using PCR with M13 primers (Table 2.3 &Table 2.4). Bands were verified using agarose electrophoresis. The template was precipitated out of the reaction mixture by mixing with 0.1 volumes of sodium acetate (3M, pH 5.2) and 2.5 volumes of ethanol *per* 1 volume of original solution. The mix was left at room temperature for 20 minutes or at -20°C overnight (~16 hours), and then centrifuged at 4°C and 13 G for 40 minutes. The liquid was removed, taking care not to disturb the almost invisible DNA pellet. The pellet was washed with 100 µL of 70% ethanol in H_20 and the centrifugation repeated for 20 minutes. The wash was removed as completely as possible and the pellet air-dried, ideally for no longer than 15 minutes. The DNA was resuspended in dH_20 and quantified using a Nanodrop.

T7 enzyme was used to transcribe DIG-labelled (Roche) antisense probes (appropriate to their orientation in the vector) *in vitro*. A volume containing 500 ng to 1 µg of template was added to 4 µL of 5x transcription buffer or 2 µL of 10x transcription buffer (Thermo Scientific), 1 µL of RNAse inhibitor (RNAsin), 2 µL DIG labelling mix (Roche), 2 µL of appropriate RNA Polymerase (Thermo Scientific), and a volume of RNAse-free H_20 to a final volume of 20 µL. The mix was prepared on ice and incubated at 37°C for 3 hours. Probe purification was performed by precipitation as above for template purification, but using dedicated RNAse-free solutions. Probes were inspected on an agarose gel as described above, mixed 1:1 with *in situ* hybridisation buffer (prepared per Appendix 2.3) to improve stability, and stored at -20°C.

2.6.4. Whole mount i*n situ* hybridization

Whole mount *in situ* hybridisation was performed as previously reported (Somorjai *et al.*, 2008; Dailey *et al.*, 2016) with the protocol detailed below. Stock reagents were prepared per Appendix 2.3, which also includes a condensed protocol. The protocol is described in brief below.

2.6.4.1. General protocol

Care was taken during the preparation and first day of the protocol to avoid contaminating the material and reagents with RNAses, which could degrade the antisense probes and their mRNA targets. Precautions included thorough cleaning of the work area, micropipettes and microscope in use, the maintenance of uncontaminated gloves, the strict use of only nuclease-free chemicals and consumables (tubes, micropipette tips, *etc.*).

Washes were performed as follows. Liquid was removed carefully with a micropipette while monitoring the embryos in the well to prevent inadvertent embryo loss, such that the well never had less than approximately 100 µL of liquid in it at any one time. It is important that the embryos never risk drying out. The removed liquid can be placed in a separate plate to check that no embryos have accidentally been discarded, or discarded directly. The next wash was then added to the well and the plate covered. Where unspecified below, washes were 400 µL and the plate was placed on an orbital shaker for the duration of the wash.

2.6.4.2. Preparation

Embryos were removed from the tubes in which they were stored in 70% ethanol (in RNAse free H_2O) at -20°C, and placed in an *in vitro* fertilization grade four-well plate (Nunclon) for inspection under a dissection microscope. Only embryos with a healthy and typical morphology for the desired developmental stage were selected. Each well contained a single stage. The embryos were rotated for 10 minutes in 70% ethanol or stored overnight at 4°C.

Embryos fixed before hatching were de-chorionated in 70% ethanol using a pair of mounted tungsten needles. The chorion was trapped against the plate bottom using one needle and gently torn open using the other, taking care not to damage the embryo. The ethanol was replaced with fresh RNAse free 70% ethanol to remove chorion debris.

2.6.4.3. First day

The embryos were washed 3 times in NaPBTw (0.9% NaCl, 0.1% Tween 20) for five minutes each. During this time, the proteinase K solution (7.5 µg/mL) and the glycine-

NaPBTw (2 mg/mL) were prepared. The proteinase K was added with staggered timing and the plates left (unrotated) for a time corresponding to their age. This duration should be determined empirically for each batch of proteinase K. When the time has elapsed, 4 µL of 10% glycine was quickly added to each well, with priority being given to the youngest embryos. The embryos were then washed with glycine-NaPBTw for five minutes, and then a post-fix is applied of 4% paraformaldehyde (PFA) in MOPS salts (100 mM MOPS [pH 7.4], 2 mM EGTA, 1 mM MgSO4) for 1 hour at room temperature without rotation. During this time, fresh 0.1M TEA (in RNAse-free water) was prepared from 1M TEA stock.

After the post-fix, the embryos were washed twice with TEA, for 1 minute and then 5 minutes. These washes were then removed (again being sure not to allow the embryos to dehydrate while they were unattended) and 3.75 µL acetic anhydride was added to 1.5mL 0.1M TEA, vortexed, and 300 µL added to each well. The plate was left stationary for 5 minutes. During this time, a second batch of acetic anhydride was prepared, this time 7.5 µL in 1.5 mL 0.1M TEA, and after the time had elapsed, 300 µL was added to the previous wash, and left stationary for a further 5 minutes. At this point the hybridisation oven and buffer were preheated to the desired hybridisation temperature (65°C in all successful *in situs* reported herein, but which must be empirically determined for each probe).

Two washes of NaPBTw were then performed for 1 minute and 5 minutes. In the latter wash, embryos were re-arranged so that embryos that are of easily distinguishable stages and types but that will be exposed to the same probe were sharing a well. It can be advisable at this stage to transfer the embryos to a new 4-well plate.

The NaPBTw was then replaced with 100 µL hybridisation buffer pre-heated to the appropriate hybridisation temperature. This was rotated for 1 minute and replaced with 200 µL pre-heated hybridisation buffer. The plates were sealed with autoclave tape and rocked in the hybridisation oven at the hybridisation temperature for 2-3 hours.

During this time, the probe was thawed on ice, and added to 200 µL pre-heated hybridisation buffer per well (0.5 µL per well in all *in situs* reported herein, but also requiring probe-specific calibration). To denature the probe, the mix was brought to 70°C in a hot block to coincide with the end of the previous (pre-hybridisation) step. The hybridisation buffer in the wells was replaced with the probe-hyb, being careful to maintain the temperature as stably as possible throughout. For this purpose, the hot-block was placed next to the microscope at the working temperature. The plate was resealed with autoclave tape and rocked in the hybridisation oven for at least 16 hours (overnight).

2.6.4.4. Second day

The probe-hyb solution can be retained and stored at -20°C for future use. It was replaced with two 5-minute and two 10-minute washes of Wash Solution (WS) 1, preheated to the hybridisation temperature. The plate was rocked in the hybridisation oven for all hot washes. Temperature fluctuations in the plate during washes were minimised using the hot-block. The WS1 was replaced by WS2 preheated to hybridisation temperature, and rocked at hybridisation temperature for 5 minutes before being transferred to rotation at room temperature for a further 10 minutes. A second wash was performed using separate room temperature WS2, for 15 minutes.

The WS2 wash was changed for room temperature WS3; this wash was replaced immediately with fresh WS3 and the plate rotated for five minutes. During this time, an RNAse mix was prepared using 1mL of WS3 for 2 µL of RNAse A (at 10 mg/mL) and 1 µL of RNAse T1.

The WS3 wash was replaced with RNAse mix and incubated at 37°C for 20 minutes. This time can be extended if necessary to reduce background noise. The RNAse mix was washed out with two 20 minute WS3 washes; care was taken to wash out as much of the RNAse mix as possible. This was followed by a 20 minute wash in WS4 and a 5 minute wash in WS5. During this latter wash, blocking solution was prepared by adding 100 µL pre-treated sheep serum to 1 mL WS5; the wash was replaced with blocking solution and the plate rotated at room temperature for 2-3 hours. At the end of this period, the antibody solution was gently defrosted and centrifuged at maximum speed for 2 minutes to remove residual particulate from pre-absorption or previous experiments. Finally, the blocking solution was replaced with 200 µL of the antibody solution, and the plate sealed with tape, and incubated at 4°C with gentle rocking for at least 16 hours.

2.6.4.5. Third day

The antibody solution was carefully removed and saved at -20°C. Reuse of antibody up to five times produced noticeable improvements in the reduction of background staining. Four 20 minute washes of NaPBTw were performed, during which the alkaline phosphatase buffers (AP) were prepared. The no-magnesium buffer (AP-) was prepared with 1 mL Tris 1M, 200 µL NaCl 5M, 50 µL 20% Tween20, and topped up to 10 mL with milliQ water. The magnesium buffer (AP+) was prepared similarly with the addition of 500 µL $MgCl_2$ and the reduction of the NaCl to 20 µL. Both solutions were filtered through a 0.22 µm PES syringe filter. Additionally, a second AP+ solution in which 50% of the water is replaced with a 20% solution of polyvinyl alcohol (PVA) (average molecular weight of 30-70K or greater) was made; for this purpose, a half-quantity of AP+ to double strength is made so that it can be filtered; *i.e.* 500 µL Tris, 10-100 µL NaCl, and 25 µL Tween20, topped up to 2.5 mL with water, filtered, and mixed with 2.5 mL PVA.

The final NaPBTw wash was replaced with 500 µL AP-. The increased volume (first wash only) was to dissolve any phosphate salts that had crystallised above the usual level of liquid. The wash was swirled and directly replaced to remove as much NaPBTw as possible. Four 10 minute AP- washes (normal) were performed. The AP- washes can be followed by four 10-minute AP+ washes or by transferring the embryos in a minimal volume of AP- into 500 µL AP+ in a fresh plate followed by rotation for 10 minutes and two/three further 10 minute AP+ washes. These steps were taken (AP- washes, large volume wash, new plate) to prevent the crystalline precipitation that can occur if traces of NaPBTw wash come into contact with AP+. During the AP+ washes, the staining solution was prepared. NBT and BCIP were warmed at 37°C for 3 minutes to ensure complete dissolution, then 3.5 µL NBT (added first and flicked to mix) and then 3.5 µL BCIP was added to 1 mL AP+ or AP+/PVA. The staining solution was mixed with gentle flicking, and stored in the dark until use.

The final AP+ wash was replaced with 200 µL of the staining solution and the plate incubated in the dark without rotation. From this point on the staining reaction was underway, and the plate contents were monitored regularly for staining. The illumination angle of the microscope can be adjusted to increase staining visibility. The signal can develop in as little as 20 minutes to 2 hours but can take days or weeks. The reaction can be slowed by cooling to 4-16°C (*e.g.* overnight) or can be boosted by incubation at 37°C (*e.g.* when experienced with specific probe/stage/experimental configuration staining times), although these modifications can induce precipitation. Embryo staining can be paused in AP-. Staining solution was changed for fresh after any overnight incubation at room temperature or when the solution took on a pinkish hue.

2.6.4.6. Experiment termination

When the staining had proceeded to the desired extent, the reaction was stopped with four 10 minute washes in AP- (rotating in the dark) followed by four 10 minute washes in NaPBTw. Embryos were then fixed in 4% PFA-PBS or PFA-MOPS for 1 hour at room temperature or preferably overnight at 4°C; the plate was not rotated and kept in the dark. The embryos were washed twice for 10 minutes in NaPBTw – at this point any attached debris was removed from the embryos by pipetting or gentle manipulation with tungsten needles. The NaPBTw was replaced with 80-95% glycerol in milliQ water, homogenised overnight by rotation, and stored at 4°C in plates for imaging.

2.6.5. Visualization

Embryos were mounted on microscope slides in 95% glycerol in H_20 and visualised with a Leitz DMRB microscope (Leica Microsystems) using Nomarski optics. A Retiga 2000R camera and the QCapture software suite (QImaging) were used for image capture. Image processing was performed in the GNU Image Manipulation Package (GIMP) and Inkscape.

3. Homeobox genes in *Spirobranchus lamarcki* operculum regeneration

The work presented in this chapter has been accepted for publication in *Genome Biology and Evolution* under the title "*A revised spiralian homeobox gene classification incorporating new polychaete transcriptomes reveals a diverse TALE class and a divergent Hox gene*" by Barton-Owen, Szabó, Somorjai and Ferrier (2018). The manuscript as submitted is included in Appendix 3.1. This chapter includes data that were omitted from the publication. To avoid unnecessary paraphrasing, some sections of text from the published manuscript are reused herein.

The transcriptome analysed in this chapter was generated by Dr Réka Szabó, and assembled and annotated by Drs Szabó and Miguel Pinheiro. I carried out the survey and classification of homeobox genes.

Note on terminology

For the purposes of clarity, genes within the homeobox orthology groups included in HomeoDB2 (Zhong, Butts, and Holland 2008; Zhong and Holland 2011) (not all of which are homeobox families *sensu* Holland 2012) are referred to herein as 'canonical' homeobox genes. Those that do not appear to belong to these groups are referred to as 'non-canonical.' The proliferation of non-canonical TALE-class homeobox genes within the Spiralia is referred to as the Spiralian TALE Expansion (abbreviated to STE) and the clades erected by Paps *et al.* (2015) are referred to as TALE clades (TALE-). The proliferation of non-canonical PRD class sequences is referred to as the PRD Expansion (abbreviated PRD-E). If a non-canonical gene does not appear to belong to a specific orthology group/clade (based on the measures used herein, *i.e.* homeodomain phylogeny), they are referred to as orphans. As genomic sampling density increases, some of these orphans will no doubt be found to belong to taxonomically-restricted orthology groups.

There is some confusion with regards to terminology used to describe genes with apparent homology to Nk1-7 genes but lacking identifiable orthology to a specific family. These genes are often referred to as (and named) 'Nk-like' or 'NKL' (*e.g.* Paps *et al.*, 2015). However, 'NKL' (Nk-like or Nk-linked) is also the name of a subclass of the ANTP class (Ferrier 2008; Hui *et al.*, 2012). To avoid potential confusion, I will refer to these genes as 'similar to Nk' until they are classified, at which point they will be named based on the taxonomic extent of detectable orthology.

As explained in section 1.3.2.1, I have elected to use the terms Spiralia and Lophotrochozoa *sensu stricto*; *i.e.*, all non-ecdysozoan Protostomia except Chaetognaths belong to the Spiralia, and the Annelida, Mollusca, Brachiopoda, Nemertea, Phoronida, and Ectoprocta/Bryozoa belong to a subclade called the Lophotrochozoa (Laumer *et al.*, 2015; Luo *et al.*, 2018) after the original definition including the Annelida, Mollusca, Brachiopoda (Articulates + Inarticulates), Phoronida and Bryozoa (Halanych *et al.*, 1995, who did not include any other spiralian taxa). This convention is not universally employed; some (*e.g.* Paps *et al.*, 2015) use Lophotrochozoa *sensu lato* to mean the entire Spiralia. This is not entirely without justification because of the variability of placement of the Ectoprocta between studies.

3.1. Introduction

3.1.1. The spiralian homeobox expansions

The first comprehensive classification of the entire homeobox complement of the genome of a member of the Lophotrochozoa was performed by Paps *et al.*, (2015) on *Crassostrea gigas*, the Pacific oyster. Of the 136 homeobox genes identified by their survey, 31 sequences defied placement in known orthology groups more specific than a class. Paps *et al.* developed a nomenclature of 19 spiralian-specific clades (based on homeodomain phylogenies) to classify these and similar genes they retrieved from the genomes of seven other Spiralia, resulting in three 'Nk-like' clades, six PRD class clades (I-VI), nine TALE class clades (I-IX), one clade each from the LIM, SIX, and CUT classes, and a number of orphans both identifiable as belonging to a homeobox class and unclassifiable. In transcriptomes covering a 38-stage time-course of development, they observed that the majority of these genes peaked in expression in early development, some in late development, and only one in the mid-developmental period (the neck of the developmental hourglass; see section 3.1.2).

Morino *et al.* (2017) cloned TALE gene sequences from the limpet *Nipponacmea fuscoviridis* and the blue coral worm *Spirobranchus kraussii* (formerly *Pomatoleios kraussii*). They did not attempt to integrate the nomenclature of Paps *et al.* (2015) into their phylogenetic analysis, instead naming a large clade containing only STE genes the 'Spiralian TALEs' (abbreviated to SPILEs, *i.e.* SPIralian taLEs). Not all STE genes belong to the SPILE clade; notably, *C. gigas* TALE5 and the TALE-I clade from the analysis of Paps *et al.* (2015) fall outside this clade in the phylogeny of Morino *et al.* (2017).

In the phylogeny of Paps *et al.* (Supplementary Figure 4 — also contains SIX, CERS, and Hnf homeodomains) TALE clades II-IX form a clade which is topologically separated from TALE-I and the TALE gene families (*Irx, Meis, Mkx, Pbx, Pknox*, and *Tgif*). This inexact replication of the topology of Morino *et al.* (2017) suggests that there is a meaningful distinction within the STE between SPILE and non-SPILE genes. One obvious explanation is that SPILE genes share a broad common orthology, while non-SPILE STEs could have independent (and possibly more recent) origins as TALE family paralogues.

Morino *et al.* also present *in situ* hybridization data for five SPILE genes from *N. fuscoviridis* and three from *S. kraussii* eggs and morulae, representing the first *in vivo* expression data for STE genes. They found that the genes are deployed in quartet-specific patterns along the animal-vegetal axis in both animals. Further, through an enviable series of microinjected morpholino and over-expression treatments of *N. fuscoviridis* eggs, they demonstrated that *NfSPILE*s *A-D* engage in complex inter-regulation, and that *NfSPILE-C* and *D* determine macromere and first quartet fate.

Given the discovery of these previously unknown, non-canonical genes involved in important ontogenic processes, Morino *et al.* hypothesised that the STE was critical to the establishment of spiralian development. This hypothesis will be reexamined in section 3.4.7 in light of the findings of the present study.

3.1.2. The developmental hourglass

The concept of the developmental hourglass, which has been influential in evolutionary developmental biology, is introduced here because it is pertinent in interpreting the available transcriptomic data concerning STE genes and other genes found in the course of the present study (*e.g.* sections 3.4.1.5, 3.4.2.1, & 3.4.3).

Early theoreticians (Baer 1828; referenced in Hall 1997) hypothesised that development would follow a pattern wherein 'general' structures (*i.e.* those shared most between taxa) would appear earlier than more specialised structures. Consequently, early developing embryos were expected to be similar between species, and diverge further as ontogenesis proceeded.

However, morphological diversity was not found to incrementally increase along vertebrate developmental time courses; instead, pronounced differences were observed in early and late development but were far less apparent in mid development (Haeckel 1874; referenced in Švorcová 2012). An idealised plot of two-dimensional 'morphospace' against time thus resembles an hourglass, after which the phenomenon was named (Duboule 1994; Raff 1996; referenced in Švorcová 2012).

The confirmation that a similar pattern is observed in other phyla (Sander 1975; Goldstein, Frisse, and Thomas 1998) led to the idea of a 'phylotypic stage' (Sander 1983; referenced in Slack, Holland, and Graham 1993) (or more accurately, a period comprising several stages) at the neck of the hourglass, a mid-developmental time window at which embryos would converge on a morphology archetypical to their phylum (but dissimilar between phyla). Early molecular data indicated that Hox genes are usually expressed in the phylotypic period, leading to the suggestion that the use of a conserved, Hox-centric, positional specification system was the synapomorphy of the Animalia (Slack, Holland, and Graham 1993): the 'zootype.'

The concept of the Hox-centric zootype has not survived the genomic age, condemned by the absence of ctenophoran and poriferan Hox genes (Ryan and Baxevanis

2007) and supplanted by many other observed genetic synapomorphies (*e.g.* Srivastava *et al.*, 2010). In contrast, the phylotypic period has retained its relevance to Evo-Devo thought because of the broad support it has received from transcriptomic studies of development, from insects and nematodes (Kalinka *et al.*, 2010; Levin *et al.*, 2012; Zalts and Yanai 2017), spiralians (Paps *et al.*, 2015; Xu *et al.*, 2016), vertebrates (Hazkani-Covo, Wool, and Graur 2005; Irie and Sehara-Fujisawa 2007; Irie and Kuratani 2011; Wang *et al.*, 2013; Hu *et al.*, 2017), and larger scale comparative studies (Domazet-Lošo and Tautz 2010; Schep and Adryan 2013; Gerstein *et al.*, 2014; Drost *et al.*, 2015; Levin *et al.*, 2016), as well as modern morphometric studies (Young *et al.*, 2014).

Very broadly, these studies support the notion that in mid-development across the Bilateria, transcriptome age and inter-species similarity increase, transcriptome diversity and temporal expression divergence decreases, and gene expression is more resistant to evolutionary change (Domazet-Lošo and Tautz 2010; Kalinka *et al.*, 2010; Irie and Kuratani 2011; Wang *et al.*, 2013; F. Xu *et al.*, 2016). Several studies report that homeobox gene expression increases in mid-development (Levin *et al.*, 2012, 2016; Schep and Adryan 2013) and expressed homeobox genes are very highly conserved (Zalts and Yanai 2017).

Although the data now seem to broadly support the concept of the phylotypic period, it has received cogent criticism throughout its lifetime (reviewed by Švorcová 2012 & Irie and Kuratani 2014). Some rejected outright the notion of the phylotypic stage (*e.g.* Richardson 1995; Richardson *et al.*, 1997; Bininda-Emonds, Jeffery, and Richardson 2003; Kalinka and Tomancak 2012). Other authors have criticised the imprecise notion of the boundaries of the phylotypic period in vertebrates (*e.g.* Irie and Kuratani 2011). Some analyses support a funnel-like model resembling von Baer's ideas more than the hourglass (Roux and Robinson-Rechavi 2008; Comte, Roux, and Robinson-Rechavi 2010; Piasecka *et al.*, 2013).

It has also been suggested that for any given comparison between species, nested hourglasses from different taxonomic levels might be at work (Irie and Sehara-Fujisawa 2007), and some analyses have found more than one neck to the hourglass (Domazet-Lošo and Tautz 2010; Levin *et al.*, 2012). Others have pointed instead to a mid-developmental transition between early and late developmental gene expression phases, albeit one significantly less conserved *between* phyla than the early and late phases themselves (Levin *et al.*, 2016). The relationship between the phylotypic period, archetypal phyletic body plan and taxonomic phyla is one of ongoing debate (Irie and Kuratani 2014; Irie 2017).

The mechanisms by which the neck of the hourglass is enforced are not yet totally resolved (Irie 2017). However, the consensus of available evidence points to pleiotropic constraints on the gene regulatory networks involved in the phylotypic period, an idea suggested by Sander (1983; referenced in Galis and Metz 2002), Duboule (1994) and Raff (1996; referenced by Irie and Kuratani 2014). In this conception, the phylotypic period sees a spike in interconnectivity in genes and regulatory networks via complex *cis*- and *trans*- regulation; a transition between simpler, all-encompassing networks in early development and more discrete, modular, and localised networks in later development, both of which would be more robust to change. Mutation of, or *cis*-regulatory change to, genes involved in this interconnected period would be expected to have a slew of *trans*-regulatory consequences and cause serious or lethal teratogenic effects.

Evidence for this constraint has been found in chordate transcriptomes (Hu *et al.*, 2017) and in population-level studies of zebrafish (Schmidt and Starck 2011) and *C. elegans* (Zalts and Yanai 2017). Support is also found in the vulnerability of the phylotypic stage in vertebrates (Galis and Metz 2002) and the pleiotropic effects of mutations to segment polarity genes in flies (Galis, van Dooren, and Metz 2002). The observations of signalling pathway and transcription factor expression enrichment in the phylotypic period (Irie and Kuratani 2011; Levin *et al.*, 2012, 2016; Schep and Adryan 2013) and the strong constraint under which these phylotypically-expressed genes evolve (*e.g.* Zalts and Yanai 2017) are also consistent with pleiotropic forces.

3.1.3. New genes, old constraints

New genes are being generated at a constant, high rate (reviewed by Tautz and Domazet-Lošo 2011). Most new genes are lost quickly, but of the ones which become established and are retained, some become coopted into development, where they can contribute to important forces in phenotypic evolution (Heyn *et al.*, 2014; also reviewed by Chen, Krinsky, and Long 2013; Kemkemer and Long 2014; McLysaght and Guerzoni

2015). These new additions to ontogenic transcriptomes have been quite consistently found to be biased against expression in the neck of the developmental hourglass, when the transcriptome age increases (Domazet-Lošo and Tautz 2010; Kalinka *et al.*, 2010; Irie and Kuratani 2011; Paps *et al.*, 2015; Xu *et al.*, 2016), though a surprisingly large contingent of recently evolved genes in phylotypic period transcriptomes from *C. gigas, Haliotis discus hannai* (Gastropoda)*,* and *Perinereis aibuhitensis* (Polychaeta) has been reported (Xu *et al.*, 2016).

New genes are a significant part of the earliest zygotic transcriptomes in mice, fish and flies (Heyn *et al.*, 2014), but the age of the early developmental transcriptome has been observed to differ, from a young age comparable with late development in fish, to much closer to the older age of the phylotypic period transcriptome in flies (Domazet-Lošo and Tautz 2010). New genes have also been found to be involved in late developmental processes, such as mollusc mantle secretomes responsible for shell formation and patterning (Aguilera, McDougall, and Degnan 2017), where marked variance is found between even relatively similar species, and in bird beak shape evolution (A. Abzhanov, pers. comm.).

However, the forces discouraging expression in this mid-developmental window don't necessarily preclude novel genes from gaining important (*i.e.* necessary for morphological normality) and essential (*i.e.* loss results in death) roles in early and late development. For example, a substantial portion of genes with a recent evolutionary origin in *Drosophila* have acquired essential roles in development, comparable to the portion of essential old genes (30% of young vs 25-35% of old — Chen, Zhang, and Long 2010). Conversely, young genes and genes with paralogues have been found to be less likely to be essential than older genes in mice (Chen *et al.*, 2012).

A substantial spike in species-restricted genes is often evident in plots of transcriptome age (*e.g.* Tautz and Domazet-Lošo 2011; Aguilera, McDougall, and Degnan 2017). Although this effect is possibly exacerbated as an artefact of sparse genome sampling, it is also an expected product of the high rate of new gene loss, which is only marginally lower than their generation rate (Tautz and Domazet-Lošo 2011). New genes seem to undergo a kind of turnover, wherein they are gained, adopted in roles (which can be important — see above and in section 3.1.4), and are replaced by even newer genes at an (evolutionarily) rapid pace. Whether or not they are retained in the long term, they are often argued to be a critical source of evolutionary novelty (Tautz and Domazet-Lošo 2011; Chen, Krinsky, and Long 2013) — for example, the molluscan radula (Hilgers *et al.*, 2018).

Novel genes have previously been detected and profiled in annelid regeneration (Myohara, Niva, and Lee 2006; Takeo, Yoshida-Noro, and Tochinai 2008, 2009). One of these, *grimp*, which was identified only in the oligochaete *Enchytraeus japonensis*, is required for mesodermal cell proliferation at the onset of blastema formation (Takeo, Yoshida-Noro, and Tochinai 2009), and others (*mino,* a EF-hand domain-containing gene, and *horu*, a gene without detectable homology) are digestive tract region markers that are remodelled as part of morphallactic regeneration (Takeo, Yoshida-Noro, and Tochinai 2008).

3.1.4. Novel and orphan homeobox genes

The entirely new genes discussed above, which have no detectable homology, are distinct from 'novel' or 'orphan' homeobox genes of the STE or the others that are described herein (*e.g.* section 3.3.6). Whereas the latter may have arisen via *de novo* transcription of previously non-coding sequence (reviewed by Tautz and Domazet-Lošo 2011; McLysaght and Guerzoni 2015; McLysaght and Hurst 2016), the presence of any known domain effectively guarantees that at least the domain-containing region of the new gene is derived from the duplication of an existing gene. However, if the precise orthology or paralogy relationships to other gene families is now undetectable, the new gene has presumably undergone substantial change in the form of exon gain/loss, exon/domain shuffling, or sequence divergence under neutral drift or positive selection (Chen, Krinsky, and Long 2013). P. W. Holland *et al.* (2017) highlighted these mechanisms of gene origination — with the mollusc STE genes of Paps *et al.* (2015) as a specific example — as an underappreciated evolutionary force.

Unlike *de novo* genes, where their biochemical function is generally unknown and not shaped by natural selection prior to their origination, these genes appear with a preexisting capacity for transcriptional regulation (though some appear to lose it; *e.g.* PRD

Clade VI homeodomains are missing 28 central residues). However, unlike homeobox genes with detectable paralogy stretching back deep in time, these genes have diverged substantially from their cryptic paralogue 'parent,' either as a result of a disruptive gene duplication event or subsequent asymmetrical change. Presumably, this divergence is accompanied by (or contingent upon) release from the constraints that governed the original gene, and the novel gene can now be expressed and used to regulate other genes in new contexts and networks depending on the degree to which regulatory environment and protein activity are preserved.

The behaviour of cryptic and highly divergent paralogues has not been systematically contrasted to *de novo* originated genes. They are generally included under the broad envelope of novel or orphan genes (Tautz and Domazet-Lošo 2011; Chen, Krinsky, and Long 2013) but obscured completely by automatic phylostratigraphic pipelines which classify genes according to oldest BLAST-detectable homology; intact homeoboxes are therefore placed in the phylostratum common to the Eukaryota. Resolving phylostratigraphy based on orthology/paralogy, rather than general homology, is prohibitively labour-intensive, but might be informative to our understanding of transcription factor evolution.

On the basis of the loosening of the constraints around their parent gene and the probable initial radical divergence*,* we might expect novel homeobox genes to have more in common (in terms of evolutionary dynamics) with *de novo* genes (see previous section) than with non-cryptic paralogues. Conversely, factors like gene length, exon number, interactivity/connectivity, and timing/uniformity/extent of expression have previously been observed to affect gene fate (Aury *et al.*, 2006; Makino, Hokamp, and McLysaght 2009; Rodgers-Melnick *et al.*, 2011; Chain, Dushoff, and Evans 2011; Satake *et al.*, 2012; Fares *et al.*, 2013; Jiang *et al.*, 2013; McGrath *et al.*, 2014) and could be variably inherited from the parent gene. To the extent that the fate of novel genes is influenced by these factors, cryptic paralogues might be expected to behave differently to genes produced by *de novo* origination.

An example of a novel homeobox gene which evolved important taxonomically-restricted ontogenic roles is *bicoid* (*bcd*), a morphogen responsible for anterior patterning only in cyclorrhaphan flies (reviewed by McGregor 2005). The *Hox3* pro-orthologue in the cyclorrhaphan ancestor duplicated and asymmetrically diverged to produce *zerknüllt* (*zen*) and *bcd*, although this relationship was unknown before 1999 (Stauber, Jäckle, and Schmidt-Ott 1999). Ancestral (but insect-specific) maternal and zygotic roles were segregated between *bcd* and *zen* respectively. *bcd* seems to have gained new *trans*-regulatory targets and, along with a cooption of *exuperantia*, new anterior mRNA localisation (Oliveira *et al.*, 2017), 'usurping' the existing *ocelliless/hunchback/caudal*-based anterior determination mechanism and resulting in changes to the *cis*-regulation of those genes (Lemke *et al.*, 2008). These changes have been suggested to allow faster development (McGregor 2005). *bcd* has been secondarily lost in some cyclorrhaphans, being partially replaced by *panish*, a taxonomically-restricted and apparently chimeric cysteine-clamp gene (Klomp *et al.*, 2015).

The fact that *bcd* could be satisfactorily identified as a *Hox3/zen* paralogue and that particulars about its evolutionary history could be reconstructed count against its direct applicability to the homeobox gene expansions discussed herein, the origins of which may be permanently obscure (see sections 3.4.6 & 3.4.7). However, it does indicate that the gene networks governing important and essential ontogenic processes in early development are susceptible to radical revision associated with the introduction of new genes, and that these changes, like the ones before them, are also not immutable.

3.1.5. Aims

I aimed to classify the homeobox content of the regenerative transcriptomes of *S. lamarcki*. In the course of identifying these sequences, several were discovered that did not appear to belong to any specific homeobox gene family or orthology group. To identify these difficult-to-classify genes, surveys of available lophotrochozoan genomes were undertaken, and the results combined with other published sequences were used to build robust homeodomain phylogenies.

3.2. Methods

The transcriptome was sampled, sequenced and assembled as described in section 2.2.1, from animals kept per section 2.1.1. A survey of the homeobox complement of the regeneration transcriptomes were carried out as described in section 2.3.2. An alignment of the transcriptomic homeoboxes, *S. lamarcki* sequences from previous studies (Kenny and Shimeld 2012; Hui 2008; McDougall *et al.*, 2011) and homeobox sequences from *B. floridae* and *T. castaneum* were used to produce a neighbour-joining phylogeny per section 2.4. On the basis of this and examination of alignments including non-homeodomain sequence, all but 10 of the 70 transcriptome sequences were identified.

To solve the identity of these final 10, four in-depth surveys and analyses were performed; of Hox and ParaHox genes, TALE-class genes, PRD-class genes, and Nk1-7, Tlx, Lbx and Msx genes. These surveys were performed as described in sections 2.3.3.1 (TALE and PRD) and 2.3.3.2 (Hox/ParaHox and Nk/Tlx/Lbx/Msx). Homeodomain alignments of these surveys were used to produce neighbour-joining, maximum likelihood, and Bayesian phylogenetic analyses. Support values for each analysis were mapped onto the Bayesian tree and the results visualised (section 2.4).

3.2.1. Cryptic species check

A search was designed to determine the extent of sequence divergence between the Plymouth population (from which the material for the developmental transcriptome and genome were collected – Kenny and Shimeld 2012; Kenny *et al.*, 2015) and the St Andrews Bay population, and to verify that the regeneration transcriptomes weren't derived from different (cryptic) species. BLASTn searches of the genome were performed with the entire regeneration transcriptome co-assembly, the developmental transcriptome, and the open reading frames (ORFs) of the homeobox-containing contigs retrieved by earlier searches.

From the transcriptomic searches, the number and lengths of contigs returning no hits were collected, as well as the identities and gaps for the top result of those with hits. For the homeoboxes, each result was manually inspected and the identities and gaps recorded.

3.3. Results

3.3.1. Cryptic species survey

The results of the whole transcriptome *versus* genome searches are presented in Table 3.1. Of the homeobox-containing sequences from the regeneration transcriptomes, the mean identity was 99.0% (3 s.f.), with a mean of 0.128 gaps (3 s.f.). No transcriptome sequence was found to be absent from the genome. Most sequences (50 of 56, ~90%) were represented by more than one genomic sequence over at least some of their length, with varying degrees of sequence identity. It was concluded on this basis that the Plymouth and St Andrews Bay populations are not separate cryptic species, and that the gene pairs discovered during my surveys were not being misidentified as paralogues instead of orthologues.

Table 3.1. Metrics from BLASTn searches of the *S. lamarcki* transcriptomes against the genome. All percentages are reported to three significant figures.

Database	Database size, # seqs	Proportion of seqs not returning a hit	Database size, # bases	Proportion of bases in seqs not returning a hit	Mean identitiy of top hit	Mean gaps in top hit, %
Trinity co-assembly – (regeneration)	360,107	24.9%	221,075,893	17.2%	95.7%	1.14%
Kenny & Shimeld, 2012 (development)	50,151	0.160%	61,261,605	0.114%	97.2%	0.735%

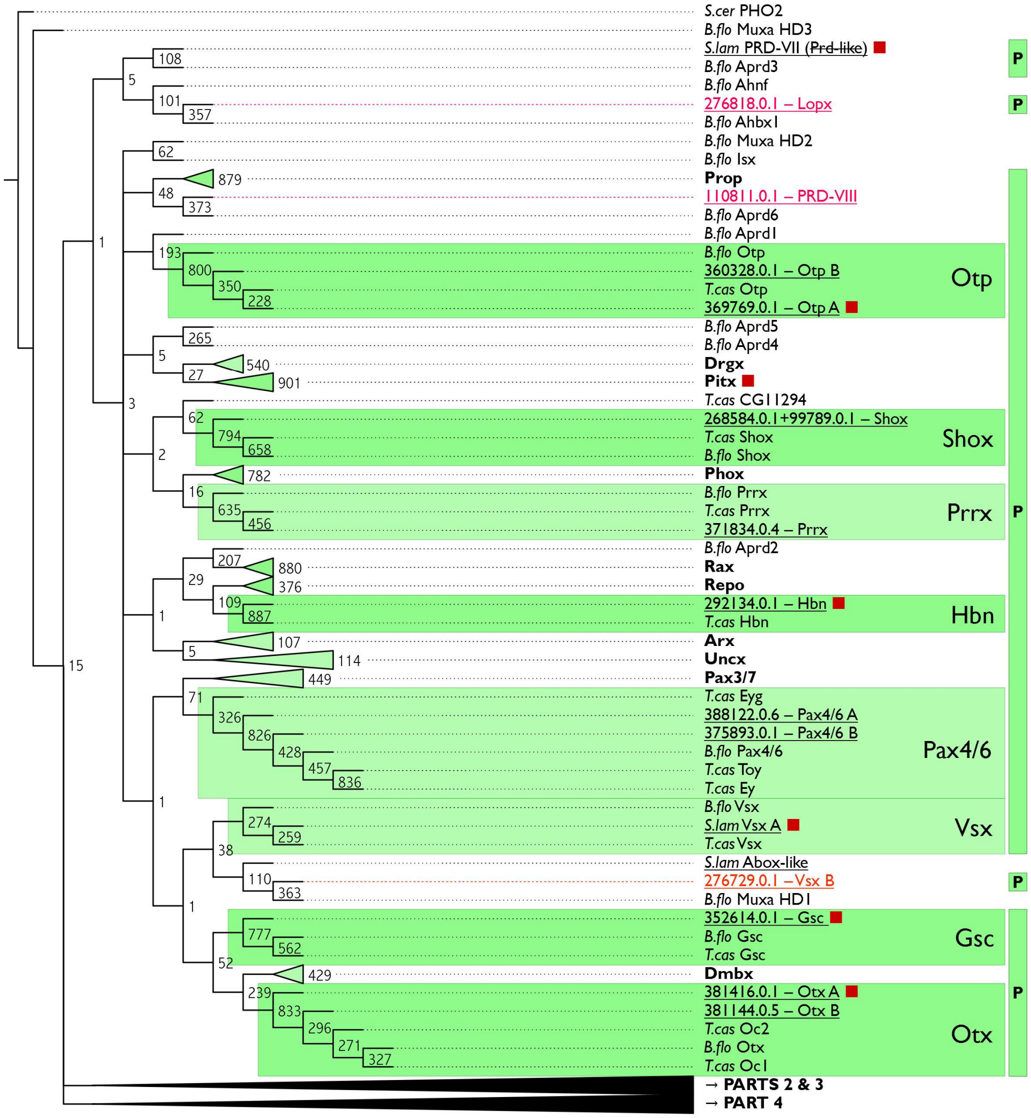
3.3.2. The homeodomain content of a regenerative transcriptome

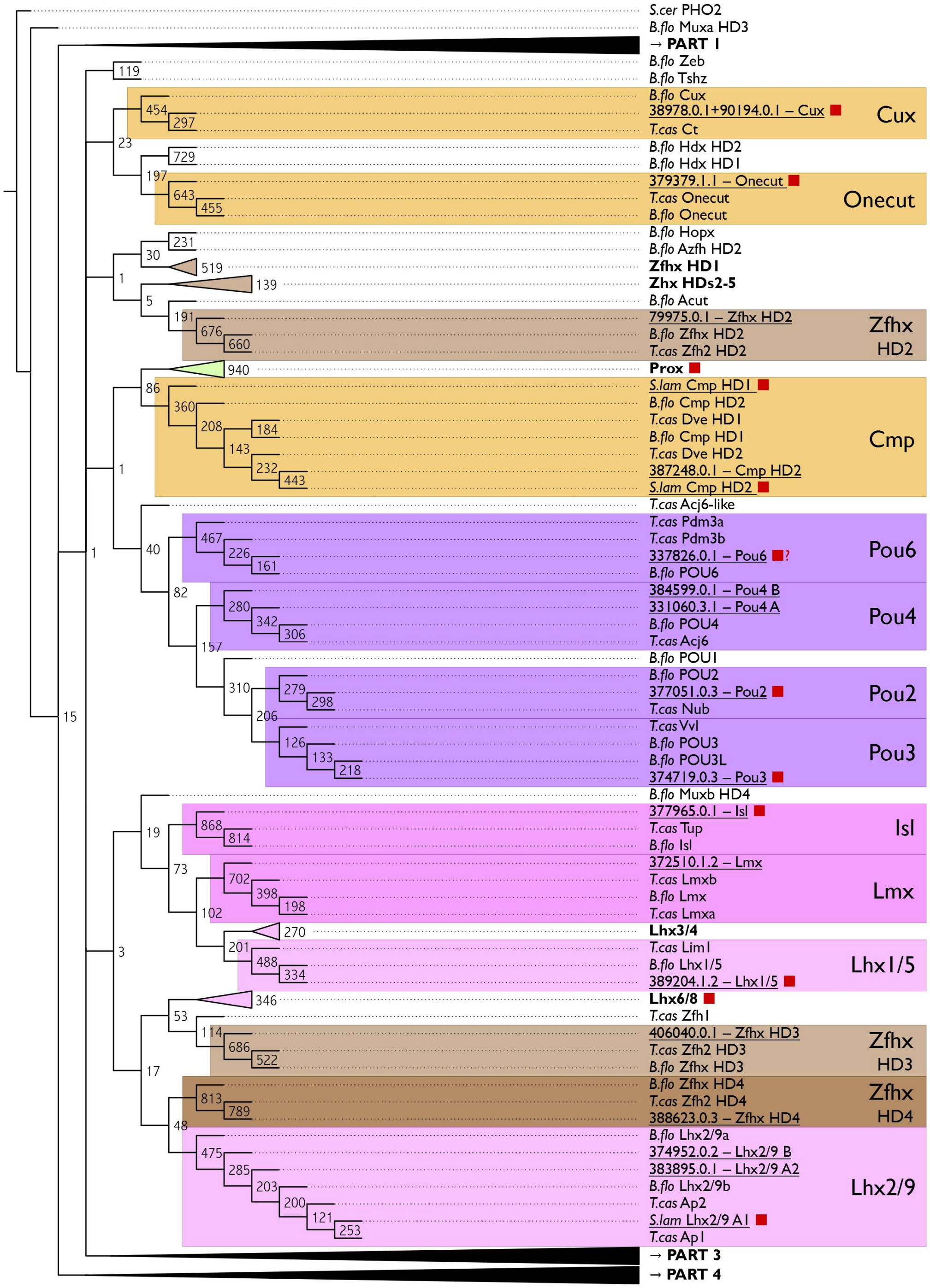
The transcriptomes of *S. lamarcki* operculum regeneration were analysed for their homeobox gene family content. A brief précis of the results is presented in Table 3.2, a summary of salient information is presented in Appendix 3.2a, and complete details are presented in Appendix 3.2b.

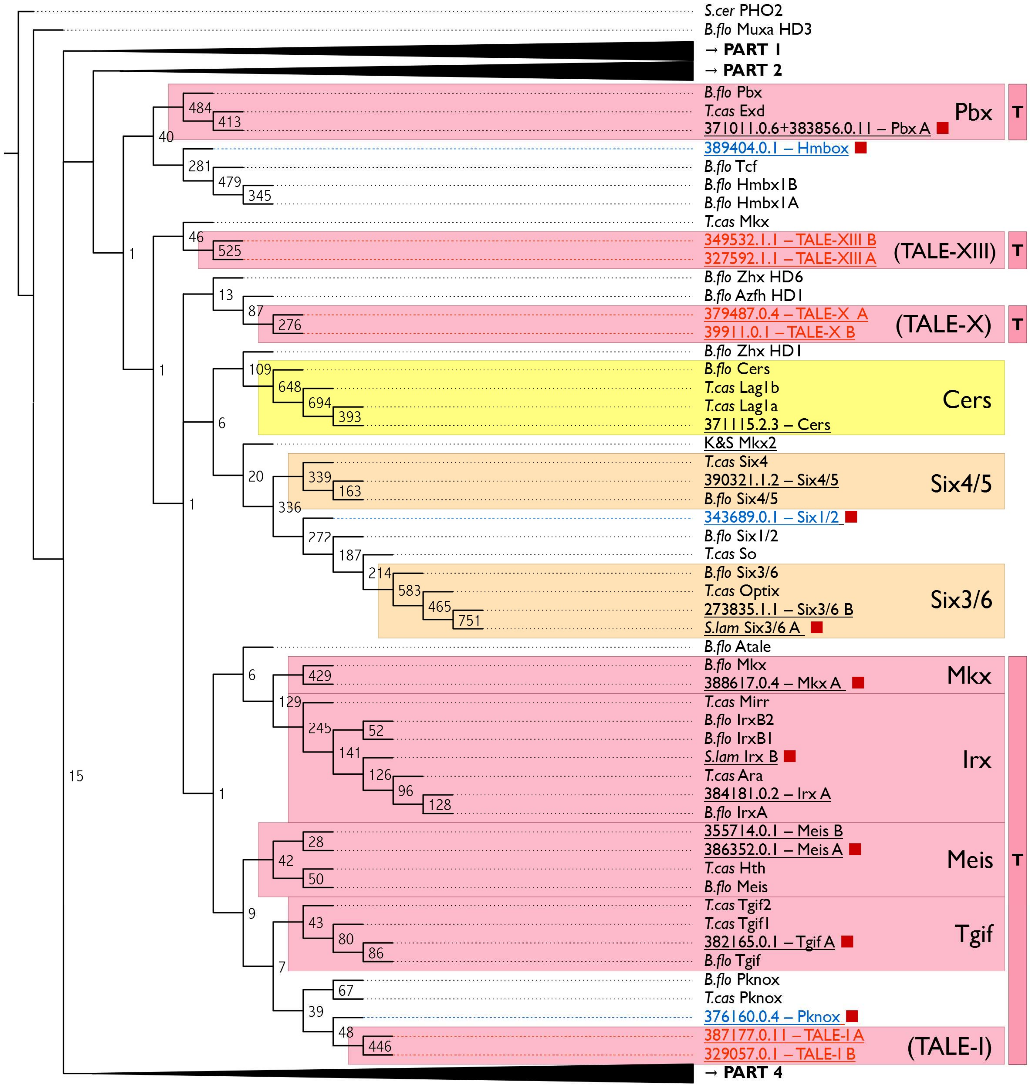
Table 3.2. Summary of homeobox-containing sequences found in the *S. lamarcki* regeneration transcriptomes. Sequences previously identified by McDougall *et al.*, (2011) are marked with a dagger, and those previously identified by Kenny & Shimeld (2012) are marked with an asterisk. Difficult-to-classify genes are marked in bold, and those belonging to gene families or clades described herein are underlined. Taken from Barton-Owen, Szabó, Somorjai, & Ferrier (2018).

CLASS	FAMILY /NAME
ANTP:	Antp
	BarH
	BarX
	Dbx*
	Dlx-a †
	Dlx-b †
	Emx A
	Emx B
	En
	Msx
	Msxlx
	Nk1a
	Nk1b
	Nk2.1a*
	Nk2.1b*
	Nk2.2b
	Nk5*
	Nk6*
	Spiro-Nk
	Tlx E
CERS:	Cers*
CUT:	Cmp*
	Cux*
	Onecut*
HNF	Hmbox*
LIM:	Isl*
	Lhx1/5*
	Lhx2/9 A2*
	Lhx2/9 B
	Lmx
(unclassified):	Lopx
ZF:	Zfhx

CLASS	FAMILY /NAME
POU:	Pou2*
	Pou3*
	Pou4 A
	Pou4 B
	Pou6
PRD:	Gsc*
	Hbn*
	Otp A*
	Otp B
	Otx A*
	Otx B
	Pax4/6 A
	Pax4/6 B
	PRD-VIII
	Prrx
	Shox
	Vsx B
SINE:	Six1/2*
	Six3/6 (B)
	Six4/5
TALE:	Irx A
	TALE-I A
	TALE-I B
	TALE-X A
	TALE-X B
	TALE-XIII A
	TALE-XIII B
	Meis A*
	Meis B
	Mkx A*
	Pbx A*
	Pknox*
	Tgif A*


S.cer PHO2
B.flo Muxa HD3
S.lam PRD-VII (~~Prd-like~~) 🟥
108
B.flo Aprd3
5
B.flo Ahnf
101
276818.0.1 – Lopx
357
B.flo Ahbx1
62
B.flo Muxa HD2
B.flo Isx
879
Prop
48
110811.0.1 – PRD-VIII
373
B.flo Aprd6
1
B.flo Aprd1
193
B.flo Otp
800
360328.0.1 – Otp B
350
T.cas Otp
228
369769.0.1 – Otp A 🟥
Otp
265
B.flo Aprd5
5
B.flo Aprd4
540
Drgx
27
901
Pitx 🟥
3
T.cas CG11294
62
268584.0.1+99789.0.1 – Shox
794
T.cas Shox
658
B.flo Shox
Shox
2
782
Phox
16
B.flo Prrx
635
T.cas Prrx
456
371834.0.4 – Prrx
Prrx
207
B.flo Aprd2
880
Rax
29
376
Repo
109
292134.0.1 – Hbn 🟥
887
T.cas Hbn
Hbn
1
107
Arx
5
114
Uncx
15
449
Pax3/7
71
T.cas Eyg
326
388122.0.6 – Pax4/6 A
826
375893.0.1 – Pax4/6 B
428
B.flo Pax4/6
457
T.cas Toy
836
T.cas Ey
Pax4/6
1
B.flo Vsx
274
S.lam Vsx A 🟥
259
T.cas Vsx
Vsx
38
S.lam Abox-like
110
276729.0.1 –Vsx B
363
B.flo Muxa HD1
1
352614.0.1 – Gsc 🟥
777
B.flo Gsc
562
T.cas Gsc
Gsc
52
429
Dmbx
239
381416.0.1 – Otx A 🟥
833
381144.0.5 – Otx B
296
T.cas Oc2
271
B.flo Otx
327
T.cas Oc1
Otx
→ PARTS 2 & 3
→ PART 4
P
P
P
P
P


S.cer PHO2
B.flo Muxa HD3
→ PART 1
B.flo Zeb
B.flo Tshz
B.flo Cux
38978.0.1+90194.0.1 – Cux
T.cas Ct
Cux
B.flo Hdx HD2
B.flo Hdx HD1
379379.1.1 – Onecut
T.cas Onecut
B.flo Onecut
Onecut
B.flo Hopx
B.flo Azfh HD2
Zfhx HD1
Zhx HDs2-5
B.flo Acut
79975.0.1 – Zfhx HD2
B.flo Zfhx HD2
T.cas Zfh2 HD2
Zfhx HD2
Prox
S.lam Cmp HD1
B.flo Cmp HD2
T.cas Dve HD1
B.flo Cmp HD1
T.cas Dve HD2
387248.0.1 – Cmp HD2
S.lam Cmp HD2
Cmp
T.cas Acj6-like
T.cas Pdm3a
T.cas Pdm3b
337826.0.1 – Pou6 ?
B.flo POU6
Pou6
384599.0.1 – Pou4 B
331060.3.1 – Pou4 A
B.flo POU4
T.cas Acj6
Pou4
B.flo POU1
B.flo POU2
377051.0.3 – Pou2
T.cas Nub
Pou2
T.cas Vvl
B.flo POU3
B.flo POU3L
374719.0.3 – Pou3
Pou3
B.flo Muxb HD4
377965.0.1 – Isl
T.cas Tup
B.flo Isl
Isl
372510.1.2 – Lmx
T.cas Lmxb
B.flo Lmx
T.cas Lmxa
Lmx
Lhx3/4
T.cas Lim1
B.flo Lhx1/5
389204.1.2 – Lhx1/5
Lhx1/5
Lhx6/8
T.cas Zfh1
406040.0.1 – Zfhx HD3
T.cas Zfh2 HD3
B.flo Zfhx HD3
Zfhx HD3
B.flo Zfhx HD4
T.cas Zfh2 HD4
388623.0.3 – Zfhx HD4
Zfhx HD4
B.flo Lhx2/9a
374952.0.2 – Lhx2/9 B
383895.0.1 – Lhx2/9 A2
B.flo Lhx2/9b
T.cas Ap2
S.lam Lhx2/9 A1
T.cas Ap1
Lhx2/9
→ PART 3
→ PART 4
119
454
297
23
729
197
643
455
231
30
519
1
139
5
191
676
660
940
86
360
208
184
143
232
443
1
40
467
226
161
1
82
280
342
306
157
310
279
298
206
126
133
218
15
19
868
814
73
702
398
198
102
270
201
488
334
3
346
53
114
686
522
17
813
789
48
475
285
203
200
121
253


S.cer PHO2
B.flo Muxa HD3
→ PART 1
→ PART 2
B.flo Pbx
T.cas Exd
371011.0.6+383856.0.11 – Pbx A
Pbx
T
389404.0.1 – Hmbox
B.flo Tcf
B.flo Hmbx1B
B.flo Hmbx1A
T.cas Mkx
349532.1.1 – TALE-XIII B
327592.1.1 – TALE-XIII A
(TALE-XIII)
T
B.flo Zhx HD6
B.flo Azfh HD1
379487.0.4 – TALE-X A
39911.0.1 – TALE-X B
(TALE-X)
T
B.flo Zhx HD1
B.flo Cers
T.cas Lag1b
T.cas Lag1a
371115.2.3 – Cers
Cers
K&S Mkx2
T.cas Six4
390321.1.2 – Six4/5
B.flo Six4/5
Six4/5
343689.0.1 – Six1/2
B.flo Six1/2
T.cas So
B.flo Six3/6
T.cas Optix
273835.1.1 – Six3/6 B
S.lam Six3/6 A
Six3/6
B.flo Atale
B.flo Mkx
388617.0.4 – Mkx A
Mkx
T.cas Mirr
B.flo IrxB2
B.flo IrxB1
S.lam Irx B
T.cas Ara
384181.0.2 – Irx A
B.flo IrxA
Irx
355714.0.1 – Meis B
386352.0.1 – Meis A
T.cas Hth
B.flo Meis
Meis
T
T.cas Tgif2
T.cas Tgif1
382165.0.1 – Tgif A
B.flo Tgif
Tgif
B.flo Pknox
T.cas Pknox
376160.0.4 – Pknox
387177.0.11 – TALE-I A
329057.0.1 – TALE-I B
(TALE-I)
→ PART 4
484
413
40
281
479
345
46
525
1
13
87
276
1
109
648
694
393
6
20
339
163
336
1
272
187
214
583
465
751
6
429
129
245
52
141
126
96
128
15
1
28
42
50
9
43
80
86
7
67
39
48
446

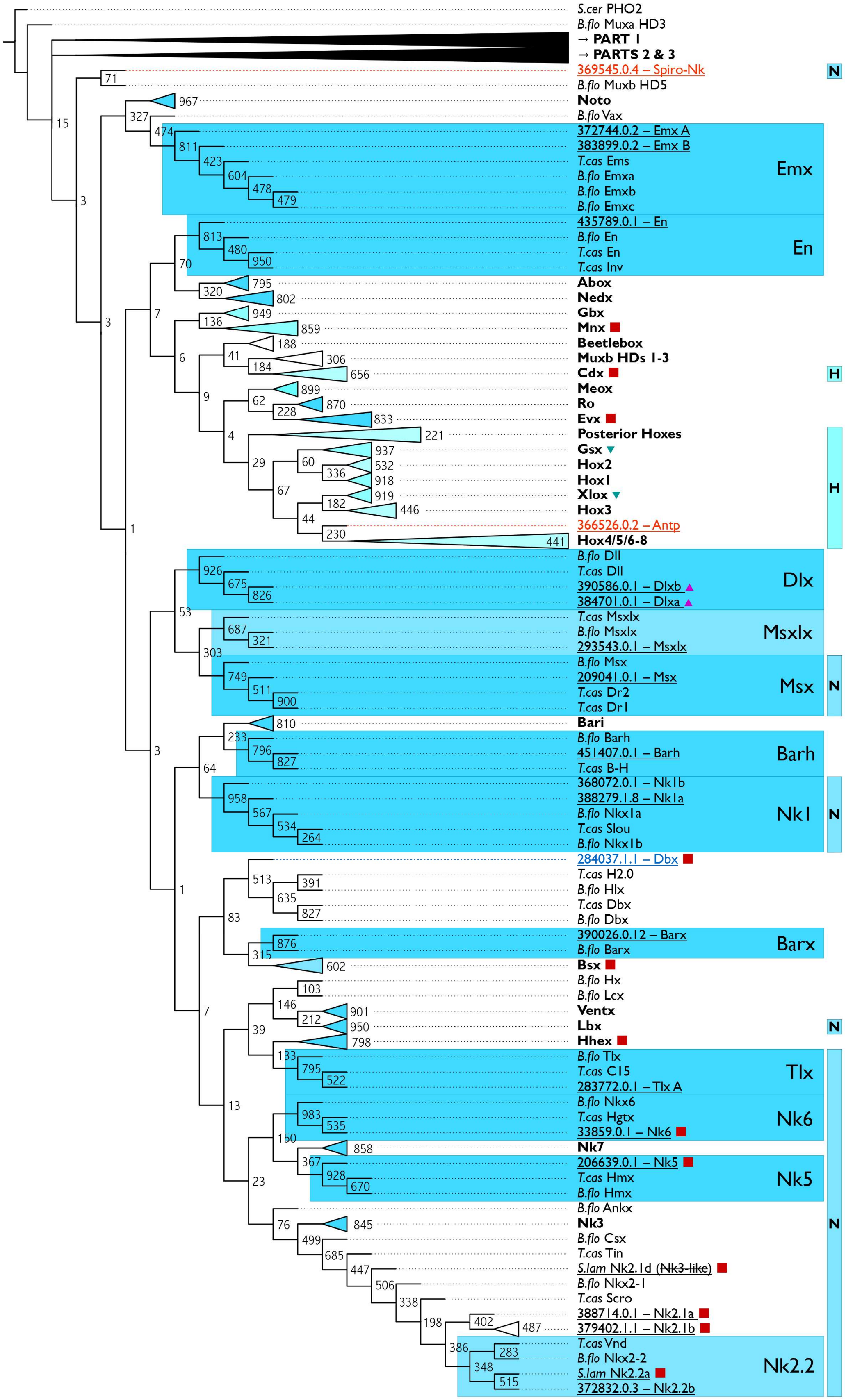

S.cer PHO2
B.flo Muxa HD3
→ PART 1
→ PARTS 2 & 3
71
369545.0.4 – Spiro-Nk
N
B.flo Muxb HD5
967
Noto
15
327
B.flo Vax
474
372744.0.2 – Emx A
811
383899.0.2 – Emx B
423
T.cas Ems
604
B.flo Emxa
478
B.flo Emxb
479
B.flo Emxc
Emx
3
435789.0.1 – En
813
B.flo En
480
T.cas En
950
T.cas Inv
70
En
795
Abox
320
802
Nedx
3
7
949
Gbx
136
859
Mnx
188
Beetlebox
41
306
Muxb HDs 1-3
6
184
656
Cdx
H
899
Meox
62
870
Ro
9
228
833
Evx
4
221
Posterior Hoxes
937
Gsx
60
532
Hox2
29
336
918
Hox1
67
919
Xlox
182
446
Hox3
44
366526.0.2 – Antp
H
230
441
Hox4/5/6-8
1
B.flo Dll
926
T.cas Dll
675
390586.0.1 – Dlxb
826
384701.0.1 – Dlxa
Dlx
53
T.cas Msxlx
687
B.flo Msxlx
321
293543.0.1 – Msxlx
Msxlx
303
B.flo Msx
749
209041.0.1 – Msx
511
T.cas Dr2
900
T.cas Dr1
Msx
N
810
Bari
233
B.flo Barh
796
451407.0.1 – Barh
827
T.cas B-H
Barh
3
64
368072.0.1 – Nk1b
958
388279.1.8 – Nk1a
567
B.flo Nkx1a
534
T.cas Slou
264
B.flo Nkx1b
Nk1
N
284037.1.1 – Dbx
513
391
T.cas H2.0
B.flo Hlx
1
635
T.cas Dbx
827
B.flo Dbx
83
390026.0.12 – Barx
876
B.flo Barx
315
Barx
602
Bsx
103
B.flo Hx
B.flo Lcx
146
901
Ventx
7
212
950
Lbx
N
39
798
Hhex
133
B.flo Tlx
795
T.cas C15
522
283772.0.1 – Tlx A
Tlx
B.flo Nkx6
983
T.cas Hgtx
535
33859.0.1 – Nk6
Nk6
13
150
858
Nk7
367
206639.0.1 – Nk5
928
T.cas Hmx
670
B.flo Hmx
Nk5
23
B.flo Ankx
76
845
Nk3
499
B.flo Csx
685
T.cas Tin
447
S.lam Nk2.1d ~~(Nk3-like)~~
506
B.flo Nkx2-1
338
T.cas Scro
198
402
388714.0.1 – Nk2.1a
487
379402.1.1 – Nk2.1b
386
T.cas Vnd
283
B.flo Nkx2-2
348
S.lam Nk2.2a
515
372832.0.3 – Nk2.2b
Nk2.2
N

Figure 3.1 (4 parts). Neighbour-joining cladogram of homeodomain sequences found in the *S. lamarcki* (<u>*S.lam*</u>) regeneration transcriptomes, against the homeodomain sequences of *Tribolium castaneum* (*T.cas*) and *Branchiostoma floridae* (*B.flo*), rooted using the *Saccharomyces cerevisiae* (*S.cer*) PHO2 homeodomain sequence. The analysis was performed in PHYLIP using 4 gamma categories and 1000 bootstraps (support values indicated). Clades are highlighted by family; clades with less than 70% support are paler. Successfully reconstructed families which do not include a sequence from the regeneration transcriptomes have been collapsed. *S. lamarcki* sequences (including in collapsed families) which were first identified by previous studies are marked with a coloured shape; sequences identified by Kenny and Shimeld (2012) are marked with a red square, those identified by Hui (2008) by a blue downwards-pointing triangle, and those identified by McDougall *et al.* (2011) by a pink upwards-pointing triangle. Regeneration transcriptome sequence IDs are abbreviated (*i.e.* comp276818_c0_seq1 is written 276818.0.1). Sequences falling outside clades of established families are coloured blue if they were identified on the basis of similarity to sequences from Kenny and Shimeld (2012) and red if they were identified on the basis of subsequent focussed analyses. Where a gene has been reclassified from Kenny and Shimeld (2012), the old classification is included but struck out. Sequences included in said analyses are marked by coloured boxes to the right-hand side: P (PRD class), Figs. 3.6 & 3.7, Table 3.6; T (TALE class), Figs 3.4 & 3.5, Tables 3.4 & 3.5; N (Nk, Msx, Tlx & Lbx families), Figs. 3.8 & 3.9, Table 3.7; H (Hox and ParaHox families), Figs 3.2 & 3.3, Table 3.3. New clade names proposed herein as a result of these analyses are marked with an asterisk. The alignment used to produce this tree is presented in Appendix 3.4a, and the complete Newick format tree is presented in Appendix 3.3a.

Seventy transcriptome sequences were identified, of which sixty could be assigned to canonical homeobox families by BLAST searches, protein sequence alignment and homeodomain phylogenetic analyses. The neighbour-joining homeodomain phylogeny used as a basis for the majority of classification is presented in Figure 3.1. Twenty-five sequences were identical or near-identical to sequences previously described by Kenny & Shimeld (2012), and two were identical or near-identical to the *Dlx-a* and *Dlx-b* sequences previously described by McDougall *et al.* (2011). Three likely belong to the same multi-homeodomain gene (*Zfhx*). Three pairs were merged based on bridging genomic or developmental transcriptomic sequence. The remaining ten could not be placed in canonical clades, and a selection of detailed analyses was performed to classify these genes and to survey the various gene duplications in *S. lamarcki*.

3.3.3. A divergent *Antp*

Among the difficult-to-classify genes was an unusual Hox/ParaHox-like gene. A broad selection of bilaterian Hox and ParaHox cluster protein sequences (details in Appendix 3.2b) were collected and aligned (Appendix 3.4b), and a partially collapsed

Bayesian phylogeny with support values added from equivalent neighbour-joining and maximum likelihood analyses was produced (Figure 3.2, Table 3.3), based on the homeodomain and ten flanking positions (five from each side of the homeodomain). Candidate *S. lamarcki* orthologues were found in the whole genome sequence (Kenny *et al.*, 2015) for all expected polychaete Hox (Fröbius, Matus, and Seaver 2008) and ParaHox (Kulakova, Cook, and Andreeva 2008; Hui *et al.*, 2009) families except *Antp* and *Post1*. Unfortunately (but not unusually for homeodomain phylogenies of Hox genes), the analyses did not place *Dfd*, *Scr*, *Antp* and *Lox4* in distinct clades, but did place the unidentified gene in this undifferentiated Hox4/5/Medial clade (Figure 3.2 (2 parts)). On the basis of this placement and consistent support excluding it from other Hox/ParaHox clades, I concluded that the unidentified gene is most probably the missing *Antp* gene.

An alignment of this putative *S. lamarcki Antp* against other lophotrochozoan Antp proteins and a broader selection of other medial Hox sequences (Figure 3.3) reveals that six residues in the homeodomain (marked by dots) are invariant across all included Hox sequences except the putative *S. lamarcki* Antp.

Table 3.3. Summary of the collapsed gene families in Figure 3.2. Clade colouration and symbols denoting sequence origin are the same as in Figure 3.2. Annelid species: *S.lam* = *Spirobranchus lamarcki*; *C.tel* = *Capitella teleta*; *A.vir* = *Alitta virens*; *H.rob* = *Helobdella robusta*; *P.dum* = *Platynereis dumerilii*. Brachiopod species: *L.ana* = *Lingula anatina*; *T.tra* = *Terebratalia transversa*. Mollusc species: *C.gig* = *Crassostrea gigas*; *L.gig* = *Lottia gigantea*; *E.sco* = *Euprymna scolopes*; *O.bim* = *Octopus bimaculoides*. Deuterostome species: *B.flo* = *Branchiostoma floridae*; *S.kow* = *Saccoglossus kowalevskii*; *S.pur* = *Strongylocentrotus purpuratus*. Cnidarian species: *N.vec* = *Nematostella vectensis*.

	lab	pb	Hox3	(Dfd)	(Scr)	Lox5	(Antp)	(Lox4)	Lox2	Post2	Post1	Pa-c	Gsx	Xlox	Cdx
S.lam	1	1	1	1	1	1	1 ◆	1	1	1	0	_	1	1	1
C.tel	1	1	1	1	1	1	1	1	1	1	1		1	1	1
H.rob	2			2	3	1	1	2	2	1				1	1
A.vir	1	1	1	1	1	1	1	1	1	1	1		1		1
P.dum	1	1	1	1	1	1	1	1	1	1	1		1	1	1
C.gig	1	1	1	1	1	1		1	1	1	1		1	1	1
L.gig	1	1	1	1	1	1	1	1	1	1	1		1	1	1
E.sco	1		1	1	1	1	1	1		1	1		1		
O.bim	1				1	1	1	1	1	1	1		1	1	
L.ana	1	1	1	1	1	1	1	1		1	1		1	1	1
T.tra	1	1	1	1	1	1	1	1		1	1				1
B.flo	1	1	1	1	1								1	1	1
S.kow	1	1	1	1	1							3	1	1	1
S.pur		1	1		1							3	1	1	1
N.vec													1		

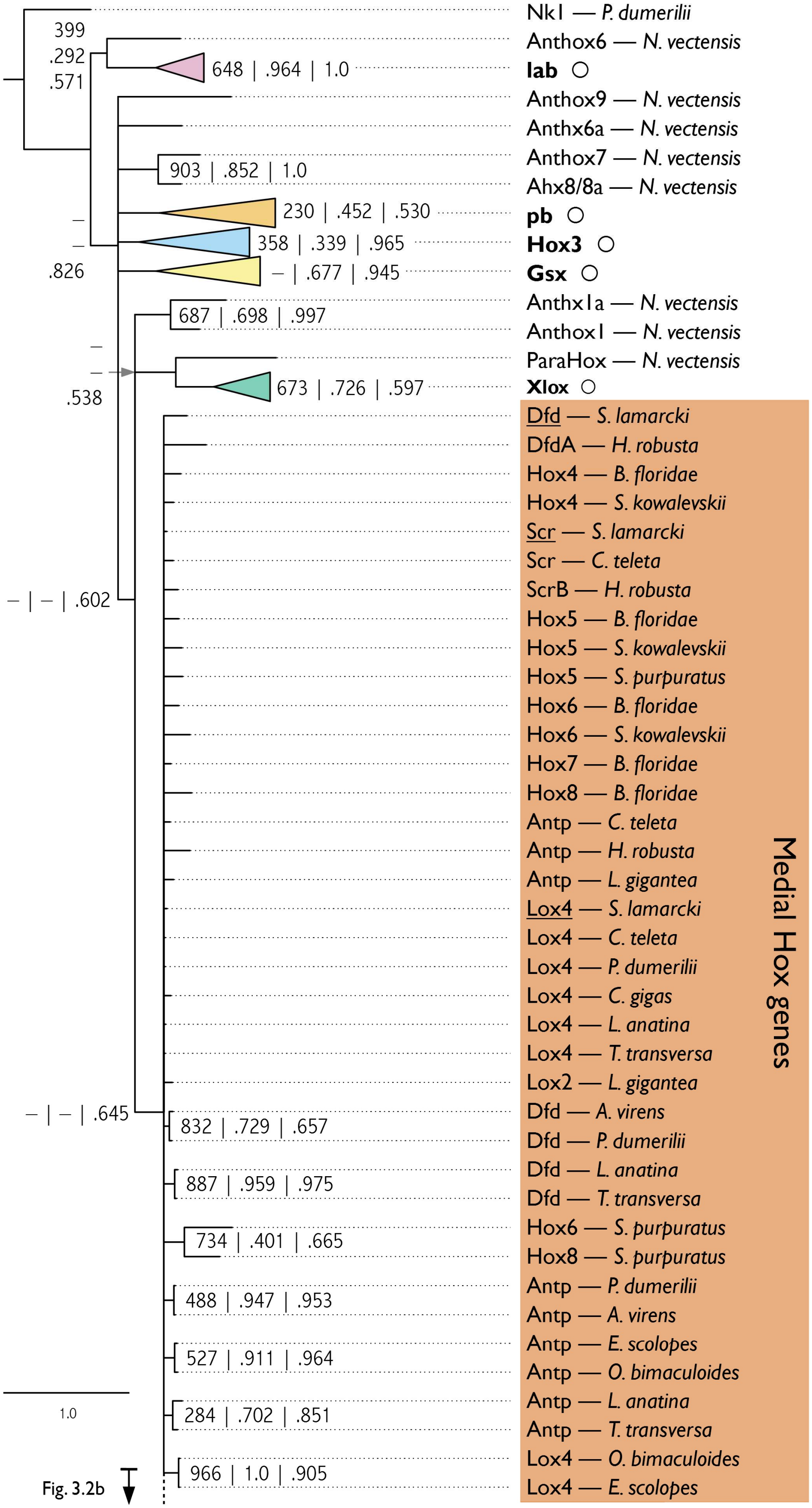

Nk1 — P. dumerilii
Anthox6 — N. vectensis
lab ○
648 | .964 | 1.0
399
.292
.571
Anthox9 — N. vectensis
Anthx6a — N. vectensis
Anthox7 — N. vectensis
Ahx8/8a — N. vectensis
903 | .852 | 1.0
pb ○
230 | .452 | .530
Hox3 ○
358 | .339 | .965
Gsx ○
– | .677 | .945
–
–
.826
Anthx1a — N. vectensis
Anthox1 — N. vectensis
687 | .698 | .997
ParaHox — N. vectensis
Xlox ○
673 | .726 | .597
–
–
.538
Dfd — S. lamarcki
DfdA — H. robusta
Hox4 — B. floridae
Hox4 — S. kowalevskii
Scr — S. lamarcki
Scr — C. teleta
ScrB — H. robusta
– | – | .602
Hox5 — B. floridae
Hox5 — S. kowalevskii
Hox5 — S. purpuratus
Hox6 — B. floridae
Hox6 — S. kowalevskii
Hox7 — B. floridae
Hox8 — B. floridae
Antp — C. teleta
Antp — H. robusta
Antp — L. gigantea
Lox4 — S. lamarcki
Lox4 — C. teleta
Lox4 — P. dumerilii
Lox4 — C. gigas
Lox4 — L. anatina
Lox4 — T. transversa
Lox2 — L. gigantea
Dfd — A. virens
Dfd — P. dumerilii
832 | .729 | .657
– | – | .645
Dfd — L. anatina
Dfd — T. transversa
887 | .959 | .975
Hox6 — S. purpuratus
Hox8 — S. purpuratus
734 | .401 | .665
Antp — P. dumerilii
Antp — A. virens
488 | .947 | .953
Antp — E. scolopes
Antp — O. bimaculoides
527 | .911 | .964
Antp — L. anatina
Antp — T. transversa
284 | .702 | .851
1.0
Lox4 — O. bimaculoides
Lox4 — E. scolopes
966 | 1.0 | .905
Fig. 3.2b
Medial Hox genes

Fig. 3.2a

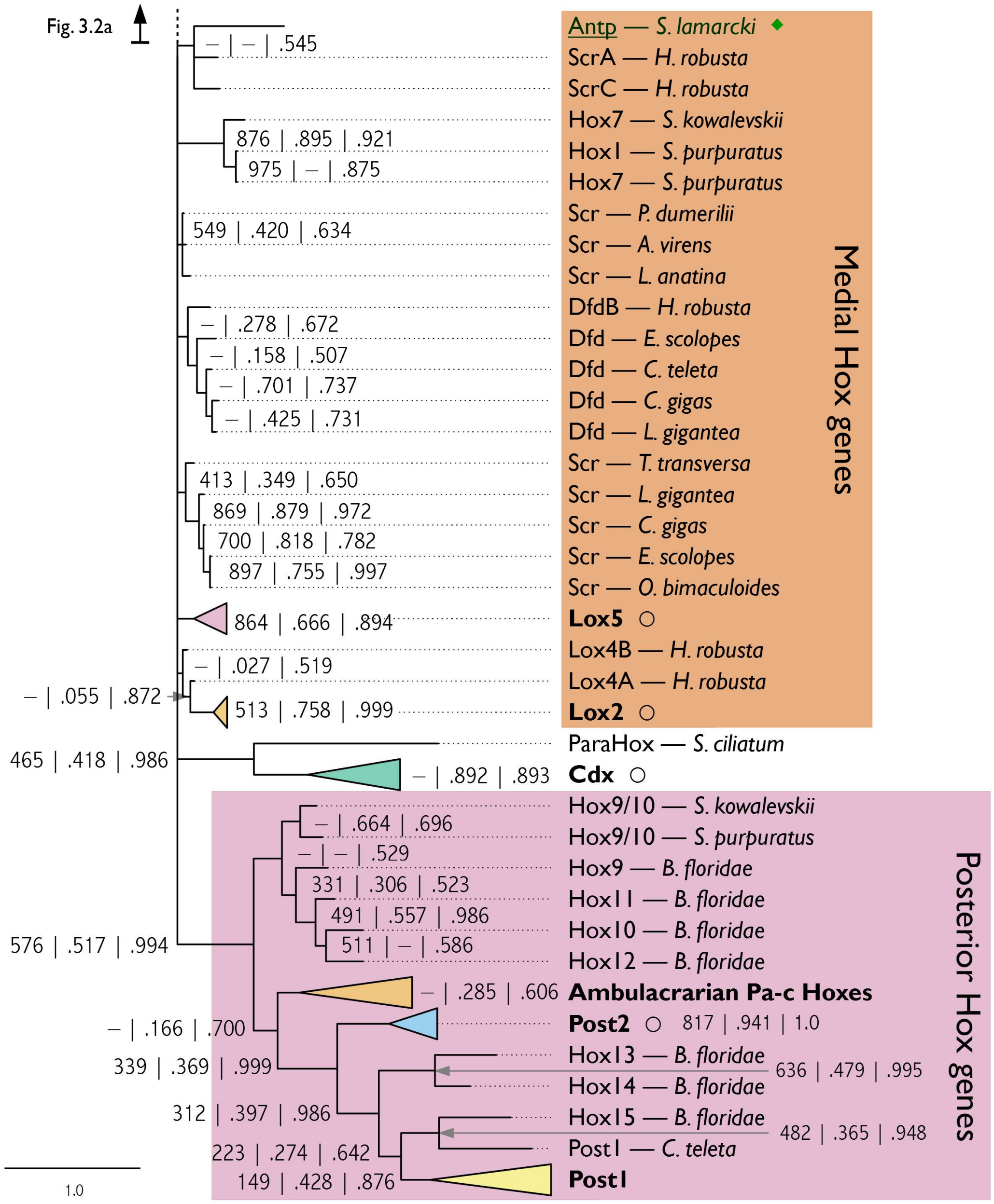

Antp — S. lamarcki ◆
– | – | .545
ScrA — H. robusta
ScrC — H. robusta
876 | .895 | .921
Hox7 — S. kowalevskii
Hox1 — S. purpuratus
975 | – | .875
Hox7 — S. purpuratus
549 | .420 | .634
Scr — P. dumerilii
Scr — A. virens
Scr — L. anatina
DfdB — H. robusta
– | .278 | .672
Dfd — E. scolopes
– | .158 | .507
Dfd — C. teleta
– | .701 | .737
Dfd — C. gigas
– | .425 | .731
Dfd — L. gigantea
Scr — T. transversa
413 | .349 | .650
Scr — L. gigantea
869 | .879 | .972
Scr — C. gigas
700 | .818 | .782
Scr — E. scolopes
897 | .755 | .997
Scr — O. bimaculoides
864 | .666 | .894
Lox5 ○
Lox4B — H. robusta
– | .027 | .519
Lox4A — H. robusta
– | .055 | .872
513 | .758 | .999
Lox2 ○
Medial Hox genes
ParaHox — S. ciliatum
465 | .418 | .986
– | .892 | .893
Cdx ○
Hox9/10 — S. kowalevskii
– | .664 | .696
Hox9/10 — S. purpuratus
– | – | .529
Hox9 — B. floridae
331 | .306 | .523
Hox11 — B. floridae
491 | .557 | .986
Hox10 — B. floridae
576 | .517 | .994
511 | – | .586
Hox12 — B. floridae
– | .285 | .606
Ambulacrarian Pa-c Hoxes
Post2 ○
817 | .941 | 1.0
– | .166 | .700
339 | .369 | .999
Hox13 — B. floridae
636 | .479 | .995
Hox14 — B. floridae
312 | .397 | .986
Hox15 — B. floridae
482 | .365 | .948
223 | .274 | .642
Post1 — C. teleta
149 | .428 | .876
Post1
Posterior Hox genes
1.0

Figure 3.2 (2 parts). Bayesian phylogeny of Hox and ParaHox homeodomains and flanking sequences from a selection of metazoan genomes, showing the basis for the identification of the divergent *S. lamarcki* Hox gene as *Antp*. Support values for each node are from neighbour-joining (out of 1000 bootstraps), maximum likelihood (proportion of 1000 bootstraps), and Bayesian (posterior probability) phylogenies (in order, separated by vertical bars or newlines). A dash indicates where a node is not present in the corresponding tree. Gene families that have been successfully reconstructed have been collapsed into coloured triangles and a summary of their contents given in Table 3.2. *S. lamarcki* sequences (all underlined) are marked with a green diamond if found in the regenerative transcriptomes, and with a black circle if only found in the genome (collapsed families only). The scale bar indicates amino acid substitutions per site. Full sequence details are included in Appendix 3.2b. The alignment used to produce this tree is presented in Appendix 3.4b, and a full version of the Newick format tree is presented in Appendix 3.3b. Annelid species: *S. lamarcki* = *Spirobranchus lamarcki*; *C. teleta* = *Capitella teleta*; *A. virens* = *Alitta virens*; *H. robusta* = *Helobdella robusta*; *P. dumerilii*/*P.dum* = *Platynereis dumerilii*. Brachiopod species: *L. anatina*/*L.ana* = *Lingula anatina*; *T. transversa* = *Terebratalia transversa*. Mollusc species: *C. gigas* = *Crassostrea gigas*; *L. gigantea* = *Lottia gigantea*; *E. scolopes* = *Euprymna scolopes*; *O. bimaculoides* = *Octopus bimaculoides*. Deuterostome species: *B. floridae* = *Branchiostoma floridae*; *S. kowalevskii* = *Saccoglossus kowalevskii*; *S. purpuratus* = *Strongylocentrotus purpuratus*. Cnidarian species: *N. vectensis* = *Nematostella vectensis*. Adapted from Barton-Owen, Szabó, Somorjai, & Ferrier (2018).

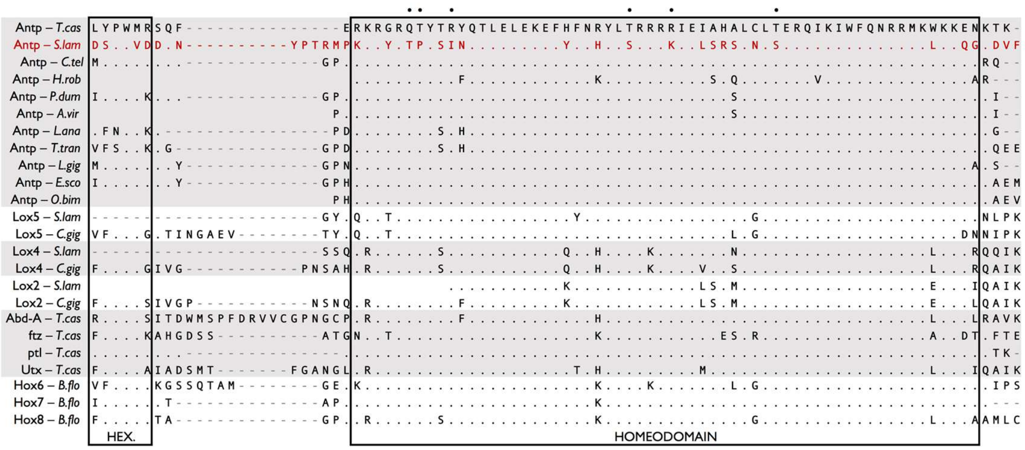


Figure 3.3. Protein sequence alignment of hexapeptide, linker, homeodomain and flanking region of medial Hox genes (Hox6-8 family) from a selection of bilaterians, demonstrating the degree of sequence divergence of *Spirobranchus* Antp (highlighted in red). Identities (full stop) are marked relative to the sequence of *Tribolium castaneum* Antp. Residue positions at which *Spirobranchus* Antp is the only variant sequence shown are marked with a black dot. Full sequence details are included in Appendix 3.2b. HEX. = hexapeptide. Annelid sequences: *S.lam* = *Spirobranchus lamarcki*; *C.tel* = *Capitella teleta*; *H.rob* = *Helobdella robusta*; *P.dum* = *Platynereis dumerilii*; *A.vir* = *Alitta virens*. Brachiopod species: *L.ana* = *Lingula anatina*; *T.tra* = *Terebratalia transversa*. Mollusc species: *C.gig* = *Crassostrea gigas*; *L.gig* = *Lottia gigantea*; *E.sco* = *Euprymna scolopes*; *O.bim* = *Octopus bimaculoides*. Insect species: *T.cas* = *Tribolium castaneum*. Deuterostome species: *B.flo* = *Branchiostoma floridae*. Taken from Barton-Owen, Szabó, Somorjai, & Ferrier (2018).

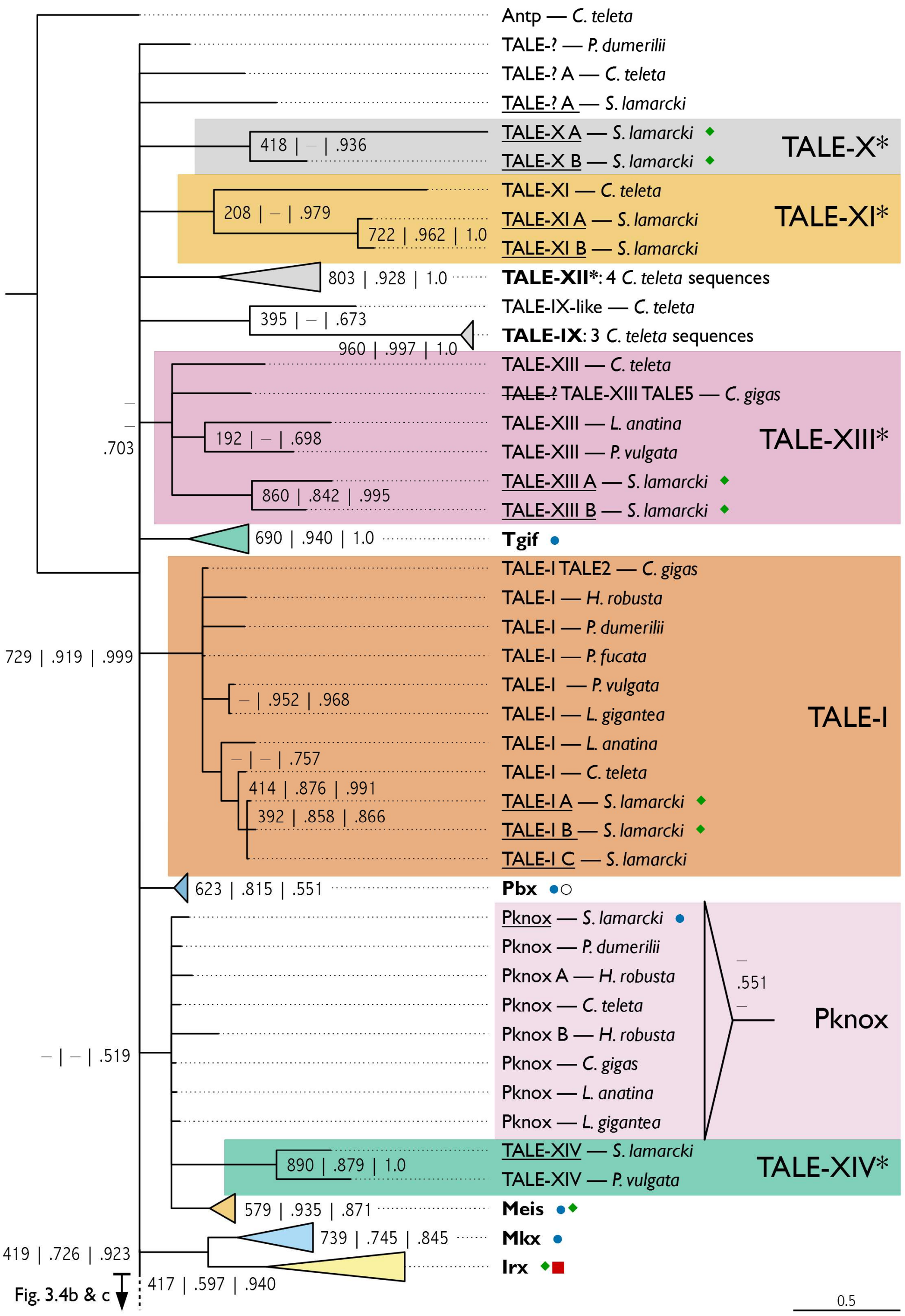
3.3.4. TALE class homeobox genes

Thirteen transcriptomic homeodomain sequences had the three amino acid loop extension diagnostic of TALE (Three Amino-acid Loop Extension) class of homeobox genes. Five of these were identical to previously described *S. lamarcki* canonical TALE-class genes: *Tgif*, *Pbx Pknox*, *Meis B*, and *Mkx1* (Kenny and Shimeld 2012). A further two of these could be classified on the basis of phylogenies as other canonical TALE-class genes: *Meis A* and *Irx A* (Figure 3.4). Finally, six sequences were not obvious homologues of canonical TALE class families.

To classify these six sequences and to confirm the identifications of the other seven, a deep recursive search for divergent TALE-class homeodomains in the available genomes of *S. lamarcki*, *C. teleta*, *H. robusta*, *P. dumerilii*, *L. anatina*, *L. gigantea*, and *P. vulgata* was performed. To these were added sequences from Paps *et al.*'s (2015) recent classification of spiralian TALE families, SPILE (Spiralian TALE) sequences from the NCBI database (Morino, Hashimoto, and Wada 2017), and canonical TALE class family sequences.

An alignment of the homeodomains was used to construct a Bayesian phylogeny with support values added from equivalent neighbour-joining and maximum likelihood analyses (Figure 3.4, Table 3.4). To accommodate all of these new and published sequences in a phylogenetically coherent framework, I propose an expansion and modification of the nomenclature of Paps *et al.* (2015), comprising nine spiralian TALE clades: TALE clades I-IX (see Table 1 in Paps *et al.*, 2015). This includes the reclassification of some members of two clades (TALE clades IV and VI), the addition of new members to five clades (TALE clades I, III, IV, VII, & VIII), and the erection of ten new clades (TALE clades X-XIX), of which one may be the product of long-branch attraction (TALE-X), five are genus-specific (TALE clades X, XII, XIV, XVI, & XIX) and one contains a previously unclassified *Crassostrea* sequence (TALE-XIII). The analysis suggests the sequence previously classified as an *Mkx* paralogue by Kenny and Shimeld (2012) belongs to TALE-XVIII. Seven sequences were found to be orphans or only weakly related to a clade. The unclassified transcriptome sequences were classed into TALE clades I, XIII, and X. A summary of the proposed changes and additions to the TALE clade classification is presented in Table 3.5.

A summary of pertinent information about each of the TALE clades is presented in Table 3.6.



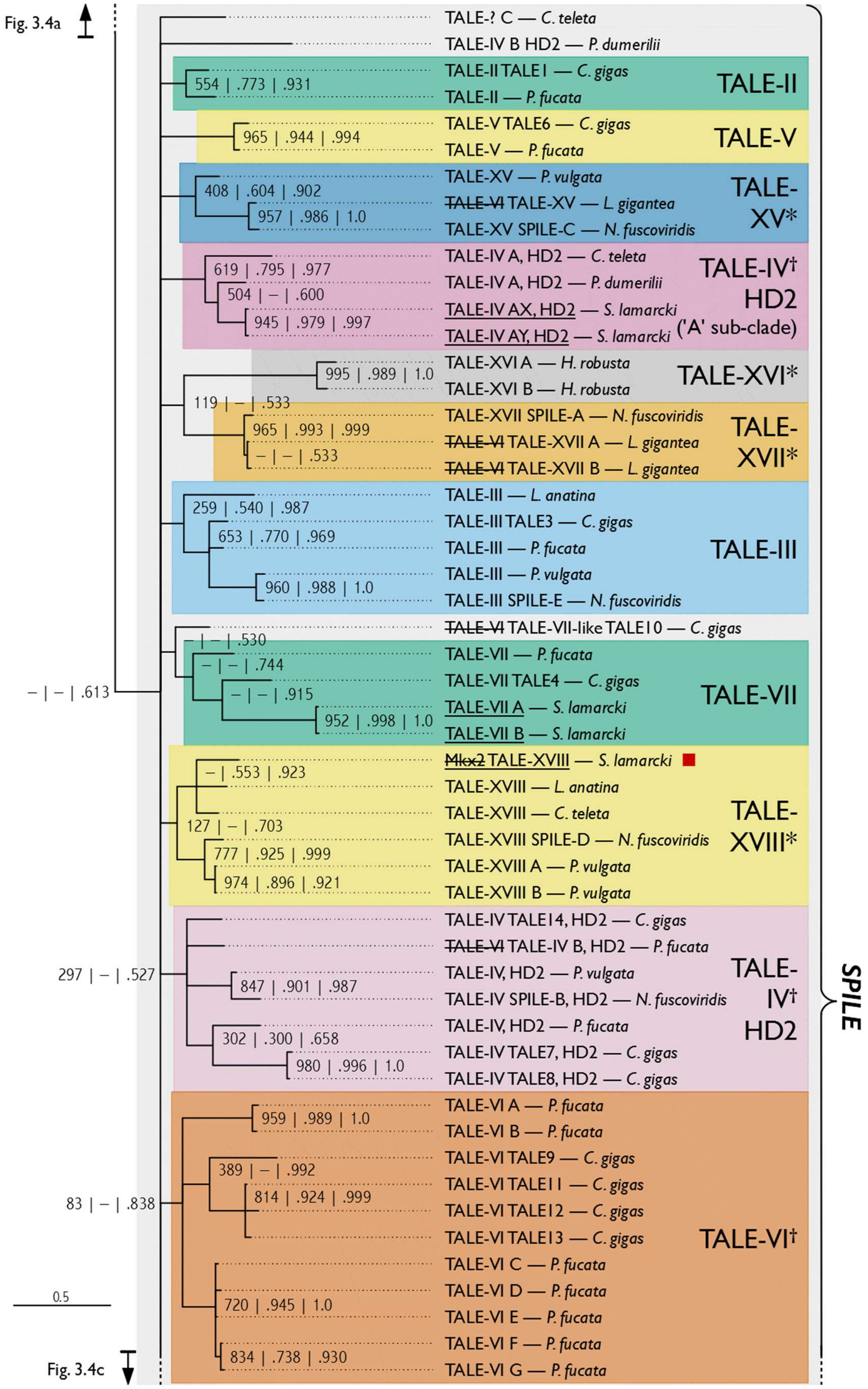


Fig. 3.4a
TALE-? C — C. teleta
TALE-IV B HD2 — P. dumerilii
554 | .773 | .931
TALE-II TALE1 — C. gigas
TALE-II — P. fucata
TALE-II
965 | .944 | .994
TALE-V TALE6 — C. gigas
TALE-V — P. fucata
TALE-V
408 | .604 | .902
957 | .986 | 1.0
TALE-XV — P. vulgata
~~TALE-VI~~ TALE-XV — L. gigantea
TALE-XV SPILE-C — N. fuscoviridis
TALE-XV*
619 | .795 | .977
504 | – | .600
945 | .979 | .997
TALE-IV A, HD2 — C. teleta
TALE-IV A, HD2 — P. dumerilii
TALE-IV AX, HD2 — S. lamarcki
TALE-IV AY, HD2 — S. lamarcki
TALE-IV† HD2 ('A' sub-clade)
995 | .989 | 1.0
119 | – | .533
TALE-XVI A — H. robusta
TALE-XVI B — H. robusta
TALE-XVI*
965 | .993 | .999
– | – | .533
TALE-XVII SPILE-A — N. fuscoviridis
~~TALE-VI~~ TALE-XVII A — L. gigantea
~~TALE-VI~~ TALE-XVII B — L. gigantea
TALE-XVII*
259 | .540 | .987
653 | .770 | .969
960 | .988 | 1.0
TALE-III — L. anatina
TALE-III TALE3 — C. gigas
TALE-III — P. fucata
TALE-III — P. vulgata
TALE-III SPILE-E — N. fuscoviridis
TALE-III
~~TALE-VI~~ TALE-VII-like TALE10 — C. gigas
– | – | .530
– | – | .744
– | – | .915
952 | .998 | 1.0
– | – | .613
TALE-VII — P. fucata
TALE-VII TALE4 — C. gigas
TALE-VII A — S. lamarcki
TALE-VII B — S. lamarcki
TALE-VII
– | .553 | .923
127 | – | .703
777 | .925 | .999
974 | .896 | .921
~~Mlx2~~ TALE-XVIII — S. lamarcki ■
TALE-XVIII — L. anatina
TALE-XVIII — C. teleta
TALE-XVIII SPILE-D — N. fuscoviridis
TALE-XVIII A — P. vulgata
TALE-XVIII B — P. vulgata
TALE-XVIII*
297 | – | .527
847 | .901 | .987
302 | .300 | .658
980 | .996 | 1.0
TALE-IV TALE14, HD2 — C. gigas
~~TALE-VI~~ TALE-IV B, HD2 — P. fucata
TALE-IV, HD2 — P. vulgata
TALE-IV SPILE-B, HD2 — N. fuscoviridis
TALE-IV, HD2 — P. fucata
TALE-IV TALE7, HD2 — C. gigas
TALE-IV TALE8, HD2 — C. gigas
TALE-IV† HD2
SPILE
83 | – | .838
959 | .989 | 1.0
389 | – | .992
814 | .924 | .999
720 | .945 | 1.0
834 | .738 | .930
TALE-VI A — P. fucata
TALE-VI B — P. fucata
TALE-VI TALE9 — C. gigas
TALE-VI TALE11 — C. gigas
TALE-VI TALE12 — C. gigas
TALE-VI TALE13 — C. gigas
TALE-VI C — P. fucata
TALE-VI D — P. fucata
TALE-VI E — P. fucata
TALE-VI F — P. fucata
TALE-VI G — P. fucata
TALE-VI†
0.5
Fig. 3.4c

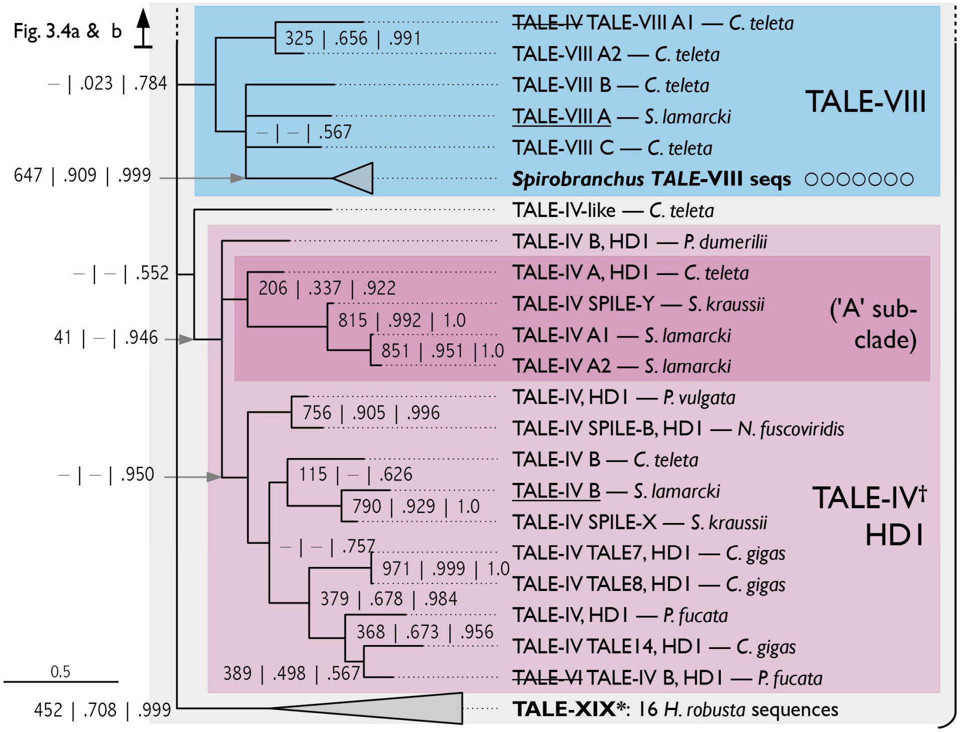


Figure 3.4 (3 parts). Bayesian phylogeny of TALE class homeodomain sequences from a selection of lophotrochozoan genomes, showing the frequent duplication of canonical TALE class genes and the basis of our proposed revision to the spiralian TALE clade (TALE-) classification (Paps *et al.*, 2015). The SPILE clade (*per* Morino, Hashimoto, and Wada 2017) is marked by a grey box and labelled bracket. Support values for each node are from neighbour joining (out of 1000 bootstraps), maximum likelihood (proportion of 1000 bootstraps), and Bayesian (posterior probability) phylogenies (in order, separated by vertical bars or new lines). A dash indicates where a node is not present in the corresponding tree. Established bilaterian gene families that have been successfully reconstructed have been collapsed into coloured triangles, and a summary of their contents is presented in Table 3.4. In some cases, new families or family subsets containing several sequences all from a single genus have also been collapsed to aid visualisation. Single genus families are highlighted in grey, but otherwise colour selection is arbitrary, and not meant to indicate a relationship except in the case of the TALE-IV clades. Similarly, paralogue lettering, where present, is not intended to consistently imply direct orthology, though where evident, direct orthologues have been lettered accordingly. *S. lamarcki* sequences (all underlined) are marked with a green diamond if found in the regenerative transcriptomes, with a red square if found in the developmental transcriptome (Kenny and Shimeld 2012), and a blue dot if found in both. Collapsed families have their *S. lamarcki* gene complement indicated nearby with the same symbols as above, with an empty black circle indicating a gene that has been found only in the genome. New gene families suggested herein are marked with an asterisk. Gene families that have gained or lost sequences from Paps *et al.* (2015) are marked with a dagger. Where a gene has been reclassified from Paps *et al.* (2015) or Kenny and Shimeld (2012), the old classification is included but struck out. Established gene families that were successfully reconstructed in the neighbour-joining and/or maximum likelihood analyses but not the Bayesian analysis are marked by a 'cartoon' clade (not to horizontal scale) and corresponding support values to the right-hand side. The scale bar indicates amino acid substitutions per site. Full sequence details are included in Appendix 3.2b. The alignment used to produce this

tree is presented in Appendix 3.4c, and the full version of the Newick format tree is presented in Appendix 3.3c. Annelid species: *S. lamarcki* = *Spirobranchus lamarcki*; *S. kraussi* = *Spirobranchus* (formerly *Pomatoleios*) *kraussi*; *C. teleta* = *Capitella teleta*; *H. robusta* = *Helobdella robusta*; *P. dumerilii* = *Platynereis dumerilii*. Brachiopod species: *L. anatina* = *Lingula anatina*. Mollusc species: *C. gigas* = *Crassostrea gigas*; *P. fucata* = *Pinctada fucata*; *L. gigantea* = *Lottia gigantea*; *N. fuscoviridis* = *Nipponacmea fuscoviridis*; *P. vulgata* = *Patella vulgata*. Insect species (only in collapsed clades): *Tribolium castaneum*, *Drosophila melanogaster*. Adapted from Barton-Owen, Szabó, Somorjai, & Ferrier (2018).

Table 3.4. Summary of the collapsed gene families in Figure 3.4. Clade colouration and symbols denoting sequence origin are the same as in Figure 3.4. Annelid species: *S.lam* = *Spirobranchus lamarcki*; *S.kra* = *Spirobranchus* (formerly *Pomatoleios*) *kraussi*; *C.tel* = *Capitella teleta*; *H.rob* = *Helobdella robusta*; *P.dum* = *Platynereis dumerilii*. Brachiopod species: *L.ana* = *Lingula anatina*. Mollusc species: *C.gig* = *Crassostrea gigas*; *L.gig* = *Lottia gigantea*; *P.vul* = *Patella vulgata*. Insect species: *T.cas* = *Tribolium castaneum*; *D.mel* = *Drosophila melanogaster*.

	Irx	Meis	Mkx	Pbx	Pknox	Tgif	TALE-XII	TALE-IX	*Spir.* TALE-VIII	TALE-XIX
S.lam	2 ◆ ■	2 ● ◆	1 ●	2 ●	1	1 ●	-	-	7	-
S.kra	-	-	-	-	-	-	-	-	1	-
C.tel	3	1	1	1	1	-	4	3	-	-
H.rob	7	1	1	4	2	-	-	-	-	16
P.dum	2	2	-	2	1	2	-	-	-	-
C.gig	4	1	1	1	1	1	-	-	-	-
L.gig	4	1	1	1	1	1	-	-	-	-
P.vul	3	-	1	-	-	1	-	-	-	-
L.ana	4	1	-	1	1	1	-	-	-	-
A.mel	-	-	-	-	-	-	-	-	-	-
D.mel	1	-	-	1	-	-	-	-	-	-
T.cas	2	-	-	1	-	-	-	-	-	-

Table 3.5. Gene-centric summary of revisions to the TALE classification system of Paps *et al.* (2015). In the Origin column, 'N' denotes that the sequence is newly discovered by this analysis, 'P' that the sequence was included in Paps *et al.*'s (2015) analysis, and 'M' that the sequences were described by Morino *et al.* (2017). *S. lamarcki* sequences marked with green diamonds were found in the regenerative transcriptomes; those marked with red squares were described by Kenny and Shimeld (2012) in their developmental transcriptome. In genes with two homeodomains, a tick indicates the presence of a homeodomain. A cross indicates the absence, either through lack of sequence coverage or apparent homeodomain degradation. 'F' indicates the presence of a truncated sequence due to lack of sequence coverage. 'W' indicates a truncated homeodomain not due to lack of sequence coverage. An unusual *H. robusta* sequence with two homeodomains is highlighted in red. Clade colouration is as in Fig 3.4. The Paps *et al.* 2015 name column refers to the identifying information given in Supplementary Files 4 & 7 in Paps *et al.* (2015), and the Original classification column to the clade to which they were assigned by that analysis. Full sequence details are included in Appendix 3.2b. Annelid species: *S. lamarcki* = *Spirobranchus lamarcki*; *S. kraussi* = *Spirobranchus* (formerly *Pomatoleios*) *kraussi*; *C. teleta* = *Capitella teleta*; *H. robusta* = *Helobdella robusta*; *P. dumerilii* = *Platynereis dumerilii*. Brachiopod species: *L. anatina* = *Lingula anatina*. Mollusc species: *C. gigas* = *Crassostrea gigas*; *P. fucata* = *Pinctada fucata*; *L. gigantea* = *Lottia gigantea*; *N. fuscoviridis* = *Nipponacmea fuscoviridis*; *P. vulgata* = *Patella vulgata*. Adapted from Barton-Owen, Szabó, Somorjai, & Ferrier (2018).

	Species	Origin	Sequence name	HD1	HD2	Paps *et al.* 2015 name	Original classif'n
I	*S. lamarcki*	N	*TALE-I A*◆ , *B*◆ , *C*			-	-
	C. teleta	P	*TALE-I*			Ctel 1513294 24 8	unchanged
	H. robusta	N	*TALE-I*			-	-
	P. dumerilii	N	*TALE-I*			-	-
	L. anatina	N	*TALE-I*			-	-
	C. gigas	P	*TALE-I TALE2*			Cgi TALE2	unchanged
	P. fucata	P	*TALE-I*			Pfuc 24948 1 11659 JP	unchanged
	L. gigantea	P	*TALE-I*			Lgig 1414665 30 1	unchanged
	P. vulgata	N	*TALE-I*			-	-
II	*C. gigas*	P	*TALE-II TALE1*			Cgi TALE1	unchanged
	P. fucata	P	*TALE-II*			Pfuc 13151 1 32296 JP/ Pfuc 13478 1 32332 JP	unchanged (HDs identical)
III	*L. anatina*	N	*TALE-III*			-	-
	C. gigas	P	*TALE-III TALE3*			Cgi TALE3	unchanged
	P. fucata	P	*TALE-III*			Pfuc 98062 1 56909 JP	unchanged
	N. fuscoviridis	M	*TALE-III SPILE-E*			-	-
	P. vulgata	N	*TALE-III*			-	-
IV	*S. lamarcki*	N	*TALE-IV A1, A2, B*	✓	✗	-	-
	S. lamarcki	N	*TALE-IV AX, AY*	F	✓	-	-
	S. kraussi	M	*TALE-IV SPILE-X, SPILE-Y*	✓	✗	-	-
	C. teleta	P	*TALE-IV A*	✓	✓	Ctel 1526117 32 9	unchanged
	C. teleta	P	*TALE-IV B*	✓	✗	Ctel 1505080 24 4	unchanged
	P. dumerilii	N	*TALE-IV B*	✓	W	-	-

	P. dumerilii	N	*TALE-IV A*	F	✓	-	-
	C. gigas	P	*TALE-IV TALE7, 8, 14*	✓	✓	Cgi TALE7, 8, 14	unchanged
	P. fucata	P	*TALE-IV A*	✓	✓	Pfuc 1892 1 66137 JP	unchanged
	P. fucata	P	*TALE-IV B*	✓	✓	Pfuc 6497 1 45448 JP	**TALE-VI**
	N. fuscoviridis	M	*TALE-IV SPILE-B*	✓	✓	-	-
	P. vulgata	N	*TALE-IV*	✓	✓	-	-
V	*C. gigas*	P	*TALE-V TALE6*			Cgi TALE6	unchanged
	P. fucata	P	*TALE-V*			Pfuc 255 1 07443 JP	unchanged
VI	*C. gigas*	P	*TALE-VI TALE9, 11-13*			Cgi TALE9, 11-13	unchanged
	P. fucata	P	*TALE-VI A*			Pfuc 1442 1 22591 JP	unchanged
	P. fucata	P	*TALE-VI B*			Pfuc 22569 1 62158 JP	unchanged
	P. fucata	P	*TALE-VI C*			Pfuc 22555 1 40373 JP	unchanged
	P. fucata	P	*TALE-VI D*			Pfuc 18402 1 40058 JP	unchanged
	P. fucata	P	*TALE-VI E*			Pfuc 10095 1 38990 JP	unchanged
	P. fucata	P	*TALE-VI F*			Pfuc 2547 1 30160 JP	unchanged
	P. fucata	P	*TALE-VI G*			Pfuc 312 1 50785 JP	unchanged
VII	*S. lamarcki*	N	*TALE-VII A, B*			-	-
	C. gigas	P	*TALE-VII TALE4*			Cgi TALE4	unchanged
	P. fucata	P	*TALE-VII*			Pfuc 6013 1 23936 JP	unchanged
VIII	*S. lamarcki*	N	*TALE-VIII A, B, C, D, E, F, G, H*			-	-
	S. kraussi	M	*TALE-VIII SPILE-Z*			-	-
	C. teleta	P	*TALE-VIII B (1-3?)*			Ctel 1505086 31 9/ Ctel 1505698 31 9/ Ctel 1499331 27 4	unchanged (HDs identical)
	C. teleta	P	*TALE-VIII A1*			Ctel 1499505 38 4	**TALE-IV**
	C. teleta	M	*TALE-VIII A2, C*			-	-
IX	*C. teleta*	P	*TALE-IX A*			Ctel 1518266 30 6	unchanged
	C. teleta	P	*TALE-IX B*			Ctel 1518128 28 9	unchanged
	C. teleta	P	*TALE-IX C*			Ctel 1502937 32 5	unchanged
X	*S. lamarcki*	N	*TALE-X A*◆ , *B*◆			-	-
XI	*S. lamarcki*	N	*TALE-XI A, B*			-	-
	C. teleta	N	*TALE-XI*			-	-
XII	*C. teleta*	N/M	*TALE-XII A1, A2, A3, B*			-	-
XIII	*S. lamarcki*	N	*TALE-XIII A*◆ , *B2*◆				
	C. teleta	N	*TALE-XIII*			-	-
	L. anatina	N	*TALE-XIII*			-	-
	C. gigas	P	*TALE-XIII TALE5*			Cgi TALE5	TALE-?
	P. vulgata	N	*TALE-XIII*			-	-
XIV	*S. lamarcki*	N	*TALE-XIV*			-	-
	P. vulgata	N	*TALE-XIV*			-	-
XV	*L. gigantea*	P	*TALE-XV*			Lgig 1419427 48 9	**TALE-VI**
	N. fuscoviridis	M	*TALE-XV SPILE-C*			-	-
	P. vulgata	N	*TALE-XV*			-	-
XVI	*H. robusta*	N	*TALE-XVI A, B*			-	-
XVII	*L. gigantea*	P	*TALE-XVII A*			Lgig 1410135 44 3	**TALE-VI**
	L. gigantea	P	*TALE-XVII B*			Lgig 1410138 39 8	**TALE-VI**

	N. fuscoviridis	M	*TALE-XVII SPILE-A*			-	-
XVIII	*S. lamarcki*	N	*TALE-XVIII*■			-	Mkx2
	C. teleta	M	*TALE-XVIII*			-	-
	L. anatina	N	*TALE-XVIII*			-	-
	N. fuscoviridis	M	*TALE-XVIII SPILE-D*			-	-
	P. vulgata	N	*TALE-XVIII A, B*			-	-
XIX	*H. robusta*	N	*TALE-XIX A*	✓	✓	-	
	H. robusta	N	*TALE-XIX B-P (15 sequences)*			-	
unclassified	*S. lamarcki*	N	*TALE-? A*			-	-
	C. teleta	M	*TALE-? A, C, TALE-IV-like, TALE-IX-like*			-	-
	P. dumerilii	N	*TALE-?*			-	-
	C. gigas	P	*TALE-VII-like TALE10*			Cgi TALE10	**TALE-VI**

Table 3.6. Clade-centric summary of various attributes of the TALE clades. In the Expression column, each symbol denotes a gene in that clade for which there is evidence of expression. Red indicates the data originate from Kenny & Shimeld (2012); blue, from Paps *et al.* (2015); purple, from Morino *et al.* (2017), and green from the present study. Mollusc embryogenesis is denoted by a triangle, annelid development by a square, and annelid regeneration by a diamond. A qualitative assessment of the confidence placed in the robustness of a clade in future analyses (based on support values and inspection of the sequences) is given in the Confidence column. Notes detailing information pertinent to the confidence rating are given in the Comments column.

<table>
<tr><th>Clade</th><th>Original de-
scription</th><th>SPILE?</th><th>Taxonomic
distribution</th><th>Loss/
retention</th><th>Gene
count</th><th>Dev'
Expr.</th><th>Support values</th><th>Confidence</th><th>Comments</th></tr>
<tr><td>TALE-I</td><td>Paps <i>et al.</i>,</td><td>N</td><td>Loph.</td><td>Widely retained</td><td>11</td><td>◆ ◆ ▲</td><td>729 | .919 | .999</td><td>High</td><td></td></tr>
<tr><td>TALE-II</td><td>D"</td><td>Y</td><td>Bivalvia</td><td>–</td><td>2</td><td>▲</td><td>554 | .773 | .931</td><td>High</td><td></td></tr>
<tr><td>TALE-III</td><td>D"</td><td>Y</td><td>Spiralia</td><td>Annelid loss</td><td>5</td><td>▲ ▲</td><td>259 | .540 | .987</td><td>High</td><td></td></tr>
<tr><td>TALE-IV</td><td>D"</td><td>Y</td><td>Spiralia</td><td>Brachiopod loss</td><td>16</td><td>▲ ▲ ▲
■ ■ ▲</td><td>HD1: 41 | – | .946
HD2: 297 | – | .527
HD2a: 619 | .795 | .977</td><td>High</td><td></td></tr>
<tr><td>TALE-V</td><td>D"</td><td>Y</td><td>Bivalvia</td><td>–</td><td>2</td><td></td><td>965 | .944 | .994</td><td>High</td><td></td></tr>
<tr><td>TALE-VI</td><td>D"</td><td>Y</td><td>Bivalvia</td><td>–</td><td>11</td><td>▲ ▲ ▲</td><td>83 | – | .838</td><td>Moderate</td><td>Lots of sequences left Paps <i>et al.</i>'s TALE-VI in this analysis</td></tr>
<tr><td>TALE-VII</td><td>D"</td><td>Y</td><td>Loph.</td><td>Widely lost</td><td>4</td><td>▲</td><td>– | – | .744</td><td>Moderate</td><td>Incomplete <i>Spirobranchus</i> sequences</td></tr>
<tr><td>TALE-VIII</td><td>D"</td><td>Y</td><td>Sedentaria</td><td>–</td><td>13-15</td><td>■</td><td>– | .023 | .784</td><td>Moderate</td><td>A lot of incomplete sequences</td></tr>
<tr><td>TALE-IX</td><td>D"</td><td>N</td><td><i>Capitella</i></td><td>–</td><td>3</td><td>–</td><td>960 | .997 | 1.0</td><td>High</td><td></td></tr>
<tr><td>TALE-X</td><td>Present study</td><td>N</td><td><i>Spiro.</i></td><td>–</td><td>2</td><td>◆ ◆</td><td>418 | – | .936</td><td>Low</td><td>The two sequences are very dissimilar – product of LBA?</td></tr>
<tr><td>TALE-XI</td><td>D"</td><td>N</td><td>Sedentaria</td><td>–</td><td>3</td><td>–</td><td>208 | – | .979</td><td>Low</td><td>The <i>C.tel</i> sequence is very dissimilar to the <i>S.lam</i> sequences</td></tr>
<tr><td>TALE-XII</td><td>D"</td><td>N</td><td><i>Capitella</i></td><td>–</td><td>4</td><td>–</td><td>803 | .928 | 1.0</td><td>High</td><td></td></tr>
<tr><td>TALE-XIII</td><td>D"</td><td>N</td><td>Loph.</td><td>Widely retained</td><td>6</td><td>◆ ◆</td><td>– | – | .703</td><td>Low</td><td>Unconvincing when looking at sequences</td></tr>
<tr><td>TALE-XIV</td><td>D"</td><td>N</td><td>Loph.</td><td>Widely lost</td><td>2</td><td>–</td><td>890 | .879 | 1.0</td><td>Moderate</td><td>Both sequences incomplete</td></tr>
<tr><td>TALE-XV</td><td>D"</td><td>Y</td><td>Gastropoda</td><td>–</td><td>3</td><td>▲</td><td>408 | .604 | .902</td><td>High</td><td></td></tr>
<tr><td>TALE-XVI</td><td>D"</td><td>Y</td><td><i>Helobdella</i></td><td>–</td><td>2</td><td>–</td><td>995 | .989 | 1.0</td><td>High</td><td></td></tr>
<tr><td>TALE-XVII</td><td>D"</td><td>Y</td><td>Gastropoda</td><td>–</td><td>3</td><td>▲</td><td>965 | .993 | .999</td><td>High</td><td></td></tr>
<tr><td>TALE-XVIII</td><td>D"</td><td>Y</td><td>Loph.</td><td>Widely retained</td><td>6</td><td>■ ▲</td><td>127 | – | .703</td><td>High</td><td></td></tr>
<tr><td>TALE-XIX</td><td>D"</td><td>Y</td><td><i>Helobdella</i></td><td>–</td><td>16</td><td>–</td><td>452 | .708 | .999</td><td>High</td><td></td></tr>
<tr><td>(TALE-?)</td><td>–</td><td>–</td><td>–</td><td>–</td><td>7</td><td>▲</td><td>–</td><td>–</td><td></td></tr>
</table>

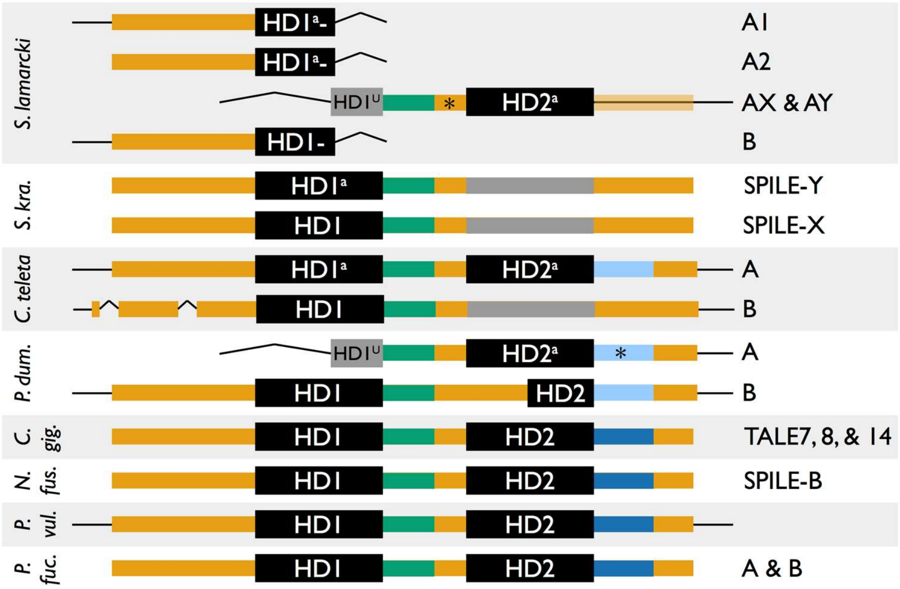


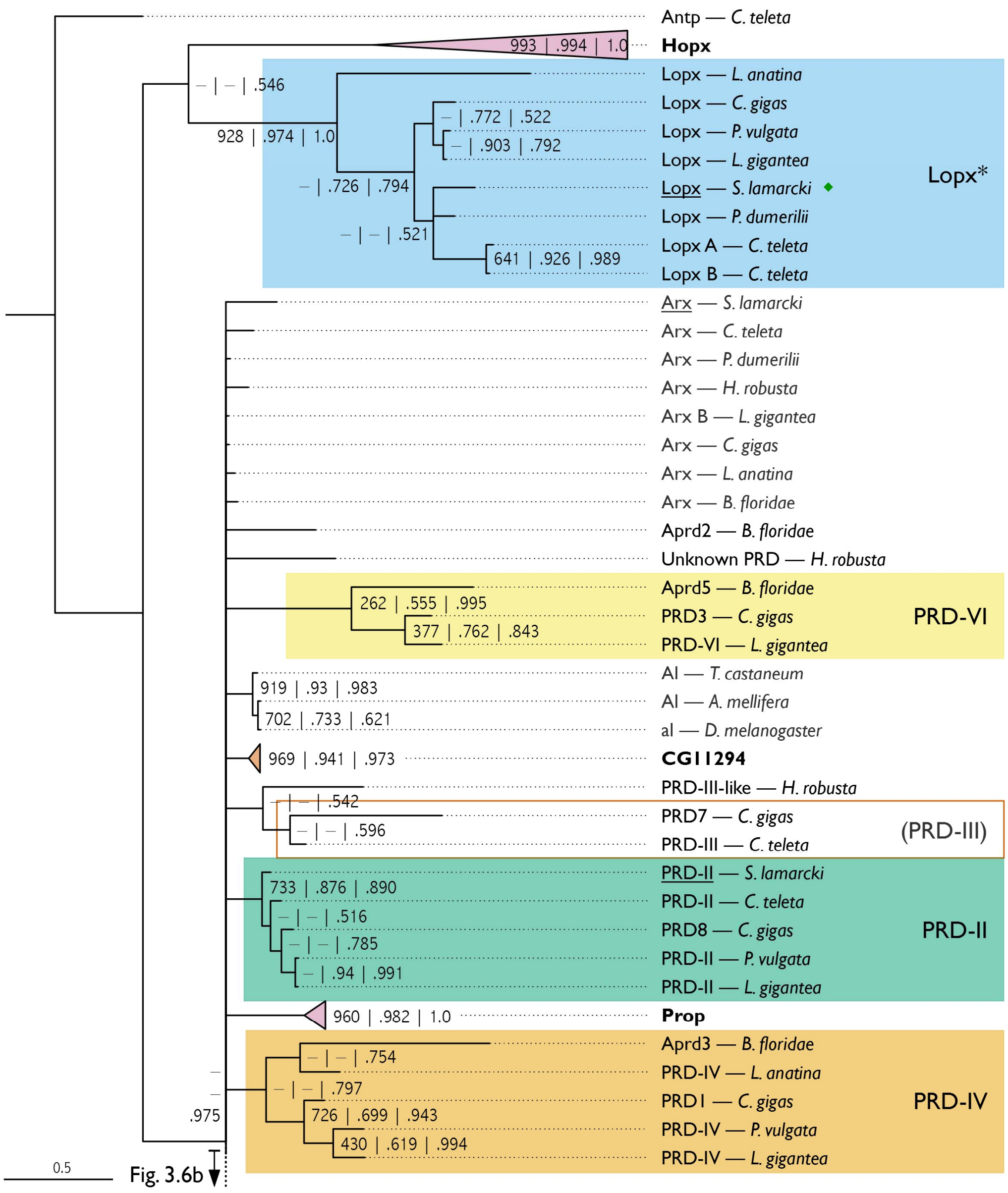
Figure 3.5. A schematic representation of the sequence fragments of TALE-IV genes, showing the evidence for genes containing two TALE-class HDs. Non-coding sequence is indicated with a thin black line. Coding sequence is indicated with a thick coloured line; semi-transparent if the extent of the exonic sequence is not easily predictable. Green and blue regions represent areas of high sequence conservation C-terminal to each of the homeodomains. Light blue colouration represents regions where the sequence is recognisably homologous to the blue region but has substantially diverged. Regions that are unusually long relative to equivalent homologous regions are marked with an asterisk. Regions with apparent homology to homeodomains but which have degraded are represented with thick grey lines. Homeodomains are represented with boxes coloured black if recognised by the NCBI Conserved Domain Search or grey otherwise. Half-size homeodomains are due to introns (*S. lamarcki* AX & AY, *P. dumerilii* A) or truncated homeoboxes (*P. dumerilii* B). Homeodomains are marked 'a' if they belong to the A/annelid-only sub-clade (see Figure 3.1) or 'U' if they were too short to be identified using the phylogeny. Not to scale. Annelid species: *S. lamarcki* = *Spirobranchus lamarcki*; *S. kraussi* = *Spirobranchus* (formerly *Pomatoleios*) *kraussi*; *C. teleta* = *Capitella teleta*; *P.dum.* = *Platynereis dumerilii*. Mollusc species: *C. gigas* = *Crassostrea gigas*; *P. fuc.* = *Pinctada fucata*; *N. fus.* = *Nipponacmea fuscoviridis*; *P. vul.* = *Patella vulgata*. Taken from Barton-Owen, Szabó, Somorjai, & Ferrier (2018).

In the course of manually inspecting sequences for alignment, it was observed that most TALE-IV sequences have two TALE-class homeodomains. The available evidence for TALE-IV gene structure is summarized in Figure 3.5.

3.3.5. PRD class homeobox genes

Ten transcriptomic sequences were identified as canonical PRD-class genes: *Prrx*, *Shox*, *Otp B*, *Otx B*, *Vsx B*, *Pax4/6 A* & *B*, and four identical or near-identical to previously described *S. lamarcki* sequences: *Gsc*, *Hbn*, *Otp A*, and *Otx A* (Kenny and Shimeld 2012). Two sequences were also identified which could not be placed in canonical PRD-class gene families. One of these was matched by BLAST to sequences that had been automatically identified as *ceh-37*, one of the *Caenorhabditis elegans* paralogues of *Otx*, but appeared to share little homology with the original *ceh-37* gene. The other was matched by BLAST searches to amphioxus *Aprd6*. To classify these genes, putative and previously identified PRD-class homeodomains were collected from *S. lamarcki*, *C. teleta*, *H. robusta*, *P. dumerilii*, *L. anatina, C. gigas*, *L. gigantea*, *P. vulgata*, *A. mellifera*, *D. melanogaster*, *T. castaneum*, and *B. floridae.* (Appendix 3.2b). These were aligned (Appendix 3.4d) and the alignment used to produce a Bayesian phylogeny with support values added from equivalent neighbour-joining and maximum likelihood analyses (Figure 3.6, Table 3.7).

This phylogeny successfully reconstructed all canonical PRD-class clades (except Arx) and the same non-canonical PRD Clades as Paps *et al.* (2015) (PRD Clades I-VI). In addition, one further clade (PRD-VII) was resolved. *S. lamarcki Prd-like* (Kenny and Shimeld 2012) is reclassified here as PRD-VII.


Antp — C. teleta
993 | .994 | 1.0
Hopx
– | – | .546
Lopx — L. anatina
Lopx — C. gigas
– | .772 | .522
Lopx — P. vulgata
928 | .974 | 1.0
– | .903 | .792
Lopx — L. gigantea
Lopx*
– | .726 | .794
Lopx — S. lamarcki
Lopx — P. dumerilii
– | – | .521
Lopx A — C. teleta
641 | .926 | .989
Lopx B — C. teleta
Arx — S. lamarcki
Arx — C. teleta
Arx — P. dumerilii
Arx — H. robusta
Arx B — L. gigantea
Arx — C. gigas
Arx — L. anatina
Arx — B. floridae
Aprd2 — B. floridae
Unknown PRD — H. robusta
Aprd5 — B. floridae
262 | .555 | .995
PRD3 — C. gigas
PRD-VI
377 | .762 | .843
PRD-VI — L. gigantea
Al — T. castaneum
919 | .93 | .983
Al — A. mellifera
702 | .733 | .621
al — D. melanogaster
969 | .941 | .973
CG11294
PRD-III-like — H. robusta
– | – | .542
PRD7 — C. gigas
– | – | .596
(PRD-III)
PRD-III — C. teleta
PRD-II — S. lamarcki
733 | .876 | .890
PRD-II — C. teleta
– | – | .516
PRD8 — C. gigas
PRD-II
– | – | .785
PRD-II — P. vulgata
– | .94 | .991
PRD-II — L. gigantea
960 | .982 | 1.0
Prop
Aprd3 — B. floridae
– | – | .754
PRD-IV — L. anatina
–
– | – | .797
–
PRD1 — C. gigas
PRD-IV
.975
726 | .699 | .943
PRD-IV — P. vulgata
430 | .619 | .994
PRD-IV — L. gigantea
0.5
Fig. 3.6b

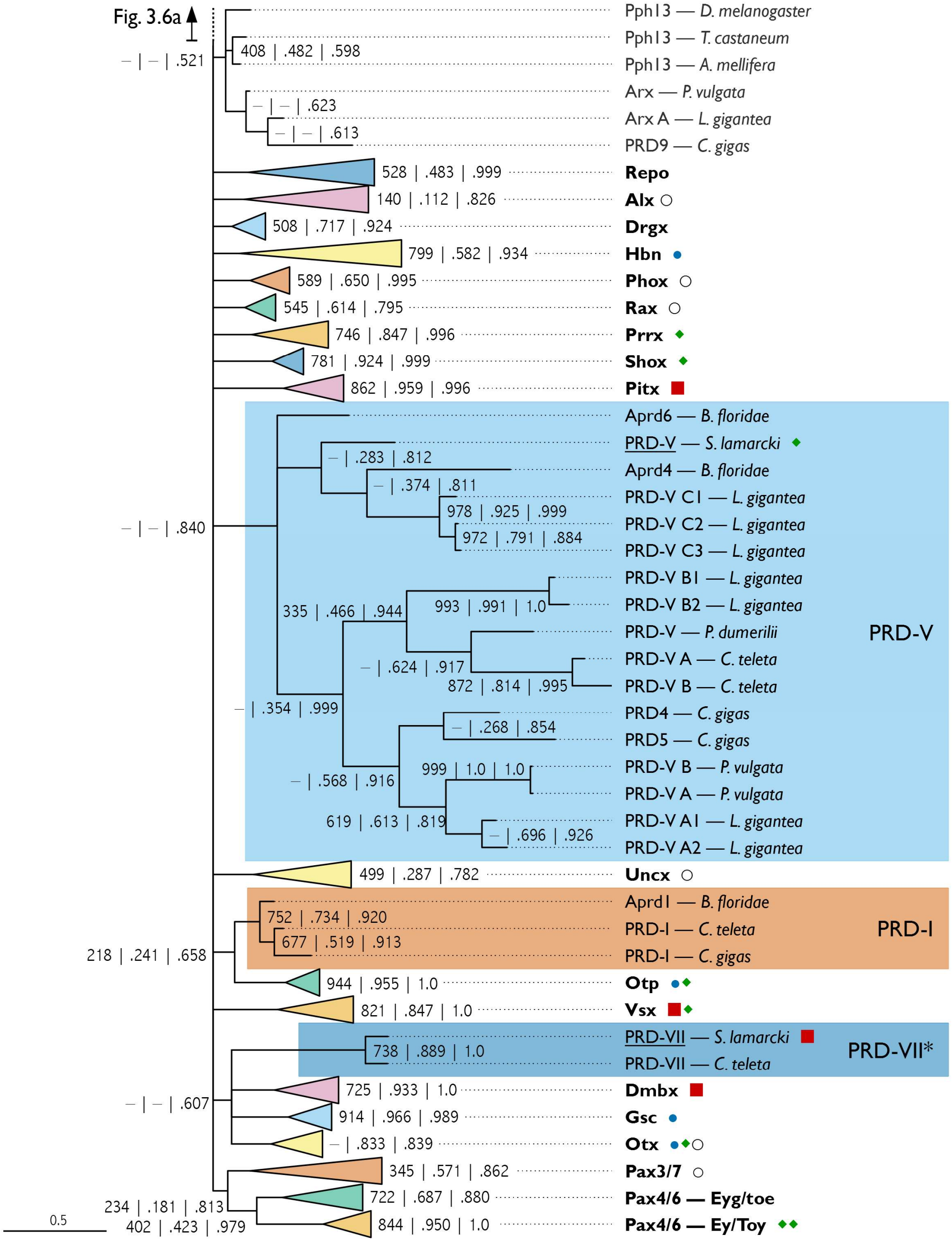


Fig. 3.6a
Pph13 — D. melanogaster
Pph13 — T. castaneum
Pph13 — A. mellifera
Arx — P. vulgata
Arx A — L. gigantea
PRD9 — C. gigas
– | – | .521
408 | .482 | .598
– | – | .623
– | – | .613
528 | .483 | .999 Repo
140 | .112 | .826 Alx ○
508 | .717 | .924 Drgx
799 | .582 | .934 Hbn ●
589 | .650 | .995 Phox ○
545 | .614 | .795 Rax ○
746 | .847 | .996 Prrx ◆
781 | .924 | .999 Shox ◆
862 | .959 | .996 Pitx ■
Aprd6 — B. floridae
PRD-V — S. lamarcki ◆
Aprd4 — B. floridae
PRD-V C1 — L. gigantea
PRD-V C2 — L. gigantea
PRD-V C3 — L. gigantea
PRD-V B1 — L. gigantea
PRD-V B2 — L. gigantea
PRD-V — P. dumerilii
PRD-V A — C. teleta
PRD-V B — C. teleta
PRD4 — C. gigas
PRD5 — C. gigas
PRD-V B — P. vulgata
PRD-V A — P. vulgata
PRD-V A1 — L. gigantea
PRD-V A2 — L. gigantea
PRD-V
– | – | .840
– | .283 | .812
– | .374 | .811
978 | .925 | .999
972 | .791 | .884
335 | .466 | .944
993 | .991 | 1.0
– | .624 | .917
872 | .814 | .995
– | .354 | .999
– | .268 | .854
999 | 1.0 | 1.0
– | .568 | .916
619 | .613 | .819
– | .696 | .926
499 | .287 | .782 Uncx ○
Aprd1 — B. floridae
PRD-I — C. teleta
PRD-I — C. gigas
PRD-I
752 | .734 | .920
677 | .519 | .913
218 | .241 | .658
944 | .955 | 1.0 Otp ●◆
821 | .847 | 1.0 Vsx ■◆
PRD-VII — S. lamarcki ■
PRD-VII — C. teleta
738 | .889 | 1.0
PRD-VII*
– | – | .607
725 | .933 | 1.0 Dmbx ■
914 | .966 | .989 Gsc ●
– | .833 | .839 Otx ●◆○
345 | .571 | .862 Pax3/7 ○
234 | .181 | .813
722 | .687 | .880 Pax4/6 — Eyg/toe
402 | .423 | .979
844 | .950 | 1.0 Pax4/6 — Ey/Toy ◆◆
0.5

Figure 3.6 (2 parts). Bayesian phylogeny of PRD class homeodomain sequences from a selection of bilaterian genomes, and the new unclassified Lopx gene family. Support values for each node are from neighbour joining (out of 1000 bootstraps), maximum likelihood (proportion of 1000 bootstraps), and Bayesian (posterior probability) phylogenies (in order, separated by vertical bars or new lines). A dash indicates where a node is not present in the corresponding tree. Established bilaterian gene families that have been successfully reconstructed have been collapsed into coloured triangles, and a summary of their contents is presented in Table 3.7. In some cases, new families or family subsets containing several sequences all from a single genus have also been collapsed to aid visualisation. Colour selection is arbitrary, and not meant to indicate a relationship. Paralogue lettering, where present, is not intended to consistently imply direct orthology, though where evident, direct orthologues have been lettered accordingly. *S. lamarcki* sequences (all underlined) are marked with a green diamond if found in the regenerative transcriptome, with a red square if found in the developmental transcriptome, and a blue dot if found in both. Collapsed families have their *S. lamarcki* gene complement indicated nearby with the same symbols as above, with an empty black circle indicating a gene that has been found only in the genome. New gene families suggested herein are marked with an asterisk. The scale bar indicates amino acid substitutions per site. Full sequence details are included in Appendix 3.2b. The alignment used to produce this tree is presented in Appendix 3.4d, and the full version of the Newick format tree is presented in Appendix 3.3d. Annelid species: *S. lamarcki* = *Spirobranchus lamarcki*; *C. teleta* = *Capitella teleta*; *H. robusta* = *Helobdella robusta*; *P. dumerilii* = *Platynereis dumerilii*. Brachiopod species: *L. anatina* = *Lingula anatina*. Mollusc species: *C. gigas* = *Crassostrea gigas*; *L. gigantea* = *Lottia gigantea*; *P. vulgata* = *Patella vulgata*. Insect species: *A. mellifera* = *Apis mellifera*; *D. melanogaster* = *Drosophila melanogaster*; *T. castaneum* = *Tribolium castaneum*. Deuterostome species: *B. floridae* = *Branchiostoma floridae*. Adapted from Barton-Owen, Szabó, Somorjai, & Ferrier (2018).

Table 3.7. Summary of the collapsed gene families in Figure 3.6. Clade colouration and symbols denoting sequence origin are the same as in Figure 3.6. Annelid species: *S.lam* = *Spirobranchus lamarcki*; *C.tel* = *Capitella teleta*; *H.rob* = *Helobdella robusta*; *P.dum* = *Platynereis dumerilii*. Brachiopod species: *L.ana* = *Lingula anatina*. Mollusc species: *C.gig* = *Crassostrea gigas*; *L.gig* = *Lottia gigantea*; *P.vul* = *Patella vulgata*. Insect species: *A.mel* = *Apis mellifera*; *D.mel* = *Drosophila melanogaster*; *T.cas* = *Tribolium castaneum*. Deuterostome species: *B.flo* = *Branchiostoma floridae*.

	Hopx	CG11294	Prop	Repo	Alx	Dmbx	Drgx	Gsc	Hbn	Phox	Prrx	Shox	Rax	Pitx	Uncx	Otp	Otx	Vsx	Pax3/7	Eyg/toe	Ey/toy
S.lam	-	-	-	-	1	1 🟥	-	1 ●	1 ●	1	1 ◆	1 ◆	1	1 🟥	1	2 ●◆	3 ●◆	2 🟥◆	1	-	2 ◆◆
C.tel	-	-	-	-	1	1	1	1	1	1	1	1	1	1	2	2	1	2	1	-	1
H.rob	-	-	-	-	1	1	1	-	1	1	2	1	1	1	1	2	3	2	2	-	2
P.dum	1	-	-	-	1	1	-	1	1	1	1	1	1	4	2	2	1	1	1	-	1
C.gig	2	-	1	1	-	1	1	1	1	-	1	1	1	1	1	1	1	2	1	1	1
L.gig	1	-	-	-	1	1	1	1	1	-	1	1	1	1	1	1	1	3	1	1	1
P.vul	1	-	-	-	1	1	1	-	-	1	-	1	1	-	1	1	1	1	2	-	1
L.ana	1	-	-	1	1	1	1	1	1	1	1	1	1	1	2	1	2	2	1	-	1
A.mel	-	1	1	1	-	-	1	1	1	1	1	1	1	1	1	1	2	1	3	1	2
D.mel	-	1	1	1	-	-	1	1	1	1	1	1	1	1	2	1	1	3	3	2	2
T.cas	-	1	-	1	-	-	1	1	1	1	1	1	1	1	1	1	2	1	3	1	2
B.flo	-	-	1	1	1	1	1	1	-	1	1	1	1	1	3	1	1	1	1	-	1

3.3.6. A novel unclassified homeobox gene family

The putative *ceh-37* genes grouped into their own strongly supported clade separate to all PRD-class gene families except the highly divergent *Hopx*. I therefore propose a new gene orthology group, named *Lopx* (LOPhotrochozoan only homeobox). An alignment of the homeodomain and some flanking sequence of these proteins against sequences with which they have previously been putatively identified, as well as a conserved motif unique to *Lopx* genes, illustrates the distinctive nature of the *Lopx* family (Figure 3.7).

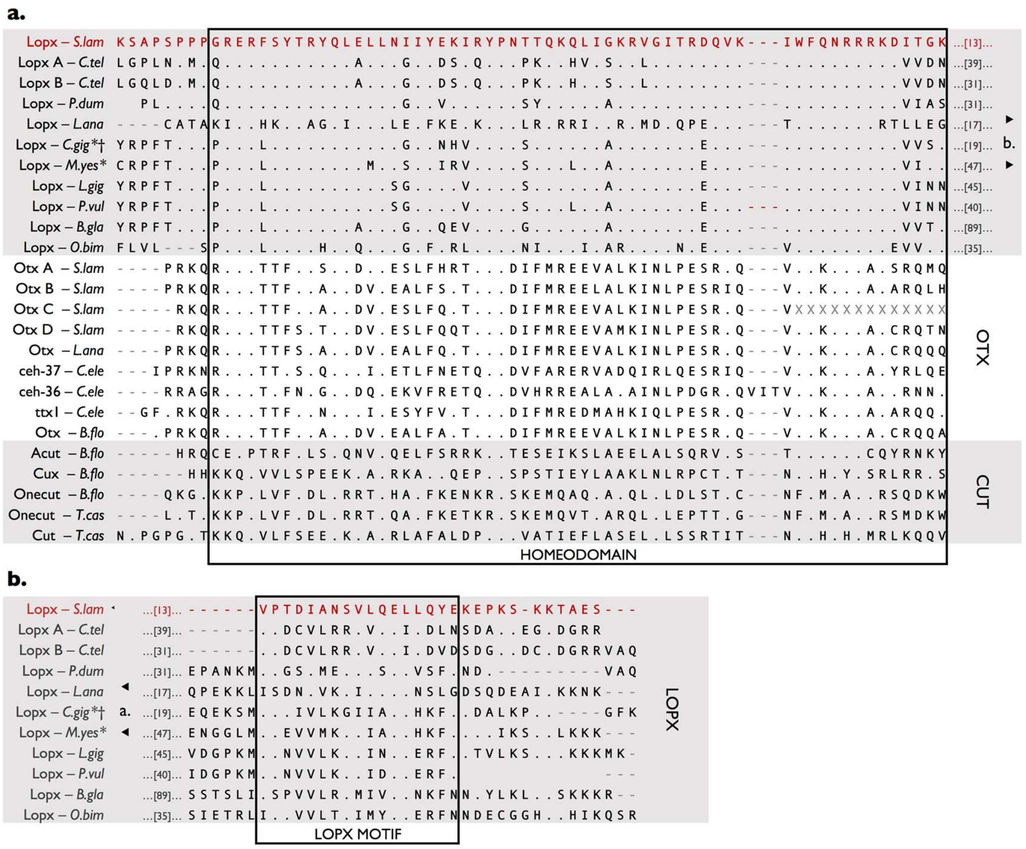
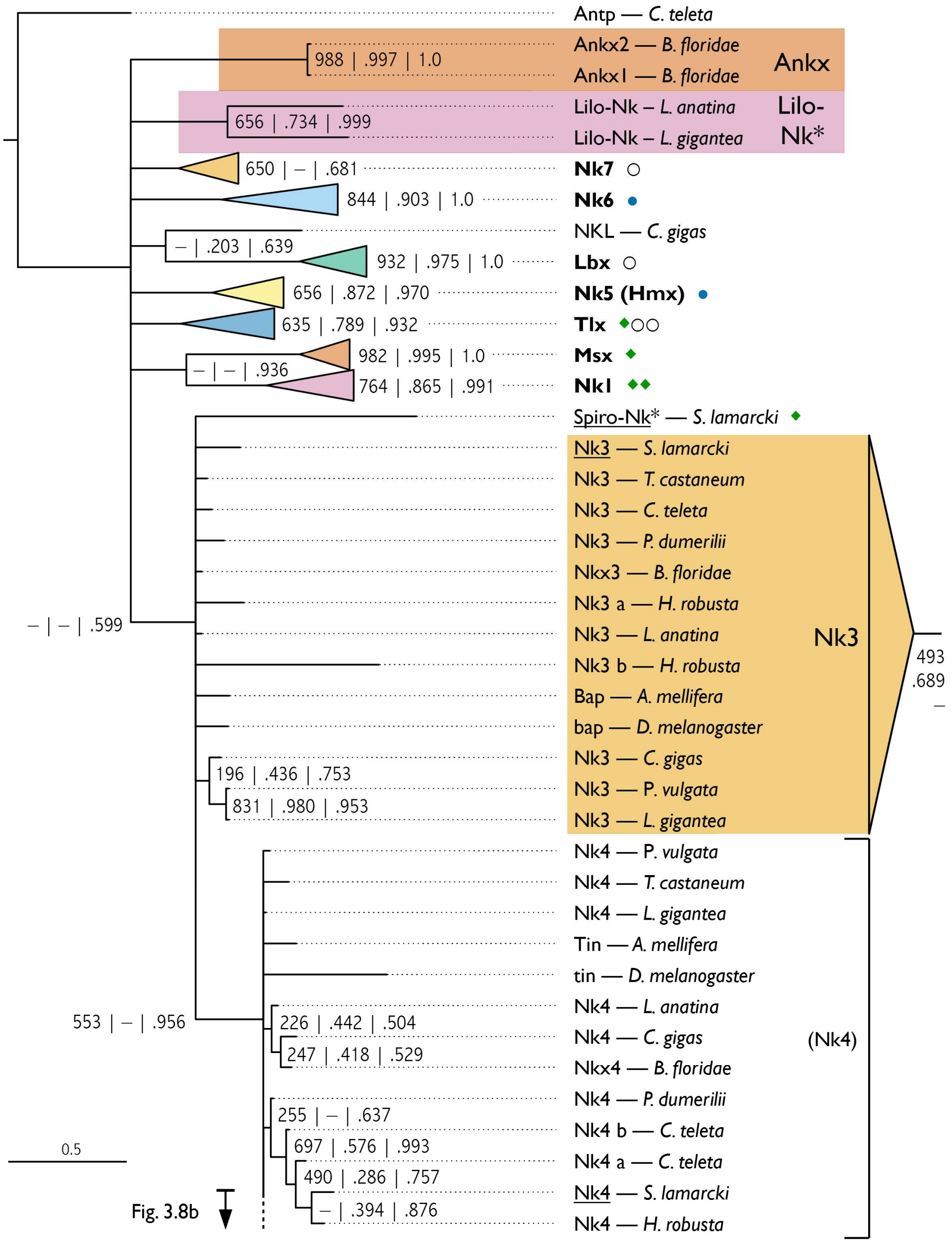


Figure 3.7. Sequence alignment of Lopx homeodomain and N-terminal flanking region (**a**) and a C-terminal conserved motif unique to Lopx proteins (**b**) from a selection of lophotrochozoan species, compared to gene families/classes that Lopx genes have been mistaken for by automatic annotation pipelines (Otx/ceh-37 - marked with asterisks) and in general homeodomain trees (CUT class - marked with dagger). Identities (full stops) are marked relative to the sequence of *Spirobranchus lamarcki* Lopx. The *S. lamarcki* Lopx sequence is highlighted in red. Full sequence details are included in Appendix 3.2b. Annelid species: *S.lam* = *Spirobranchus lamarcki*; *C.tel* = *Capitella teleta*; *P.dum* = *Platynereis dumerilii*. Brachiopod species: *L.ana* = *Lingula anatina*. Mollusc species: *C.gig* = *Crassostrea gigas*; *L.gig* = *Lottia gigantea*; *P.vul* = *Patella vulgata*; *M.yes* = *Mizuhopecten yessoensis* (syn. *Patinopecten yessoensis*); *B.gla* = *Biomphalaria glabrata*; *O.bim* = *Octopus bimaculoides*. Ecdysozoan species: *C.ele* = *Caenorhabditis elegans*; *T.cas* = *Tribolium castaneum*. Deuterostome species: *B.flo* = *Branchiostoma floridae*. Taken from Barton-Owen, Szabó, Somorjai, & Ferrier (2018).

3.3.7. Nk, Msx, Lbx & Tlx families

Seven sequences from the transcriptomes were identified as members of canonical Nk families: *Nk1a*, *Nk1b*, *Nk2.2b* and four identical or nearly identical to previously described *S. lamarcki* sequences: *Nk2.1a*, *Nk2.1b*, *Nk5* and *Nk6* (Kenny and Shimeld 2012). An eighth sequence was also identified as similar to Nk genes, but could not be placed in a canonical family. To classify the known sequences and profile Nk family gene duplication in *S. lamarcki*, putative and previously identified *Nk1-7*, *Msx*, *Lbx* and *Tlx* homeodomain sequences were collected from *S. lamarcki*, *C. teleta, H. robusta, P. dumerilii, L. anatina, C. gigas, L. gigantea, P. vulgata, A. mellifera, D. melanogaster, T. castaneum,* and *B. floridae*, (Appendix 3.2b) including the non-canonical *C. gigas* NKL gene and the amphioxus *Ankx* genes. An alignment of these homeodomains (Appendix 3.4e) was used to produce a Bayesian phylogeny with support values added from equivalent neighbour-joining and maximum likelihood analyses (Figure 3.8, Table 3.8). All clades except *Nk2.1*, *Nk3*, and *Nk4* were successfully reconstructed. The analysis does not suggest a common origin of all divergent lophotrochozoan Nk genes except those from *L. anatina* and *L. gigantea*, leading to the name *Lilo-Nk* (*i.e. Lingula-Lottia* Nk). Although the unidentified *Spirobranchus Nk* gene is located close to the *Nk3* family members in Figure 3.8, it has a clearly different sequence (Figure 3.9); it was therefore named *Spiro-Nk.* The phylogeny also indicates that *S. lamarcki Nk3-like* (Kenny and Shimeld 2012) should be reclassified as an *Nk2.1* paralogue (*Nk2.1d*).


Antp — C. teleta
988 | .997 | 1.0
Ankx2 — B. floridae
Ankx1 — B. floridae
Ankx
656 | .734 | .999
Lilo-Nk – L. anatina
Lilo-Nk – L. gigantea
Lilo-Nk*
650 | – | .681
Nk7 ○
844 | .903 | 1.0
Nk6 ●
NKL — C. gigas
– | .203 | .639
932 | .975 | 1.0
Lbx ○
656 | .872 | .970
Nk5 (Hmx) ●
635 | .789 | .932
Tlx ◆○○
982 | .995 | 1.0
Msx ◆
– | – | .936
764 | .865 | .991
Nk1 ◆◆
Spiro-Nk* — S. lamarcki ◆
Nk3 — S. lamarcki
Nk3 — T. castaneum
Nk3 — C. teleta
Nk3 — P. dumerilii
Nkx3 — B. floridae
Nk3 a — H. robusta
– | – | .599
Nk3 — L. anatina
Nk3
Nk3 b — H. robusta
493
.689
–
Bap — A. mellifera
bap — D. melanogaster
Nk3 — C. gigas
196 | .436 | .753
Nk3 — P. vulgata
831 | .980 | .953
Nk3 — L. gigantea
Nk4 — P. vulgata
Nk4 — T. castaneum
Nk4 — L. gigantea
Tin — A. mellifera
tin — D. melanogaster
Nk4 — L. anatina
553 | – | .956
226 | .442 | .504
Nk4 — C. gigas
(Nk4)
247 | .418 | .529
Nkx4 — B. floridae
Nk4 — P. dumerilii
255 | – | .637
Nk4 b — C. teleta
697 | .576 | .993
Nk4 a — C. teleta
0.5
490 | .286 | .757
Nk4 — S. lamarcki
Fig. 3.8b
– | .394 | .876
Nk4 — H. robusta

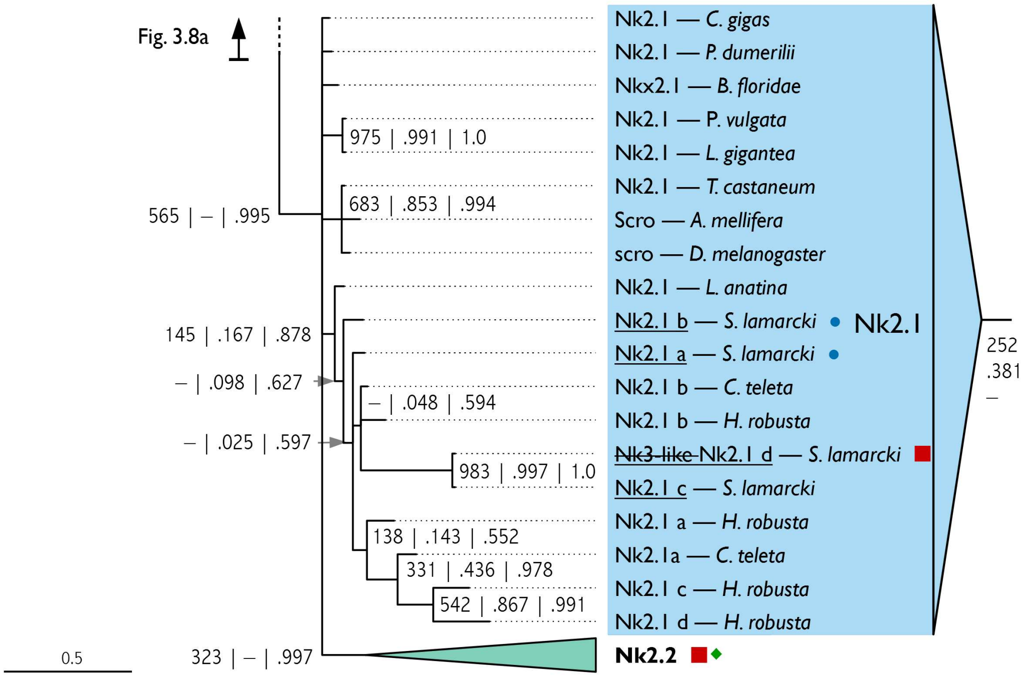
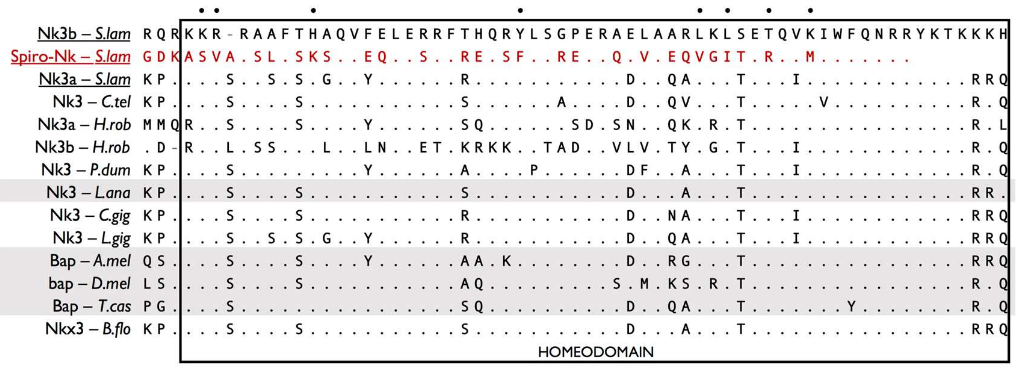


Figure 3.8 (2 parts). Bayesian phylogeny of Nk, Msx, Tlx and Lbx homeodomain sequences from a selection of bilaterian genomes, showing the various *S. lamarcki* gene duplications and the Spiro-Nk orphan. Support values for each node are from neighbour joining (out of 1000 bootstraps), maximum likelihood (proportion of 1000 bootstraps), and Bayesian (posterior probability) phylogenies (in order, separated by vertical bars or new lines). A dash indicates where a node is not present in the corresponding tree. Established bilaterian gene families that have been successfully reconstructed have been collapsed to coloured triangles, and a summary of their contents is presented in Table 3.8. In some cases, new families or family subsets containing several sequences all from a single genus have also been collapsed to aid visualisation. Colour selection is arbitrary, and not meant to indicate a relationship. Paralogue lettering, where present, is not intended to consistently imply direct orthology, though where evident, direct orthologues have been lettered accordingly. *S. lamarcki* sequences (all underlined) are marked with a green diamond if found in the regenerative transcriptome, with a red square if found in the developmental transcriptome, and a blue dot if found in both. Collapsed families have their *S. lamarcki* gene complement indicated nearby with the same symbols as above, with an empty black circle indicating a gene that has been found only in the genome. New gene families suggested herein are marked with an asterisk. Where a gene has been reclassified from Kenny and Shimeld (2012), the old classification is included but struck out. Established gene families that were successfully reconstructed in the neighbour joining and/or maximum likelihood analyses but not the Bayesian analysis are marked by a 'cartoon' clade (not to horizontal scale) and corresponding support values to the right. The scale bar indicates amino acid substitutions per site. Full sequence details are included in Appendix 3.2b. The alignment used to produce this tree is presented in Appendix 3.4e, and the full version of the Newick format tree is presented in Appendix 3.3e. Annelid species: *S. lamarcki* = *Spirobranchus lamarcki*; *C. teleta* = *Capitella teleta*; *H. robusta* = *Helobdella robusta*; *P. dumerilii* = *Platynereis dumerilii*. Brachiopod species: *L. anatina* = *Lingula anatina*. Mollusc species: *C. gigas* = *Crassostrea gigas*; *L. gigantea* = *Lottia gigantea*; *P. vulgata* = *Patella vulgata*. Insect species: *A. mellifera* = *Apis mellifera*; *D. melanogaster* = *Drosophila melanogaster*; *T. castaneum* = *Tribolium castaneum*. Deuterostome species: *B. floridae* = *Branchiostoma floridae*. Adapted from Barton-Owen, Szabó, Somorjai, & Ferrier (2018).

Table 3.8. Summary of the collapsed gene families in Figure 3.8. Clade colouration and symbols denoting sequence origin are the same as in Figure 3.8. Annelid species: *S.lam* = *Spirobranchus lamarcki*; *C.tel* = *Capitella teleta*; *H.rob* = *Helobdella robusta*; *P.dum* = *Platynereis dumerilii*. Brachiopod species: *L.ana* = *Lingula anatina*. Mollusc species: *C.gig* = *Crassostrea gigas*; *L.gig* = *Lottia gigantea*; *P.vul* = *Patella vulgata*. Insect species: *A.mel* = *Apis mellifera*; *D.mel* = *Drosophila melanogaster*; *T.cas* = *Tribolium castaneum*. Deuterostome species: *B.flo* = *Branchiostoma floridae*.

	Nk1	Nk2.1	Nk2.2	Nk3	(Nk4)	Nk5	Nk6	Nk7	Tlx	Lbx	Msx
S.lam	2 ◆◆	4 ●●■	2 ■◆	1	1	1 ●	1 ●	1	3 ◆	1	1 ◆
C.tel	2	2	2	1	2	2	1	1	1	1	1
H.rob	4	4	5	2	1	4	3	1	2	2	1
P.dum	1	1	1	1	1	1	1	1	1	1	1
C.gig	1	1	1	1	1	2	1	1	2	1	1
L.gig	2	1	1	1	1	1	1	1	2	1	3
P.vul	2	1	1	1	1	1	1	1	2	1	2
L.ana	2	1	1	1	1	1	1	1	1	1	1
A.mel	1	1	1	1	1	1	1	1	1	1	2
D.mel	1	1	1	1	1	1	1	1	1	2	1
T.cas	1	1	1	1	1	1	1	1	1	1	2
B.flo	2	1	1	1	1	1	1	1	1	1	1



```
                       ••      •              •            • •  •  •
Nk3b – S.lam      RQR KKR-RAAFTHAQVFELERRFTHQRYLSGPERAELAARLKLSETQVKIWFQNRRYKTKKKH
Spiro-Nk – S.lam  GDK ASVA.SL.SKS..EQ..S..RE.SF..RE..Q.V.EQVGIT.R..M.......
Nk3a – S.lam      KP. ...S..S.S.G..Y......R...........D..QA...T...I............RRQ
Nk3 – C.tel       KP. ...S....S...........S......A....D..QV...T.....V..........R.Q
Nk3a – H.rob      MMQ R..S....S....Y......SQ......SD.SN..QK.R.T................R.L
Nk3b – H.rob      .D- R..L.SS...L..LN..ET.KRKK..TAD..VLV.TY.G.T...I............R.Q
Nk3 – P.dum       KP. ...S.........Y......A....P......DF..A...T...I............R.Q
Nk3 – L.ana       KP. ...S....S...........S...........D...A...T................RR.
Nk3 – C.gig       KP. ...S....S...........R...........D..NA...T...I............RRQ
Nk3 – L.gig       KP. ...S..S.S.G..Y......R...........D..QA...T...I............RRQ
Bap – A.mel       QS. ...S....S....Y......AA.K........D..RG...T................RRQ
bap – D.mel       LS. ...S....S...........AQ.........S.M.KS.R.T................R.Q
Bap – T.cas       PG. ...S................SQ..........D..QA...T.......Y.........R.Q
Nkx3 – B.flo      KP. ...S....S...........S...........D...A...T................RRQ
                                              HOMEODOMAIN
```

Figure 3.9. Alignment of the Spiro-Nk homeodomain (red) and three preceding positions against assorted Nk3 gene family members, demonstrating that Spiro-Nk does not belong to the Nk3 family. Positions in the homeodomain which are invariant in the sampled Nk3s but different in Spiro-Nk are marked with a dot above the alignment. Annelid species: *S.lam* = *Spirobranchus lamarcki*; *C.tel* = *Capitella teleta*; *H.rob* = *Helobdella robusta*; *P.dum* = *Platynereis dumerilii*. Brachiopod species: *L.ana* = *Lingula anatina*. Mollusc species: *C.gig* = *Crassostrea gigas*; *L.gig* = *Lottia gigantea*. Insect species: *A.mel* = *Apis mellifera*; *D.mel* = *Drosophila melanogaster*; *T.cas* = *Tribolium castaneum*. Deuterostome species: *B.flo* = *Branchiostoma floridae*.

3.4. Discussion

Annelid genome evolution is considered to be generally conservative relative to various other animal lineages (Raible *et al.*, 2005; Hui *et al.*, 2009, 2012; Ferrier 2012). It was therefore unexpected to find such a surprising diversity of non-canonical and difficult-to-classify homeobox genes in the *S. lamarcki* regenerating opercular transcriptomes, including six TALE class genes, a PRD class gene, a gene similar to Nk, a divergent *Hox* gene, and another unclassifiable gene, in addition to the 56 other genes belonging to 46 known homeobox families (Table 3.2). To classify these genes, an in-depth survey of the TALE, PRD, Nk, and *Hox* content of the *S. lamarcki* genome (Kenny *et al.*, 2015) was performed and complemented with surveys of the available genomes of a number of other Lophotrochozoa and published sequences.

This section will primarily address the non-canonical and difficult-to-classify sequences. The canonical homeobox gene families found in the transcriptomes are discussed in Chapter 6.

3.4.1. The Spiralian TALE Expansion

In this analysis, 6 non-canonical TALE sequences from the *S. lamarcki* regenerative transcriptomes and 1 from the developmental transcriptome (Kenny and Shimeld 2012) were analysed with 18 sequences from the genome (Kenny *et al.*, 2015) and a broad sampling of sequences from other lophotrochozoan genomes. The 9 TALE clades of Paps *et al.* (2015) were successfully reconstructed, although the reclassification of some sequences modified their constituents. In addition, 10 new clades were named (TALE clades X-XIX).

The integration of the Paps *et al.* (2015) and Morino *et al.* (2017) nomenclatures is discussed in the next section (section 3.4.2.3), followed by some observations about the properties of the various TALE clades (section 3.4.2.4). Some general observations about the dynamics of the Spiralian TALE expansions (STE) are made in section 3.4.2.5. Some important methodological considerations to ongoing efforts to profile the STE are discussed in section 3.4.2.6, and the current state of knowledge about the biology of STE genes is discussed in section 3.4.2.7. Some priorities for STE study and some potential tools for the time-efficient survey of STEs in new and existing genomes are discussed are section 3.4.6.

3.4.1.1. The SPILE nomenclature

The phylogenetic analysis of Morino *et al.* (2017, Figure 1) indicated a distinction between a monophyletic clade (with a support value of 82%) containing the majority of STE genes, and another containing all canonical TALE families and a selection of other STE genes. This STE monophylum is reconstructed in the Bayesian analysis presented in Figure 3.4, although it was not exactly reconstructed in the concurrent neighbour-joining or maximum likelihood analyses, and the clade received a posterior probability of only 0.613. TALE clades I and X-XIV are non-SPILE clades, and TALE clades II-VIII and XV-XIX are SPILE clades (see TALE-IX discussion below). Of these, sequences from TALE clades X, XI, XIV, XVI and XIX appear only in the current analysis, so are unverified in their placement.

The SPILE clade is less obvious but nonetheless detectable in the tree of Paps *et al.* (2015, Supplementary Figure 4) in the topological proximity of TALE-I and *C. gigas* TALE5 (now TALE-XIII) to the canonical TALE families and the clade containing TALE clades II-VIII + Cers, though the topology is disrupted by the inclusion of SINE, CERS and Hnf class homeodomains. However, the analysis of Paps *et al.* (2015) contradicts Figure 3.4 by placing TALE-IX among the apparent SPILE genes. Unfortunately, TALE-IX sequences were omitted from the analysis of Morino *et al.* (2017). TALE-IX was placed among the non-SPILE genes by my analysis, and is referred to as non-SPILE in summaries presented herein.

The establishment of the SPILE nomenclature is perhaps unfortunate given that the STE genes outside the SPILE clade have as much claim to the designation ‘SPIralian taLEs’ as those within it. A name reflecting their (presumably) common gene ancestry — for example, the Spiralian TALE Orthology Group — might have been preferable, but has not been adopted herein to promote continuity with past research.

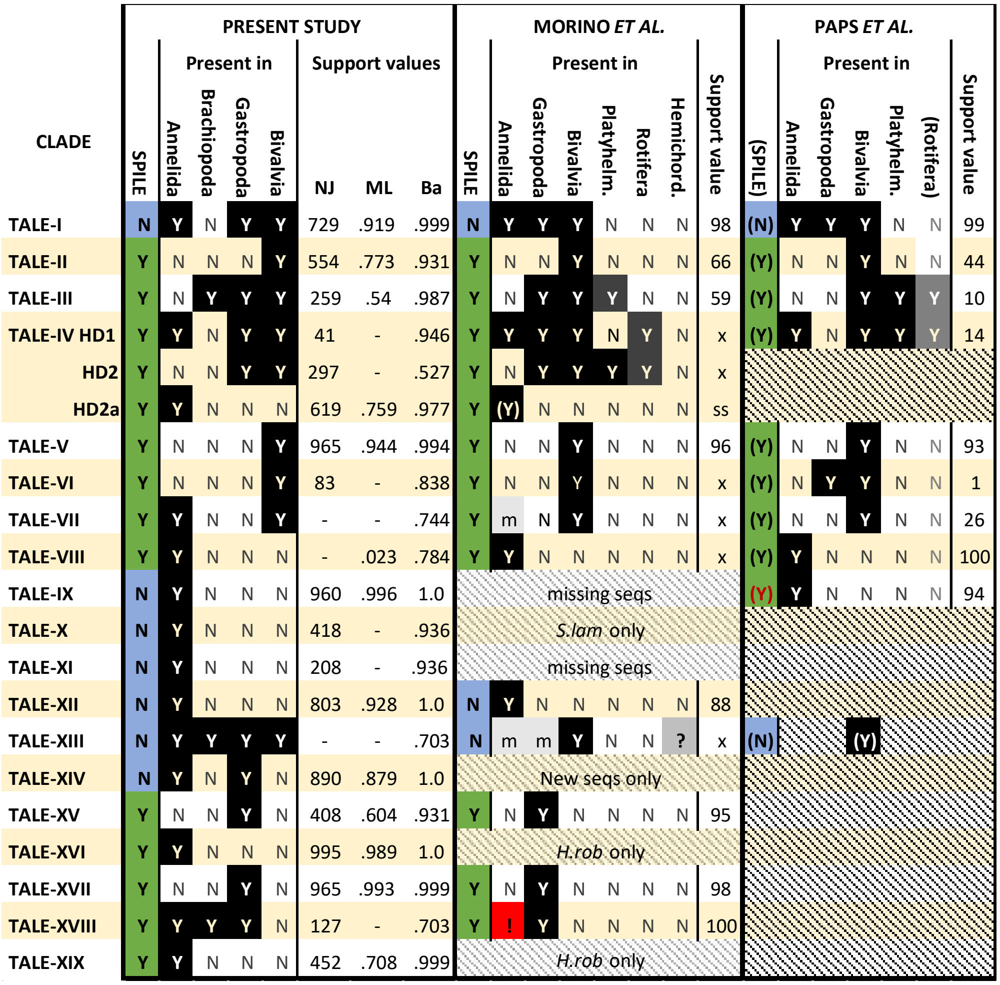
3.4.1.2. Comparison with previous datasets

Tabular comparisons between the present study, Morino *et al.* (2017), and Paps *et al.* (2015) are presented below. A summary of the presence/absence, SPILE placement, taxonomic distribution and associated support values for each TALE clade is given in Table 3.9. A summary of the SPILE and non-SPILE gene count from each surveyed genome is given in Table 3.10.

Full details of the sequences used in this analysis, and how they correspond to the names used in Paps *et al.* (2015) and Morino *et al.* (2017) are included in Appendix 3.2b. A graphical comparison of the TALE clade constituents and their identity between analyses is presented in Appendix 3.5.

The searches and analyses in the present study were performed before the publication of the work of Morino *et al.* (2017), and consequently sequences retrieved by their surveys but missed by mine were not integrated into the analyses presented herein. These include eight homeodomain sequences from *Lottia gigantea*, one from *Pinctada fucata,* and two from *C. teleta*.

Table 3.9. Comparison of cladistic and topological data from the present study, Morino *et al.* (2017), and Paps *et al.* (2015). In the 'Present in' columns, parentheses indicate that the clade is represented by a single gene in that analysis, 'm' that the necessary data are missing from this analysis, '?' when the data are insufficient to determine if the sequences belong to the clade (as is the case with the possible hemichordate TALE-XIII), and '!' when the sequence has been misplaced. In the analysis of Paps *et al.* (2015), rotiferan sequences were indicated as present in tree notations but weren't actually present in the tree. The analysis of Paps *et al.* also did not reconstruct a SPILE clade, but a comparable distinction is visible in their topology. Support values derive from the following phylogenetic analysis methodologies in the present study: NJ = neighbour joining; ML = maximum likelihood; Ba = Bayesian.



	PRESENT STUDY								MORINO *ET AL.*								PAPS *ET AL.*						
		Present in				Support values				Present in								Present in					
CLADE	SPILE	Annelida	Brachiopoda	Gastropoda	Bivalvia	NJ	ML	Ba	SPILE	Annelida	Gastropoda	Bivalvia	Platyhelm.	Rotifera	Hemichord.	Support value	(SPILE)	Annelida	Gastropoda	Bivalvia	Platyhelm.	(Rotifera)	Support value
TALE-I	N	Y	N	Y	Y	729	.919	.999	N	Y	Y	Y	N	N	N	98	(N)	Y	Y	Y	N	N	99
TALE-II	Y	N	N	N	Y	554	.773	.931	Y	N	N	Y	N	N	N	66	(Y)	N	N	Y	N	N	44
TALE-III	Y	N	Y	Y	Y	259	.54	.987	Y	N	Y	Y	Y	N	N	59	(Y)	N	N	Y	Y	Y	10
TALE-IV HD1	Y	Y	N	Y	Y	41	-	.946	Y	Y	Y	Y	N	Y	N	x	(Y)	Y	N	Y	Y	Y	14
HD2	Y	N	N	Y	Y	297	-	.527	Y	N	Y	Y	Y	Y	N	x							
HD2a	Y	Y	N	N	N	619	.759	.977	Y	(Y)	N	N	N	N	N	ss							
TALE-V	Y	N	N	N	Y	965	.944	.994	Y	N	N	Y	N	N	N	96	(Y)	N	N	Y	N	N	93
TALE-VI	Y	N	N	N	Y	83	-	.838	Y	N	N	Y	N	N	N	x	(Y)	N	Y	Y	N	N	1
TALE-VII	Y	Y	N	N	Y	-	-	.744	Y	m	N	Y	N	N	N	x	(Y)	N	N	Y	N	N	26
TALE-VIII	Y	Y	N	N	N	-	.023	.784	Y	Y	N	N	N	N	N	x	(Y)	Y	N	N	N	N	100
TALE-IX	N	Y	N	N	N	960	.996	1.0	missing seqs								(Y)	Y	N	N	N	N	94
TALE-X	N	Y	N	N	N	418	-	.936	*S.lam* only														
TALE-XI	N	Y	N	N	N	208	-	.936	missing seqs														
TALE-XII	N	Y	N	N	N	803	.928	1.0	N	Y	N	N	N	N	N	88							
TALE-XIII	N	Y	Y	Y	Y	-	-	.703	N	m	m	Y	N	N	?	x	(N)			(Y)			
TALE-XIV	N	Y	N	Y	N	890	.879	1.0	New seqs only														
TALE-XV	Y	N	N	Y	N	408	.604	.931	Y	N	Y	N	N	N	N	95							
TALE-XVI	Y	Y	N	N	N	995	.989	1.0	*H.rob* only														
TALE-XVII	Y	N	N	Y	N	965	.993	.999	Y	N	Y	N	N	N	N	98							
TALE-XVIII	Y	Y	Y	Y	N	127	-	.703	Y	!	Y	N	N	N	N	100							
TALE-XIX	Y	Y	N	N	N	452	.708	.999	*H.rob* only														

Table 3.10. Comparison of the number of sequences retrieved from each genome by the present study, Morino *et al.* (2017), and Paps *et al.* (2015). Numbers in brackets indicate that the sequences have been taken directly from the preceding analysis, not from an independent search. A range indicates the presence of multiple identical homeodomains (*e.g. C. teleta*) or partial homeodomains which are likely but not certainly parts of the same homeodomain (see Figure 3.5). Paps *et al.*'s analysis did also not reconstruct a SPILE clade, but comparable distinction is visible in their topology; for the purposes of the Paps *et al.* analysis tally, *C. teleta* TALE-IX sequences have been counted as SPILEs (marked with asterisk).

		PRESENT		MORINO *ET AL.*		PAPS *ET AL.*	
PHYLUM/CLASS	SPECIES	SPILE	NON-SPILE	SPILE	NON-SPILE	SPILE	NON-SPILE
ANNELIDA	*Spirobranchus lamarcki*	14-16	11	-	-	-	-
	Spirobranchus kraussi	(3)	(0)	3	0	-	-
	Capitella teleta	10-12	12	11-14	5	10-13	1*
	Helobdella robusta	19	1	-	-	0?	1?
	Platynereis dumerilii	1	2	-	-	-	-
BRACHIOPODA	*Lingula anatina*	2	2	-	-	-	-
GASTROPODA	*Lottia gigantea*	3	1	9-12	1	3	1
	Patella vulgata	5	3	-	-	-	-
	Nipponacmea fuscoviridis	(5)	(0)	5	0	-	-
BIVALVIA	*Pinctada fucata*	(13)	(1)	13	1	13-14	1
	Crassostrea gigas	(12)	(2)	(12)	(2)	12	2
PLATYHELMINTHES	*Echinococcus multilocularis* or *granularis*	-	-	3	1	3	0
	Clonorchis sinensis	-	-	-	-	1	0
	Schistosoma mansoni	-	-	-	-	2	0
ROTIFERA	*Adineta vaga*	-	-	6	0	2	0
HEMICHORDATA	*Saccoglossus kowalevskiii*	-	-	0	1	-	-
CEPHALOCHORDATA	*Branchiostoma floridae*	-	-	(0)	(1)	-	-
CNIDARIA	*Nematostella vectensis*	-	-	0	1	-	-

3.4.1.3. General observations

S. lamarcki and *C. teleta* are vying for the most STE-rich species yet surveyed, each having at least 25 sequences. In *S. lamarcki*, most of the sequences are concentrated in TALE clades containing sequences from other species, with only three orphan/*S.lamarcki*-only TALE sequences (five if TALE-XI is rejected), whereas the *C. teleta* complement is more diverse, with at least 11 such sequences (12 without TALE-XI, more if the long-branch rejects from Morino *et al.* are included).

The diversity of *S. lamarcki* and *C. teleta* STE sequences is not found throughout the annelids. Assuming that the TALE clades are reliable indicators of orthology and that the surveys have been exhaustive, *Platynereis dumerilii* and *Helobdella robusta* both seem to have undergone major STE loss, retaining only TALE-I (both) and TALE-IV (*P. dumerilii* only). *H. robusta* appears to have undergone a recent proliferation of 18 sequences in two *H. robusta*-only clades (TALE clades XVI and XIX, both SPILE). These gains and losses in could be related to the unusual dynamism of its genome (Simakov *et al.*, 2013), which seems to also be reflected in its non-STE homeobox complements (*c.f.*

Table 3.3, Table 3.4, Table 3.7, & Table 3.8). In contrast, *P. dumerilii* is the exemple *par excellence* of the generally conservative nature of annelid genome evolution (Raible *et al.*, 2005; Hui *et al.*, 2009, 2012; Ferrier 2012; Zantke *et al.*, 2014). Greater taxon sampling (see section 3.4.6.1) and data on STE synteny are needed to understand the dynamics of STE evolution in annelids.

The pattern of gene retention across the Lophotrochozoa as a whole is similarly heterogenous. With the assumption that the TALE clades are orthology groups, the lophotrochozoan ancestor must have possessed at least TALE clades I, III, IV, VII, XIII, XIV, and XVIII, but no species yet surveyed has representatives of all these clades. Furthermore, rarely do these losses seem to be common to whole classes or phyla; with the taxonomic resolution currently available, most gene loss seems to have occurred relatively recently and piecemeal between species.

The extent to which the STE should be seen as a truly spiralian-wide phenomenon is debatable. The vast majority of the expansion seems to have occurred in the Lophotrochozoa; thus far STE sequences belonging to non-lophotrochozoan spiralians number only ten from five genomes, compared to at least 115 sequences from the nine surveyed lophotrochozoan genomes (excluding *S. kraussi* and *N. fuscoviridis*). All but two of the TALE clades are restricted to the Lophotrochozoa, and so far, a single orphan (*Echinococcus multilocularis TALEHD4*, Morino, Hashimoto, and Wada 2017) is the only indicator that any TALE clades without lophotrochozoan genes exist.

The extent to which the TALE clade nomenclature reflects real orthology groups is unclear. It is probable that many of the TALE clades are comprised of orthologues, but only one — TALE-IV — has been examined in any depth beyond the homeodomain sequence. The presence of two homeodomains, and the conserved regions C-terminal to each (Figure 3.5) both provide support beyond phylogenetic analyses of the homeodomain for the notion that these genes belong to a distinct orthology group stretching at least as far as the base of the Lophotrochozoa, but very probably beyond (Paps *et al.*, 2015; Morino, Hashimoto, and Wada 2017) (see below).

The evolutionary history of the STE is difficult to even speculatively reconstruct from the current paucity of information. The SPILE clade, apparently initially containing only TALE clades III and IV, was present in the common ancestor of the Rotifera, Platyhelminthes, and Trochozoa. Five more TALE clades (listed above), including three non-SPILE clades, originated in the common lophotrochozoan ancestor. In the annelids and molluscs, the pace of TALE expansion (from both SPILE and non-SPILE origins) has been by far the most rapid and flexible, with frequent duplication, rapid divergence to the point where the original paralogy is mostly undetectable beyond the SPILE/non-SPILE distinction, and seemingly unconstrained losses.

TALE-IV

On the basis of the evidence for orthology of domain structure and of conserved motifs outside the homeodomain (Figure 3.5), TALE-IV represents a genuine orthology group which stretches at least as far back as the Lophotrochozoa. *Adineta vaga* TALEHDs 1 & 2 and 3/5 & 4/6 (3 and 5 are identical, as are 4 and 6) (Morino, Hashimoto, and Wada 2017) lie only 3-4 kilobase pairs apart from one another, but the annotation (Flot *et al.*, 2013) places them all in separate transcripts. Sequence comparison (Appendix 3.4f) also indicates that the non-homeobox conserved motifs aren't present in the *A. vaga* sequences. TALEHDs 1, 3 & 5 more closely resemble lophotrochozoan TALE-IV HD2s than TALEHDs 2, 4 & 6 resemble lophotrochozoan TALE-IV HD1s. Further work is necessary to determine if the annotations are correct and if any other indicators of orthology can be detected in *A. vaga* and other non-lophotrochozoan Spiralia. From these data, it might be possible to reconstruct some of the specifics of the event that potentially merged two adjacent homeobox genes.

Within the Lophotrochozoa, the sequences have a non-straightforward topology that necessitates explanation. Specifically, the annelids seem to have two paralogues; in one, the second homeodomain has diverged from the mollusc sequence into a completely separate clade (labelled 'A' in Figure 3.4 & Figure 3.5), and in the other the second homeodomain has been degraded. Although evidence of the original homeodomain is visible in alignments, the degraded homeodomain regions are dissimilar to one another, suggesting independent loss events. In *S. kraussi*, the homeodomain of the former paralogue (the non-A-type) has also been degraded. *P. dumerilii* also has an orphan gene (*TALE-IV-like*) that has been placed beside or within TALE-IV in both the present study and by Morino *et al.* (2017), but with low support values.

Curiously, a *H. robusta* sequence (*TALE-XIX A*) also seems to have acquired a second homeobox independently of the presumed TALE-IV pro-orthologue. A multi-homeobox state has not previously been observed for any TALE class genes, and is only rarely seen in some other animal homeobox gene classes, such as *Hdx* (POU class), *dve*/*Compass* (CUT class), *Zfhx* and *Zhx*/*Homez* (ZF class), *Muxa* and *Muxb* (orphan genes in amphioxus), and *Dux* genes in mammals (PRD class) (Booth and Holland 2007; Takatori *et al.*, 2008; Zhong and Holland 2011).

3.4.1.4. Methodological concerns (or: it's too late topologise)

Search saturation

Three surveys (Paps *et al.*, 2015; Morino, Hashimoto, and Wada 2017; present study) of the *C. teleta* genome have all retrieved an overlapping but different set of STE homeodomains; that is, although each researcher may have saturated the searches performed using their particular methodology, the survey efforts as a whole are in principle incomplete. One possible explanation is the query pool; some of the most divergent homeodomains may be 'out of reach' of most queries. Many of the missed *C. teleta* sequences in question are *C. teleta*-specific and some are extremely long-branch, indicating that these sequences are missed because they are not represented well enough in the query sequences. As a result, species-restricted, non-lophotrochozoan, and highly divergent homeodomains may be systematically under-detected by previous BLAST search methodologies.

One solution to this problem is to use and maintain as wide a query pool as possible, and for previously surveyed genomes to be re-surveyed as the query pool is broadened. Unfortunately, these deep recursive searches are extremely time-consuming. Proposals for some automative tools to reduce the unnecessary human workload are presented in section 3.4.9.1.

Topology

Many of the TALE clades discussed herein have been reconstructed in either or both of the previous published phylogenetic analyses (Paps *et al.*, 2015; Morino, Hashimoto, and Wada 2017), as well as in either or both of the neighbour-joining and maximum likelihood phylogenies performed on the same dataset as the presented Bayesian phylogeny. The reconstruction of a clade between analyses including different sets of sequences and between multiple types of phylogenetic analysis on the same dataset is perhaps as good an indicator of the degree of confidence which should be placed in a clade as the support values it receives from any one analysis.

However, several of the clades inspire rather less confidence for various reasons; they're as yet unreconstructed in other analyses (*e.g.* TALE clades X, XI, XIV, XVI & XIX), their sequences are more heterogeneous than other clades (*e.g.* TALE-X), or their composition has changed between analyses (e.g TALE-VI). As aptly demonstrated by TALE-VI, the discovery of more sequences is likely to disrupt the topology presented here.

Nomenclature

The non-arbitrary and coherent determination of what nodes should be dubbed TALE clades when sequence diversity is so high and apparent orthologue retention is so unevenly distributed is a problem for the study of the STE. Although the tree presented in Morino *et al.* (2017, Figure 1) almost completely recapitulates the topology presented here (Figure 3.4), it would not be possible to non-arbitrarily impose a comparable nomenclature, and if the nomenclature used herein is mapped onto their tree, the clades seem to be chosen incoherently. This is presumably why Morino *et al.* (2017) elected to erect the

SPILE nomenclature and not attempt to integrate the TALE clade nomenclature of Paps *et al.* (2015).

In contrast, the Bayesian analysis presented in this analysis (Figure 3.4) collapses the (presumably uninformative) topology that separates informative clades into large polytomies with relatively little 'nesting'. In most cases, the clades appearing proximally from these polytomies represent meaningful distinctions, including the well-established canonical TALE families. Therefore, with the application of the criteria that a node must either be supported in all three parallel analyses (NJ, ML, Bayesian) or have received a support value of at least 70% in the Bayesian tree to be classified as an TALE clade, the node/clade determinations in this analysis are on a basis comparable to the canonical families.

Another issue with nomenclature may materialise as and when the clades named herein are disrupted in future analyses: specifically, that the addition of more sequences will pull previously described clades apart to the point that the old nomenclature becomes confusing and misleading. This problem can be partly ameliorated by the maintenance of a clear, explicit and open record of the relation between the latest nomenclature and previous iterations (*e.g.* Table 3.5, Appendix 3.5) If changes to the nomenclature are made during the course of analysis and publication, it would be preferable if these changes were reflected in all published material (TALE clade numbering is inconsistent between the manuscript, Supplementary Figure 4, and Supplementary Figure 7 of Paps *et al.*, 2015).

3.4.1.5. STE biology

Expression data are currently restricted to two members of the genus *Spirobranchus* (present study, Kenny and Shimeld 2012; Morino, Hashimoto, and Wada 2017), the bivalve *C. gigas* (Paps *et al.*, 2015; F. Xu *et al.*, 2016), and the gastropod *N. fuscoviridis* (Morino, Hashimoto, and Wada 2017). A tabular summary is presented in Table 3.11. Normalised developmental expression data of the 14 *C. gigas* STE genes, taken from Xu *et al.* (2016), are presented in Figure 3.10.

Early development

The preponderance of expression data indicates most STE genes are expressed almost exclusively between the zygote and morula stages. Transcriptomic data from Paps *et al.* (2015) and Xu *et al.* (2016) provide evidence for the expression of *C. gigas* TALEs 1-4 and 7-14 in early development, with very little expression of most genes after the gastrula (Figure 3.10; Paps *et al.*, 2015, Figure 3).

Morino *et al.* (2017) produced the first *in situ* hybridisation images of STE genes and both knock-down and over-expression data for a selection of their *N. fuscoviridis* SPILE genes. Their data indicate that SPILE genes are expressed in quartet-specific domains in the early development of both molluscs and annelids, and (at least in molluscs) engage in complex inter-regulation of one another in the specification of cell fate along the animal-vegetal axis. *SPILE-D* (TALE-XVIII) specifies the apical structure, and knock-downs lost both the apical tuft and later expression of *Otp*, a first quartet derivative marker. *SPILE-C* (TALE-XV) is sufficient to determine macromere (3Q) cell fate.

The gene orthology inference from this study (discussed in section 3.4.2.6) could help interpret the data of Morino *et al.* (2017). For example, *N. fuscoviridis SPILE-B*, which is expressed in the 2nd quartet at 16 cells and then the vegetal pole (3rd quartet and macromere) at 32 cells, has well-supported orthology (being members of TALE-IV) to *S. krausii SPILE*s *X* and *Y* - deployed in the animal pole at the same stages.

Table 3.11. Summary of biological data concerning STE genes from the present study, Kenny & Shimeld (2012), Paps *et al.* (2015) and Morino *et al.* (2017). Abbreviations: Loph. = Lophotrochozoa (*sensu stricto*); Cg = *Crassostrea gigas*; Sl = *Spirobranchus lamarcki*; Nf = *Nipponacmea fuscoviridis*; Sk = *Spirobranchus kraussi*; 0d = mature unregenerating opercular transcriptome; 2d = 2 day regenerating opercular transcriptome; 6d = 6 day regenerating opercular transcriptome; mtr = maternal; 4/8/16/32c = 4/8/16/32 cell stage; troch. = trochophore stage; 2q = second quartet; veg. = vegetal pole; anim. = animal pole; nucl = nuclear (ubiquitous); OE = over-expression, KD = morpholino knock-down. In the Morino *et al.* column, changes to expression are marked by an arrow (->) in both the stages and regions columns; the stage ranges and regions before and after the arrows correspond to one another respectively. Where KD is greyed out, it was not found to have an effect.

CLADE	SPILE?	ORIGIN/ RESTRICTION	Morino *et al.* 2017				Paps *et al.* 2015.		Kenny & Shimeld 2012		Present study	
			GENE	STAGES	REGION	FUNC	GENE	STAGES	GENE	STAGES	GENE	READ COUNTS 0d|2d|6d
TALE-I	**N**	**LOPH.**					Cg TALE2	2c			Sl TALE-I A Sl TALE-I B	3|11|52 14|4|95
TALE-II	**Y**	**BIVALVIA**					Cg TALE1	egg				
TALE-III	**Y**	**SPIRALIA**	Nf SPILE-E	mtr-32c	ubiquitous		Cg TALE3	egg				
TALE-IV	**Y**	**SPIRALIA**	Nf SPILE-B Sk SPILE-X Sk SPILE-Y	16c -> 32c 16c-32c 16c-32c	2q -> veg. animal mixed anim.	OE+KD	Cg TALE7 Cg TALE8 Cg TALE14	early morula early morula blastula				
TALE-V	**Y**	**BIVALVIA**					Cg TALE6	juvenile				
TALE-VI	**Y**	**BIVALVIA**					Cg TALE9 Cg TALE11 Cg TALE12 Cg TALE13	early morula early morula early morula early morula				
TALE-VII	**Y**	**LOPH.**					Cg TALE4 Cg TALE10?	early morula earlu morula				
TALE-VIII	**Y**	**SEDENTARIA**	Sk SPILE-Z	8c-32c	vegetal							
TALE-IX	**N**	***CAPITELLA***										
TALE-X	**N**	***SPIRO-BRANCHUS***									Sl TALE-X A Sl TALE-X B	3306|1743|725 3|3|1
TALE-XI	**N**	**SEDENTARIA**										
TALE-XII	**N**	***CAPITELLA***										
TALE-XIII	**N**	**LOPH.**					Cg TALE5	juvenile			Sl TALE-XIII A Sl TALE-XIII B	13|0|6 6|0|9
TALE-XIV	**N**	**LOPH.**										
TALE-XV	**Y**	**GASTROPODA**	Nf SPILE-C	16c-32c	vegetal	OE+KD						
TALE-XVI	**Y**	***HELOBDELLA***										
TALE-XVII	**Y**	**GASTROPODA**	Nf SPILE-A	4c-8c -> 16c-32c	nucl. -> veg.	KD						
TALE-XVIII	**Y**	**LOPH.**	Nf SPILE-D	mtr-8c -> 16c-32c	ubiq. -> veg.	KD+OE			Sl TALE-XVIII A	24h-72h		
TALE-XIX	**Y**	***HELOBDELLA***										

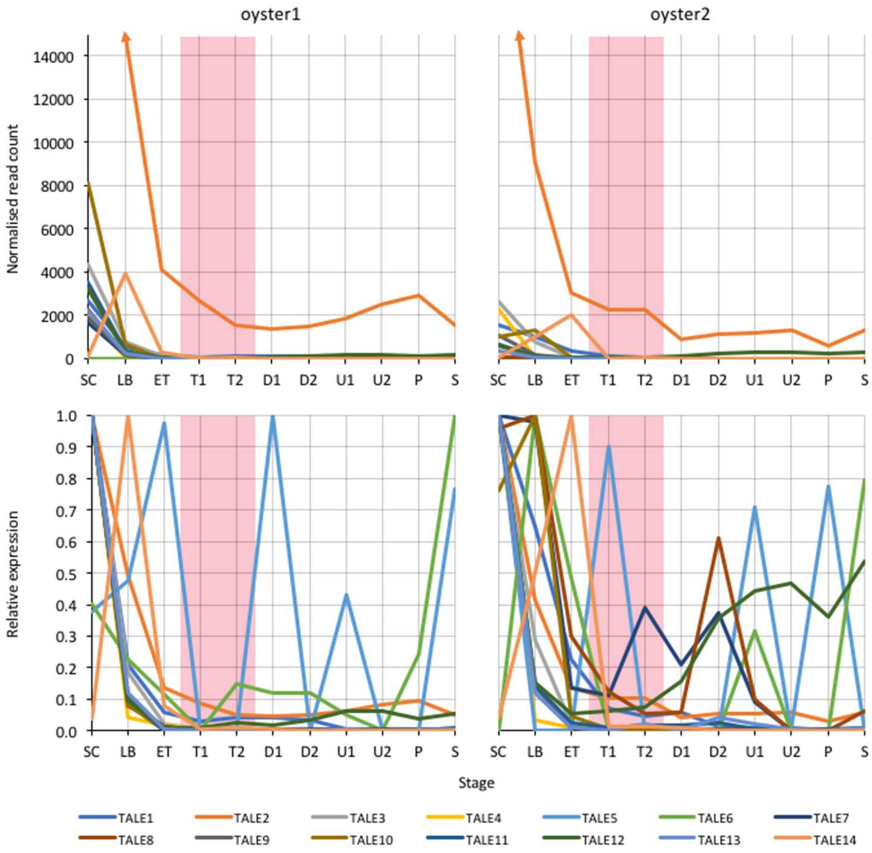


Figure 3.10. Normalised developmental expression data for the 14 *Crassostrea gigas* STE genes described by Paps *et al.*, (2015), using data from Xu *et al.*, (2016), Supplementary Table S2. Read counts were normalised against the total read size of the SC transcriptome for each replicate. The pink bar is used to indicate the trochophore stage, after Xu *et al.*, SC = single cell (presumed); LB = late blastula (presumed); ET = Early Trochophore (presumed); T1/2 = Trochophore 1/2; D1/2 = D-shaped larvae 1/2; U1/2 = Umbo 1/2; P = Pediveliger; S = Spat.

*SPILE*s *D* and *B* belong to lophotrochozoan-wide (TALE-XVIII) and spiralian-wide (TALE-IV) clades respectively, but *SPILE*s *A* and *C* seem to have no non-gastropod orthologues. These genes are actually quite ancient — the most recent common ancestor of the Patellogastropoda (to which all gastropods sampled herein belong) was dated to 156–297 million years ago (Nakano and Ozawa 2007), and of the Gastropoda as a whole to approximately 350-500 million years ago (Zapata *et al.*, 2014) — but are nonetheless taxonomically restricted.

Although Morino *et al.* (2017) invite the comparison between SPILEs and Hox genes, it seems likely based on these observations that the evolutionary dynamics of these genes represents the polar opposite to the deep coding sequence, syntenic, and regulatory conservation of Hox cluster genes, instead being typified by rapid gene turnover and flexibility of deployment. Only the expansion of taxonomic sampling of *in situ* hybridisation and functional data will be able to elucidate the aptitude of this comparison.

Similarly, it seems likely that evidence to support the hypothesis of Morino *et al.* (2017) that the STE event was critical for the origination of spiralian development would be obscured by the rapid, flexible evolution and turnover of these genes. The available evidence of orthologue deployment and function (*i.e. NfSPILE-B* vegetal pole expression vs *SkSPILE*s *X* and *Y* animal pole expression, see above) and the roles of taxonomically-restricted clades (*i.e. NfSPILE-C,* a member of the gastropod-only TALE-XV, specifies the macromeres)*,* already support the suspicion that very little signal will survive from deep evolutionary time.

Mid to late development

Paps *et al.* (2015) highlighted the general paucity of STE expression in the trochophore stage, relating this to the evolutionary developmental hourglass (introduced in section 3.1.2). The emerging picture of apparently extreme evolutionary flexibility of the STE genes seems likely to predispose them against deployment in a highly constrained phylotypic stage.

Intriguingly, a *Spirobranchus* gene identified as a divergent *Mkx* paralogue (*Mkx2*) by Kenny and Shimeld (2012) but reclassified here as a member of TALE-XVIII, derives from a transcriptome of 24 to 72 hour embryos, which (given that the developmental timing of the Plymouth and St Andrews Bay populations of *S. lamarcki* are not different), are trochophores (McDougall *et al.*, 2006).

Xu *et al.* (2016) found that the transcriptome age indices reach their peak at the late trochophore stages of *C. gigas,* the Pacific abalone *Haliotis discus hannai*, and the polychaete sand worm *Perinereis aibuhitensis*, and the transcriptome diversity index reached its minimum in the *C. gigas* trochophore. However, the patterns of relative expression of genes of different phylostrata are complex, and genes in the 2 most recent phylostrata — Bivalvia- and *C. gigas*- specific — were highly expressed.

The BLAST-based phylostratigraphy pipeline of Xu *et al.* (2016) placed *C. gigas* TALEs 1-14 in PS2, the Eukaryote phylostratum. This placement doesn't represent a failure of these pipelines insofar as that they are explicitly designed to (approximately) determine the *de novo* origin of taxonomically-restricted genes via the first known instance of broad homology, and not to distinguish the emergence and divergence of cryptic paralogues. However, it does indicate that global expression pattern analyses can entirely miss the biologically meaningful and critical developmental information found in transcription factor evolution. Paps *et al.* (2015) point to their homeobox-centric candidate gene approach as a factor behind the strength of their phylotypic signal.

Another possibility to explain the expression of *S. lamarcki* TALE-XVIII is that the purported phylotypic restraints to the trochophore don't actually apply particularly strongly to this system. Firstly, the phylotypic period is not as unambiguously defined in annelids as in some other phyla. Xu *et al.* (2016) place it in the late trochophore, and this stage was also identified as the point of the mid-developmental transition in *P. dumerilii* (Levin *et al.*, 2016). However, it has previously been suggested to be the segmenting metatrochophore (Slack 2003), and if simple enumeration of expressed Hox genes may be taken as a phylotypic indicator (as is often assumed: *e.g.* Xu *et al.*, 2016), the *P. dumerilii* data would point to the metatrochophore or nectochaete stages (Kulakova *et al.*, 2007). *S. lamarcki* is the only member of the Sedentaria sampled herein which actually retains a trochophore; *C. teleta* larval development omits it (Hill and Boyer 2001) and the Hirudinea develop directly (Purschke 2002).

Secondly, regardless of phylotypology, the trochophore of *S. lamarcki* is already thought to be unusual in the apparent dearth of any Hox gene deployment, a feature of larval development of other annelids (Irvine and Martindale 2000; Peterson *et al.*, 2000; M. Kulakova *et al.*, 2007; Fröbius, Matus, and Seaver 2008; Steinmetz *et al.*, 2011) (see further discussion in section 3.4.4).

The data of Paps *et al.* (2015) data indicate that the strongest expression of any STE gene in the *C. gigas* trochophore is *TALE5* (TALE-XIII, non-SPILE) during the Trochophore 3 and 4 stages, though *TALE5* is expressed rather patchily throughout development and peaks in the juvenile. The data from Xu *et al.* (2016) indicate that *TALE2*

(TALE-I, non-SPILE) expression, though very high in early development (the most reads of any STE gene), continues through the trochophore. This trend is less obvious but evident in the data of Paps *et al.* (2015), which also indicate late developmental expression of *TALE12* (absent from the data of Xu *et al*, 2016).

Adult and regeneration

TALE5 (TALE-XIII) and *TALE6* (TALE-V) were reported by Paps *et al.* (2015) to be most expressed in the juvenile, though they both also have developmental expression.

In the *S. lamarcki* regenerative transcriptomes, six STE sequences are reported: TALE-I A and B, TALE-X A and B, and TALE-XIII A and B. Insofar as expression information can be reliably inferred from un-replicated read counts, TALE-X A is relatively highly expressed (5,774 total reads) while all the others are expressed relatively little (particularly TALE-X B, with seven reads). No sequence is most highly expressed in early regeneration, but several peak in late regeneration (*e.g.* TALE-I A and B). All are apparently expressed in mature tissue. None of the sequences share a common expression topology (*i.e.* have their expression maxima and minima in the same stages).

With the information currently available, it seems possible that the non-SPILE STE genes are more likely to be expressed in adults, in apparent contrast with the propensity for SPILE deployment in early development. However, this impression is almost completely based on the data from *S. lamarcki* — the only mature adult/regenerative expression data considered — and so can only be taken as tentative.

3.4.2. The PRD Expansion

The PRD expansion (PRD-E) is at a much more modest scale than the TALE expansion, currently totalling just 33 spiralian sequences (and six *B. floridae* sequences). Paps *et al.* (2015) erected a nomenclature system for describing their spiralian data comprising of six clades (I-VI), which this analysis has successfully reconstructed. In addition, I named one more clade, PRD-VII.

Unlike the STE, there is no evidence of a 'meta-clade' suggesting a common origin event of more than one of the clades. However, this analysis indicates that some of the

PRD clades (namely clades I, IV, VI and V) include cephalochordate *Aprd* genes. Although further work is needed to determine if any degree of meaningful orthology is detectable between cephalochordates and lophotrochozoans (see below), this result raises the possibility that the most recent common ancestor of the two groups — probably the ancestor of all Bilateria — possessed a number of PRD Clade pro-orthologues which, being lost in most deuterostomes and the Ecdysozoa, were previously unidentified as homeobox families. This possibility should be treated with caution until it has been better substantiated.

Some of the PRD-E genes have extremely unusual and possibly non-functional homeodomains. *C. gigas* PRD3 appears to have lost 28 residues from the centre of the homeodomain, corroborated in another mollusc Clade VI sequence (*L. gigantea* PRD-VI) but not present in *B. floridae Aprd5* (also placed in PRD Clade VI). This major homeodomain modification is evidence for molluscan orthology but does not weigh in the favour of cephalochordate inclusion. *C. gigas* PRD7 (but not *C. teleta* PRD-III, both members of PRD clade III) has apparently lost/replaced the exon after the conserved intron at position 46/47, although the strong possibility remains that this is the result of an inaccurate gene model (CGI_10025814). Although these divergent homeodomains are still recognisable as such by the NCBI CDS (Marchler-Bauer *et al.*, 2011, 2015), phylostratigraphy by Xu *et al.*, (2016) placed PRD3 in PS10 (Bivalvia) and PRD7 in PS4 (Metazoa).

3.4.2.1. PRD-E biology

The sole *S. lamarcki* PRD-E gene expressed in opercular regeneration, PRD-V, is not found in either the uncut/mature tissue or in late/6dpf regeneration, and is present as only 19 reads in the early/2dpf regenerative transcriptome. A PRD-E gene was also found in a transcriptome of 24-72 hour *S. lamarcki* trochophore development, named PRD-like by Kenny and Shimeld (2012) and renamed PRD-VII herein.

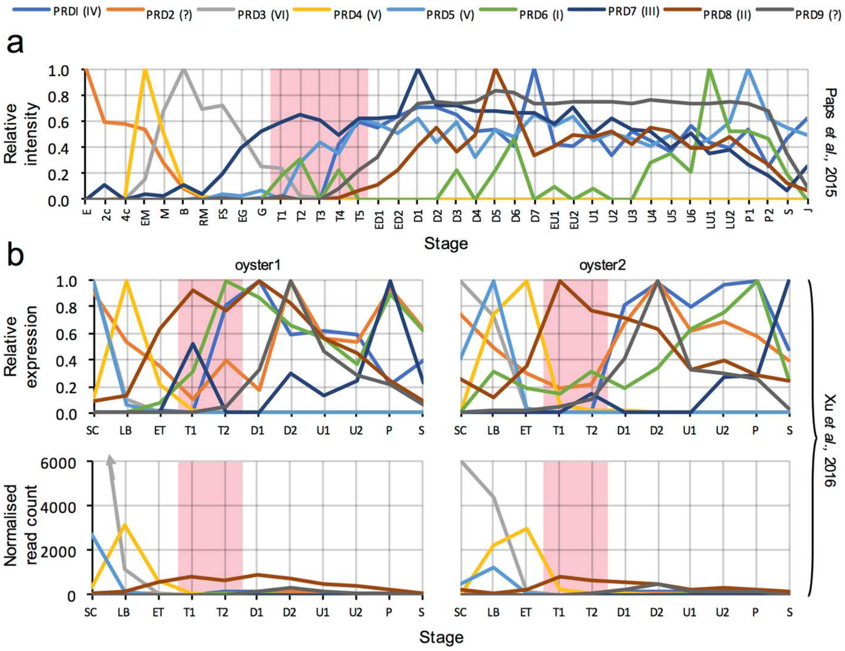


Figure 3.11. Relative developmental expression data for *C. gigas* PRD1-9 genes from **(a)** Paps *et al.* (2015), Figure 3, and **(b)** Xu *et al.* (2016), Supplementary Table S2. For **(a)**, relative colour intensity was measured from the heat map of Paps *et al.* (2015). For **(b)**, read counts were normalised against the total read size of the SC transcriptome for each replicate, and then the relative expression level taken. The pink bar is used to indicate the trochophore stage, after Paps *et al.* and Xu *et al.* In **(a)**: E = egg; 2/4c = 2/4 cell, EM = early morula; M = morula; B = blastula; RM = rotatory movement; FS = free swimming; EG = early gastrula; T1-5 = Trochophore 1-5; ED1-2 = early D-shaped larvae 1-2; D1-7 = D-shaped larvae 1-7; EU1-2 = early umbo 1-2; U1-6 = umbo 1-6; LU1-2 = late umbo 1-2; P1-2 = Pediveliger 1-2; S = Spat; J = juvenile. In **(b)**: SC = single cell (presumed); LB = late blastula (presumed); ET = early trochophore (presumed); T1-2 = trochophore 1-2; D1-2 = D-shaped larvae 1-2; U1-2 = umbo 1-2; P = pediveliger; S = spat.

Data from *C. gigas* development, derived from Paps *et al.* (2015) and Xu *et al.* (2016), are presented in Figure 3.11. The normalised read count from Xu *et al.* (2016), indicates that the PRD-E genes share the broad pattern of the highest recorded expression being in the pre-trochophore stages with the STE genes, but the relative expression graphs in Figure 3.11a and b indicates that many more PRD-E genes peak in post-trochophore development than STE genes do. PRD8 (PRD Clade II), the only *C. gigas* gene with a known *S. lamarcki* (presumed) orthologue, is expressed in the trochophore (according to the data of Xu *et al,* 2016 only) and D-shaped larvae (according to both studies).

3.4.3. Lopx

One homeodomain sequence (comp276818_c0_seq2) from the transcriptome could not be definitively identified nor classified (to a homeobox gene family or class) on the basis of BLAST searches and the large neighbour-joining tree (Figure 3.1). Its closest matches were a *L. gigantea* sequence automatically annotated as *ceh-37*, one of *Caenorhabditis elegans'* three *Otx* paralogues, and another from the scallop *Mizuhopecten yessoensis* (syn. *Patinopecten yessoensis*) annotated as *OTX1-like*. A similar gene was present in the analysis of Paps *et al.* (2015), which they named *Cgi_Hbx_2* but did not place in a class.

This gene was included in the PRD class analysis because of its apparent similarity to *Otx,* but it does not possess a PRD domain. My analysis (Figure 3.6) places the genes on a polytomy with the unusual *Hopx* gene (Chen *et al.*, 2002) but outside the rest of the PRD class. *Hopx* has been previously placed in the PRD class on the basis of its (relative) sequence similarity to *Gsc* and *Pax6,* the identity of residues in the *Hopx* homeodomain which are invariant in PRD-class homeodomains (Q12, L16, F20, P26, W48, and F49, taking into account a one residue insertion present in *Hopx* between residues 23-24) and the presence of a conserved intron between residues 46/47 (Holland, Booth, and Bruford 2007).

The *Otx*-like sequence, while possessing the conserved intron position (found on *S. lamarcki* poma61 genome nodes 1248323 and 1030675) as well as Q12, L16, P26, W48, and F49, does not have F20 (instead it has Y20). Therefore, I do not contend that this sequence belongs to the PRD class, but place it among the other unclassified homeobox genes. Some of the PRD-E genes described in this chapter do not meet these criteria either, but have been previously placed in the PRD class by Paps *et al.* (2015).

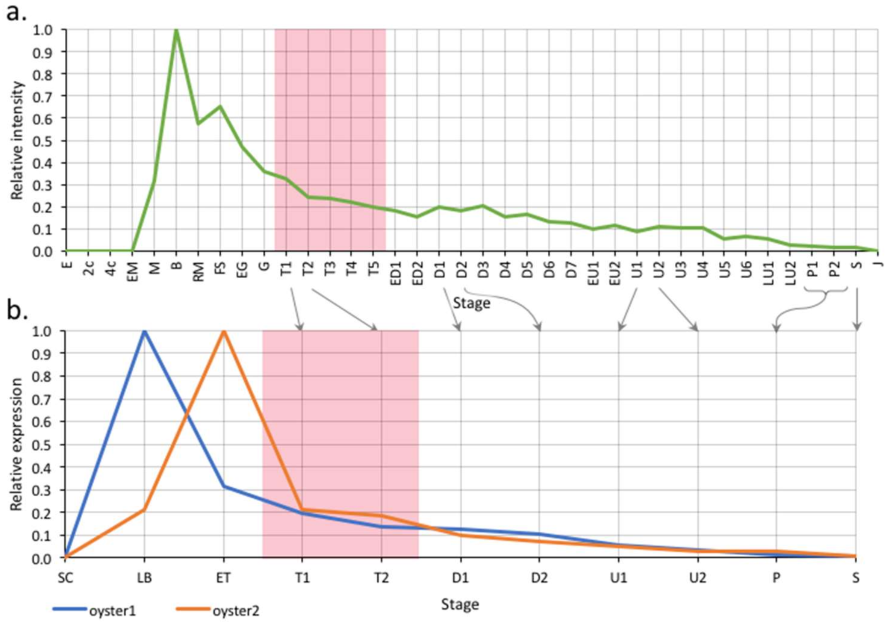


Figure 3.12. Relative developmental expression data for *C. gigas Lopx* (CGI_10016179) from **(a)** Paps *et al.* (2015, Figure 3), and **(b)** Xu *et al.* (2016, Supplementary Table S2). For **(a)**, relative colour intensity was measured from Paps *et al.*'s heat map. For b., read counts were normalised against the total read size of the SC transcriptome for each replicate, and then the relative expression level taken. The pink bar is used to indicate the trochophore stage, after Paps *et al.* (2015) and Xu *et al.* (2016). In **(a)**: E = egg; 2/4c = 2/4 cell, EM = early morula; M = morula; B = blastula; RM = rotatory movement; FS = free swimming; EG = early gastrula; T1-5 = Trochophore 1-5; ED1-2 = early D-shaped larvae 1-2; D1-7 = D-shaped larvae 1-7; EU1-2 = early umbo 1-2; U1-6 = umbo 1-6; LU1-2 = late umbo 1-2; P1-2 = Pediveliger 1-2; S = Spat; J = juvenile. In **(b)**: SC = single cell (presumed); LB = late blastula (presumed); ET = early trochophore (presumed); T1-2 = trochophore 1-2; D1-2 = D-shaped larvae 1-2; U1-2 = umbo 1-2; P = pediveliger; S = spat. Where directly equivalent stages exist between **(a)** and **(b)**, they are indicated with grey arrows.

Related sequences were collected from among the Lophotrochozoa, including from *C. teleta*, *P. dumerilii, L. anatina, C. gigas, M. yessoensis, L. gigantea, Biomphalaria glabrata* (a ram's horn snail) and *Octopus bimaculoides.* I conclude on the basis of the data in Figure 3.6 and Figure 3.7 that this group of genes represents a hitherto undescribed gene family, which I name *Lopx* (LOPhotrochozoan only homeoboX).

Insofar as the un-replicated read count data from the *S. lamarcki* regenerative transcriptome can be taken as a measure of expression, *Lopx* appears to be constitutively but only weakly expressed in the adult operculum (17 reads), increases in early regeneration (49 reads), and falls back to almost pre-regenerative levels in late regeneration (21 reads). Paps *et al.*'s (2015, Figure 3) and Xu *et al.*'s (2016, Supplementary Table S2) data from *C. gigas* embryogenesis (Figure 3.12) indicate that *Lopx* expression spikes in the blastula, is immediately reduced in the rotatory movement stage, and then decreases approximately linearly from the free-swimming pre-gastrula on, though one of Xu *et al.*'s replicates places the spike in the early trochophore. It does not appear to be expressed in the juvenile.

Beyond these rudimentary data, *Lopx* is currently a complete enigma. Its position outwith the established homeobox gene classes makes it unique among the novel homeobox genes described herein. *In vivo* investigation of the spatiotemporal expression patterning and function of *Lopx* should be a priority in any ongoing efforts to understand the roles of new and taxonomically-restricted homeobox genes in lophotrochozoan development and regeneration.

3.4.4. Antp and the Hox cluster

The Hox-like gene found in the transcriptomes is likely to be the *S. lamarcki Antp* orthologue. It forms part of the group of *Hox4/5/6-8* genes that do not resolve into individual orthology groups in Figure 3.2 (or in comparable phylogenetic analyses by others), and is excluded from the anterior Hox genes, *Lox2* and *Lox5*, the posterior Hox clade, and the ParaHox gene clades by relatively strong support values (lab, 648/.964/1.0; pb, 230/.452/.530; Hox3, 358/.339/.965; Lox2, 513/.758/.999; Lox5, 864/.666/.894; posterior Hox genes, 576/.517/.994; Cdx, –/.892/.893; Gsx, –/.677/.945; Xlox, 673/.726/.597). Orthologues were identified for all expected annelid Hox and ParaHox genes (Fröbius, Matus, and Seaver 2008; Kulakova, Cook, and Andreeva 2008) except *Antp* and *Post1* (*i.e. lab, pb, Hox3, Dfd, Scr, Lox5, Lox4, Lox2*, & *Post2*; *Gsx, Xlox,* & *Cdx*). Although it is not impossible for it to be a divergent *Dfd/Scr/Lox5/Lox4/Lox2* paralogue, it is simpler to posit instead that the missing *Antp* orthologue has drastically diverged. As is evident in Figure 3.3, the divergence is considerable; even within the highly-conserved homeodomain, *S. lamarcki Antp* varies in six residues that are otherwise mostly invariant in the entire medial Hox group (*Hox6-8*). It has diverged outside the homeodomain to the point that little similarity is obvious with other Hox genes.

This divergent *Antp* sequence is the only expressed Hox sequence yet detected in *S. lamarcki*, including the 24-72 hour trochophore transcriptome (Kenny and Shimeld 2012). Expression of a broad spectrum of Hox genes has been previously reported in other annelids, including the embryogenesis of *Chaetopterus* (Irvine and Martindale 2000; Peterson *et al.*, 2000), nereids (M. Kulakova *et al.*, 2007; Steinmetz *et al.*, 2011), *C. teleta* (Fröbius, Matus, and Seaver 2008) and *H. robusta* (Kourakis and Martindale 2001; Gharbaran, Aisemberg, and Alvarado 2012; Gharbaran, Alvarado, and Aisemberg 2014), and in caudal regeneration in nereids (Novikova *et al.*, 2013; Pfeifer, Dorresteijn, and Fröbius 2012) and *C. teleta* (de Jong and Seaver 2016), as well as being a well-established part of expected regenerative transcriptomic activity (K. C. Wang, Helms, and Chang 2009; Novikova *et al.*, 2016).

The absence of Hox expression from any previously-surveyed *S. lamarcki* context except the divergent *Antp* in regeneration is therefore surprising. Apart from *Antp* and the loss of *Post1* — potentially not unusual for Sedentaria (Barucca, Canapa, and Biscotti 2016) — the *S. lamarcki* Hox complement are not noticeably atypical, derived, or divergent in their coding sequences (*c.f.* Appendix 3.4b), indicating (insofar as is possible to tell from this metric alone) that they are presumably still used in some context and that *S. lamarcki* has not so completely dispensed or modified its Hox gene deployment that they are no longer constrained.

It is an intriguing possibility that aspects of *S. lamarcki* biology like their unusually poor capacity for caudal regeneration compared to many other annelids (Bely, Zattara, and Sikes 2014) and their blastema-less opercular regeneration (Szabó and Ferrier 2014) could be in some way related to their unusual Hox deployment. Continuing to try to detect their expression, confirming their absence from embryogenesis and larval development, and profiling *Antp* involvement in opercular regeneration are important avenues in unravelling these puzzling phenomena.

3.4.5. *Spiro-Nk*

The Nk1-7/Msx/Lbx/Tlx phylogenetic analysis (Figure 3.8) did not place all the unusual similar-to-Nk genes of *S. lamarcki*, *C. gigas, L. anatina,* and *L. gigantea* in a single clade, although the latter two were placed together. These genes were dubbed *Lilo-Nk* (*Lingula-Lottia Nk*). If they do share a common origin with the sequences from *S. lamarcki* (dubbed *Spiro-Nk*) and *C. gigas* (which retains its name from Paps *et al.* 2015, *NKL*), it is now invisible in the present data. The discovery of new similar-to-Nk sequences from species not surveyed herein could draw these genes into a single clade. However, given the apparent flexibility in *Nk1-7* paralogue number observed among the Sedentaria (*c.f.* Table 3.8), it is likely that *Spiro-Nk* is a divergent, cryptic paralogue of a canonical Nk gene (the topology of the Bayesian tree in Figure 3.8 suggesting *Nk3*).

Although the un-replicated read counts of the opercular regenerative transcriptomes are not particularly reliable as a measure of expression levels, it does appear as if the expression of *Spiro-Nk* is constitutive (34 reads in the mature operculum) and increases in regeneration (138 reads in early and 99 reads in late regeneration).

3.4.6. The dynamics of non-canonical homeobox gene gain and loss

A cladogram of the Bilateria, on which has been mapped the minimum necessary gain and loss events to explain the distribution of (mostly) non-canonical genes in the taxa surveyed herein, is presented in Figure 3.13.

In trying to interpret the topology of the gene gain and loss events, we must be aware of several caveats. The first is that the terminal branches do not consistently represent the same taxonomic levels: in the Mollusca, for example, the data from *P. vulgata* and *L. gigantea* have been collapsed into the Gastropoda, and likewise with *C. gigas* and *P. fucata* to the Bivalvia. The Brachiopoda are only represented by a single species. The Platyhelminthes and Rotifera were not surveyed herein and rely on the synthesis of information from elsewhere (Paps *et al.*, 2015 and Morino, Hashimoto, and Wada 2017).

The cladogram also necessarily assumes that the surveys of the genomes were exhaustive and that the PRD, TALE and Nk clades used herein all represent orthology groups, not just the ones in which orthology has been established (Post1, Lopx, & TALE-IV). Gene divergence, in which sequences diverge so extensively that any signal of specific orthology/paralogy is lost, is now invisible and therefore all instances are recorded as gene loss/gain events.

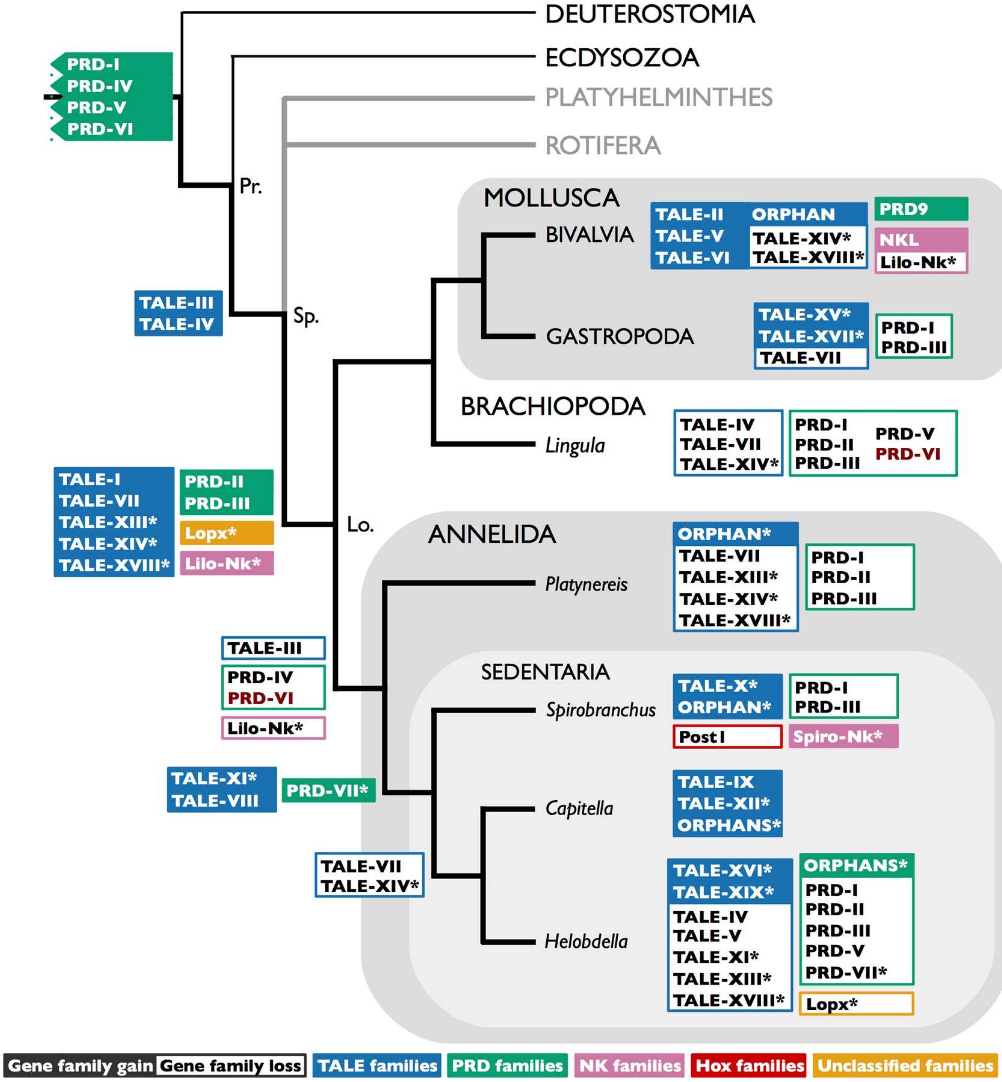


Figure 3.13. Summary of the minimum gene family gain/loss events necessary to explain the distribution of orthology groups observed. Annelid-centric cladogram of the Bilateria, summarising the minimum gene family gain and loss events necessary to explain the pattern of gene presence and absence in the species surveyed, for TALE class genes (blue), PRD class genes (green), Nkx genes (pink), Hox genes (red), and unclassified genes (orange). White text on a coloured background indicates a putative gene gain event; black text on a white background with a coloured border indicates a putative gene loss event. The only gain or loss event influenced by the internal topology of the Lophotrochozoa is marked in dark red (*i.e.* PRD-VI). New gene families suggested herein are marked with an asterisk. Clades not sampled in these analyses are marked with grey lines. Clades from which sequences were included but not extensively surveyed in my work and with severely limited taxonomic sampling are marked by a thin black line. The Protostomia, Spiralia, and Lophotrochozoa clade nodes are marked Pr., Sp., and Lo. respectively. The topology of the cladogram is adapted from data in Weigert *et al.* (2014) and Luo *et al.* (2018), and the position of some gain/loss events from Paps *et al.* (2015). Clades not sampled here or in Paps *et al.* (2015) (including Phoronida, Nemertea, Entoprocta and Gastrotricha) have been omitted to aid comparison with Paps *et al.* (2015) Figure 4. For collapsed clades with more than one sampled species (*i.e.* Bivalvia and Gastropoda), gene

gains are marked if they have been found in any of the species in that group, but gene losses marked only if they have not been identified in any. Canonical or previously described families are only marked for *S. lamarcki.* Taken from Barton-Owen, Szabó, Somorjai, & Ferrier (2018).

With these caveats in mind, the most striking pattern visible in the tree is the major concentration of gene gain and particularly loss events in the terminal branches; that is, it appears as though these events are not evenly distributed across phyletic levels. If it was indeed the case that genes with ancient lophotrochozoan or spiralian origins were being lost with greater frequency since the most recently sampled nodes than in earlier animals, it might require special explanation. However, it seems more likely that the impression given by Figure 3.13 is biased by one of the following factors.

The first is that this impression is exacerbated by the caveats listed above. Some clades — particularly TALE clades VII and XIV, which each contain only one *Spirobranchus* and one mollusc sequence, and in neither of which is vested the highest confidence (Table 3.6) — are frequently lost, and if these are not orthologue groups, the impression of terminal branch gene loss would be reduced. That the entire Bivalvia and Gastropoda are collapsed also means that these terminal branches themselves represent ancestral clades rather than species, particularly with regards to gene loss (gene gain was marked for any species within, but gene loss only for all). Rather curiously, there are no gene gain events that are synapomorphic to either of the well-represented phyla, the Mollusca and Annelida.

The density of taxa sampled herein is the highest (at least within the Lophotrochozoa) of any study thus far, but could nonetheless be too sparse to reliably determine the age of gene losses. For instance, the impression of gene loss is disproportionately contributed to by *L. anatina, P. dumerilii,* and *H. robusta,* which seem to have undergone extensive gene loss. Broader taxon sampling could easily indicate that these species are anomalously prone to loss.

Another factor is that, although it is preferable from Occam's point of view to posit a single later gene gain than an earlier gene gain followed by one or more losses, it is not necessarily true that this is what happened. Some of the more recent taxonomically-restricted clades — *e.g.* TALE clades II, V, VI, XV or XVII — could have been acquired earlier and lost in multiple clades since then, which, while multiplying the number of loss events, would for the most part move them farther up the cladogram. Frequent loss events are clearly common among these non-canonical genes.

The predominance of gene richness of the TALE/PRD non-canonical expansions belongs to the Lophotrochozoa. Although the lophotrochozoan focus of this study has undoubtedly skewed this impression, previous studies (Paps *et al.*, 2015; Morino, Hashimoto, and Wada 2017) found substantially fewer genes in the non-lophotrochozoan spiralian genomes (see Appendix 3.5). Determining whether a meaningful link between lophotrochozoan evolutionary developmental biology and TALE/PRD expansions can be drawn will require a great deal more of information about the evolution and function of the latter.

3.4.7. The roles and evolutionary importance of new homeobox genes

In early development

Given the functional evidence presented by Morino *et al.* (2017) (*e.g. NfSPILE-D*, knock-down of which causes loss of the apical structures in *N. fuscoviridis*), it is likely that some of these non-canonical genes have adopted important or essential ontogenic roles in early development. Paps *et al.* (2015) report that the majority of new homeobox genes (STEs, PRD-Es, and others) from *C. gigas* also peak in expression during early development (22 of 33 genes, *i.e.* 67%) — almost the inverse of the pattern they report for canonical homeobox genes (27 of 101 genes, 27%). This potential bias towards early expression (and the possible bifurcation between STE/early development and PRD-E/late developmental) is interesting and not easily explainable.

Homeobox genes and transcription factors in general are usually found to be deployed less in early (Levin *et al.*, 2012; Zalts and Yanai 2017) and late (Levin *et al.*, 2012; Piasecka *et al.*, 2013; Zalts and Yanai 2017) developmental transcriptomes than in mid-development. Schep and Adryan (2013) report a consistent under-representation of homeobox genes amongst other transcription factors in chordate and ecdysozoan early development, indicating that this period might even be somehow adverse to homeobox gene deployment. Alternatively, its flexibility (Heyn *et al.*, 2014) could allow the production of evolutionary novelty via the introduction of new homeobox gene regulatory networks and functions.

Understanding the diversity of the early developmental roles of non-canonical homeobox genes and how they relate to the apparently extreme patterns of gene gain, divergence, and loss, could provide not only indispensable information about the evo-devo of the Spiralia but also an insight into the evolution of homeobox GRNs.

In regeneration

A surprising diversity of these non-canonical and difficult-to-classify homeobox genes was also found expressed in regeneration, a post-embryonic ontological process (see General Introduction). The read counts of these genes are presented in Table 3.12. Most are more highly expressed (insofar as can be gleaned from these data, which are very limited) in one or both of the regenerative stages than they are in mature tissue, with the exceptions of TALE-XIII A and the two TALE-X genes. All the genes except PRD-V are expressed at some level in the mature operculum tissue.

The operculum is an evolutionarily novel modification of a single radiole from the sabellid radiolar crown, restricted to the Serpulidae but not universal amongst them (Bok *et al.*, 2017). Based on other studies of novel morphological traits (*e.g.* Babonis, Martindale, and Ryan 2016; Hilgers *et al.*, 2018; others reviewed by McLysaght and Guerzoni 2015), it is reasonable to hypothesise that novel/taxonomically-restricted genes were involved in the evolution of the operculum; it might even be the case that the new and divergent genes classified herein are among these. All of the STE genes detected in the regenerating operculum have also either duplicated in *Spirobranchus* since the *Spirobranchus*/*Capitella* split (TALE clades I & XIII), or are *Spirobranchus*-restricted (TALE-X), a potential indicator of interesting recent roles in evolutionary change. Discovering and comparing the genetic bases of opercular development and regeneration — particularly with regard to patterning or modification of radiolar developmental programmes — could offer a potentially fascinating insight into the relationship between the two ontogenic processes.

Table 3.12. Adjusted read counts of unusual genes from the transcriptomes of *S. lamarcki* operculum regeneration. dpa = days post amputation. 0dpa is from mature operculum tissue. Cells are coloured based on their fold change from the 0dpa count: light red, ≥5x; dark red, ≥10x.

Gene	Family	Adjusted read counts			
		0dpa	2dpa	6dpa	total
Antp	ANTP-HOXL	34	49	109	192
Spiro-Nk	ANTP-NKL	34	138	99	271
Lopx	Unclassified	17	49	21	87
PRD-VIII	PRD	0	19	0	19
TALE-I A	TALE	3	11	52	66
TALE-I B	TALE	14	4	95	113
TALE-XIII A	TALE	13	0	6	19
TALE-XIII B	TALE	6	0	9	15
TALE-X A	TALE	3306	1743	725	5774
TALE-X B	TALE	3	3	1	7

3.4.8. Other homeobox expansions

The general pattern of homeobox gene evolution is conservative. It is usually possible to phylogenetically detect gene orthology across hundreds of millions of years based purely on the 60-63 residue-long sequence of the homeodomain, and in the case of many genes, it is possible to trace the orthology group back to the bilaterian ancestor, in which case it is referred to as a gene family (Holland 2012). In some cases, homology is detectable beyond this point; for example, the TALE class was present in the common ancestor of plants and animals (Burglin 1997).

However, since the evolutionarily important homeobox expansions that occurred in the cniderian-bilaterian and bilaterian ancestors (Ryan *et al.*, 2006; Putnam *et al.*, 2007), which produced and populated most of the bilaterian homeobox classes, homeobox evolution seems to have been restrained, probably by various evolutionary forces relating to their roles in developmental regulatory control (see General Introduction). However, they have not been static, and small-scale duplications (*c.f.* Table 3.2, Chapter 4), large clustered paralogue arrays (e.g. the Obox, Rux, and Dux loci; Rajkovic et al. 2002; MacLean and Wilkinson 2010; Leidenroth and Hewitt 2010; Zhong and Holland 2011), whole genome duplication and preferential homeobox ohnologue retention (Huminiecki and

Heldin 2010), radical apomorphy of developmental programme (Ruvkun and Hobert 1998; Aboobaker and Blaxter 2003; Edvardsen *et al.*, 2005), and heterogenous paralogue and family loss (Butts, Holland, and Ferrier 2010; Zhong and Holland 2011; Mendivil Ramos, Barker, and Ferrier 2012) have all contributed to the variations on the ancestral homeobox complement present in the genomes of modern species.

Consequently, clade-specific homeobox genes at all taxonomic levels are commonplace, as is sequence divergence extreme enough that genes cannot even be placed within a family, or even a class. Within this context, *Lopx, Spiro-* and *Lilo-Nk,* and even the divergent *Antp* (and *Post1* loss in the context of annelid Hox retention — Barucca, Canapa, and Biscotti 2016) are not unusual. However, the TALE (and to a lesser extent, the PRD) expansion first described in Paps *et al.* (2015) and elaborated in Morino *et al.* (2017) and the present study are very uncommon, if not unprecedented. Perhaps the closest are the Aprd genes in cephalochordates, although my analysis (Figure 3.4, section 3.4.2) suggests a link with the PRD-E genes. Their scope, in terms of the number of genes, their unusual diversity and flexibility, and the phyletic, developmental, and morphological diversity of the phyla involved (particularly annelids and molluscs — Giribet 2008) all distinguish these instances from previously reported homeobox expansions.

3.4.9. Future work

3.4.9.1. STE and PRD-E surveying

Taxon sampling

Ongoing efforts to survey the novel homeobox genes in the TALE and PRD classes will benefit as taxon sampling of spiralian genomes improves. However, as with much of bioinformatics, analysis of the already available data is very far from complete, and genomes at interesting points in the Spiralia await surveying.

Within Annelida, these include *Hermodice carunculata* (Mehr *et al.*, 2015), a member of the un-sampled Amphinomediae, and the already-performed homeobox survey of the earthworm *Eisenia fetida* (Zwarycz *et al.*, 2016). *E. fetida,* as a non-hirudinean member of the Clitellata, is well-positioned to elucidate the origin of the unusual TALE clades of *H. robusta*, and integrating data from *H. carunculata* and *E. fetida* can improve our understanding of annelid TALE diversity. Within Mollusca, the genome of *Octopus bimaculoides* (Albertin *et al.*, 2015) also allows an important expansion of the survey into an un-sampled clade (Cephalopoda).

Further sampling of the non-lophotrochozoan Spiralia is also likely to be instructive. Previous surveys have retrieved sequences from the trematode platyhelminthes *Clonorchis sinensis* and *Schistosoma mansoni* (Paps *et al.*, 2015), the cestode platyhelminthes *Echinococcus granulosus* (Paps *et al.*, 2015) and *Echinococcus multilocularis* (Morino, Hashimoto, and Wada 2017), and the rotifer *Adineta vaga* (Morino, Hashimoto, and Wada 2017), but the picture of the TALE clades in these groups is fragmentary, and as yet the picture of diverse and robust gene-family-like clades emerging from the lophotrochozoan-only sequence phylogenies in this analysis is not shared by non-lophotrochozoans. Ensuring the complete saturation of searches in the above genomes and improving sampling including in the turbellarian platyhelminthes (Wasik *et al.*, 2015) and the recently released nemertean and phoronid genomes (Luo *et al.*, 2018), will help in this effort. Given the degree of sophistication as a model system offered by *Schmidtea mediterranea* (see below), the newly-released genome assembly would also be a worthwhile inclusion (Grohme *et al.*, 2018).

The taxonomic coverage of transcriptomes is greater than of complete nuclear genomes, and surveying the available bryozoan (Wong et al. 2014), chaetognath (Marlétaz *et al.*, 2008), cycliophoran (Neves and Strempel 2016) and rotifer (Hanson *et al.*, 2013) transcriptomes could also be worthwhile. Searching any available spiralian transcriptomic time-courses (*e.g.* the developmental transcriptomic time-courses from Xu *et al.*, 2016; Levin *et al.*, 2016) and sequence read archives would also be useful for identifying priority targets in an *ex silico* candidate gene approach (see section 3.4.6.2).

Automation

A substantial amount of work is necessary to get a comprehensive grip on the diversity of divergent, orphan and cryptic homeobox genes in the Spiralia, even if this only involved the analyses of existing genomes and transcriptomes suggested above (and recursive re-analyses of old genomes based on new finds). Continuing to explore new genetic data for their unusual homeobox gene complements will have diminishing returns in terms of important novel data unless significant time reductions can be made.

The obvious bottleneck in the pipeline from genome search to finished gene classification is human involvement, pointing towards the development of automative tools and pipelines as the solution. No single tool would in isolation provide major time savings, but the following suggestions are tools which could be scripted in Python with relative ease:

Given the XML format results of BLAST searches of a genome using the complete query pool, merge redundant hits and present for inspection only unique hits aligned against their best match from the queries. Easily achievable using BioPython's (Cock *et al.*, 2009) NCBIXML module.

Given a set of sequence names, hit locations, and translation frames derived from BLAST searches, retrieve the sequence, translate it in the desired frame, and trim it to the apparent ORF containing the target homeodomain. Moderately easily achievable using plain Python or BioPython's SeqIO module.

Given a set of sequences (retrieved by the above tools), categorise their contents to allow easy inspection, automatic filtering of spurious hits, and putative classification. This would probably need to involve several filters:

- Compare to sequences in the query pool to eliminate exact matches to previously identified sequences from the same genome (for re-surveying genomes as the query pool expands). Achievable in a variety of ways, including Python's Levenshtein distance module or BLAST.
- Perform a local CDD search (Marchler-Bauer *et al.*, 2011) to determine if the sequence contains a recognisable homeodomain. Using a custom reference database containing only the homeodomain pattern would presumably cut down on unnecessary computation.
- Retrieve the 60 or 63 residue length of the homeodomain using the CDD search results, bearing in mind that the CDD frequently underestimates the size of the homeodomain.

- Putatively classify homeodomains using a machine learning approach. In Python, the PCP-ML package/module (Eickholt and Wang 2014) could be used to convert previously classified homeodomain sequences into the non-categorical data necessary for the supervised learning techniques implemented in Scikit-learn (Pedregosa *et al.*, 2011). It should be possible to easily distinguish TALEs from non-TALEs and canonical TALE families from non-family clades. For the well-supported, member-rich TALE families and clades, it should be possible to putatively predict gene identity with a good degree of accuracy.

These tools would alleviate a significant portion of the human time necessary to survey divergent homeodomains from new genomes. If the query pool was a part of a broader, well-maintained database, it would be relatively simple to automate the generation of homeodomain alignments suitable for the tree-building pipeline. Beyond that, the individual processes detailed above could plausibly be chained together using a bioinformatics workflow tool like Galaxy (Goecks, Nekrutenko, and Taylor 2010) Taverna (Oinn *et al.*, 2004), or dedicated bioinformatics scripting languages like Bpipe (Sadedin, Pope, and Oshlack 2012). Beyond even this, a well-designed search pipeline could be generalised to comprehensively survey the entire homeobox content of genomes and transcriptomes.

The tools and pipelines would require a substantial degree of flexibility to cope with the 'messiness' in the data which a human analyst can deal with without problem — for example, the intron often found in the last third of the homeodomain. Resilience to this kind of variability would likely be a substantial hurdle in making robust tools.

Ultimately, however, the curation of these data will always benefit from human input. It would be difficult to develop tools sophisticated enough that they could have recognised, for example, the mollusc PRD Clade VI homeodomains (which are missing 28 residues), or that the TALE-IV members frequently have two homeodomains or the degraded remnants of two, but not also allow a uselessly large proportion of spurious data through. Tools that efficiently integrate relevant data (*e.g.* from multiple BLAST searches) and present them to the researcher for manual curation could therefore be more useful than complete but unreliable pipelines.

In silico query generation

If the time burden of inspecting the results of a very large query pool was successfully minimised using the tools outlined above, a computational approach could also be enlisted to help overcome another current limitation of the search process: the query pool. It seems likely (see above) that highly divergent homeodomains exist outside of the ‘space’ covered by searches with previously identified ones, but within the ‘space’ of recognisable homeodomains.

To escape this space, hypothetical homeodomains could be generated *de novo* to use as queries. Various ways of achieving this are possible, but the easiest would involve using the probability distribution of residues for each position in known homeodomains to semi-randomly generate new ones (easily achievable in Python using the NumPy package — specifically numpy.random.choice). Queries generated this way would not actually cover all recognisable homeodomain-space, though a (customisable) degree of latitude for originality could be given to the algorithm to allow the introduction of novelty.

Beyond the homeodomain

Phylogenetic proximity based on homeodomain sequence alone is not considered sufficient evidence of orthology. Consequently, further work is necessary to satisfactorily establish which TALE clades represent real orthology groups beyond the evidence for at least lophotrochozoan TALE-IV orthology. Given the variable quality of the available genomes, syntenic analyses and even construction of more complete gene models is often difficult. It should nonetheless be possible to detect conserved motifs outside but close to the homeodomain, as was the case in TALE-IV. There is a slight chance that some of the STE or PRD-E genes could be clustered, particularly the SPILE genes. However, given the extreme evolutionary flexibility of the genes, it would be surprising if strong patterns were observable between taxa.

With suitably chosen pairs of sequences, it might be possible to detect the selective conditions under which some STE or PRD-E genes are evolving, perhaps distinguishing between positive selection, negative selection, and genetic drift (reviewed by Booker, Jackson, and Keightley 2017). These data could help elucidate the unusual evolutionary dynamics of the TALE expansion.

3.4.9.2. Antp and the Hox cluster

The *S. lamarcki* genome assembly is currently too fragmented to determine whether the Hox genes are clustered; the sequences classified in this analysis are all on separate genomic nodes, detailed in Table 3.13. For comparison, the intact portion of the *C. teleta* Hox cluster (*lab* to *Lox4*) spans about 250,000 base pairs (Fröbius, Matus, and Seaver 2008, Figure 2), more than 10x the sum length of the *S. lamarcki* nodes.

Table 3.13. Locations of the Hox genes in the *S. lamarcki* genome assembly (poma61).

Gene	Node ID	Node Length (bp)
pb	1305075	3358
lab	104376	2448
Hox3	2185155	5111
Dfd	407738	625
Scr	2197135	1034
Lox5	6606774	437
Antp	1021697	6474
Lox4	1015850	502
Lox2	4400779	330
Post2	1345801	2130

However, the advent of known *S. lamarcki* Hox nucleotide sequences would help the design of specific sequence primers for genomic walks around the loci. An existing *S. lamarcki* phage library with an average insert size of 16-17 kilobases was insufficient for reconstructing a complete ParaHox cluster (Hui 2008, although these efforts were prevented from working by repeat sequences), meaning it might be necessary to construct a genomic library with larger inserts. Genomic walking has previously been used to establish the presence (*e.g.* Garcia-Fernàndez and Holland 1994) and dissolution (*e.g.* Seo *et al.*, 2004) of Hox clustering.

3.4.9.3. *Ex silico* work

As illustrated by the work of Morino *et al.* (2017), data gathered *in vivo* are extremely important for understanding the roles of the homeobox genes. Measures of relative expression level from transcriptomes — or even from quantitative real-time PCR — fail to capture indispensable information about the spatiotemporal expression patterning, an important aspect of homeobox paralogue evolution. Whole mount *in situ* hybridisation is necessary to visualise these patterns. Beyond that, morpholino (or RNAi) and overexpression manipulations like those performed by Morino *et al.* (2017) are necessary to unravel the functional roles of these genes.

The scale of the homeobox expansion in *S. lamarcki* alone (~30 genes) means that even qPCR assays of expression would be time-consuming. Therefore, a candidate gene approach to prioritise targets for cloning and WMISH will be necessary. The obvious first choices would be the genes found to be expressed in regeneration (TALE clades I A & B, X A & B, XIII A & B, PRD-V, *Lopx* and *Spiro-Nk*) and development (TALE-XVIII and PRD-VII). Sequences from the same clades as the SPILE genes of Morino *et al.* (2017) (*i.e.* TALE clades IV, VIII and XVIII) would also allow a comparison to be developed between the closely related *Spirobranchus* species and the more distant *Spirobranchus*/*Nipponacmea* comparison.

It would be unwise to try to understand such a heterogenous set of genes with such a narrow sampling of taxa. Therefore, WMISH and functional manipulations in other systems are desirable; *C. teleta* and *C. gigas* are prime candidates. It is unfortunate that *P. dumerilii,* the best developed lophotrochozoan model for functional manipulations (Backfisch *et al.*, 2013, 2014; Simakov *et al.*, 2013; Bannister *et al.*, 2014; Zantke *et al.*, 2014), should apparently have undergone such an atypically broad STE loss. RNA interference has previously been performed in *Pinctada* spp. (Yan *et al.*, 2014; Zhao *et al.*, 2016), *C. gigas* (Huvet *et al.*, 2012), and planarians including *Schmidtea mediterranea*, (Alvarado and Newmark 1999; Oviedo *et al.*, 2008a; Rouhana *et al.*, 2013), the best-established planarian model (Oviedo *et al.*, 2008b) and a promising non-lophotrochozoan model, despite not yet being surveyed (see above).

The non-canonical genes described herein are not the only priorities for further *ex silico* investigation; the expression and function of *Antp* in operculum regeneration is particularly interesting (see section 3.4.4). The profiling of the homeodomain content of the opercular transcriptomes also offers a starting-point for candidate gene investigations of various processes happening during regeneration; for example, the possibly conserved roles of *Msx* and *Dlxa/Dlxb* in the biomineralisation of the opercular plate, given their role in mammal odontogenesis (Lézot *et al.*, 2000). The relationship of the deeply conserved roles of *Msx* in regeneration and blastema formation (Mannini *et al.*, 2008) to the blastema-less regeneration of *S. lamarcki* also warrants further investigation. Another possible candidate for investigation would be the relationship between *Pax4/6* (*c.f.* Halder, Callaerts, and Gehring 1995; Gehring and Ikeo 1999; Manousaki Tereza *et al.*, 2011) and the regeneration of the radiolar eyes likely present on the rim of the *S. lamarcki* opercular cup (Bok *et al.*, 2017). The potential and conserved regenerative roles of canonical homeobox families are discussed further in Chapter 6.

3.5. Conclusions

Homeobox genes are instrumental in the orchestration of a huge variety of developmental mechanisms, including in regeneration and biomineralisation. The operculum regeneration transcriptomes contain a broad selection of canonical ANTP-, CUT-, LIM-, POU-, PRD-, SINE- and TALE-class genes, many of them as multiple paralogues. Additionally, I report the expression of a surprising number of novel homeobox genes, including a previously unidentified homeobox gene family (*Lopx*), members of rapid taxonomically-restricted homeobox expansions with cryptic paralogy (TALE clades IA & B, XA & B, XIIIA & B, and PRD-V) and highly divergent canonical homeobox genes (*Antp* and *Spiro-Nk*). This diversity of divergent homeobox genes, considered in combination with the absence of expression of some expected gene families (*i.e.* other *Hox* genes), indicates that *S. lamarcki* is unusual compared to previous surveys of regeneration. Further surveys of expression in new regenerative models are necessary to determine whether the *S. lamarcki* operculum is an isolated example of divergence or represents a previously hidden but widespread diversity of homeobox deployment in regeneration.

The historical study of the deep homology of homeobox gene families, and the relations between ancient sequence, syntenic, regulatory, and functional conservation, have been of cardinal importance to the understanding of animal ontology and evolution produced by the field of Evo-Devo. However, the Spiralia seem to possess an unprecedented diversity of relatively unconstrained and taxonomically-restricted homeobox genes in addition to the expected complement of bilaterian homeobox families. Understanding what these genes do, why they are gained and lost so readily, and why they diverge so quickly in the meantime, could help elucidate why the Spiralia are so phyletically and morphologically diverse (Giribet 2008).

4. Homeobox genes in *Branchiostoma lanceolatum* regeneration

4.1. Introduction

Regeneration in vertebrates proceeds *via* the production of a blastema, a collection of populations of undifferentiated, proliferative cells that accumulate in a mass underlying the wound epithelium (see section 1.1.2.3). These masses are contributed to by multipotent but lineage-restricted stem cells deriving from sources including dedifferentiated mature myotubes and satellite cells. The blastema grows and eventually patterns and differentiates into the replacement tissue. Homeobox genes are known to be important to many of these processes, from wound healing (section 1.2.4.2) to dedifferentiation and satellite cell recruitment, maintenance and proliferation (section 1.2.4.3), to axial patterning of the replacement tissues (section 1.2.4.4).

Amphioxus belong to the basal-most chordate clade (cephalochordates), which split from the vertebrate + tunicate lineage approximately 550-578 MYA (section 1.4.2). Since that time, they have undergone relatively conservative evolution, and possess the best extant representative of the genome of the ancestral chordate (section 1.4.3); in particular, they have not lost any homeobox gene families, though they have gained several homeobox gene duplications (section 1.4.4). Amphioxus can regenerate their post-anal tails as adults in a process that closely resembles vertebrate structure regeneration (section 1.4.1; Somorjai *et al.*, 2012; Somorjai, Escrivà, and Garcia-Fernàndez 2012), a capability that is relatively rare in chordates. As such, they are of evolutionary developmental and medical interest.

Two homeobox genes, *Msx* and *Pax3/7*, have previously been found to be expressed in amphioxus tail regeneration (Somorjai *et al.*, 2012), consistent with previously described roles in vertebrate regeneration. *Msx* has conserved roles in vertebrates in inducing and maintaining dedifferentiation of multinucleate mature muscle fibres underlying the wound site into mononucleate cells (Kumar *et al.*, 2004; Odelberg, Kollhoff, & Keating, 2000) that can be major contributors to the blastema (Echeverri, Clarke, and Tanaka 2001; Echeverri and Tanaka 2002). *Pax3/7* paralogues in vertebrates are expressed in a population of satellite cells that derive from muscle progenitor cells and reside quiescently in mature muscle until injury reactivates these cells to proliferate and migrate to the blastema. In some species these cells are indispensable for regeneration (Lepper, Conway, and Fan 2009; Sambasivan *et al.*, 2011) while others rely on dedifferentiation (Sandoval-Guzmán *et al.*, 2014).

The known roles of *Msx* and *Pax3/7* are primarily in the recruitment and maintenance of a population of stem-like cells from adult tissues with which the lost tissue can be replaced – a regenerative process without a direct analogue in development. In contrast, the processes of patterning these replacement tissues in later regeneration are analogous to the ontogenic processes that produced the original. However, the extent to which regeneration is actually a direct recapitulation of these processes is a topic of ongoing research. Because of their conserved roles in controlling cellular behaviour and structural development, studying the deployment of homeobox genes is an ideal tool to illuminate the extent (or otherwise) of homology between these two processes.

4.1.1. Aims

Drs Somorjai and Dailey produced transcriptomes of the uncut/mature and Stage 2 (*sensu* Somorjai *et al.*, 2012)/14 days post-amputation regenerating tissue, including the blastema. I aimed to retrieve the homeobox gene content of these transcriptomes, and identify the genes present in the mature and regenerating samples.

4.2. Methods

A transcriptome was prepared of the complete mature and regenerating tissues of the post-anal tail of *B. lanceolatum*. Adult animals were collected and maintained as described in section 2.1.2. The experimental conditions, amputation protocol, total RNA extraction and assembly and mapping pipeline are described in section 2.2.2 and in Dailey

(2017). Homeobox gene searches were performed as described in section 2.3.2. Genes were identified based on BLAST searches and manual and MAFFT (Katoh, Rozewicki, and Yamada 2017; Katoh and Standley 2013) alignments made in Jalview 2.x.

4.3. Results

The initial homeobox searches were performed on a *de novo* assembly of *B. lanceolatum* mature, regenerative, and developmental (Oulion *et al.*, 2012) transcriptomes (section 4.3.1). With the advent of advance availability of the *B. lanceolatum* genome (courtesy of the European Amphioxus Genome Consortium; Marletaz *et al.*, in press), the mature and regenerating transcriptomes were mapped against the genomic predicted transcript database. The assembly work in both cases was performed by S. Dailey. Predicted transcripts corresponding to homeobox genes were identified from the predicted transcript database (Appendix 4.1), and the mature and regenerative read counts retrieved from the transcriptome mapping (section 4.3.2).

4.3.1. Homeobox genes from the *de novo* transcriptome assembly

26 isotigs, 31 singletons (*i.e.* unassembled reads) from the uncut transcriptome, and 31 singletons from the regenerating transcriptome were identified by the initial BLAST searches as containing homeobox gene sequence. The details of these sequences, their identification, basis for their identification, and normalised uncut and regenerative read counts are presented in Table 4.1. The alignment files used in the identification of these genes are included in Appendix 4.2.

Table 4.1. Homeobox genes found in the *de novo* transcriptome assembly. Sequences with no reads in the regenerating sample are coloured grey. Sequences with more than one read which are up-regulated in regeneration are emboldened. Numbers in the totals column are coloured in a red to green spectrum depending on their read count, so that poorly represented genes are coloured red. In the fold change column, values are coloured with white to green (>1) or white to purple (<-1) spectra, such that sequences with many more reads in regeneration are coloured bright green and sequences with many fewer reads are coloured bright purple. Genes which appear *de novo* in regeneration are marked with an up arrow (⇑) and genes which appear in the mature tissue but not regeneration, by a down arrow (⇓). Read rows are marked grey when no expression is observed in the regenerative transcriptome.

CLASS	FAMILY	GENE	Matches	Basis of ID	Uncut	Regen	Total	Fold change
ANTP	Cdx	Cdx	RegenGJ02JL4UK	Identical HD	0	1	1	⇑
	Evx	Evxa	RegenGJ02G8I17	Alignment	0	1	1	⇑
	Gbx	**Gbx**	isotig29024	Identical HD	21	34	55	1.62
	Hox1	Hox1	UncutGJ01AK8RQ, UncutGJ01ERRZY	Identical HD + alignment	2	0	2	⇓
	Hox3	Hox3	RegenGJ02JZFXU, UncutGJ01A4Z0Y	Alignment	1	1	2	-1.00
	Hox4	**Hox4**	isotig35935	Identical HD	5	16	21	3.20
	Hox5	Hox5	UncutGJ01DP5JF	Alignment	1	0	1	⇓
	Hox6-8	Hox6	isotig36337, UncutGJ01C18YT	Identical HD	18	5	23	-3.60
	Hox9-13(15)	Hox9	isotig23779	Alignment	81	56	137	-1.45
		Hox11	RegenGJ02H7F0N	Identical HD	0	1	1	⇑
		Hox12	isotig20727, RegenGJ02FMLYQ, RegenGJ02FYH3B, RegenGJ02G3985, RegenGJ02HO8A7, RegenGJ02HPOJ3, UncutGJ01A9P5B, UncutGJ01BHKA8, UncutGJ01BYNY7, UncutGJ01C4XIC, UncutGJ01CIOLE, UncutGJ01EQIL5	Alignment	60	83	143	1.38
		Hox13	RegenGJ02GQ1H3	Identical HD	0	1	1	⇑
		Hox14	isotig20006, isotig20012	Identical HD + alignment	17	15	32	-1.13
		Hox15	RegenGJ02J1QC9	Identical HD	0	1	1	⇑
	Meox	Meox	RegenGJ02JTX06	Alignment	0	1	1	⇑
	Mnx	Mnxb	RegenGJ02HV213	Identical HD + alignment	0	1	1	⇑
	BarH	BarH	isotig36167	Alignment	19	14	33	-1.36
	Dlx	**Dll**	isotig34199	Identical HD	212	234	446	1.10
	Emx	EmxB	isotig28854, UncutGJ01DNPVA, UncutGJ01E3716	Alignment	23	12	35	-1.92
		EmxC	isotig35164, UncutGJ01B8LLW	Identical HD + alignment	25	13	38	-1.92

	En	En	UncutGJ01ACY9S, UncutGJ01ARDZ1, UncutGJ01D7K6K	Identical HD + alignment	3	0	3	⇓
	Hhex	Hhex	UncutGJ01DBXSU	Identical HD	1	0	1	⇓
	Lbx	**Lbx**	isotig36831	Alignment	6	8	14	1.33
	Msx	**Msx**	isotig35362	Identical HD	14	18	32	1.29
	Nedx	Nedxb	isotig19417	Identical HD	3	1	4	-3.00
	Nk3	**Nkx3**	isotig35111	Identical HD + alignment	6	17	23	2.83
	Nk4	Nkx4	RegenGJ02HFE62	Alignment	0	1	1	⇑
	Nk6	Nkx6	UncutGJ01CKRVP, UncutGJ01DTFZA	Identical HD + alignment	2	0	2	⇓
	Ventx	Vent2	isotig40164, RegenGJ02GWKF1, RegenGJ02H8D96	Alignment	0	11	11	⇑
CERS	Cers	Cers	RegenGJ02IHNN0	Alignment	0	1	1	⇑
CUT	Acut	Acut	UncutGJ01CMHT6	Alignment	1	0	1	⇓
HNF	Hmbhox	Hmbox1A	RegenGJ02HEVAO	Identical HD	0	1	1	⇑
LIM	Isl	Isl	UncutGJ01BHDWO	Identical HD	1	0	1	⇓
	Lhx1/5	Lhx1/5	UncutGJ01EM0RZ	Alignment	1	0	1	⇓
	Lhx2/9	Lhx2/9b	RegenGJ02IN2DN	Alignment	0	1	1	⇑
POU	Pou3	**Pou3**	isotig34454, UncutGJ01BY8TH, UncutGJ01EJQE6	Identical HD + alignment	30	38	68	1.27
PRD	Alx	Alx	isotig21241, isotig21243	Identical HD	0	2	2	⇑
	Dmbx	Dmbx	isotig37000	Alignment	7	1	8	-7.00
	Pax3/7	**Pax3/7b**	isotig29738	**Chapter 5**	6	21	27	3.50
	Pitx	Pitx	isotig34677	Alignment	15	10	25	-1.50
	Prrx	**Prrx**	RegenGJ02GZ9AQ, RegenGJ02JQECO, isotig35735	Alignment	2	44	46	22.00
	Repo	**Repo**	RegenGJ02I4CTG, RegenGJ02ISBP6, UncutGJ01C8EVP	Identical HD + alignment	1	2	3	2.00
	Shox	Shox	RegenGJ02GGV5G	Identical HD	0	1	1	⇑
SINE	Six1/2	**Six1/2**	isotig34510	Identical HD	52	75	127	1.44
	Six3/6	**Six3/6**	isotig34745	Identical HD	15	58	73	3.87
	Six4/5	Six4/5	RegenGJ02GFQB3, UncutGJ01CT51G	Alignment	1	1	2	-1.00
TALE	Irx	IrxA	RegenGJ02F5SW0, RegenGJ02I1G7W	Alignment	0	2	2	⇑
		IrxB	RegenGJ02IVP8A, UncutGJ01BVJU6	Alignment	1	1	2	-1.00
		IrxC	UncutGJ01CKRUQ, UncutGJ01DIUE9	Alignment	2	0	2	⇓
	Mkx	Mkx	isotig28784	Alignment	29	27	56	-1.07
	Pbx	Pbx	RegenGJ02G3K3O, RegenGJ02HE28F	Identical HD	0	2	2	⇑
	Pknox	Pknox	RegenGJ02IC149, UncutGJ01BHPL7	Alignment	1	1	2	-1.00

4.3.2. Homeobox sequences in the transcriptome read map

Using BLAST searches and alignments to *B. floridae* sequences retrieved from HomeoDB2 and the Joint Genome Institute, I identified the transcript models best corresponding to homeobox genes in the *B. lanceolatum* genome assembly. These are presented in Appendix 4.1. The alignments used for identification are included in Appendix 4.2. Four homeobox genes (*Hox4, Lcx, Nedxa,* and *Atale*; coloured pink in Appendix 4.1 and Table 4.2) were found not to be represented amongst the predicted transcripts.

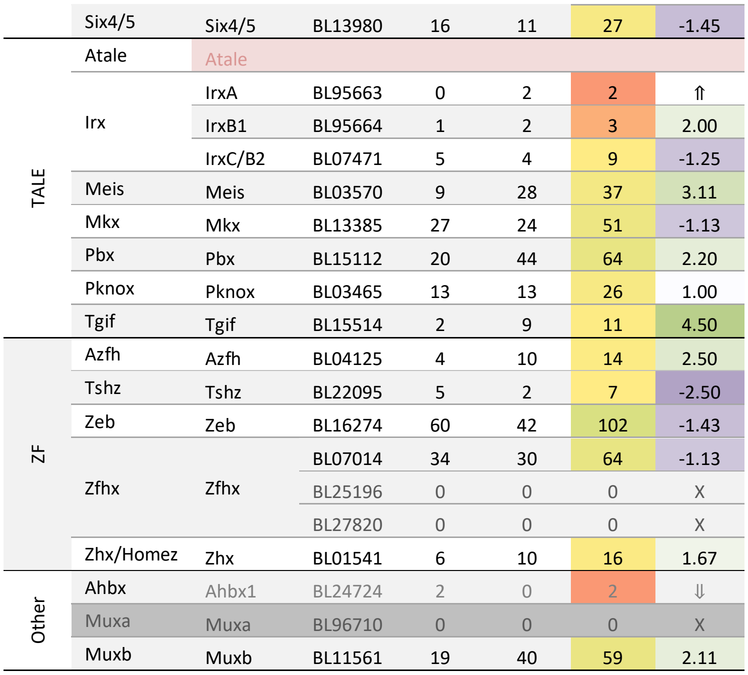
The transcriptomic reads were mapped against the genomic predicted transcripts and normalised. The read counts are presented in Table 4.2.

Table 4.2. Read counts of predicted genomic transcripts identified as the best matches for homeobox genes (Appendix 4.1). Formatting as in Table 4.1. For details on the two *Pax3/7* paralogues included, see Chapter 5. Colouration & notation per Table 4.1. Genes not represented amongst the predicted transcripts are coloured pink.

Class	Family	Gene	Ref. Seq.	Uncut (Norm)	Regen (Norm)	Total hits (norm)	Fold Change
ANTP	Cdx	Cdx	BL12756	0	56	56	⇑
	Evx	Evxa	BL08778	4	6	10	1.50
		Evxb	BL02337	0	0	0	X
	Gbx	Gbx	BL21529	19	30	49	1.58
	Gsx	Gsx	BL15948	0	0	0	X
	Hox1	Hox1	BL12289	3	0	3	⇓
	Hox2	Hox2	BL01409	1	0	1	⇓
	Hox3	Hox3	BL14546	7	21	28	3.00
	Hox4	Hox4					
	Hox5	Hox5	BL11265	2	26	28	13.00
	Hox6-8	Hox6	BL01497	3	1	4	-3.00
		Hox7	BL02690	4	8	12	2.00
		Hox8	BL01142	2	1	3	-2.00
	Hox9-13(15)	Hox9	BL02747	78	50	128	-1.56
		Hox10	BL02721	0	0	0	X
			BL25449	0	0	0	X
		Hox11	BL22794	56	70	126	1.25
		Hox12	BL01764	23	24	47	1.04
		Hox13	BL11259	18	26	44	1.44
		Hox14	BL11262	94	117	211	1.24
		Hox15	BL06042	35	30	65	-1.17
	Meox	Mox	BL03350	0	0	0	X
	Mnx	Mnxa	BL01012	0	0	0	X
		Mnxb	BL24614	0	1	1	⇑
	Pdx	Xlox	BL11413	0	0	0	X

	Abox	Abox	BL05106	0	0	0	X
	Ankx	Ankx	BL05986	0	2	2	⇑
	Barhl	Barh	BL05807	17	12	29	-1.42
	Bari	Bari	BL01266	1	0	1	⇓
	Barx	Barx	BL18096	0	0	0	X
	Bsx	Bsx	BL06267	0	0	0	X
	Dbx	Dbx	BL04439	0	0	0	X
	Dlx	Dll	BL02953	197	196	393	-1.01
	Emx	Emxa	BL04899	5	14	19	2.80
		Emxb	BL12193	20	11	31	-1.82
		Emxc	BL03198	20	8	28	-2.50
	En	En	BL18701	4	2	6	-2.00
	Hhex	Hhex	BL21396	1	0	1	⇓
	Hlx	Hlx	BL17526	0	0	0	X
	Hx	Hx	BL01814	66	35	101	-1.89
	Lbx	Lbx	BL25553	0	1	1	⇑
			BL11564	0	0	0	X
	Lcx	Lcx					
	Msx	Msx	BL24026	0	0	0	X
	Msxlx	Msxlx	BL17549	0	0	0	X
	Nedx	Nedxa					
		Nedxb	BL00186	5	6	11	1.20
	Nk1	Nkx1a	BL14817	0	0	0	X
		Nkx1b	BL08182	0	0	0	X
	Nk2.1	Nkx2-1	BL15126	0	0	0	X
	Nk2.2	Nkx2-2	BL19775	0	0	0	X
	Nk3	Nkx3	BL01669	6	14	20	2.33
	Nk4	Nkx4	BL01674	4	9	13	2.25
	Nk5/Hmx	Hmx	BL00743	1	1	2	1.00
	Nk6	Nkx6	BL21899	8	23	31	2.88
	Nk7	Nkx7	BL12992	1	1	2	1.00
	Noto	Not	BL04895	0	0	0	X
	Ro	Ro	BL16282	1	1	2	1.00
	Tlx	Tlx	BL95610	1	0	1	⇓
	Vax	Vax	BL50426	0	0	0	X
	Ventx	Vent1	BL06974	0	1	1	⇑
		Vent2	BL14812	0	8	8	⇑
CERS	Cers	Cers	BL11720	39	42	81	1.08
CUT	Acut	Acut	BL72646	2	0	2	⇓
	Cmp	Compass	BL18389	0	1	1	⇑
	Cux	Cux	BL21463	11	10	21	-1.10
	Onecut	Onecut	BL14590	4	2	6	-2.00
HNF	Ahnfx	Ahnf	BL28317	3	6	9	2.00
	Hmbox	Hmbox1A	BL13813	6	13	19	2.17
		Hmbox1B	BL11896	2	2	4	1.00
	Hmbox?	Hmbx-l 1	BL08673	19	14	33	-1.36
		Hmbx-l 2	BL12473	0	2	2	⇑
		Hmbx-l 3	BL01804	28	21	49	-1.33
	Hnf1	Tcf	BL16767	0	0	0	X
LIM	Isl	Isl	BL14099	2	4	6	2.00

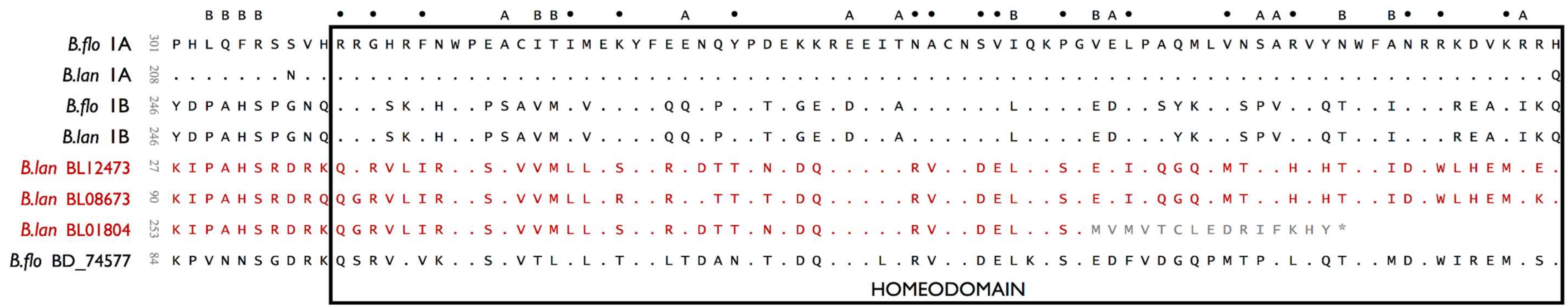
	Lhx1/5	Lhx1/5	BL00881	5	2	7	-2.50
	Lhx2/9	Lhx2/9-a	BL09173	0	2	2	⇑
		Lhx2/9-b	BL19723	2	1	3	-2.00
			BL18410	1	3	4	3.00
	Lhx3/4	Lhx3/4	BL17252	2	0	2	⇓
	Lhx6/8	Lhx6/8	BL14754	0	0	0	X
	Lmx	Lmx	BL19671	0	2	2	⇑
POU	Hdx	Hdx	BL00246	5	3	8	-1.67
	Pou1	POU1	BL23543	1	0	1	⇓
	Pou2	POU2	BL05378	0	0	0	X
	Pou3	POU3	BL16866	26	35	61	1.35
		POU3L	BL05589	0	0	0	X
	Pou4	POU4	BL20738	2	1	3	-2.00
			BL20735	0	0	0	X
	Pou6	POU6	BL02913	9	29	38	3.22
PRD	Alx	Alx	BL06191	10	57	67	5.70
			BL00778	0	0	0	X
	AprdA	Aprd1	BL22260	0	0	0	X
	AprdB	Aprd2	BL19810	0	1	1	⇑
	AprdC	Aprd3	BL03074	0	0	0	X
	AprdD	Aprd4	BL03518	0	0	0	X
	AprdD	Aprd5	BL02873	0	0	0	X
	AprdE	Aprd6	BL03811	0	0	0	X
	Arx	Arx	BL10864	3	2	5	-1.50
	Dmbx	Dmbx	BL21886	10	14	24	1.40
	Drgx	Drgx	BL03270	0	1	1	⇑
	Gsc	Gsc	BL18678	0	1	1	⇑
	Hopx	Hopx	BL41687	2	9	11	4.50
	Isx	Isx	BL06515	0	0	0	X
	Otp	Otp	BL13404	6	8	14	1.33
	Otx	Otx	BL18685	4	0	4	⇓
	Pax3/7	Pax3/7a	BL95937	0	0	0	X
		Pax3/7b	BL32034	0	0	0	X
			BL95936	0	5	5	⇑
	Pax4/6	Pax4/6	BL13922	3	3	6	1.00
			BL05977	43	26	69	-1.65
	Phox	Phox	BL07358	0	0	0	X
	Pitx	Pitx	BL12801	13	9	22	-1.44
	Prop	Prop	BL07414	0	0	0	X
	Prrx	Prrx	BL09897	2	24	26	12.00
	Rax	Rax	BL10535	6	3	9	-2.00
	Repo	Repo	BL04482	1	4	5	4.00
	Shox	Shox	BL04730	7	8	15	1.14
	Uncx	UncxA	BL07183	3	2	5	-1.50
		UncxB	BL22975	0	0	0	X
		UncxC	BL09739	0	0	0	X
	Vsx	Vsx	BL02827	1	6	7	6.00
P.	Prox	Prox	BL13719	55	43	98	-1.28
SINE	Six1/2	Six1/2	BL08389	47	63	110	1.34
	Six3/6	Six3/6	BL08388	12	42	54	3.50



	Six4/5	Six4/5	BL13980	16	11	27	-1.45
TALE	Atale	Atale					
	Irx	IrxA	BL95663	0	2	2	⇑
		IrxB1	BL95664	1	2	3	2.00
		IrxC/B2	BL07471	5	4	9	-1.25
	Meis	Meis	BL03570	9	28	37	3.11
	Mkx	Mkx	BL13385	27	24	51	-1.13
	Pbx	Pbx	BL15112	20	44	64	2.20
	Pknox	Pknox	BL03465	13	13	26	1.00
	Tgif	Tgif	BL15514	2	9	11	4.50
ZF	Azfh	Azfh	BL04125	4	10	14	2.50
	Tshz	Tshz	BL22095	5	2	7	-2.50
	Zeb	Zeb	BL16274	60	42	102	-1.43
	Zfhx	Zfhx	BL07014	34	30	64	-1.13
			BL25196	0	0	0	X
			BL27820	0	0	0	X
	Zhx/Homez	Zhx	BL01541	6	10	16	1.67
Other	Ahbx	Ahbx1	BL24724	2	0	2	⇓
	Muxa	Muxa	BL96710	0	0	0	X
	Muxb	Muxb	BL11561	19	40	59	2.11

4.3.2.1. Previously undescribed homeobox genes predicted in the genome

The sequences marked Hmbox-like (BL08673, BL12473, & BL01804) are not clear orthologues of the amphioxus *Hmbox1A* or *Hmbox1B* genes. The Conserved Domain Search (Marchler-Bauer *et al.*, 2015) does identify an HNF domain in residues 12-61 of BL12473 and residues 151-224 of BL01804, hence their placement in the HNF class. The top BLASTp match of BL12473 against the nr database is a *B. floridae* draft genome predicted transcript (BRAFLDRAFT_74577) although the match is poor by the standards of *B. lanceolatum*/*B. floridae* orthologue conservation (*i.e.* only 56% with a region of good alignment with BL12473, whereas >80% is more expected). Three other BRAFLDRAFT sequences are the next top hits (220804, 249692, & 249685) before the *B. belcheri* predicted Hmbox1B sequence. An alignment of the homeodomains and ten N-terminal positions of the *B. floridae* Hmbox1 paralogues, their corresponding *B. lanceolatum* predicted gene models (BL13813 & BL11896), the three unknown genes (BL08673, BL12473, & BL01804) and BRAFLDRAFT_74577 are presented in Figure 4.1. These genes are referred to as Hmbox-like (Hmbx-l) in subsequent tables.



```
                     BBBB     • •  •    A BB•  •   A  •      A  A••  ••B  • BA•     • AA•  B  B• •   •A  
B.flo IA       301 PHLQFRSSVH RRGHRFNWPEACITIMEKYFEENQYPDEKKREEITNACNSVIQKPGVELPAQMLVNSARVYNWFANRRKDVKRRH 385
B.lan IA       208 .......N.. ..........................................................................Q 292
B.flo IB       246 YDPAHSPGNQ ...SK.H..PSAVM.V....QQ.P..T.GE.D..A......L....ED..SYK..SPV..QT..I...REA.IKQ 330
B.lan IB       246 YDPAHSPGNQ ...SK.H..PSAVM.V....QQ.P..T.GE.D..A......L....ED...YK..SPV..QT..I...REA.IKQ 330
B.lan BL12473   27 KIPAHSRDRK Q.RVLIR..S.VVMLL.S..R.DTT.N.DQ.....RV..DEL..S.E.I.QGQ.MT..H.HT..ID.WLHEM.E. 111
B.lan BL08673   90 KIPAHSRDRQ QGRVLIR..S.VVMLL.R..R..TT.T.DQ.....RV..DEL..S.E.I.QGQ.MT..H.HT..ID.WLHEM.K. 174
B.lan BL01804  253 KIPAHSRDRK QGRVLIR..S.VVMLL.S..R.DTT.N.DQ.....RV..DEL..S.MVMVTCLEDRIFKHY*             323
B.flo BD_74577  84 KPVNNSGDRK QSRV.VK..S.VTL.L.T..LTDAN.T.DQ...L.RV..DELK.S.EDFVDGQPMTP.L.QT..MD.WIREM.S. 168
                                                        HOMEODOMAIN
```

Figure 4.1. Alignment of the undescribed homeodomains from the *B. lanceolatum* genome and ten N-terminal positions, against the *B. floridae* Hmbox genes (Hmbox1A, JGI:105752; Hmbox1B, JGI:105751) and the corresponding *B. lanceolatum* gene models (Hmbox1A, BL13813: Hmbox1B, BL11896, highlighted red), three unidentified Hmbox-like *B. lanceolatum* genomic predicted transcripts, and the top BLAST hit, *B. floridae* BRAFLDRAFT_74577. A region of non-homeodomain sequence, presumably the result of inaccurate prediction, has been coloured grey in BL01804. Positions are marked above with a bullet if the two known Hmbox1 paralogues bear more similarity to one another than to the Hmbox-like sequences, with an 'A' if the Hmbox-like sequences more closely resemble Hmbox1A, and 'B' if the Hmbox-like sequences more closely resemble Hmbox1B.

4.3.3. Comparison of homeobox content analyses

Substantial discrepancies are observed between the homeobox gene pool retrieved from the *de novo* transcriptome assembly (section 4.3.1) and the transcriptome read mapping (section 4.3.2). Sequences from 52 homeobox genes were detected in the *de novo* assembly, compared to the 97 identified from the genome read map. Two genes (*Mox* and *Msx*) were detected in the *de novo* assembly but did not have reads mapped to their predicted genomic transcript. One gene (*Hox4*) was detected in the *de novo* assembly but did not have a corresponding predicted genomic transcript against which to map reads. A comparison between the homeobox gene sequences retrieved from the *de novo* transcriptome assembly and the genome mapping is presented in Table 4.3.

Table 4.3. Comparison of the homeobox genes detected via the transcriptome map and *de novo* transcriptome assembly, with a qualitative measure of the congruency between the two methods (right block). In regulation, ticks indicate that the two methods show the same apparent difference in read counts, with brackets if either method produced only a single read. Crosses indicate that the methods contradict, with brackets indicating that the gene was not detected by at least one method. In read count, ticks indicate that the read count is different by no more than a factor of two, tildes indicate that the read count is different by a factor of two to five, and crosses that the read counts differ by more than a factor of five. Bracketed crosses indicate that the sequence was not detected by at least one method. C = CERS; P= PROS.

	GENE IDENTITIES		TRANSCRIPTOME MAPPING				*DE NOVO* TRANSCRIPTOME				CONGRUENCY	
CLASS	FAMILY	GENE	Uncut (Norm)	Regen (Norm)	Total (norm)	Fold Change	Uncut (norm)	Regen (norm)	Total (norm)	Fold change	Regulation	Read count
ANTP	Cdx	**Cdx**	0	56	56	⇑	0	1	1	⇑	(✓)	X
	Evx	**Evxa**	4	6	10	1.50	0	1	1	⇑	(✓)	X
	Gbx	**Gbx**	19	30	49	1.58	21	34	55	1.62	✓	✓
	Hox1	Hox1	3	0	3	⇓	2	0	2	⇓	✓	✓
	Hox2	Hox2	1	0	1	⇓	0	0	0	X	(X)	(X)
	Hox3	Hox3	7	21	28	3.00	1	1	2	1.00	X	X
	Hox4	Hox4					5	16	21	3.20	-	-
	Hox5	Hox5	2	26	28	13.00	1	0	1	⇓	(X)	X
	Hox6-8	Hox6	3	1	4	-3.00	18	5	23	-3.60	✓	X
		Hox7	4	8	12	2.00	0	0	0	X	(X)	(X)
		Hox8	2	1	3	-2.00	0	0	0	X	(X)	(X)
	Hox9-13(15)	Hox9	78	50	128	-1.56	81	56	137	-1.45	✓	✓
		Hox11	56	70	126	1.25	0	1	1	⇑	(✓)	X
		Hox12	23	24	47	1.04	60	83	143	1.38	~	~
		Hox13	18	26	44	1.44	0	1	1	⇑	(✓)	X
		Hox14	94	117	211	1.24	17	15	32	-1.13	X	X
		Hox15	35	30	65	-1.17	0	1	1	⇑	X	X
	Meox	Mox	0	0	0	X	0	1	1	⇑	(X)	(X)
	Mnx	**Mnxb**	0	1	1	⇑	0	1	1	⇑	(✓)	✓
	Ankx	Ankx	0	2	2	⇑	0	0	0	X	(X)	(X)
	Barhl	Barh	17	12	29	-1.42	19	14	33	-1.36	✓	✓
	Bari	Bari	1	0	1	⇓	0	0	0	X	(X)	(X)
	Dlx	Dll	197	196	393	-1.01	212	234	446	1.10	X	✓
	Emx	Emxa	5	14	19	2.80	0	0	0	X	(X)	(X)
		Emxb	20	11	31	-1.82	23	12	35	-1.92	✓	✓
		Emxc	20	8	28	-2.50	25	13	38	-1.92	✓	✓
	En	En	4	2	6	-2.00	3	0	3	⇓	✓	✓
	Hhex	Hhex	1	0	1	⇓	1	0	1	⇓	(✓)	✓
	Hx	Hx	66	35	101	-1.89	0	0	0	X	(X)	(X)
	Lbx	**Lbx**	0	1	1	⇑	6	8	14	1.33	(✓)	X
	Msx	Msx	0	0	0	X	14	18	32	1.29	(X)	(X)
	Nedx	Nedxb	5	6	11	1.20	3	1	4	-3.00	X	~
	Nk3	**Nkx3**	6	14	20	2.33	6	17	23	2.83	✓	✓
	Nk4	**Nkx4**	4	9	13	2.25	0	1	1	⇑	(✓)	X
	Nk5/Hmx	Hmx	1	1	2	1.00	0	0	0	X	(X)	(X)
	Nk6	Nkx6	8	23	31	2.88	2	0	2	⇓	X	X

	Nk7	Nkx7	1	1	2	1.00	0	0	0	X	(X)	(X)
	Ro	Ro	1	1	2	1.00	0	0	0	X	(X)	(X)
	Tlx	Tlx	1	0	1	⇓	0	0	0	X	(X)	(X)
	Ventx	Vent1	0	1	1	⇑	0	0	0	X	(X)	(X)
		Vent2	0	8	8	⇑	0	11	11	⇑	✓	✓
C.	Cers	**Cers**	39	42	81	1.08	0	1	1	⇑	(✓)	X
CUT	Acut	Acut	2	0	2	⇓	1	0	1	⇓	(✓)	✓
	Cmp	Compass	0	1	1	⇑	0	0	0	X	(X)	(X)
	Cux	Cux	11	10	21	-1.10	0	0	0	X	(X)	(X)
	Onecut	Onecut	4	2	6	-2.00	0	0	0	X	(X)	(X)
HNF	Ahnfx	Ahnf	3	6	9	2.00	0	0	0	X	(X)	(X)
	Hmbox	**Hmbox1A**	6	13	19	2.17	0	1	1	⇑	(✓)	X
		Hmbox1B	2	2	4	1.00	0	0	0	X	(X)	(X)
	Hmbox?	Hmbx-l1	19	14	33	-1.36	0	0	0	X	(X)	(X)
		Hmbx-l2	0	2	2	⇑	0	0	0	X	(X)	(X)
		Hmbx-l3	28	21	49	-1.33	0	0	0	X	(X)	(X)
LIM	Isl	Isl	2	4	6	2.00	1	0	1	⇓	(X)	X
	Lhx1/5	Lhx1/5	5	2	7	-2.50	1	0	1	⇓	(✓)	X
	Lhx2/9	Lhx2/9-a	0	2	2	⇑	0	0	0	X	(X)	(X)
		Lhx2/9-b	3	4	7	1.33	0	1	1	⇑	(✓)	X
	Lhx3/4	Lhx3/4	2	0	2	⇓	0	0	0	X	(X)	(X)
	Lmx	Lmx	0	2	2	⇑	0	0	0	X	(X)	(X)
POU	Hdx	Hdx	5	3	8	-1.67	0	0	0	X	(X)	(X)
	Pou1	POU1	1	0	1	⇓	0	0	0	X	(X)	(X)
	Pou3	**POU3**	26	35	61	1.35	30	38	68	1.27	✓	✓
	Pou4	POU4	2	1	3	-2.00	0	0	0	X	(X)	(X)
	Pou6	POU6	9	29	38	3.22	0	0	0	X	(X)	(X)
PRD	Alx	**Alx**	10	57	67	5.70	0	2	2	⇑	✓	X
	AprdB	Aprd2	0	1	1	⇑	0	0	0	X	(X)	(X)
	Arx	Arx	3	2	5	-1.50	0	0	0	X	(X)	(X)
	Dmbx	Dmbx	10	14	24	1.40	7	1	8	-7.00	X	~
	Drgx	Drgx	0	1	1	⇑	0	0	0	X	(X)	(X)
	Gsc	Gsc	0	1	1	⇑	0	0	0	X	(X)	(X)
	Hopx	Hopx	2	9	11	4.50	0	0	0	X	(X)	(X)
	Otp	Otp	6	8	14	1.33	0	0	0	X	(X)	(X)
	Otx	Otx	4	0	4	⇓	0	0	0	X	(X)	(X)
	Pax3/7	**Pax3/7b**	0	5	5	⇑	6	21	27	3.50	✓	X
	Pax4/6	Pax4/6	46	29	75	-1.59	0	0	0	X	(X)	(X)
	Pitx	Pitx	13	9	22	-1.44	15	10	25	-1.50	✓	✓
	Prrx	**Prrx**	2	24	26	12.00	2	44	46	22.00	✓	✓
	Rax	Rax	6	3	9	-2.00	0	0	0	X	(X)	(X)

	Repo	**Repo**	1	4	5	4.00	1	2	3	2.00	√	√
	Shox	**Shox**	7	8	15	1.14	0	1	1	⇑	(√)	X
	Uncx	UncxA	3	2	5	-1.50	0	0	0	X	(X)	(X)
	Vsx	Vsx	1	6	7	6.00	0	0	0	X	(X)	(X)
P.	Prox	Prox	55	43	98	-1.28	0	0	0	X	(X)	(X)
SINE	Six1/2	**Six1/2**	47	63	110	1.34	52	75	127	1.44	√	√
	Six3/6	**Six3/6**	12	42	54	3.50	15	58	73	3.87	√	√
	Six4/5	Six4/5	16	11	27	-1.45	1	1	2	1.00	X	X
TALE	Irx	**IrxA**	0	2	2	⇑	0	2	2	⇑	√	√
		IrxB1	1	2	3	2.00	1	1	2	1.00	X	√
		IrxC/B2	5	4	9	-1.25	2	0	2	⇓	√	~
	Meis	Meis	9	28	37	3.11	0	0	0	X	(X)	(X)
	Mkx	Mkx	27	24	51	-1.13	29	27	56	-1.07	√	√
	Pbx	**Pbx**	20	44	64	2.20	0	2	2	⇑	√	X
	Pknox	Pknox	13	13	26	1.00	1	1	2	1.00	√	X
	Tgif	Tgif	2	9	11	4.50	0	0	0	X	(X)	(X)
ZF	Azfh	Azfh	4	10	14	2.50	0	0	0	X	(X)	(X)
	Tshz	Tshz	5	2	7	-2.50	0	0	0	X	(X)	(X)
	Zeb	Zeb	60	42	102	-1.43	0	0	0	X	(X)	(X)
	Zfhx	Zfhx	34	30	64	-1.13	0	0	0	X	(X)	(X)
	Zhx/Homez	Zhx	6	10	16	1.67	0	0	0	X	(X)	(X)
Other	Ahbx	Ahbx1	2	0	2	⇓	0	0	0	X	(X)	(X)
	Muxb	Muxb	19	40	59	2.11	0	0	0	X	(X)	(X)

4.4. Discussion

4.4.1. Interpreting read count as indicative expression levels

Statistical methodologies have been developed for the robust detection of differential expression in transcriptomes (*e.g.* Marioni *et al.*, 2008; Oshlack, Robinson, and Young 2010; Finotello and Di Camillo 2015; Conesa *et al.*, 2016; Das, Shyamal, and Durica 2016; Łabaj and Kreil 2016). Basic filtering like normalization of raw transcriptomic database sizes has been performed on this dataset, but more sophisticated techniques to distinguish technical (*i.e.* caused by measurement error) and biological variation (Finotello and Di Camillo 2015) rely on the availability of technical replicates, which are not available in the case of this dataset.

Therefore, although the read counts from maps of the *de novo* assembly and the genome predicted transcripts are reported herein and referred to as a tentative measure of their expression, they should not be taken as a reliable measure of differential expression, particularly for genes with small read counts, from which most genes profiled herein suffer.

4.4.2. Discrepancies between the *de novo* assembly and genome read mapping

Major differences exist between the homeobox sequence data from the *de novo* assembly and from the genome read mapping (section 4.3.3). Several plausible explanations for these differences exist. The first is that the predicted transcripts contain substantially more sequence data than the queries used to retrieve putative homeobox sequences from the *de novo* transcriptomes, including (predicted) non-translated regions. However, it may be that not all this sequence belongs to the real gene transcripts. Therefore, the expression of more genes, and more reads per gene, is expected, though the reliability of the data are lower.

As well as the possibility of predicted transcripts containing inaccurate sequence, they can also miss out 5', 3' and central portions of the actual transcript. Incomplete predicted transcript sequence is a plausible explanation for the detection of three genes (*Hox4*, *Mox*, & *Msx*) in the *de novo* assembly but not the genome read map.

4.4.3. Potential differential expression of homeobox genes in regeneration

A summary of the genes represented by more reads in the regenerating transcriptome than the mature transcriptome, and by more than one read in either the *de novo* assembly or the transcript read map, is presented in Table 4.4. These represent the genes indicated by this analysis that may be upregulated in regeneration.

Table 4.4. Summary of the genes with more reads in the regeneration transcriptome than the mature transcriptome (and which have more than a single read) in either the *de novo* assembly or the genome read mapping. C = CERS. Colouration per Table 4.1.

GENE IDENTITIES			GENOME MAPPING			*DE NOVO* TRANSCRIPTOME		
CLASS	FAMILY	GENE	Uncut (Norm)	Regen (Norm)	Fold Change	Uncut (norm)	Regen (norm)	Fold change
ANTP	Cdx	Cdx	0	56	⇑	0	1	⇑
	Evx	Evxa	4	6	1.50	0	1	⇑
	Gbx	Gbx	19	30	1.58	21	34	1.62
	Hox3	Hox3	7	21	3.00	1	1	1.00
	Hox4	Hox4				5	16	3.20
	Hox5	Hox5	2	26	13.0	1	0	⇓
	Hox6-8	Hox7	4	8	2.0	0	0	X
	Hox9-13(15)	Hox11	56	70	1.25	0	1	⇑
		Hox12	23	24	1.04	60	83	1.38
		Hox13	18	26	1.44	0	1	⇑
		Hox14	94	117	1.24	17	15	-1.13
	Ankx	Ankx	0	2	⇑	0	0	X
	Emx	Emxa	5	14	2.80	0	0	X
	Lbx	Lbx	0	1	⇑	6	8	1.33
	Msx	Msx	0	0	X	14	18	1.29
	Nedx	Nedxb	5	6	1.20	3	1	-3.00
	Nk3	Nkx3	6	14	2.33	6	17	2.83
	Nk4	Nkx4	4	9	2.25	0	1	⇑
	Nk6	Nkx6	8	23	2.88	2	0	⇓
	Ventx	Vent2	0	8	⇑	0	11	⇑
ZF	Azfh	Azfh	4	10	2.50	0	0	X
	Zhx/Homez	Zhx	6	10	1.67	0	0	X
Other	Muxb	Muxb	19	40	2.11	0	0	X

GENE IDENTITIES			GENOME MAPPING			*DE NOVO* TRANSCRIPTOME		
CLASS	FAMILY	GENE	Uncut (Norm)	Regen (Norm)	Fold Change	Uncut (norm)	Regen (norm)	Fold change
C	Cers	**Cers**	39	42	1.08	0	1	⇑
HNF	Ahnfx	Ahnf	3	6	2.00	0	0	X
	Hmbox?	Hmbx-l2	0	2	⇑	0	0	X
LIM	Lhx2/9	Lhx2/9-a	0	2	⇑	0	0	X
		Lhx2/9-b	3	4	1.33	0	1	⇑
	Lmx	Lmx	0	2	⇑	0	0	X
POU	Pou3	POU3	26	35	1.35	30	38	1.27
	Pou6	POU6	9	29	3.22	0	0	X
PRD	Alx	Alx	10	57	5.70	0	2	⇑
	Dmbx	Dmbx	10	14	1.40	7	1	-7.00
	Hopx	Hopx	2	9	4.50	0	0	X
	Otp	Otp	6	8	1.33	0	0	X
	Pax3/7	Pax3/7b	0	5	⇑	6	21	3.50
	Prrx	Prrx	2	24	12.0	2	44	22.0
	Repo	Repo	1	4	4.00	1	2	2.00
	Shox	Shox	7	8	1.14	0	1	⇑
	Vsx	Vsx	1	6	6.00	0	0	X
SINE	Six1/2	Six1/2	47	63	1.34	52	75	1.44
	Six3/6	Six3/6	12	42	3.50	15	58	3.87
TALE	Irx	IrxA	0	2	⇑	0	2	⇑
		IrxB1	1	2	2.00	1	1	1.00
	Meis	Meis	9	28	3.11	0	0	X
	Pbx	Pbx	20	44	2.20	0	2	⇑
	Tgif	Tgif	2	9	4.50	0	0	X

In the genomic read mapping, of the transcripts best corresponding to the 133 previously described amphioxus homeobox genes and one new paralogue (see chapter 5), 84 were detected in the mature transcriptome and 89 in the regenerating transcriptome, with ten found only in the mature transcriptome and 15 only in the regenerating transcriptome. Of the 74 detected in both, 38 were represented by more reads in the regenerating transcriptome, and 30 by fewer reads. In total, 1327 reads from the mature transcriptome were mapped to homeobox gene transcripts, and 1609 reads from the regenerating transcriptome.

These numbers consistently indicate, insofar as un-replicated read counts of predicted transcripts can be taken to be reliable (see section 4.4.1), a modest increase in overall homeobox gene activity during regeneration. The potential roles in of a variety of key homeobox genes in amphioxus regeneration are discussed in detail in Chapter 6.

4.4.4. A new *Pax3/7* paralogue

One sequence from the *de novo* transcriptome assembly (isotig29738) was identified as a *Pax3/7* gene (Table 4.1) but in the C-terminus more closely resembled the sequence of the published *B. belcheri Pax3/7* than the *B. branchiostoma* or *B. floridae* sequences. This sequence is the subject of a detailed investigation in Chapter 5.

4.4.5. Future work

4.4.5.1. Transcriptome reanalysis

Given the several failings of the predicted gene transcript set (see Appendix 4.1), perhaps more useful would be the construction of a manually curated set of *B. lanceolatum* homeobox gene transcript models. Given that many homeobox genes are well profiled in one or more of the *Branchiostoma* model species, producing putative *B. lanceolatum* gene models should not be particularly onerous (at least for the coding sequence). This was not performed herein because of time constraints. The mapping pipeline previously used (gsMapper) eliminates reads that match more than one query, so it may be advisable to replace the predicted transcripts in Appendix 4.1 with the curated models and continue with this ‘recombinant’ database.

4.4.5.2. Improvements to transcriptome data

The transcriptomes analysed herein are suitable to suggest candidate homeobox genes by their presence/absence, but are not suitable for a rigorous analysis of differential gene expression because of the lack of replicates (section 4.4.1). To achieve this, it would be necessary to have at least three technical replicates of each sample.

The transcriptomes also include only a single regenerative timepoint, whereas greater depth of temporal sampling could allow greater insight into the expression of homeobox genes during different regenerative processes (healing, blastema cell recruitment, blastema patterning, differentiation, etc). Ideally, such a transcriptomic dataset would not omit one of mature tissue, so that comparisons could be made to tissue not undergoing regeneration. Finally, 454 pyrosequencing produces many fewer reads than other next-gen sequencing technologies, which is evident in a comparison of the *B. lanceolatum* and *S. lamarcki* datasets (see Chapter 6). Future improvements to transcriptomic technologies (discussed in Chapter 6) will also be a boon to the study of homeobox genes in regeneration.

4.4.5.3. Questions arising from the genome

Several homeobox genes were found to match more than one predicted model (Appendix 4.1), including *Hox10*, *Lbx*, *Lhx2/9b*, *Pou4*, *Alx*, *Pax3/7*, *Pax4/6*, and *Zfhx*. It would be worthwhile to check whether these models are from genuinely separate loci and represent hitherto undescribed gene duplications or are just a product of inaccurate haplotype merging or transcript prediction. Although the investigation was pursued because of the pre-existing interest in *Pax3/7* in regeneration (Somorjai *et al.*, 2012) rather than the presence of multiple genomic transcript models, *Pax3/7* (which is represented by three gene models) was found to have duplicated in the cephalochordate ancestor (see Chapter 5).

There are also three similar predicted transcripts (BL08673, BL12473, & BL01804, automatically annotated as Hmbox1) which were putatively placed in the HNF clade based on the predicted presence of an HNF domain in two of the transcripts. These genes have been referred to as Hmbox-like herein because of their apparent relation based on BLAST searches and annotation (section 4.3.2.1). These putative genes require manual inspection of their loci, curation of their models, searches of the available amphioxus (and broader deuterostome) genomes, phylogenetic work to determine their relationship to known HNF genes, and eventually could profit from more detailed work (*e.g. in situ* hybridisation).

The retrieval of homeobox gene transcript models from the genome was done by BLAST searching for known genes, and no attempt was made to detect and retrieve previously-undescribed homeobox genes in the predicted transcripts. Some (*i.e.* the Hmbox-like genes, *c.f.* above, section 4.3.2.1) were retrieved because of their automatic annotation. However, it might be productive to do search for all homeodomain sequences in a version of the predicted transcript database or whole genome from which the ones identified in Appendix 4.1 had been eliminated to ensure that there are not any more previously undescribed *B. lanceolatum* homeobox genes.

4.4.5.4. Candidate homeobox genes for future study

The genes identified in Table 4.4 may be taken as a pool from which to draw potential candidate genes for future work. Among the particularly interesting genes for future work are the Hox (*Hox3*, *Hox4*, *Hox5*, *Hox7*, *Hox11*, *Hox12*, *Hox13*, and *Hox14*) and ParaHox genes (*Cdx*), and other genes previously found to have roles in regeneration in other systems, including *Emx* (Monaghan *et al.*, 2012), *Msx* (Echeverri and Tanaka 2002; Somorjai *et al.*, 2012), *Pax3/7* (Konstantinides and Averof 2014; Somorjai *et al.*, 2012; Tanaka *et al.*, 2016; Wang and Simon 2016), *Prrx* (Lehrberg and Gardiner 2015; Satoh *et al.*, 2011), and *Meis* (Mercader, Tanaka, and Torres 2005). The known roles of these genes in regeneration, and their potential relevance to ongoing studies of cephalochordate tail regeneration, are discussed in Chapter 6.

However, the absence of detection of a gene from this list should not necessarily exclude it from being a candidate gene; the current transcriptome only measures mature and 14 dpa tissue, leaving a substantial period of early regeneration unsampled. Moreover, overall expression level may be a misleading proxy for the roles of transcription factors in biological tissues; for example, a ubiquitously-expressed gene in mature tissue could switch to a highly specific, localised spatiotemporal pattern with fewer transcripts in regeneration, which could register as 'downregulation' while missing crucial biological information (see section 6.2 & Figure 6.7). Genes of interest like *Pou2*, which was not detected in either sample or dataset but has ancient roles in reprogramming cells to a pluripotent state (Tapia *et al.*, 2012), should still be considered for future work.

Future avenues of study could start with semi-quantitative real-time PCR (qPCR) profiling of expression levels in the mature tissue and regenerating time-points to confirm the putative expression patterns seen in the transcriptomic data. Mature and regenerating cDNA material has been collected for this purpose and used for pilot experiments (Dailey 2017). These could be followed up with *in situ* hybridization experiments on intact and sectioned regenerating tails. Future experiments requiring methodological validation could include pharmacological manipulation of signalling pathways and, if the technique were to become accessible, RNA interference of gene expression. These potential avenues of research are more comprehensively discussed in section 5.4.3.

4.5. Conclusions

Homeobox transcription factors are a key part of the regulatory networks that orchestrate regeneration. The deep conservation of homeobox genes, resulting in detectable orthology throughout the Bilateria and beyond, makes them ideal for understanding how these networks and the genes used in them evolve. Here, I have identified the homeobox gene component of transcriptomes of mature and regenerating *B. lanceolatum* post-anal tail tissue and found diverse and extensive deployment of these genes. Two methods were employed; one, a *de novo* assembly, produced fewer genes but are rigorously identified; the second, a read map of the transcriptomes against predicted transcripts from the genome, is substantially more sensitive but less reliable. These results indicate a modest increase in overall homeobox gene activity in regenerating tissue, consistent with deployment in roles controlling regeneration. A significant finding was the presence of a previously undescribed *Pax3/7* paralogue, which is studied in detail in Chapter 5. This analysis can recommend candidate genes for future studies on homeobox gene function in regeneration, a selection of which are discussed in depth in Chapter 6.

5. *Pax3/7* duplication in cephalochordates

The work presented in this chapter has been published in *Scientific Reports* under the title "*Pax3/7 duplicated and diverged independently in amphioxus, the basal chordate lineage,*" by Barton-Owen, Ferrier, and Somorjai (2018). The final manuscript as submitted is included in Appendix 5.5. This chapter includes data that were omitted from the publication. To avoid unnecessary paraphrasing, some sections of text from the published manuscript are reused herein.

5.1. Introduction

The *Pax3/7/D* gene family, also called Pax group III genes, are highly conserved transcription factors belonging to both the paired and homeobox superfamilies, and within these to the Pax homology group of the PRD homeobox class (Holland, Booth, and Bruford 2007). The family was present in a single copy in the eumetazoan, bilaterian, and chordate ancestors, but has undergone several known clade-specific duplications. It is present as *Pax-D* in cnidarians (1-4 copies); *paired*, *gooseberry*, and *gooseberry-neuro* in arthropods; *Pax3/7* in lophotrochozoans (1-2 copies) and tunicates; and *Pax3* and *Pax7* ohnologues in vertebrates. No echinoderm *Pax3/7/D* orthologue has yet been identified (Howard-Ashby *et al.*, 2006).

Pax3/7 genes contain four domains. These comprise (in order from the N-terminus), the paired domain, a 126 residue DNA-binding domain named after the *Drosophila Pax3/7* orthologue *paired* (Bopp *et al.*, 1986; Treisman, Harris, and Desplan 1991) but found throughout the Metazoa (Breitling and Gerber 2000; Vorobyov and Horst 2006) and common to almost all PRD class homeobox genes. Also present is an Engrailed Homology 1 (EH1) motif (Smith and Jaynes 1996) (a.k.a the octopeptide), which is a short Groucho-interacting sequence (Tolkunova *et al.*, 1998) found in many metazoan transcription factor superfamilies (Copley 2005), possibly as the result of convergent sequence evolution (Shimeld 1997). C-terminal to the EH1 motif is the paired-type homeodomain, and, after a substantial, multi-exon linker region, the 21 residue Paired-type Homeodomain Tail (PHT) (Vorobyov and Horst 2006). In terms of domain structure and completeness, *Pax3/7* is the most similar amongst the Pax genes of extant animals to the hypothetical Proto-Pax.

The *Pax3/7* family has ancient and deeply conserved roles in bilaterian nervous system development, as observed in arthropods (Davis, D'Alessio, and Patel 2005), spiralians (Seaver *et al.*, 2012; Navet *et al.*, 2017), tunicates (Wada *et al.*, 1997) and vertebrates (Thompson and Ziman 2011; Monsoro-Burq 2015); specifically, in the specification of lateral neural plate borders (Li *et al.*, 2017). The family may also have ancient roles in bilaterian myogenesis, a process in which they act in vertebrates (Buckingham and Relaix 2015), nematodes (Yi, Bumbarger, and Sommer 2009), and crustaceans (Konstantinides and Averof 2014) but not in insects or spiralians. *Pax3/7* genes have also acquired important lineage-specific functions, including as paired-rule genes in arthropod segmentation (Davis, D'Alessio, and Patel 2005).

5.1.1. *Pax3/7* in vertebrates

The vertebrate *Pax3/7* paralogues*, Pax3* and *Pax7,* are the result of the 2 rounds of WGD in the vertebrate ancestor. *Pax3* and *Pax7* are highly conserved in sequence. They perform similar and partially overlapping roles in somitogenesis and the development of the neural plate, tube, and crest, which diverge as development proceeds (Frédéric Relaix *et al.*, 2004; Thompson *et al.*, 2008). They regulate one another interdependently, although the precise configuration is variable between vertebrate taxa (Maczkowiak *et al.*, 2010; Agoston *et al.*, 2012). Their functions are more obviously divergent in embryonic and adult muscle development; Pax3 interacts with fewer transcription regulating sites than Pax7 by a factor of ten, and interacts with homeodomain motifs with a lower affinity (Soleimani *et al.*, 2012). *Pax3* and *Pax7* expression also confers different properties to satellite cells (Yang *et al.*, 2016).

5.1.1.1. *Pax3/7* in vertebrate nervous system development

The neurogenic roles of *Pax3* and *Pax7* have been extensively studied in vertebrates (reviewed by Meulemans and Bronner-Fraser 2004; Sauka-Spengler and Bronner-Fraser 2006, 2008; Holland 2009; Thompson and Ziman 2011; Monsoro-Burq 2015). In vertebrate neurulation, a region of the dorsal ectoderm flattens, thickens, and differentiates into neuroectoderm, becoming the neural plate. In the intermediate levels of BMP found at the lateral edges of the neural plate, Wnt and Fgf signalling (reviewed by Groves and LaBonne 2014) acts on *Pax3*, *Msx1,* and *Zic1*, which specify a region of tissue called the neural plate border. The expression of orthologues of these genes at the lateral neural borders is probably anciently conserved throughout the bilaterians (Li *et al.*, 2017). The neural plate folds laterally to bring the neural plate borders in contact with each other, forming the neural tube, and the non-neural ectoderm closes dorsally over the merged neural plate border tissue (Figure 5.1).

The olfactorian ancestor evolved the capacity to produce a novel migratory neural progenitor from the neural plate border (Abitua *et al.*, 2012; Stolfi *et al.*, 2015). In the vertebrate lineage, perhaps aided by WGD events in its recent past, these cells gained pluripotency. These migratory, pluripotent neural crest cells undergo an epithelial-mesenchymal transition, travel throughout the body, and produce a wide variety of cell types including the glia and neurons of the autonomic and sensory nervous systems, a substantial portion of facial cartilage and bone, and pigment, neurosecretory, smooth muscle, and other mesenchymal cells (Schlosser 2008; Groves and LaBonne 2014). Neural crest formation is dependent on *Pax3* and *Pax7* (Monsoro-Burq, Wang, and Harland 2005; Basch, Bronner-Fraser, and García-Castro 2006; Hong and Saint-Jeannet 2007). Migrating cells continue to require *Pax3*/*Pax7* expression, which also controls exit from their proliferative pluripotent state (reviewed by Monsoro-Burq 2015).

Vertebrates also derive another, independent population of migratory progenitor cells from the neural plate border; the cranial placodes, which produce various sensory cells, neurons, glial cells and secretory and neurosecretory cells, including the lateral line system in non-amniote vertebrates, accessory sensory structures including the lens, and the inner ear. Like the neural crest, the specification of the preplacodial ectoderm also requires *Pax3* (Hong and Saint-Jeannet 2007), and placode-derived cells continue to express *Pax3* after fate specification (Baker, Stark, and Bronner-Fraser 2002).

5.1.1.2. *Pax3/7* in vertebrate somitogenesis and myogenesis

Pax3/7 genes also have important and essential roles in vertebrate myogenesis and muscle regeneration, another role that may predate the protostome/deuterostome split (Konstantinides and Averof 2014).

In vertebrates, dorsal paraxial mesoderm either side of the nascent notochord is partitioned progressively into somites, repetitive anteroposterior segments of epithelial tissue. *Pax3* expression in the paraxial mesoderm predates somitogenesis and persists into the somites. As the somites mature, they become partitioned into lineage compartments, and *Pax3* expression becomes restricted to the dermomyotome in the dorsal somite (Hammond *et al.*, 2007; Magli *et al.*, 2013), which gives rise to dorsal muscle and dermis, vascular cells, and brown fat (Buckingham 2017).

Pax3+ and *Pax3*+/*Pax7*+ cells in the dermomyotome form a migratory, proliferative, multipotent progenitor cell population that produces the developing skeletal muscle in the trunk and limbs. *Pax3* and *Pax7* are critical regulators of the GRN that controls the survival, migration, proliferative self-renewal, and cell fate of the progenitor cells. Down-regulation is associated with a cell cycle exit and the onset of myogenic terminal differentiation (reviewed by Buckingham 2007; Buckingham and Relaix 2007; Buckingham and Vincent 2009; Buckingham and Relaix 2015; Buckingham 2017).

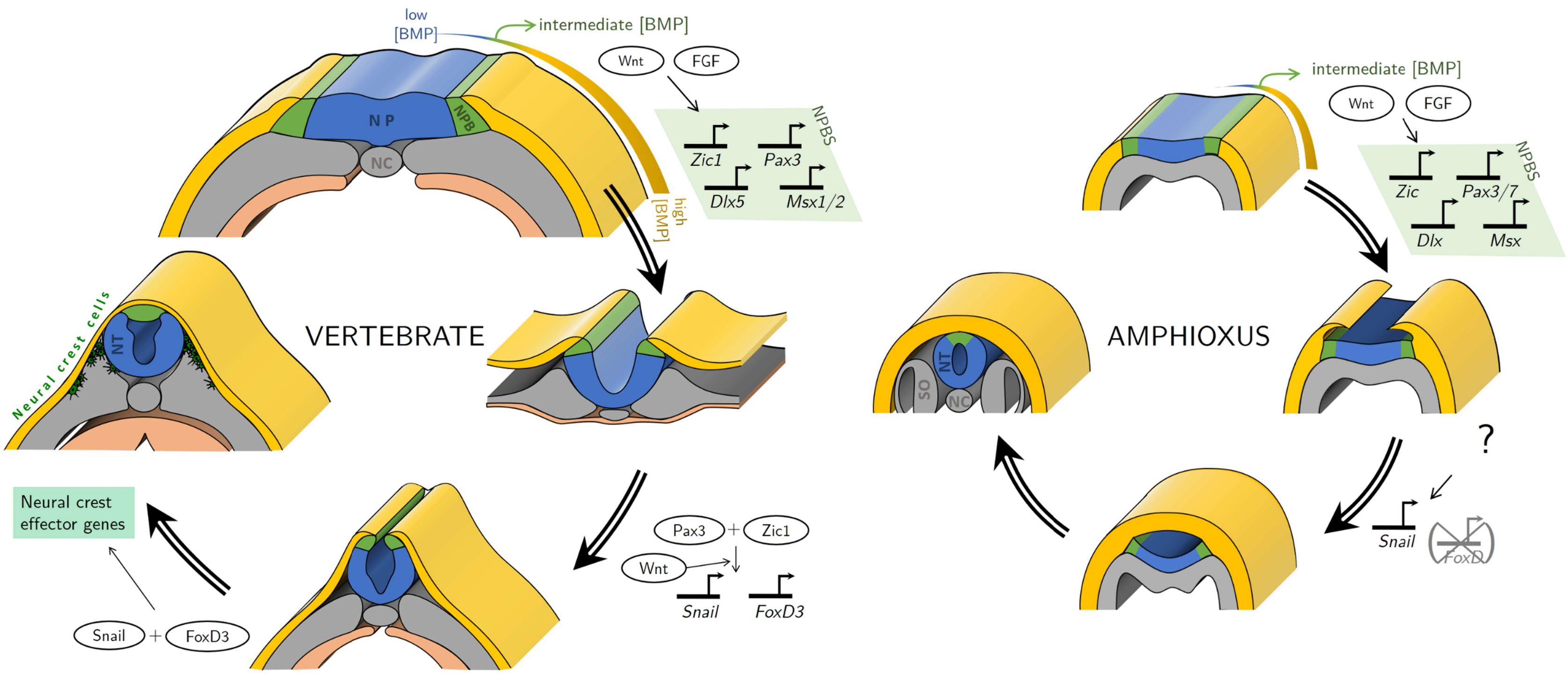


Figure 5.1. Illustration of the neurulation in vertebrates and amphioxus. Epidermal tissue is indicated in yellow; neural plate tissue is indicated in blue; neural plate border tissue is indicated in green. Mesoderm is indicated in grey; endoderm in peach. NP = neural plate; NPB = neural plate border; NPBS = neural plate border specifiers. Adapted from Meulemans & Bronner-Fraser, 2004; Sauka-Spengler & Bronner-Fraser, 2008 and Holland, Vaudet & Schubert, 2004, using information from Yu *et al.*, 2008.

Pax3/7 ***in regeneration***

In late development, a portion of the *Pax3*/*Pax7*-positive progenitor cells take up a position on muscle fibres, and become satellite cells which reside quiescently in adult muscle tissue (Relaix *et al.*, 2005; Chen, Lin, and Slack 2006; Morrison *et al.*, 2006), maintained by *Pax7* expression and with heterogeneous behaviour modified by *Pax3* (Yang *et al.*, 2016). In mammals these cells lie between the basal lamina and the myofiber sarcolemma (Yin, Price, and Rudnicki 2013; Thomas, Engler, and Meyer 2015), but in urodele amphibians (the only tetrapods capable of adult regeneration) satellite cells are separately partitioned (Morrison *et al.*, 2006). Muscle injury or growth induces this quiescent population into activity in the form of self-renewal and the production of myogenic progenitors which can repair existing damaged myofibers or make new ones; in this regenerative role, they are indispensable (Lepper, Partridge, and Fan 2011; Sambasivan *et al.*, 2011). In the event of major tissue excision in vertebrates with regenerative capacity, *Pax7+* satellite cells proliferate, migrate and contribute to the regenerative blastema of larval urodeles (Morrison *et al.*, 2006; Morrison, Borg, and Simon 2009) and neotenic adults but not metamorphotic adults, which utilise myotube dedifferentiation instead (Sandoval-Guzmán *et al.*, 2014; Tanaka *et al.*, 2016).

5.1.2. ***Pax3/7*** **in cephalochordates**

AmphiPax3/7 was first described in *B. floridae* in 1999 by Holland *et al.*, and its significance as an outgroup for the study of vertebrate neurogenic and myogenic evolution was discussed from the outset. *AmphiPax3/7* has since been studied in numerous contexts and several species (Wang *et al.*, 2005; Kozmik *et al.*, 2007; Short and Holland 2008; Somorjai *et al.*, 2008; Yu *et al.*, 2008; Chen *et al.*, 2010; Wang, Zhong, and Wang 2010; Somorjai *et al.*, 2012; Paixão-Côrtes, Salzano, and Bortolini 2013; Kaji *et al.*, 2016). It has been described as appearing in expression domains consistent with sharing a set of ancestral functional roles with vertebrates; most significantly, expression in the neural plate border at the onset of neurulation (Figure 5.1), in the somites, in the later development of the nervous system and larval musculature (Holland *et al.*, 1999; Kozmik *et al.*, 2007), and a persistently high level of expression in the adult segmental muscles (Chen *et al.*, 2010). This adult muscle expression is probably due to a population of *Pax3/7*+ cells that reside peripherally between the muscle and the basal lamina, which are enriched in the regenerative blastema and may play an important role in the regenerative capacity of amphioxus; smaller and more regeneration-competent animals show greater Pax3/7 expression. This population of satellite-like cells is very probably homologous to vertebrate myogenic satellite cells (Somorjai *et al.*, 2012).

AmphiPax3/7 has been described in *B. lanceolatum* (Somorjai *et al.*, 2008), where its expression was found to be similar to that in *B. floridae. AmphiPax3/7* was also described in *B. belcheri* (Wang *et al.*, 2005), and it was noted that the orthologue belonging to *B. belcheri* has a divergent 3'/C-terminal sequence (Wang *et al.*, 2010). qPCR was used to measure *AmphiPax3/7* expression in a variety of different *B. belcheri* developmental stages and adult organs (Chen *et al.*, 2010). Studies have also established the expression of *AmphiPax3/7* as part of the conserved *Pax-Six-Eya-Dach* network in cerebral vesicle development (Kozmik *et al.*, 2007). Five alternative splicing variants of *AmphiPax3/7* were described in *B. floridae* (Short & Holland, 2008).

5.1.3. Aims

In the course of surveying the homeobox gene content of the tail regeneration transcriptome of *B. lanceolatum* (Chapter 4), I discovered a sequence with strong identity to the C-terminal sequence of the previously reported (and hitherto thought to be divergent) *B. belcheri* Pax3/7 protein (Wang *et al.*, 2005), and which did not resemble the C-terminal sequence of the previously-reported *B. floridae* or *B. lanceolatum* Pax3/7 orthologues (Holland *et al.*, 1999; Somorjai *et al.*, 2008). Further investigation in the available *Branchiostoma* genomes revealed that two *Pax3/7* paralogues are present in cephalochordates. In this chapter, I aimed to describe the differential evolution of the paralogues, including their loci, protein sequence, and developmental expression patterns in *Branchiostoma* and *Asymmetron lucayanum.*

5.2. Methods

During the homeobox gene content survey of the *B. lanceolatum* regeneration transcriptome (Chapter 4), the *Pax3/7* contig was observed to be more similar to the previously described divergent *B. belcheri Pax3/7* orthologue than either the *B. floridae* or *B. lanceolatum Pax3/7* orthologue. Subsequent searches of available *Branchiostoma* genome assemblies (see Appendix 2.1) indicated that *Pax3/7* is present as a pair of paralogues in *Branchiostoma* sp. The first, the paralogue discovered earliest in *B. floridae* (and subsequently in *B. lanceolatum*), was dubbed *Pax3/7a*, and the other, the paralogue discovered later in *B. belcheri*, was dubbed *Pax3/7b*. Searches were executed with BLAST and results aligned with MAFFT and inspected and manually edited in Jalview. Full details of the tools used are presented in section 2.3.1, Table 2.1.

Transcriptomic evidence was sought using BLAST searches of available amphioxus transcriptomes and SRAs (detailed in Appendix 2.1). A manual assembly of *Asymmetron lucayanum Pax3/7* genes was performed using data from a sequence read archive (accession number SRX437623) on the basis of BLAST searches using *Branchiostoma Pax3/7* genes. The available *A. lucayanum* transcriptomic and genomic data was not adequate to construct a complete *A. lucayanum Pax3/7b* sequence.

Previous descriptions of *B. floridae AmphiPax3/7* (*i.e. Pax3/7a*) (Holland *et al.*, 1999) had indicated the presence of a start codon at the beginning of the PRD-domain containing exon. Although this exon is in parts identical between *Pax3/7a* and *Pax3/7b*, the genomic assemblies do not support the presence of a start codon in this position. Therefore, a BLAST search of unassembled transcriptomic reads was undertaken, using the region 5' to the beginning of the exon as a query. This was used to identify the presence of an 'exon 0' in *Pax3/7a*. Details of exon and domain boundaries of other chordate *Pax3/7* orthologues were taken from the gene entries in the NCBI and/or predicted using the NCBI Conserved Domain Search (Marchler-Bauer *et al.*, 2011, 2015). These data (Appendix 5.3) were compared to determine if *Pax3/7a*'s extra exon was the result of an exon gain in *Pax3/7a* or an exon loss in *Pax3/7b*.

Protein sequences of Pax3/7 orthologues from *C. teleta, C. gigas, T. castaneum, Saccoglossus kowalevskii, Halocynthia roretzi, Ciona intestinalis, Petromyzon marinus, Scyliorhinus torazame, Danio rerio, Python bivittatus, Gallus gallus, Mus musculus,* and *Homo sapiens* were retrieved from the NCBI database (accession numbers given in Appendix 5.4a). The full sequences were aligned against the *Branchiostoma* and *Asymmetron* Pax3/7 orthologues using MAFFT (Katoh and Standley 2013) and the alignment viewed and manually edited in Jalview (Waterhouse *et al.*, 2009). The alignment was rooted against *B. floridae* Pax4/6 and is presented in Appendix 5.4b. A phylogenetic analysis of this alignment was made using the methodology described in section 2.4.

A VISTA visualisation of an AVID alignment of the genomic *Pax3/7* locus from the available cephalochordate genomes was performed per section 2.5.1. Previously identified conserved non-coding elements (CNEs) were mapped against the cluster per section 2.5.2. A 'VISTA-like' analysis was performed per section 2.5.3 to compare the *Pax3/7a* locus to the *Pax3/7b* locus. An *in silico* micro-RNA prediction pipeline was performed on predicted *Pax3/7* 3' untranslated regions (UTRs) as described in section 2.5.4.

To visualise developmental gene expression, an *in situ* probe was designed to target the divergent 3' ends of the paralogues. Cloning, probe synthesis, whole mount *in situ* hybridisation and visualisation were performed as described in sections 2.6.5-7 using *B. lanceolatum* and *Asymmetron lucayanum* material (cDNA and embryos) collected per sections 2.6.1-4.

5.3. Results

5.3.1. Cephalochordates have two *Pax3/7* paralogues

The homeobox survey of the transcriptome of *B. lanceolatum* regeneration described in Chapter 4 uncovered a sequence (isotig29738) that was identified on the basis of BLAST searches as belonging to the Pax3/7 family. Alignment of the isotig with previously published AmphiPax3/7 sequences (Holland *et al.*, 1999; Wang *et al.*, 2005; Somorjai *et al.*, 2008) revealed that the sequence more closely resembled the previously described *B. belcheri* Pax3/7 than either the *B. lanceolatum* or *B. floridae* homologue.

Searches of the available *Branchiostoma* genomes (Appendix 2.1) revealed that the Pax3/7 gene family is present as two paralogues; *Pax3/7a*, named as such for being the first to be described (in *B. floridae*, Holland *et al.*, 1999; and subsequently in *B. lanceolatum*, Somorjai *et al.*, 2008), and *Pax3/7b*, subsequently discovered only in *B. belcheri* (W. Wang *et al.*, 2005). *Pax3/7a* and *Pax3/7b* lie adjacent to one another in the same orientation, separated by approximately 10 kilobases (Figure 5.2). Their proximity and generally conserved gene structure indicates that they are probably the result of a small-scale tandem duplication.

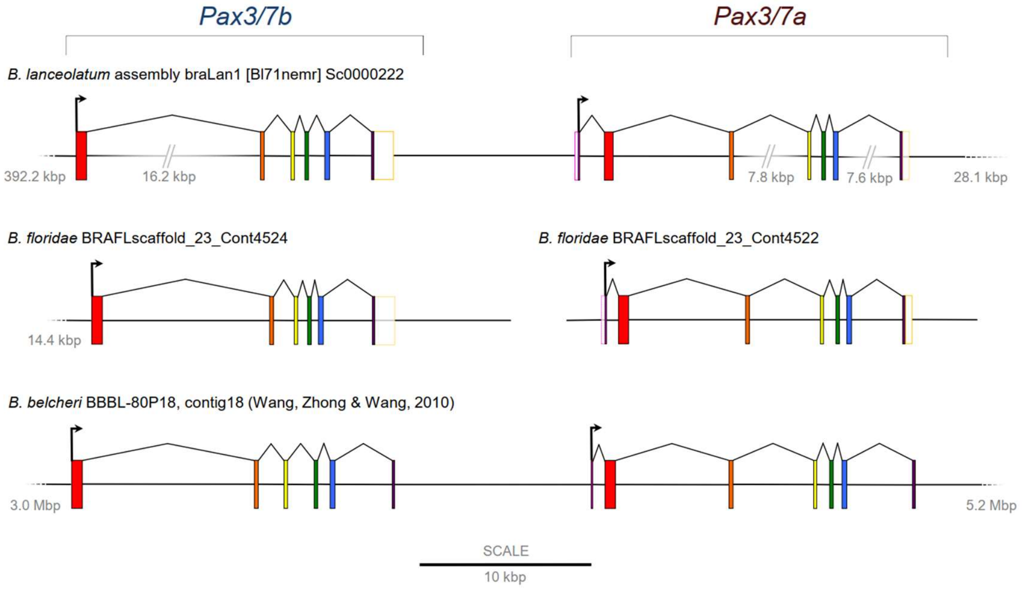

Pax3/7b
Pax3/7a
B. lanceolatum assembly braLan1 [Bl71nemr] Sc0000222
392.2 kbp
16.2 kbp
7.8 kbp
7.6 kbp
28.1 kbp
B. floridae BRAFLscaffold_23_Cont4524
B. floridae BRAFLscaffold_23_Cont4522
14.4 kbp
B. belcheri BBBL-80P18, contig18 (Wang, Zhong & Wang, 2010)
3.0 Mbp
5.2 Mbp
SCALE
10 kbp

Figure 5.2. **Gene models of the *Branchiostoma* Pax3/7 loci**. Exons are marked with coloured boxes with black; introns are represented by angular lines joining the exons. Expressed non-coding regions (indicated by transcriptomic or direct sequencing data) are marked by solid magenta (5') or gold (3') boxes with coloured lines, or semi-transparent boxes and dashed lines where predicted based on homology. Predictions were not performed on *B. belcheri*. Scale bar = 10 kbps.

Searches of an *A. lucayanum* transcriptome and the genome (details in Appendix 2.1) also revealed that *Pax3/7a* and *Pax3/7b* are present in a homologous locus in *Asymmetron,* the earliest-branching extant chordate lineage (Kon *et al.*, 2007; Igawa *et al.*, 2017), indicating the duplication occurred in the common ancestor of all extant cephalochordates. The sixth exon of *Pax3/7b* could not be recovered from the databases used, and the *A. lucayanum* Pax3/7 locus in general suffers from incomplete coverage.

5.3.1.1. The *Pax3/7* locus

An mVISTA analysis of the cephalochordate *Pax3/7* locus reveals high non-coding sequence conservation between species, particularly upstream of the two genes (Figure 5.3a), including >90% identity among the *Branchiostoma* over the majority of the ~74 kbp window shown in Figure 5.2. 84 elements from the *B. floridae*/*A. lucayanum* conserved non-coding element library (Yue *et al.*, 2016) were found within 20 kbps of the cephalochordate *Pax3/7* locus, covering approximately 12% of the non-coding sequence in this window. No CNE from this database was found to reoccur in this window, implying divergence in the *cis*-regulatory landscapes between the two paralogues.

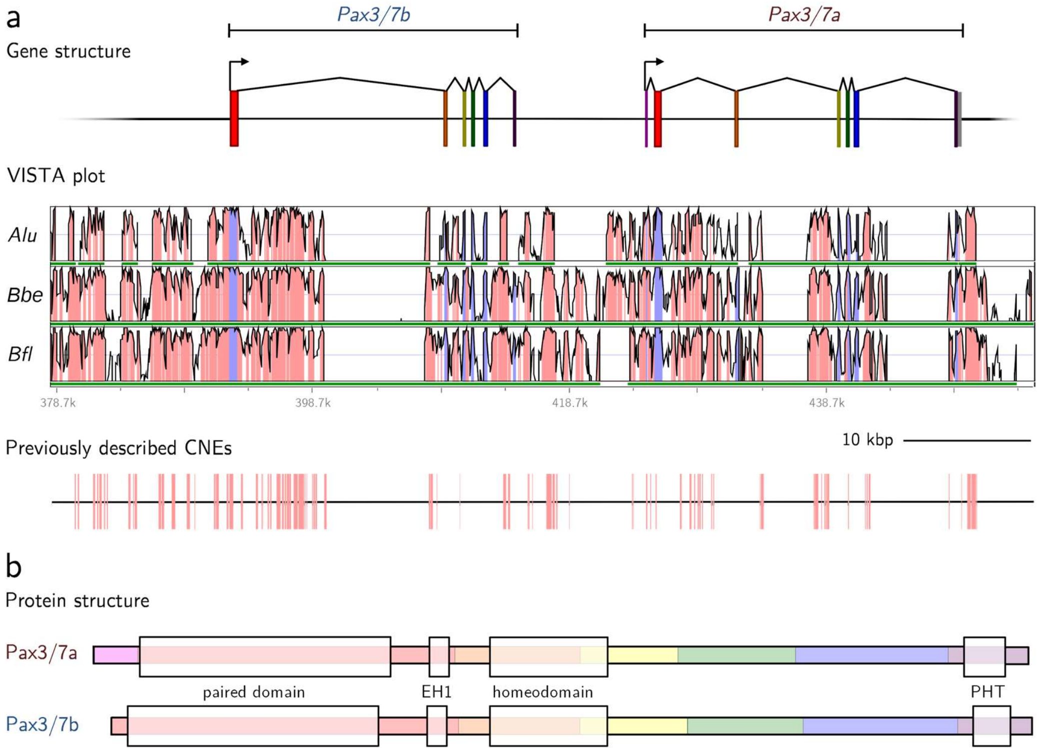


Figure 5.3. **The structure and conservation of the cephalochordate *Pax3/7* genes**. (a) The amphioxus *Pax3/7* locus, showing the gene models (top), a corresponding VISTA plot against other amphioxus species (middle) and a map of *Asymmetron*/*Branchiostoma* conserved non-coding elements from Yue *et al.*, 2016 (bottom) on the *B. lanceolatum* genome scaffold. *Alu* = *A. lucayanum*; *Bbe* = *B. belcheri*; *Bfl* = *B. floridae*. In the VISTA plot, the horizontal axis indicates position on the *B. lanceolatum* genomic scaffold; green rounded bars indicate coverage by the genome of the labelled species, and vertical axis indicates percent identity in a 45bp rolling window, with a range of 50% to 100%. Pink colouration indicates regions exceeding the threshold of 90%, while blue indicates exonic sequence. Coverage is indicated by green bars below the plot. Details of the scaffolds used in the VISTA analysis are reported in Appendix

2.4. Scale bar = 10,000 base pairs. (b) Protein structure of the Pax3/7 paralogues in amphioxus. Each exon is highlighted with a colour corresponding its colour in (a). Conserved domains are indicated with light boxes; the paired domain, the EH1 domain (also known as the octopeptide motif), the homeodomain, and the Paired-type Homeodomain Tail (Vorobyov and Horst 2006). Adapted from Barton-Owen, Ferrier & Somorjai (2018).

The *B. lanceolatum* Pax3/7 locus has large inserts, in the first intron of *Pax3/7b* (~7.6 kbps) and the third (~3.6 kbps) and sixth (~4.7 kbps) introns of *Pax3/7a* relative to all other cephalochordates (visible as gaps in Figure 5.3a). These represent the largest divergence from the prototypical cephalochordate Pax3/7 locus detected with the current data (the *Asymmetron* genomic data do not completely cover the locus).

The **Pax3/7** *loci have almost no non-coding conservation between paralogues*

The *Pax3/7a* and *Pax3/7b* loci were compared using a Python script that used a VISTA-like rolling window to analyse alignments of the *Pax3/7* loci (see section 2.5.3). A MAFFT alignment of 25,000 randomly generated nucleotides produced a mean similarity of 35.9% with a standard deviation of 7.74% in a 75 position rolling window. The VISTA-like plots are presented in Figure 5.4.

No noticeable similarity outside of the expected range was found within the loci except minor spikes just after the second shared exon and before the third (Figure 5.4b). The mean identity of the *Pax3/7* loci alignment (30.8%, SD = 16.8%) was overall lower than the random nucleotide alignment, probably the result of aligning introns of very different lengths.

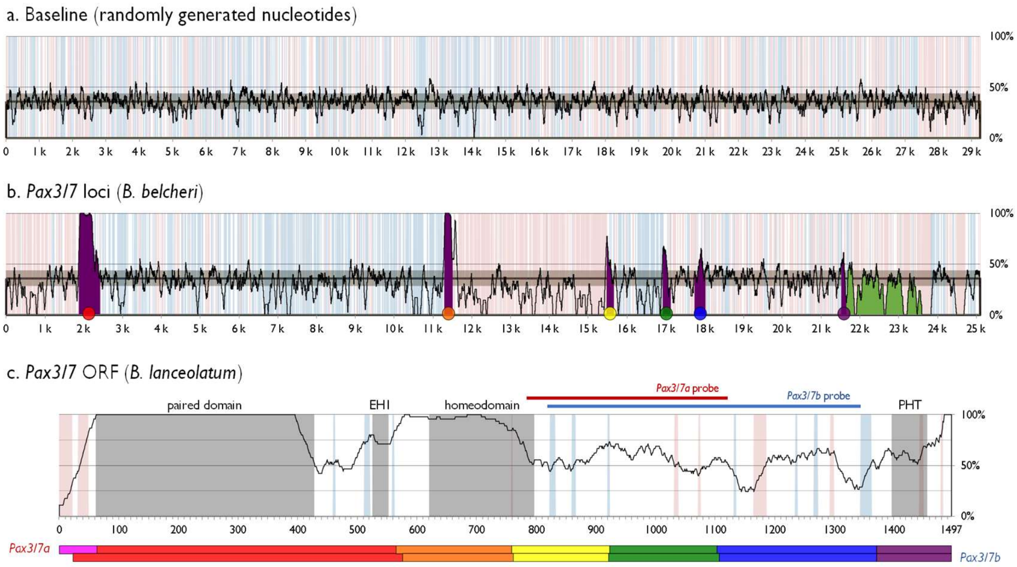


Figure 5.4. **VISTA-like visualisations of similarity of various alignments**. (a) Baseline reference of a MAFFT alignment of 25,000 random nucleotides. Gaps in the alignment are indicated by faint vertical blue and red bands. The mean (35.9%) is indicated with a horizontal line and one standard deviation (7.74%) either side is represented with dark brown shading. (b) Alignment of the *Pax3/7a* and *Pax3/7b* loci of *Branchiostoma belcheri*. Shared exons (*i.e.* not including *Pax3/7a* exon 1) are marked in dark purple, with their relationship to the exons in Figs 5.1 and 5.2 and (c) marked by coloured spots at the base; the longest known extent of 3' UTR sequence is marked in green. The mean ± standard deviation of the random baseline marked as in (a). (c) Alignment of the open reading frames of *Branchiostoma lanceolatum Pax3/7a* and *Pax3/7b*. Domains are marked with dark boxes. The exon structure is indicated below with coloured boxes (corresponding to (b) and Figs. 5.1 and 5.2). The locations of the *Pax3/7a* and *Pax3/7b in situ* hybridisation probes are indicated with coloured bars above. The rolling window length is 75 positions in (a) and (b), and 45 positions in (c). In (b) and (c), blue vertical bars represent gaps in *Pax3/7a* sequence (*i.e.* only *Pax3/7b* covers this region) and red bars represent gaps in *Pax3/7b*. N.B. the identity falling at the 3' end of the domains is an artefact of the rolling window, not because these areas are not identical.

MicroRNA targets in the Pax3/7a *and* Pax3/7b *3' UTRs*

The original *B. floridae Pax3/7(a)* nucleotide sequence (AF165886.1, Holland *et al.*, 1999) has a 3' untranslated region (UTR) 380 nucleotides in length, terminating in a poly-A tail. Unassembled reads (n=545) were retrieved by BLAST search from a *B. floridae* developmental transcriptomic sequence read archive (SRR1952655) and indicate support for the transcription of genomic sequence 3' to the last exon of at least 2,030 nucleotides in length.

The *B. lanceolatum* regenerative transcriptomic contig for *Pax3/7b* (isotig29738) possesses a UTR 1,138 nucleotides in length; a similar length (1,116 nucleotides) is supported by reads (n=219) from the *B. floridae* developmental transcriptomic sequence read archive.

Four *in silico* microRNA target prediction tools (MiRanda, RNAhybrid, PITA and FindTAR) were used in parallel on the predicted 3' UTRs of cephalochordate *Pax3/7*s, after the methodology of Candiani *et al.*, (2011) (section 2.5.4). This analysis was not completed. A summary of target sites predicted by at least three of these tools in the *B. floridae* and *B. lanceolatum* predicted 3' UTRs are presented in Table 5.1 (*Pax3/7a*) and Table 5.2 (*Pax3/7b*).

5.3.1.2. *Pax3/7* paralogue differential evolution

The gene structures of *Pax3/7a* and *Pax3/7b* are presented in Figure 5.3b. Homology is detectible at the nucleotide level between *Pax3/7a* exon 2-7 and *Pax3/7b* exon 1-6; *Pax3/7b* does not have an exon homologous to *Pax3/7a* exon 1. On the basis of a summary comparison of the structures of other deuterostome *Pax3/7* genes (presented in Table 5.3; full version presented in Appendix 5.3), I conclude that the presence of an initial exon that does not contain the start of the paired domain is ancestral to deuterostome and chordate *Pax3/7* orthologues. Therefore, cephalochordate *Pax3/7b* has lost its initial exon.

Table 5.1. **Predicted target sites in the 3' UTR of *Pax3/7a* from *B. floridae* and *B. lanceolatum*.** The analysis was performed on the 3' UTR (380 nucleotides) from a published *B. floridae* sequence (AF165886.1, Holland *et al.*, 1999) and the homologous region of the *B. lanceolatum* genome (392 nucleotides), using miRanda (Enright *et al.*, 2003), RNAhybrid (Rehmsmeier *et al.*, 2004), PITA (Kertesz *et al.*, 2007), and FindTar (Ye *et al.*, 2008). Only equivalent or similar sites predicted by at least three of the tools used are included in this summary. Salient details from the reports of each of the four tools are included in the table. Abbreviations: Q = Query, R = Reference (MiRanda); t = target, m = miRNA (RNAhybrid).

<table>
<tr><th rowspan="2">miRNA</th><th rowspan="2">Species</th><th rowspan="2">Target location</th><th colspan="4">Outputs</th></tr>
<tr><th>MiRanda</th><th>RNAhybrid</th><th>PITA</th><th>FindTAR</th></tr>
<tr><td>bfl-miR-7a</td><td><i>B. floridae</i></td><td>313-337</td><td>Q 3' uaCUAUUUA- - - GUGAUCAGAAGGa 5'
||||:|| :||| ||||||
R 5' at GATAGATTAGTACTTATCTTCCc 3'
Score: 86 | Energy: -20.02 kCal/Mol</td><td>t 5' G AGU UA C 3'
AUGAUAGAUU ACU UCUUCC
UACUAUUUAG UGA AGAAGG
m 3' UC A 5'
mfe: -23.9 kcal/mol | p-value: 0.602393</td><td>-</td><td>UACUAUUUAGUGAU- - - CAGAAGGA
||||||:||:|*||****||||||*
ATGATAGATTAGTACTTATCTTCCC
score: 44 | energy: -22.3</td></tr>
<tr><td>bfl-miR-373</td><td><i>B. lanceolatum</i></td><td>18-50</td><td>-</td><td>t 5' C U C CUCAUCUAA A 3'
UCAUCC CAU GC AAAAUCAAG
AGUAGG GUG UG UUUUAGUUU
m 3' AU U C 5'
mfe: -21.7 kcal/mol | p-value: 0.615001</td><td>Position: 42
Seed: 8:0:1
dGduplex: -16.2
dGopen: -1.92
ddG: -14.27</td><td>AUAGUAGGGU- GUUGUUUUAGUUUC
|||:|:||*|*|*|||||||||:*
CATCGCCTCATCTAAAAAATCAAGA
score: 34 | energy: -20.3</td></tr>
<tr><td rowspan="2">bfl-miR-4026-5p-1</td><td><i>B. floridae</i></td><td>253-280</td><td>Q 3' acugcAGUAGUGA- - AUGGU- ACUGu 5'
||||||:| ||||| ||||
R 5' gaaaaTCATCATTAATACCAGTGACa 3'
Score: 90 | Energy: -22.00 kCal/Mol</td><td>t 5' G UG AAAA AA G C 3'
G G UCAUCAUU UACCA UGACA
C C AGUAGUGA AUGGU ACUGU
m 3' A UG 5'
mfe: -27.8 kcal/mol | p-value: 0.583331</td><td>-</td><td>AC- - UGCAGUAGUGA- - AU- GGUACUGU
||**|**||||||:|**||*|*:|||||
TGGAAAATCATCATTAATACCAGTGACA
score: 28 | energy: -24.4</td></tr>
<tr><td><i>B. lanceolatum</i></td><td>267-293</td><td>Q 3' acugcAGUAGUGA- - AUGGU- ACUGu 5'
||||||:| ||||| ||||
R 5' gaaaaTCATCATTAATACCAGTGACa 3'
Score: 90 | Energy: -22.00 kCal/Mol</td><td>t 5' A AGGUGGAAAA AA G C 3'
GACG UCAUCAUU UACCA UGACA
CUGC AGUAGUGA AUGGU ACUGU
m 3' A 5'
mfe: -28.8 kcal/mol | p-value: 0.578804</td><td>-</td><td>AC- - UGCAGUAGUGA- - AU- GGUACUGU
||**|**||||||:|**||*|*:|||||
TGGAAAATCATCATTAATACCAGTGACA
score: 28 | energy: -24.4</td></tr>
<tr><td>bfl-miR-4028-3p-1</td><td><i>B. floridae</i></td><td>8-26</td><td>-</td><td>t 5' U A CU C 3'
GC AGCAAC CAUCAU
CG UUGUUG GUAGUA
m 3' G UU U 5'
mfe: -20.2 kcal/mol | p-value: 0.618871</td><td>Position: 15
Seed: 8:1:0
dGduplex: -12.2
dGopen: -2.06
ddG: -10.13</td><td>GCGUU- GUUGUUGUAGUAU
:||||*||||**||||||*
TGCAAGCAACCTCATCATC
score: 31 | energy: -20.1</td></tr>
<tr><td>bfl-miR-4028-3p-2</td><td><i>B. floridae</i></td><td>7-29</td><td>Q 3' ggaGUU- GUUGUUGUAGUAAGGa 5'
||| |||| |||||| |:
R 5' gt gCAAGCAACCTCATCATCCTc 3'
Score: 77 | Energy: -20.51</td><td>t 5' U GUGCA CU C 3'
CC AGCAAC CAUCAU CCU
GG UUGUUG GUAGUA GGA
m 3' AG UU A 5'
mfe: -22.4 kcal/mol | p-value: 0.610014</td><td>Position: 21
Seed: 8:1:1
dGduplex: -14.9
dGopen: -3.1
ddG: -11.79</td><td>-</td></tr>
<tr><td>bfl-miR-4108-5p</td><td><i>B. floridae</i></td><td>23-40</td><td>Q 3' ucAGGAGAUGAAGUGUGu 5'
||||||: ||||:|
R 5' caTCCTCTGAATCACGCc 3'
Score: 74 | Energy: -20.02 kCal/Mol</td><td>t 5' A AA C 3'
UCCUCUG UCACGC
AGGAGAU AGUGUG
m 3' UC GA U 5'
mfe: -23.8 kcal/mol | p-value: 0.600155</td><td>-</td><td>UCAGGAGAUGAAGUGUGU
|*||||||:**||||:|*
A- TCCTCTGAATCACGCC
score: 26 | energy: -23.5</td></tr>
<tr><td>bfl-miR-4121-3p</td><td><i>B. lanceolatum</i></td><td>333-355</td><td>-</td><td>t 5' A C UA A 3'
AUUAG ACU UAAUUCCAAC
UAGUU UGG GUUGAGGUUG
m 3' AA 5'
mfe: -20.0 kcal/mol | p-value: 0.621498</td><td>Position: 347
Seed: 8:1:1
dGduplex: -14.4
dGopen: -3.1
ddG: -11.29</td><td>UAGUU- UGGGU- - UGAGGUUGAA
||:|:*||::|**|:|||||||**
ATTAGCACTTATAATTCCAACAG
score: 27 | energy: -18.2</td></tr>
<tr><td>bfl-miR-4155</td><td><i>B. floridae</i></td><td>352-372</td><td>-</td><td>t 5' U C A 3'
GUGGAAACAG CAAAAAA
CGUCUUUGUC GUUUUUU
m 3' UU AC 5'
mfe: -20.4 kcal/mol | p-value: 0.619652</td><td>Position: 364
Seed: 8:1:0
dGduplex: -15.9
dGopen: -4.02
ddG: -11.87</td><td>CGUCUUUGUC- - - UUGUUUUUUAC
|::||||||||***||*|||||||**
GTGGAAACAGCCAAAAAAAAAAAAA
score: 39 | energy: -20.0</td></tr>
</table>

Table 5.2. Predicted target sites in the 3' UTR of *Pax3/7b* from *B. floridae* and *B. lanceolatum*. The analysis was performed on the 3' UTR (1,138 nucleotides) from the *B. lanceolatum* regenerative transcriptome (isotig29738) and the homologous region of the *B. floridae* genome (1,115 nucleotides), using miRanda (Enright *et al.*, 2003), RNAhybrid (Rehmsmeier *et al.*, 2004), PITA (Kertesz *et al.*, 2007), and FindTar (Ye *et al.*, 2008). Only equivalent or similar sites predicted by at least three of the tools used are included in this summary. Salient details from the reports of each of the four tools are included in the table. Abbreviations: Q = Query, R = Reference (MiRanda); t = target, m = miRNA (RNAhybrid).

<table>
<tr><th rowspan="2">miRNA</th><th rowspan="2">Species</th><th rowspan="2">Target location</th><th colspan="4">Outputs</th></tr>
<tr><th>MiRanda</th><th>RNAhybrid</th><th>PITA</th><th>FindTAR</th></tr>
<tr><td>bfl-miR-19c</td><td><i>B. floridae</i></td><td>906-940</td><td>Q 3' uuGUCGAACGUCC-------CC----UAAACGUGu 5'
|||| |:||| || | |||:||
R 5' aaCAGCCAGTAGGAAGCGAAGGTACCAATTGTACg 3'
Score: 62 | Energy: -21.98 kCal/Mol</td><td>t 5' A CA AAGCGAA UACCAA U 3'
AACAGC GUAGG GG UUGUACG
UUGUCG CGUCC CC AACGUGU
m 3' AA UA 5'
mfe: -24.5 kcal/mol | p-value: 0.613522</td><td>-</td><td>UUGUCGAACGUCC-------CC----UAAACGUGU
||||||**|:|||*******||****|*|||:||:
AACAGCCAGTAGGAAGCGAAGGTACCAATTGTACG
score: 27 | energy: -27.2</td></tr>
<tr><td>bfl-miR-100</td><td><i>B. floridae</i></td><td>978-1008</td><td>Q 3' ugUGUUC--AAGCC------UAGAUGCCCAa 5'
||||| |||| :|: |:||||
R 5' t aACAAGGCGTCGGCGCCACGTTAATGGGTt 3'
Score: 67 | Energy: -21.93 kCal/Mol</td><td>t 5' A GCG CG CACGUA A 3'
ACAAG UCGG C AUGGGUU
UGUUC AGCC G UGCCCAA
m 3' UG A UA A 5'
mfe: -24.6 kcal/mol | p-value: 0.612652</td><td>-</td><td>UGU-GUUCA--AGCC------UAGAUGCCCAA
|:|*||||***||||******:|:*|:|||||
ATAACAAGGCGTCGGCGCCACGTTAATGGGTT
score: 22 | energy: -22.8</td></tr>
<tr><td></td><td><i>B. lanceolatum</i></td><td>997-1027</td><td>Q 3' ugUGUUC--AAGCC------UAGAUGCCCAa 5'
||||| ||||| :|: |:||||
R 5' t aACAAGCCTTCGGCGCCACGTTAATGGGTt 3'
Score: 75 | Energy: -20.58 kCal/Mol</td><td>t 5' G CC CG CACGUUA A 3'
AUACAAG UUCGG C AUGGGUU
UGUGUUC AAGCC G UGCCCAA
m 3' UA A 5'
mfe: -25.6 kcal/mol | p-value: 0.60837</td><td>-</td><td>UGU-GUUC--AAGCC------UAGAUGCCCAA
|:|*||||**|||||******:|:*|:|||||
ATAACAAGCCTTCGGCGCCACGTTAATGGGTT
score: 30 | energy: -21.6</td></tr>
<tr><td>bfl-miR-430d</td><td><i>B. floridae</i></td><td>83-107</td><td>-</td><td>t 5' A CAA AC A 3'
CAACAA CA AGCAUCAAA
GUUGUU GU UCGUGGUUU
m 3' UUA CC 5'
mfe: -20.6 kcal/mol | p-value: 0.630354</td><td>Position: 99
Seed: 8:1:1
dGduplex: -15.1
dGopen: -1.45
ddG: -13.64</td><td>GUUGUUUUAGUU--CGUGGUUUCC
||||||*|*|||**|||:||||**
CAACAACAACAACAGCATCAAAAT
score: 44 | energy: -19.2</td></tr>
<tr><td>bfl-miR-4017-3p</td><td><i>B. floridae</i></td><td>283-308</td><td>Q 3' agaCUGUGG-UAUGU--ACGUGAAAu 5'
|||||| ||||| ||:|| ||
R 5' gt aGACACCTATACAACTGTACATTa 3'
Score: 88 | Energy: -20.99 kCal/Mol</td><td>t 5' A U AC A 3'
GACACC AUACA UGUAC
CUGUGG UAUGU ACGUG
m 3' AGA AAAU 5'
mfe: -26.0 kcal/mol | p-value: 0.606321</td><td>Position: 300
Seed: 8:1:1
dGduplex: -20.9
dGopen: -5.89
ddG: -15</td><td>-</td></tr>
<tr><td>bfl-miR-4026-5p-1</td><td><i>B. lanceolatum</i></td><td>17-39</td><td>Q 3' acUGCAGUAGUGAAUGGUACugu 5'
:||||| || ||||||
R 5' ggGCGTCAGGACAAACCATGcaa 3'
Score: 79 | Energy: -25.49 kCal/Mol</td><td>t 5' G GG AA A 3'
GGCGUCA AC ACCAUG CA
CUGCAGU UG UGGUAC GU
m 3' A AG AA U 5'
mfe: -27.5 kcal/mol | p-value: 0.599807</td><td>-</td><td>ACUGCAGUA--GU---GAAU-----GGUACUGU
*|:|||||***||***|*|:******:||||:|
GGGCGTCAGGACAAACCATGCAACCATATGATA
score: 7 | energy: -26.9</td></tr>
<tr><td>bfl-miR-4028-3p-1</td><td><i>B. floridae</i></td><td>85-102</td><td>Q 3' gcGUUGUUGUUGUAGUAu 5'
|||||||||||| |||
R 5' aaCAACAACAACAGCATc 3'
Score: 94 | Energy: -21.02 kCal/Mol</td><td>t 5' A A 3'
CAACAACAGCAUCA
GUUGUUGUUGUAGU
m 3' GC AU 5'
mfe: -25.3 kcal/mol | p-value: 0.606655</td><td>Position: 94 97
Seed: 8:1:0 8:1:1
dGduplex: -16.4 -20.7
dGopen: -1.13 -1.03
ddG: -15.26 -19.66</td><td>-</td></tr>
<tr><td>bfl-miR-4028-3p-2</td><td><i>B. lanceolatum</i></td><td>457-474</td><td>-</td><td>t 5' G UUU A 3'
CCUCAAU UCAUUCC
GGAGUUG AGUAAGG
3' UUGUUGU A 5'
mfe: -22.8 kcal/mol | p-value: 0.620563</td><td>Position: 466
Seed: 8:1:0
dGduplex: -17.9
dGopen: -4.59
ddG: -13.3</td><td>GGAGUUGUUGUUGUAGUAAGGA
||||||:**:****|||||||*
CCTCAATTTT----TCATTCCA
score: 14 | energy: -21.8</td></tr>
<tr><td>bfl-miR-4028-5p-7</td><td><i>B. floridae</i></td><td>85-102</td><td>Q 3' uuGUUGUUGUUGUAGUAa 5'
|||||||||||| |||
R 5' aaCAACAACAACAGCATc 3'
Score: 94 | Energy: -22.20 kCal/Mol</td><td>t 5' C A 3'
AACAACAACAGCAUCA
UUGUUGUUGUUGUAGU
m 3' AA 5'
mfe: -27.8 kcal/mol | p-value: 0.595090</td><td>Position: 94 97
Seed: 8:1:0 8:1:1
dGduplex: -18.5 -22.7
dGopen: -1.13 -1.03
ddG: -17.36 -21.66</td><td>-</td></tr>
<tr><td>bfl-miR-4028-5p-9</td><td><i>B. floridae</i></td><td>85-102</td><td>Q 3' gaGUUGUUGUUGUAGUAa 5'
|||||||||||| |||
R 5' aaCAACAACAACAGCATc 3'
Score: 94 | Energy: -20.98 kCal/Mol</td><td>t 5' A A 3'
CAACAACAGCAUCA
GUUGUUGUUGUAGU
m 3' GA AA 5'
mfe: -26.0 kcal/mol | p-value: 0.603414</td><td>Position: 94 97
Seed: 8:1:0 8:1:1
dGduplex: -17.2 -21.4
dGopen: -1.13 -1.03
ddG: -16.06 -20.36</td><td>-</td></tr>
<tr><td>bfl-miR-4099-5p</td><td><i>B. lanceolatum</i></td><td>928-953</td><td>Q 3' uaGGUCUAGC--AG-UACCACGGUUa 5'
|||| | | | | |||||||||
R 5' agCCAGTTGGAAGCGAAGGTGCCAAt 3'
Score: 85 | Energy: -25.59 kCal/Mol</td><td>t 5' G GAAG AA U 3'
CCAG UUG CG GGUGCCAAU
GGUC AGC GU CCACGGUUA
3' UA U A A 5'
mfe: -26.0 kcal/mol | p-value: 0.606567</td><td>-</td><td>UAGGUCUAGCA--G-UACCACGGUUA
|*||||*|*|***|*|*||||||||||
AGCCAGTTGGAAGCGAAGGTGCCAAT
score: 17 | energy: -25.1</td></tr>
<tr><td>bfl-miR-4121-5p</td><td><i>B. floridae</i></td><td>781-799</td><td>Q 3' gaGU-UGGGUUGAGGUUGUg 5'
|| |||||: |||:|||
R 5' acCAGACCCAG-TCCGACAc 3'
Score: 82 | Energy: -24.82 kCal/Mol</td><td>t 5' A U 3'
GACCCAG UCCGACAC
UUGGGUU AGGUUGUG
m 3' GAG G 5'
mfe: -27.5 kcal/mol | p-value: 0.597166</td><td>Position: 791
Seed: 8:1:1
dGduplex: -19.7
dGopen: -5.86
ddG: -13.83</td><td>-</td></tr>
<tr><td></td><td><i>B. lanceolatum</i></td><td>797-819</td><td>Q 3' gaGUUGGGUUG----AGGUUGUg 5'
|:||| :|| |||:|||
R 5' t tCGACCAGACCCTGTCCGACAc 3'
Score: 77 | Energy: -22.47 kCal/Mol</td><td>t 5' U A CCUG U 3'
UCGACC GAC UCCGACAC
AGUUGG UUG AGGUUGUG
m 3' G G 5'
mfe: -26.5 kcal/mol | p-value: 0.602026</td><td>Position: 811
Seed: 8:1:1
dGduplex: -18.2
dGopen: -7.97
ddG: -10.22</td><td>GAGUUGGGUUG----AGGUUGUG
:||:|||*:||****|||:||||
TTCGACCAGACCCTGTCCGACAC
score: 24 | energy: -24.3</td></tr>
</table>

Previous published versions of *Pax3/7a* (*e.g.* EEN66816.1, from Holland *et al.*, 1999) reconstructed the start of the gene as the start of *Pax3/7b*; this was the result of the complete nucleotide identity of the *Pax3/7a* and *Pax3/7b* paired domain (see below) leading the researchers to assume they were from the same gene. The initial exon was recovered by a search of raw unassembled transcriptomes from *B. lanceolatum* (SRR2057056 and SRR2164904) and *A. lucayanum* (SRR1138336), using an excerpt from the genomic sequence 5' of the second *Pax3/7a* exon as a query. The current gene model is supported by 30 reads from *B. lanceolatum* and 86 from *A. lucayanum* that overlap the exon junction.

Table 5.3. **Exon structures of Pax3/7 genes from vertebrates and non-vertebrate deuterostomes**, showing that the presence of an initial exon that does not contain the start of the paired domain (*c.f.* the exon lengths and paired box position columns) is probably ancestral to the deuterostomes. Paired and homeodomain positions are as predicted by the NCBI Conserved Domain Search. *M. musculus* = *Mus musculus*; *H. sapiens* = *Homo sapiens*; *G. gallus* = *Gallus gallus*; *A. mississippiensis* = *Alligator mississippiensis*; *D. rerio* = *Danio rerio*; *C. milii* = *Callorhinchus milii*; *L. japonicum* = *Lethenteron japonicum*; *S. kowalevskii* = *Saccoglossus kowalevskii*; *B. lanceolatum* = *Branchiostoma lanceolatum*; *C. intestinalis* = *Ciona intestinalis*.

<table>
<thead>
<tr><th></th><th>SPECIES</th><th>ACCESSION</th><th>GENE</th><th>ISO-FORM</th><th># EXONS</th><th>EXON LENGTHS, AA</th><th>PRD BOX POS'N</th><th>HOM. DOM. POS'N</th></tr>
</thead>
<tbody>
<tr><td rowspan="7"><i>Pax3</i></td><td><i>M. musculus</i></td><td>AAH48699.1</td><td>3</td><td>b</td><td>9</td><td>28|78|43|45|68|55|71|82|11*</td><td>34-159</td><td>222-275</td></tr>
<tr><td><i>H. sapiens</i></td><td>NP_852124.1</td><td>3</td><td>3</td><td>8</td><td>28|78|43|45|68|55|71|89*</td><td>34-159</td><td>222-275</td></tr>
<tr><td><i>G. gallus</i></td><td>BAB85652.1</td><td>3</td><td>-</td><td>6</td><td>partial chromosome 47|68|55|71|82|11*</td><td>34-161</td><td>222-275</td></tr>
<tr><td><i>A. mississippiensis</i></td><td>XP_006266132.1</td><td>3</td><td>X2</td><td>9</td><td>28|78|43|45|68|55|71|82|11*</td><td>34-159</td><td>222-275</td></tr>
<tr><td><i>D. rerio</i></td><td>NP_571352.1/AAC41253.1</td><td>3a</td><td>-</td><td>8</td><td>28|78|43|48|66|55|73|116*</td><td>34-159</td><td>223-276</td></tr>
<tr><td><i>D. rerio</i></td><td>NP_001315326.1</td><td>3b</td><td>-</td><td>9</td><td>28|78|42|51|66|55|62|73|11*</td><td>34-158</td><td>225-278</td></tr>
<tr><td><i>C. milii</i></td><td>XP_007887991.1</td><td>3</td><td>-</td><td>13</td><td>56|66|58|19|14|77|43|45|66|55|71|82|11*</td><td>220-346</td><td>405-458</td></tr>
<tr><td rowspan="8"><i>Pax7</i></td><td><i>M. musculus</i></td><td>NP_035169.1</td><td>7</td><td>-</td><td>9</td><td>28|78|43|43|66|55|67|82|38*</td><td>34-159</td><td>218-271</td></tr>
<tr><td><i>H. sapiens</i></td><td>NP_002575.1/CAA65522.1</td><td>7</td><td>1</td><td>8</td><td>28|78|43|45|66|55|67|136*</td><td>34-163</td><td>220-273</td></tr>
<tr><td><i>G. gallus</i></td><td>NP_990396.1</td><td>7</td><td>-</td><td>10</td><td>28|78|43|43|66|55|66|22|82|38*</td><td>34-159</td><td>218-271</td></tr>
<tr><td><i>A. mississippiensis</i></td><td>XP_014457309.1</td><td>7</td><td>-</td><td>9</td><td>46|43|43|66|55|66|82|38*</td><td></td><td>157-210</td></tr>
<tr><td><i>D. rerio</i></td><td>XP_009304561.1/CAM12909.1</td><td>7a</td><td>X1</td><td>9</td><td>28|78|43|43|66|55|65|83|38*</td><td>34-161</td><td>218-271</td></tr>
<tr><td><i>D. rerio</i></td><td>NP_001139621.1</td><td>7b</td><td>-</td><td>9</td><td>28|78|52|43|65|55|65|83|38*</td><td>34-170</td><td>226-279</td></tr>
<tr><td><i>C. milii</i></td><td>XP_007896063.1</td><td>7</td><td>-</td><td>9</td><td>28|78|43|43|67|55|58|82|38*</td><td>34-161</td><td>219-272</td></tr>
<tr><td><i>L. japonicum</i></td><td>KE993866.1:1</td><td>7</td><td>-</td><td>9</td><td>18|79|43|49|69|54|68|90|39*</td><td>24-149</td><td>215-268</td></tr>
<tr><td rowspan="4"><i>Pax3/7</i></td><td><i>S. kowalevskii</i></td><td>XP_006823827.1</td><td>3B-I</td><td></td><td>6</td><td>16|162|66|55|51|90*</td><td>17-141</td><td>202-254</td></tr>
<tr><td><b><i>B. lanceolatum</i></b></td><td><b>MF979121</b></td><td><b>3/7a</b></td><td></td><td><b>7</b></td><td>23| 162 |65|49|60|79|42*</td><td><b>24-151</b></td><td><b>207-260</b></td></tr>
<tr><td><b><i>B. lanceolatum</i></b></td><td><b>MF979122</b></td><td><b>3/7b</b></td><td></td><td><b>6</b></td><td>| 178 |62|54|59|79|39*</td><td><b>9-136</b></td><td><b>197-250</b></td></tr>
<tr><td><i>C. intestinalis</i></td><td>XP_018669403.1</td><td>3/7</td><td></td><td>10</td><td>15|55|49|38|75|75|85|95|96|39*</td><td>16-143</td><td>192-245</td></tr>
</tbody>
</table>

Paired domain conservation and conversion

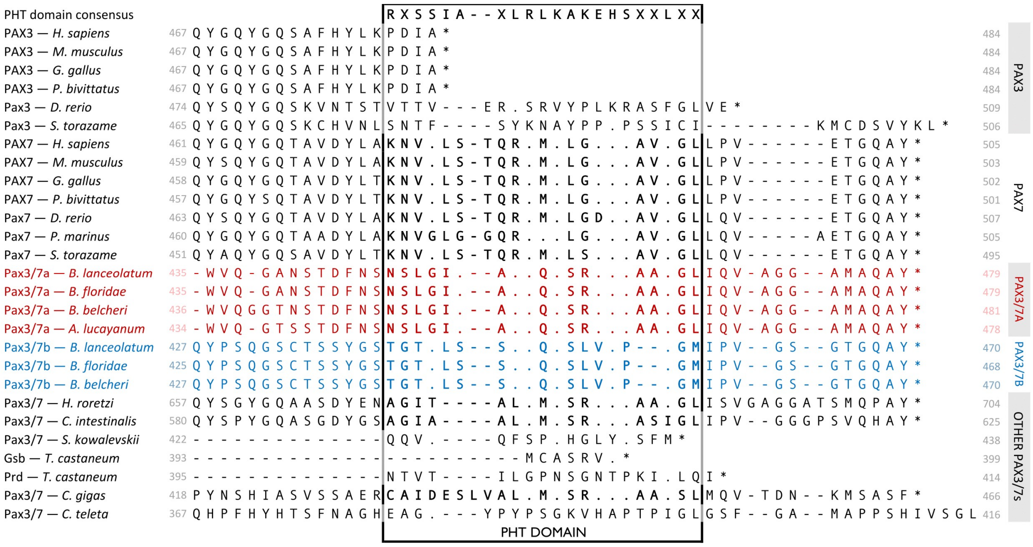
The paired domains of *Pax3/7a* and *Pax3/7b* are almost identical, sharing extremely strong conservation at the nucleotide level (*c.f.* Figure 5.5c). If codons with inconsistencies between the various sources of sequence data (genome assembly, transcriptome or BAC clone) are discounted, no within-species difference exists between the nucleotide sequences of the paired domain *Pax3/7a* and *Pax3/7b*. Two sites in *B. belcheri* and one site in *B. floridae* (highlighted in Appendix 5.1) were found to have within-species similarity between the Pax3/7 paralogues but between species differences with all other (amphioxus) Pax3/7 orthologues; *i.e.* within the paired domain, *B. belcheri Pax3/7a* is more similar to *B. belcheri Pax3/7b* than it is to *B. lanceolatum*/*B. floridae Pax3/7a* and more similar than *B. belcheri Pax3/7b* is to *B. lanceolatum*/*B. floridae Pax3/7b*, and the same has been observed in *B. floridae*.

EH1 domain conservation

The cephalochordate Pax3/7s have an EH1 domain (also called the Octapeptide or TN) which differs in its initial residue only. The Pax3/7a and Pax3/7b homeodomains are identical except for the 60th position, which is glutamine in *Pax3/7a* and alanine in *Pax3/7b*. The homeodomains are flanked by fourteen consecutive identical residues in the N-terminal direction and four in the C-terminal direction. Pax3/7a and Pax3/7b also have a PHT (Paired-type Homeodomain Tail) domain (Vorobyov and Horst 2006). An alignment of the C-termini of a selection of vertebrate, non-vertebrate deuterostome, and protostome Pax3/7 sequences, including the PHT domain, is presented in Figure 5.5. Pax3/7a has a more prototypical PHT domain than Pax3/7b. An amino acid alignment of Pax3/7a and Pax3/7b is presented in Appendix 5.2, highlighting the paired, EH1, homeo- and PHT domains, as well as the location of the *Pax3/7* probes used for *in situ* hybridisation (section 5.3.5).

5.3.1.3. *Pax3/7* duplicated independently in the cephalochordates

The amphioxus Pax3/7 protein sequences were aligned against Pax3/7 sequences from vertebrates (*H. sapiens*, *M. musculus*, *G. gallus*, *P. bivittatus*, *D. rerio*, *S. torazame* and *P. marinus*), tunicates (*H. roretzi* and *C. intestinalis*), a hemichordate (*S. kowalevskii*), and members of the Protostomia (*T. castaneum, C. teleta* and *C. gigas*). Details of the sequences in the alignment, and the alignment itself is presented in Appendix 5.4. This alignment was used to produce Bayesian, maximum likelihood and neighbour-joining trees, rooted against *B. floridae* Pax4/6. Support values from the maximum likelihood and neighbour-joining trees were mapped onto equivalent nodes in the Bayesian tree, which is presented in Figure 5.6. Vertebrate Pax3 and Pax7 are not more closely related to either Pax3/7a or Pax3/7b than to each other, indicating that the cephalochordate Pax3/7 duplication event occurred independently in the common ancestor of the extant cephalochordates.



```
PHT domain consensus                                        R X S S I A - - X L R L K A K E H S X X L X X
PAX3 — H. sapiens            467 Q Y G Q Y G Q S A F H Y L K P D I A *                                                              484
PAX3 — M. musculus           467 Q Y G Q Y G Q S A F H Y L K P D I A *                                                              484
PAX3 — G. gallus             467 Q Y G Q Y G Q S A F H Y L K P D I A *                                                              484
PAX3 — P. bivittatus         467 Q Y G Q Y G Q S A F H Y L K P D I A *                                                              484
Pax3 — D. rerio              474 Q Y S Q Y G Q S K V N T S T V T T V - - - E R . S R V Y P L K R A S F G L V E *                    509
Pax3 — S. torazame           465 Q Y G Q Y G Q S K C H V N L S N T F - - - - S Y K N A Y P P . P S S I C I - - - - - - - - K M C D S V Y K L *  506
PAX7 — H. sapiens            461 Q Y G Q Y G Q T A V D Y L A K N V . L S - T Q R . M . L G . . . A V . G L L P V - - - - - - E T G Q A Y *  505
PAX7 — M. musculus           459 Q Y S Q Y G Q T A V D Y L A K N V . L S - T Q R . M . L G . . . A V . G L L P V - - - - - - E T G Q A Y *  503
PAX7 — G. gallus             458 Q Y G Q Y G Q T A V D Y L T K N V . L S - T Q R . M . L G . . . A V . G L L P V - - - - - - E T G Q A Y *  502
PAX7 — P. bivittatus         457 Q Y G Q Y S Q T A V D Y L T K N V . L S - T Q R . M . L G . . . A V . G L L P V - - - - - - E T G Q A Y *  501
Pax7 — D. rerio              463 Q Y S Q Y G Q T A V D Y L A K N V . L S - T Q R . M . L G D . . A V . G L L Q V - - - - - - E T G Q A Y *  507
Pax7 — P. marinus            460 Q Y G Q Y G Q T A A D Y L A K N V G L G - G Q R . . . L G . . . A V . G L L Q V - - - - - A E T G Q A Y *  505
Pax7 — S. torazame           451 Q Y A Q Y G Q S A V D Y L T K N V . L S - T Q R . M . L S . . . A V . G L L Q V - - - - - - E T G Q A Y *  495
Pax3/7a — B. lanceolatum     435 - W V Q - G A N S T D F N S N S L G I . - - A . . Q . S R . . . A A . G L I Q V - A G G - - A M A Q A Y *  479
Pax3/7a — B. floridae        435 - W V Q - G A N S T D F N S N S L G I . - - A . . Q . S R . . . A A . G L I Q V - A G G - - A M A Q A Y *  479
Pax3/7a — B. belcheri        436 - W V Q G G T N S T D F N S N S L G I . - - A . . Q . S R . . . A A . G L I Q V - A G G - - A M A Q A Y *  481
Pax3/7a — A. lucayanum       434 - W V Q - G T S S T D F N S N S L G I . - - A . . Q . S R . . . A A . G L I Q V - A G G - - A M A Q A Y *  478
Pax3/7b — B. lanceolatum     427 Q Y P S Q G S C T S S Y G S T G T . L S - - S . . Q . S L V . P - - . G M I P V - - G S - - G T G Q A Y *  470
Pax3/7b — B. floridae        425 Q Y P S Q G S C T S S Y G S T G T . L S - - S . . Q . S L V . P - - . G M I P V - - G S - - G T G Q A Y *  468
Pax3/7b — B. belcheri        427 Q Y P S Q G S C T S S Y G S T G T . L S - - S . . Q . S L V . P - - . G M I P V - - G S - - G T G Q A Y *  470
Pax3/7 — H. roretzi          657 Q Y S G Y G Q A A S D Y E N A G I T - - - - A L . M . S R . . . A A . G L I S V G A G G A T S M Q P A Y *  704
Pax3/7 — C. intestinalis     580 Q Y S P Y G Q A S G D Y G S A G I A - - - - A L . M . S R . . . A S I G L I P V - - G G G P S V Q H A Y *  625
Pax3/7 — S. kowalevskii      422 - - - - - - - - - - - - - - Q Q V . - - - - Q F S P . H G L Y . S F M *                                438
Gsb — T. castaneum           393 - - - - - - - - - - - - - - - - - - - - - - - - M C A S R V . *                                      399
Prd — T. castaneum           395 - - - - - - - - - - - - - - N T V T - - - - I L G P N S G N T P K I . L Q I *                          414
Pax3/7 — C. gigas            418 P Y N S H I A S V S S A E R C A I D E S L V A L . M . S R . . . A A . S L M Q V - T D N - - K M S A S F *  466
Pax3/7 — C. teleta           367 Q H P F H Y H T S F N A G H E A G . - - - Y P Y P S G K V H A P T P I G L G S F - - G A - - M A P P S H I V S G L  416
```

Figure 5.5. **An alignment of the C-termini of Pax3/7 sequences** from a selection of vertebrate, non-vertebrate deuterostome, and protostome species, illustrating the conservation of the Pax3/7 PHT. The overall consensus sequence for the PHT, as determined by Vorobyov and Horst (2006) is given at the top of the PHT domain box, and all identities (marked with a full stop) are relative to this consensus. The PHT domain box is grey in sequences which lack a PHT domain. The *Asymmetron lucayanum* Pax3/7b sequence does not have coverage of the C-terminus, hence its omission. *H. sapiens* = *Homo sapiens*; *M. musculus* = *Mus musculus*; *G. gallus* = *Gallus gallus*; *P. bivittatus* = *Python bivittatus*; *D. rerio* = *Danio rerio*; *S. torazame* = *Scyliorhinus torazame*; *B. lanceolatum/floridae/belcheri* = *Branchiostoma lanceolatum/floridae/belcheri*; *A. lucayanum* = *Asymmetron lucayanum*; *H. roretzi* = *Halocynthia roretzi*; *C. intestinalis* = *Ciona intestinalis*; *S. kowalevskii* = *Saccoglossus kowalevskii*; *T. castaneum* = *Tribolium castaneum*; *C. gigas* = *Crassostrea gigas*; *C. teleta* = *Capitella teleta.*

5.3.2. *Pax3/7* expression in the cephalochordates

Whole mount *in situ* hybridisation (WMISH) was used to visualise *Pax3/7* paralogue expression patterns in amphioxus development. Probes were designed to target the divergent 3' ends of *Pax3/7a* and *Pax3/7b* to avoid paralogue cross-reactivity (see Figure 5.4, Appendix 5.2). The *Pax3/7a* probe targeted a region with 54.5% similarity (with 33 gaps) to the equivalent region of *Pax3/7b* (aligned with MAFFT) and the

Pax3/7b probe targeted a region with 51.1% similarity (with 72 gaps) to *Pax3/7a*. Probe locations in the transcripts are marked in Figure 5.4c. WMISH was performed on *B. lanceolatum* embryos from mid-gastrula (stage G5) to larvae (L2) (Figure 5.7 & Figure 5.8) and on *A. lucayanum* embryos (Figure 5.9). These results indicate that the expression of the paralogues differs during embryogenesis.

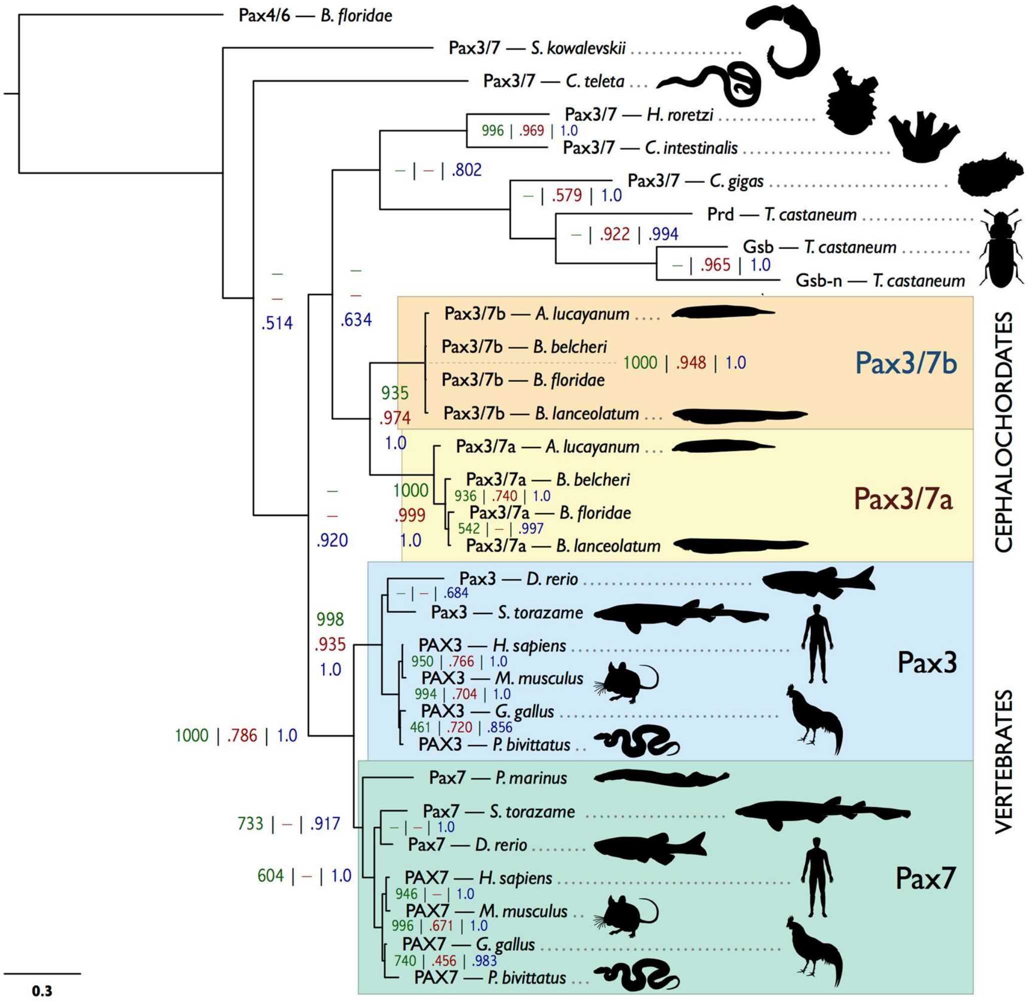


Figure 5.6. **Bayesian tree of Pax3/7 genes**. Support values are presented as follows: bootstraps out of 1000 from PHYLIP (dark green) | bootstraps out of 1.0 from equivalent nodes from maximum likelihood (dark red) | posterior probabilities from equivalent nodes from a Bayesian analysis (dark blue). Absence of an equivalent node in the corresponding analysis is indicated by a dash. The alignment and accession numbers of all included sequences are reported in Appendix 5.4. The scale bar in the lower left corner indicates amino acid substitutions per site. *S. kowalevskii* = *Saccoglossus kowalevskii*; *C. teleta* = *Capitella teleta*; *H. roretzi* = *Halocynthia roretzi*; *C. intestinalis* = *Ciona intestinalis*; *C. gigas* = *Crassostrea gigas*; *T. castaneum* = *Tribolium castaneum*; *A. lucayanum* = *Asymmetron lucayanum*; *B.*

lanceolatum/floridae/belcheri = *Branchiostoma lanceolatum/floridae/belcheri*; *D. rerio* = *Danio rerio*; *S. torazame* = *Scyliorhinus torazame*; *H. sapiens* = *Homo sapiens*; *M. musculus* = *Mus musculus*; *G. gallus* = *Gallus gallus*; *P. bivitattus* = *Python bivitattus*; *P. marinus* = *Petromyzon marinus*. Figure taken from Barton-Owen, Ferrier, & Somorjai (2018).

5.3.2.1. *Branchiostoma lanceolatum*

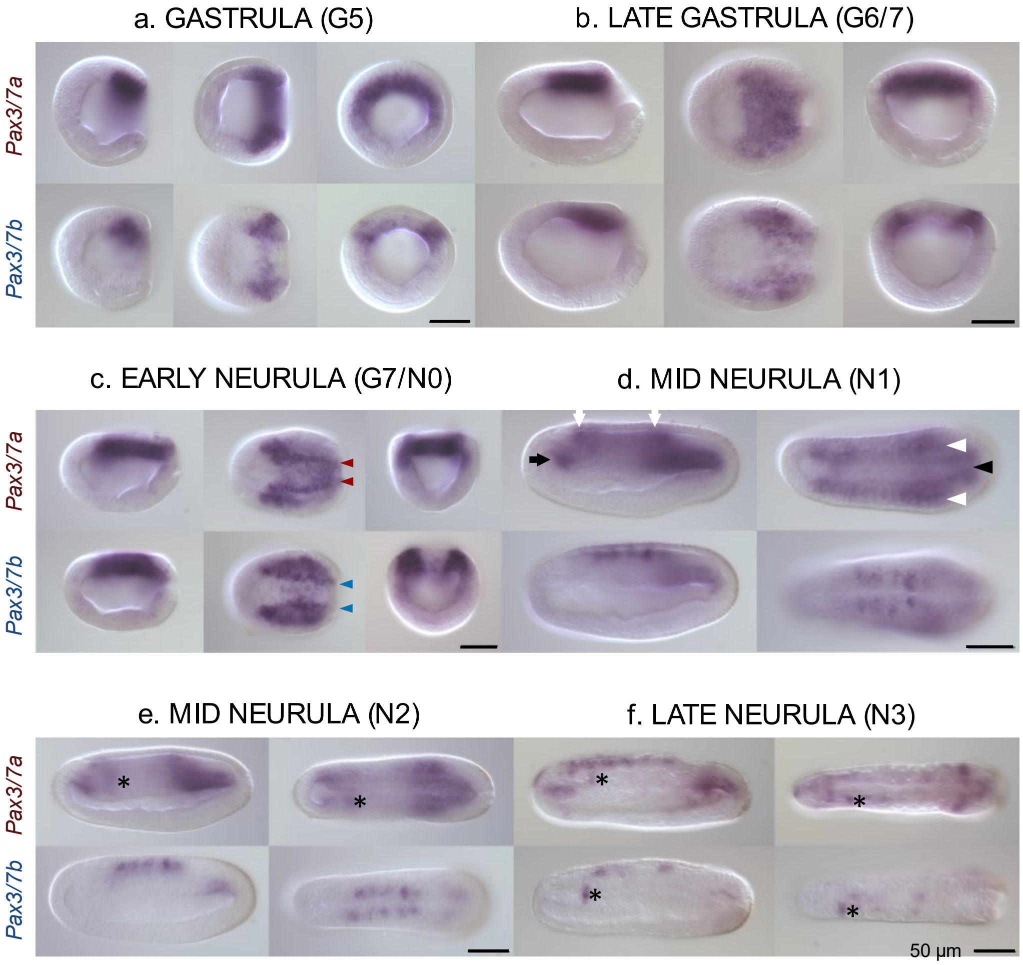


Figure 5.7. **Expression of *Pax3/7a* and *Pax3/7b* in *B. lanceolatum* early development**, visualised with whole mount *in situ* hybridisation with *Pax3/7a*-specific probe (top rows of image blocks) and *Pax3/7b*-specific probe (bottom rows of image blocks). Embryos are oriented laterally, dorsally, and blastoporally (G5-7 only) in a left-to-right order. Dorsal and lateral views are presented with the anterior facing left. (a) 10 hours post fertilization (hpf) gastrula (G5). (b) 12 hpf late gastrula (G6/7). (c) 14 hpf early neurula (G7/N0). (d) 16 hpf mid neurula (N1). (e) 21 hpf mid neurula (N2). (f) 24 hpf late neurula (N3). Red and blue arrowheads (c): differential patterning in the neural plate border. Black arrow(d): anterior mesodermal tissue expression. White arrows (d): anterior and posterior ends of the neural tube. White arrowheads (d) postero-lateral somatic tissue. Black arrowhead (d) postero-medial notochord tissue. Asterisk: sinistral expression domain found in both paralogues and *A. lucayanum* (immediately anterior to mark), Scale bars = 50 micrometres. Adapted from Barton-Owen, Ferrier & Somorjai (2018).

Expression of **Pax3/7a**

In mid-gastrulae (G5, Figure 5.7a), *Pax3/7a* is expressed in a semicircular band in the dorsal endoderm of the blastoporal lip. This expression pattern remains relatively diffuse in the late gastrula (G6/7, Figure 5.7b), but by the early neurula (N0, Figure 5.7c) the expression domain has become condensed into lines running symmetrically either side of the midline (Figure 5.7c, red arrowheads) with enlarged anterior patches, though weak expression persists throughout the posterior. By the hatchling neurula (N1, Figure 5.7d), *Pax3/7a* has diffuse expression with greater concentration in five indistinct, bilaterally symmetrical areas; the anterior mesodermal tissue (black arrow), the anterior end and posterior of the neural tube (white arrows), the postero-lateral somitic tissue (white arrowheads), and the postero-medial notochord tissue (black arrowhead). These expression domains continue with little change through to the mid neurula (N2, Figure 5.7e), except for the appearance of a distinct domain of asymmetrical *Pax3/7a* expression in the anterior (marked throughout by an asterisk placed just posteriorly), which is consistently absent or very weak on the right side. In the late neurula (N3, Figure 5.7f), *Pax3/7a* expression has become condensed into the anterior and posterior mesodermal regions, and into the left anterior somite. Patchy and granular neural regions of expression have also appeared. The asymmetrical domain persists into the early larva (L1, Figure 5.8a) while the other domains of expression are substantially reduced such that only a few anterior neural and the posterior mesodermal domains are present. This pattern continues in the later L1 and L2 larvae (Figure 5.8b & c), with faint, patchy neural expression reappearing in the latter stage.

Expression of **Pax3/7b**

In mid-gastrulae (G5, Figure 5.7a), *Pax3/7b* is expressed in smaller lateral patches in the dorsum in both germ layers, in contrast to *Pax3/7a*, which is restricted to the endoderm. This lateral expression pattern continues in the late gastrula (G6/7, Figure 5.7b); by the early neurula (N0, Figure 5.7c) the interior lateral borders of the expression domain have become strongly resolved (Figure 5.7c, blue arrowheads), though weak medial expression continues. *Pax3/7b* expression overlaps with *Pax3/7a* in the posterior regions

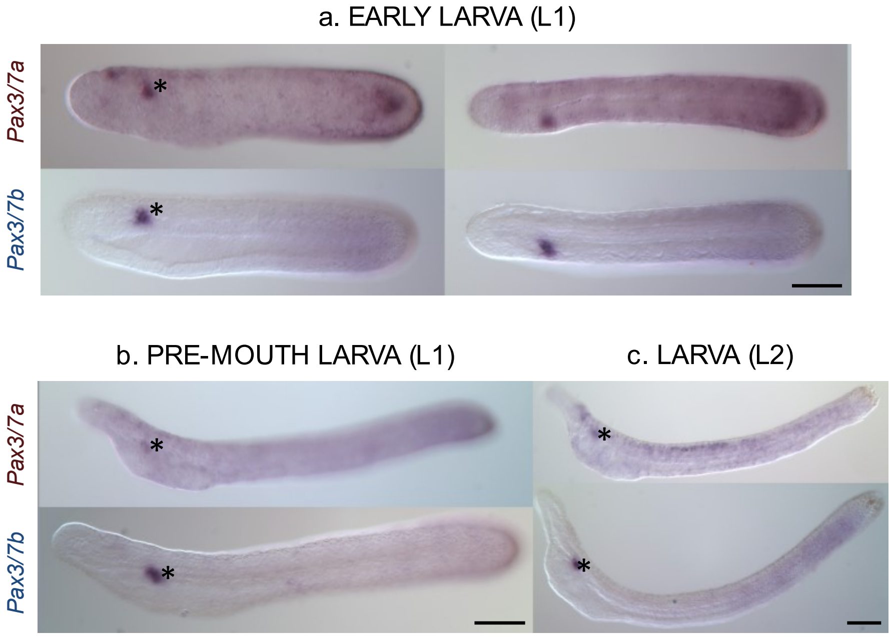


Figure 5.8. **Expression of *Pax3/7a* and *Pax3/7b* in *B. lanceolatum* later development**, visualised by whole mount *in situ* hybridization of *Pax3/7a*-specific probe (top row of image blocks) and *Pax3/7b*-specific probe (bottom row of image blocks). Views are presented with anterior to the left, and are in left-to-right order, lateral (all stages) and dorsal (early larvae only). (a) 30 hpf early larva (L1). (b) 36 hpf pre-mouth larva (L1). (c) 48 hpf larva (L2). Asterisk: common sinistral domain of expression. Scale bars = 50 micrometers. Adapted from Barton-Owen, Ferrier & Somorjai (2018).

but with a much weaker signal. By the N1 stage (Figure 5.7d), in contrast to *Pax3/7a*, there are five distinct, symmetrical domains of *Pax3/7b* expression in the dorsolateral neural tube. These spots are flanked at their anterior and posterior limits by the weaker, more diffuse regions of *Pax3/7a* expression (white arrows). This expression pattern continues with little change through to the mid neurula (N2, Figure 5.7e). By stage N3 (Figure 5.7f), the neural regions of expression are reduced in size and number, retaining only the two anterior-most and posterior-most spots, while the strong asymmetrical domain of expression in the left anterior somite previously distinguished by *Pax3/7a* expression is now also labeled by *Pax3/7b* (asterisks). This domain persists with strong expression into the early larva (L1, Figure 5.8a) while expression ceases elsewhere in the later L1 and L2 larvae (Figure 5.8b & c).

5.3.2.2. *Asymmetron lucayanum*

Whole mount *in situ* hybridisation was also performed on *A. lucayanum* early and mid neurulae, using probes covering an identical region of the *A. lucayanum Pax3/7a* and *Pax3/7b* transcripts (Figure 5.4). The experimental protocol for *in situ* hybridisation has not yet been optimised for *A. lucayanum*, and consequently the results are lesser in quality than those for *B. lanceolatum*, particularly having a less favourable signal-to-background ratio. The images are presented in Figure 5.9.

The *A. lucayanum* embryos are not directly comparable in developmental stage: the *A. lucayanum* early neurula (N1, Figure 5.9a) is more elongated than the *B. lanceolatum* early neurula (G7/N0, Figure 5.7c) but less than the mid neurula stage (N1, Figure 5.7d). The N2 stages of *A. lucayanum* and *B. lanceolatum* (Figure 5.7e and Figure 5.9b) seem approximately equivalent but the internal physiology of the *A. lucayanum* mid neurula is much less visible.

The patterns for *Pax3/7a* and *Pax3/7b* are not directly equivalent to those from *B. lanceolatum*. In the *A. lucayanum* early neurula (N1, Figure 5.9a), *Pax3/7a* seems to have diffuse expression throughout the dorsal mid-section. Although it is less expressed along the centre of the neural plate, it is broadly expressed throughout the mesoderm. In contrast, *Pax3/7b* is expressed in restricted symmetrical domains in the neural plate border.

In the mid-neurula (N2, Figure 5.9b), *Pax3/7a* is expressed in two bilaterally symmetrical domains of expression in the mid-anterior. *Pax3/7b* retains vestiges of the symmetrical, partitioned neural plate border expression, and is strongly expressed in the left anterior domain seen in *B. lanceolatum Pax3/7a* and *Pax3/7b* (Figure 5.7e & f; Figure 5.8; marked with an asterisk posteriorly throughout).

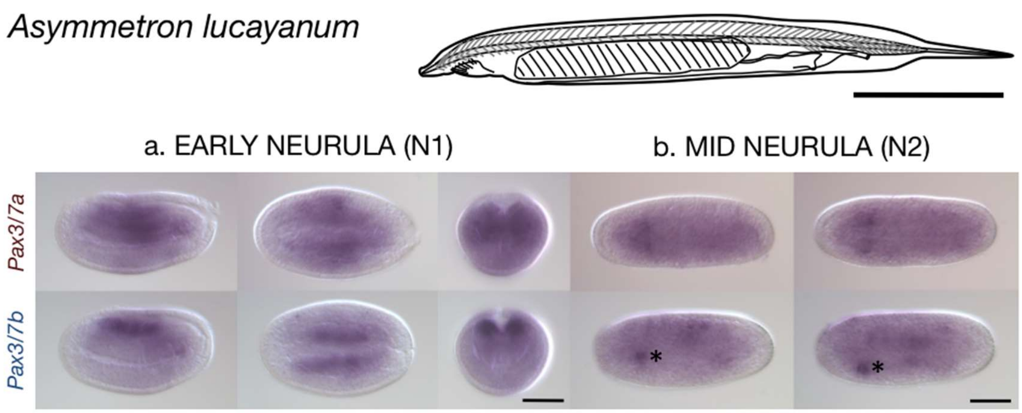


Figure 5.9. Expression of *Pax3/7a* and *Pax3/7b* in *A. lucayanum* early and mid neurulae. Top: Illustrative line drawing of adult *A. lucayanum*, adapted from Andrews, 1893. Scale bar ≈ 5 mm. Bottom: Whole mount *in situ* hybridisation images of *Pax3/7a*-specific probe (top row of block) and *Pax3/7b*-specific probe (bottom row of block) in *A. lucayanum* embryos. Views are presented in the left-to-right order: lateral, dorsal, and blastoporal (early neurula only). Lateral and dorsal views are oriented with the anterior to the left. (a) Early neurula, N1, 12 hpf. (b) Mid neurula, N2, 16 hpf. Asterisks mark to the immediate anterior the sinistral domain of expression found in both paralogues and in *A. lucayanum*. Scale bars = 50 micrometres.

5.4. Discussion

Gene duplication is an important mechanism in evolution, providing a potent source of new genetic material on which evolution can act outside the constraints on single-copy genes. Transcription factors stand out as a particularly important subset of retained and adapted paralogous genes. Paralogue divergence includes subfunctionalisation and neofunctionalisation of binding specificity and motif recognition, upstream regulatory control, and cofactor interaction, which all provide opportunities for more intricate spatio-temporal expression control and the potential for the generation of novel gene regulatory networks and morphology (Voordeckers, Pougach, and Verstrepen 2015).

The two rounds of whole genome duplication (2R-WGD) at the base of the vertebrate lineage (Putnam *et al.*, 2008) provided an ample source of stoichiometrically-balanced raw genetic material, possibly facilitating the elaboration of vertebrate novelties including the head, neural crest, and neurogenic placodes (Gans and Northcutt 1983; Kassahn *et al.*, 2009). In contrast, cephalochordate genomes bear no indications of paleopolyploidy events (Putnam *et al.*, 2008; Holland *et al.*, 2008; Huang *et al.*, 2014), and share more similarities in terms of architecture and gene content with the chordate ancestral genome than other extant chordate clades (Louis, Roest Crollius, and Robinson-Rechavi 2012). Cephalochordates therefore have many fewer paralogues than vertebrates, though both RNA-mediated and DNA-mediated duplications have been described. Among the latter, homeobox genes are most numerous; paralogues have been found in *Evx* (Ferrier *et al.*, 2001; Minguillón *et al.*, 2002), *Emx* (Minguillón *et al.*, 2002; N. A. Williams and Holland 2000; Takatori *et al.*, 2008), *Mnx*, *Vent*, *Nk1*, *Nedx*, *Uncx*, *Lhx2/9*, *Irx*, *Pou3* (Takatori *et al.*, 2008), and *Hox9-15* (Holland *et al.*, 2008; Feiner *et al.*, 2011), many of which are the result of small-scale tandem duplications. Of these, only *Vent1* and *Vent2* have been the subject of detailed functional assays, which established their *cis*- and trans-regulation in the amphioxus dorsoventral patterning regulatory network (Kozmikova *et al.*, 2011) and their expression in pharmacologically manipulated embryos (Kozmikova *et al.*, 2013).

5.4.1. Cephalochordate *Pax3/7* evolution

Data presented herein from three species of *Branchiostoma* and *Asymmetron lucayanum*, a representative of the earliest branching of the extant amphioxus genera, support the idea that tandem gene duplication may have been an important mechanism for generating cell type diversity in the cephalochordate ancestor. I report that amphioxus possess two paralogues of *Pax3/7*, a gene notable for its functions in neural plate border specification, its vertebrate roles in neural crest and placode specification, and for its involvement in somitogenesis, myogenesis and the population of regenerative muscle satellite cells possibly common to all bilaterians (Konstantinides and Averof 2014). I confirm that this duplication predates the modern cephalochordate radiation but post-dates the divergence from other chordates, implying that the chordate ancestor had a single copy.

One of my key findings is that *Pax3/7a* and *Pax3/7b* diverged symmetrically but heterogeneously between duplication and the cephalochordate radiation (Figure 5.6). They share very strong nucleotide sequence conservation, complete amino acid sequence identity in the paired domain, and very strong conservation in the EH1/Octapeptide motif and homeodomain. In contrast, they have diverged substantially in the linker regions, the N-terminus (where *Pax3/7b* seems to have lost an exon) and the four exons of the C-terminus.

The paralogues have changed little since their divergence, both in coding sequence and local CNEs; of the pair, *Pax3/7a* has changed more since the *Asymmetron/Branchiostoma* speciation events, indicating it might be under slightly relaxed selection, but has a more prototypical PHT domain (Vorobyov and Horst 2006) (Figure 5.5), while *Pax3/7b* is more conserved among species. Pronounced evolutionary asymmetry is common amongst tandem paralogues (reviewed by Holland *et al.*, 2017), for instance, in *AmphiEvx*; however, examples in which asymmetry is not observed have also been documented (*e.g. AmphiEmx*).

Although cephalochordates are considered to be slow-evolving, the pattern we observe in paralogue divergence is also consistent with the recent estimate that the crown cephalochordate node dates to only 38.8-46.0 million years ago (MYA), in contrast to previous results placing it ~120-250 MYA (see Igawa et al. 2017). Based on their calibration date of the cephalochordate/Olfactores split approximately 550 MYA, the duplication, fixation, fate-determination and preservation phases of paralogue evolution (see Innan and Kondrashov 2010) all occurred in the ~500 MYA interval during which no evident radiation occurred. Comparatively rapid change and quicker preservation is considered typical of tandem duplications (Voordeckers, Pougach, and Verstrepen 2015), although as *Pax3/7* genes are transcription factors involved in development, and specifically neurogenesis (Roux, Liu, and Robinson-Rechavi 2017), their sequence and expression domain change may have been severely constrained.

Symmetry in paralogue evolution

Symmetry of sequence evolution rate between paralogues is considered indicative of subfunctionalisation (Yampolsky and Bouzinier 2014). The evolutionary trajectory of cephalochordate *Pax3/7* duplicates, based on the symmetry of sequence change evident in Figure 5.6, seems to accord with the duplication-degeneration-complementation (DDC) model of Force *et al.* (1999) or the specialisation model of Hughes (1994). According to these models, the duplicated pair, under relaxed purifying selection, accumulates either mutations that complementarily degrade (DDC) or improve (specialisation) their capacity to perform subsets of their pre-duplication function, until the loss of either paralogue is deleterious. The only non-duplicated chordate or deuterostome outgroups for ancestral *Pax3/7* function are found in the tunicates and hemichordates. However, both groups have a highly divergent *Pax3/7* sequence, and the former of which has a very derived genome and morphology. Consequently, it is difficult to determine the exact set of ancestral functions of the *Pax3/7* pro-orthologue in the chordate ancestor. Nevertheless, a conserved role in neural border specification is highly probable, given enrichment of *Pax3/7* in lateral neuroblasts in a number of bilaterians (Li *et al.*, 2017).

The* Pax3/7 *tandem duplication and locus evolution

The cephalochordate *Pax3/7* locus is consistent with a DNA-mediated small-scale, tandem duplication of 20+ kilobases. An attempt was made to detect homology between the *Pax3/7a* and *Pax3/7b* loci using a script that implemented a rolling window of identity summation on a genomic alignment, producing a VISTA-like visualisation (Figure 5.4). Almost no non-coding nucleotide similarity was retrieved for the *Pax3/7a*/*Pax3/7b* comparison, indicating the almost complete divergence of the loci. Given their duplication potentially occurred ~550 MY (see section 1.4.2, Figure 1.11, & Igawa *et al.*, 2017), this is expected, but prevents detection of further details about the duplication event, including its precise extent and whether the exon loss of *Pax3/7b* was the result of incomplete duplication.

Since the cephalochordate radiation responsible for all extant cephalochordate genera, recently dated to 38.8-46.0 MYA (Igawa *et al.*, 2017) but previously placed at 120-360 MYA (Kon *et al.*, 2007; Yue *et al.*, 2014), the *Pax3/7* duplication locus has been tightly conserved between species (Figure 5.3a), but has nonetheless undergone divergence. The largest post-speciation divergence is in *B. lanceolatum,* where insertions of several kilobases have occurred in the first introns of *Pax3/7b* and the third and last introns of *Pax3/7a* (Figure 5.2).

MicroRNA targets in the* Pax3/7 *3' UTRs

The miRNA survey presented herein is limited by the available data concerning the extent of the *Pax3/7a* and *Pax3/7b* 3' UTRs. Holland *et al.* (1999) cloned the entirety of the 3' UTR (including the poly(A) tail) of *B. floridae Pax3/7a*, although data from SRAs suggests that a much longer region might be transcribed (see section 5.3.1.1). Alignment of the published *B. floridae Pax3/7a* 3' UTR against the *B. lanceolatum* genome allowed the prediction of a similarly-sized homologous 3' UTR. For *Pax3/7b*, the 1,138 bp region after the first in-frame stop codon in isotig29738 was taken as a predicted 3' UTR and used to predict a homologous *B. floridae* 3' UTR. Both UTR predictions would benefit from experimental validation.

Distinct targets of six microRNAs were detected in the *B. floridae Pax3/7a* UTR and three in the predicted *B. lanceolatum Pax3/7a* UTR; of these, one target (bfl-miR-4026-5p-1) was found in both (Table 5.1). Targets of seven microRNAs were detected in the predicted *B. floridae Pax3/7b* UTR, and five in the predicted *B. lanceolatum Pax3/7b* UTR; of these, two targets (bfl-miR-100 and bfl-miR-4121-5p) were found in both (Table 5.2). Three targets (bfl-miR-4026-5p-1, bfl-miR-4028-3p-1, bfl-miR-4028-3p-2) were found in a *Pax3/7a* and a *Pax3/7b* UTR. Seven miRNAs with *Pax3/7* 3' UTR targets have previously been reported to be expressed in amphioxus development (Zhou *et al.*, 2012). A summary of these data is presented in Table 5.4.

MicroRNA-mediated regulation of vertebrate *Pax3* and *Pax7* in muscle development and pathology has been extensively characterised, focussing on miR-1 (Chen *et al.*, 2010; Hirai *et al.*, 2010; Goljanek-Whysall *et al.*, 2011; Li *et al.*, 2012) and miR-206 (Hirai *et al.*, 2010; Goljanek-Whysall *et al.*, 2011; Dey, Gagan, and Dutta 2011; Li *et al.*, 2012; Liu *et al.*, 2012; Hanna *et al.*, 2016). Involvement of the vertebrate- or mammal-specific (Heimberg *et al.*, 2008) miR-27 (Crist *et al.*, 2009), miR-431 (Wu *et al.*, 2015) and miR-486 (Dey, Gagan, and Dutta 2011) has also been reported.

The apparent lack of conserved microRNA control of *Pax3/7* between vertebrates and cephalochordates is consistent with previous studies finding very little conservation of the regulation of specific genes by specific miRNAs across long evolutionary timescales (Chen and Rajewsky 2006; Xu *et al.*, 2013), despite the deep conservation of miRNAs and 3' UTR motifs themselves (Chen and Rajewsky 2006). Given the evolutionary flexibility of the miRNA/specific target relationship, miRNA target prediction in cephalochordate *Pax3/7*s is more likely to be a useful tool in assaying differential paralogue and post-speciation orthologue regulatory evolution within the cephalochordates than between cephalochordates and other deuterostomes. The possible finding that three targets are shared between *Pax3/7a* and *Pax3/7b* (which potentially started diverging millions of years earlier than the origin of vertebrates) is significant in this light.

Table 5.4. A summary of the presence of microRNA targets in the 3' UTRs of *Pax3/7a* and *Pax3/7b* of *B. lanceolatum* and *B. floridae,* and of the developmental expression data reported by Zhou *et al.*, 2011 (where present). *B. flo* = *B. floridae*; *B. lan* = *B. lanceolatum.* Light blue indicates an absence of expression; the purple-to-yellow spectrum indicates the expression intensity (low to high), adapted from Figure 5 of Zhou *et al.* (2012).

Name			Species, Gene				Reported expression				
Family	Orientation	Number	B.flo Pax3/7a	B. lan Pax3/7a	B.flo Pax3/7b	B.lan Pax3/7b	Egg	Gastrula	Neurula	Larva	Adult
7a			✓				🟪	🟦	🟦	🟦	🟦
19c					✓		🟦	🟦	🟦	🟪	🟦
100					✓	✓					
373				✓							
430d					✓		🟨	🟨	🟨	🟨	🟧
4017	-3p				✓						
4026	**-5p**	**-1**	✓	✓		✓					
4028	-5p	-7			✓		🟪	🟪	🟪	🟧	🟨
	-5p	-9					🟦	🟪	🟫	🟦	🟦
	-3p	**-1**	✓		✓						
	-3p	**-2**	✓			✓					
4099	-5p										
4108	-5p		✓								
4121	-5p				✓	✓					
	-3p			✓							
4155			✓								

Gene conversion in the paired domain

The complete nucleotide sequence conservation within the paired domains of *Pax3/7a* and *Pax3/7b* suggests that this 375 nucleotide region is subject to gene conversion, a process by which a genomic region is uni-directionally overwritten by homologous sequence via various responses to double-strand breaks (reviewed by Chen *et al.*, 2007). Gene conversion is an important mechanism in shaping paralogue fate (reviewed by Innan and Kondrashov 2010), where it can exert a homogenising force between paralogues that produces concerted evolution and minimises divergence (Teshima and Innan 2004; Mano and Innan 2008). The regions involved are usually short, rarely above 1 kb in vertebrates (Chen *et al.*, 2007), but the literature rarely addresses examples of exons or, as in this case, specific domains, being subject to localised gene conversion within the context of otherwise divergent paralogues. However, a relevant example is found in the *engrailed*-family genes of the desert locust *Schistocerca gregaria,* which have undergone region-specific gene conversion (including of the homeobox) in the midst of divergent 5' and 3' sequence (Peel, Telford, and Akam 2006).

There is some indication that gene conversion has occurred since the speciation events that separate the sampled *Branchiostoma* species, in the form of two sites in *B. belcheri* and one in *B. floridae* at which there is greater similarity between paralogues than between orthologues (see section 5.3.1). These could be the product of assembly error if the genomic reads used were short enough that not all of them contained sequence that unambiguously identified them as belonging to either *Pax3/7a* or *Pax3/7b*, which could produce an *in silico* 'gene conversion' effect that obscured dimorphism as well as indications of real gene conversion. The *B. belcheri* genomic assembly pathway was engineered to overcome its unusually extreme polymorphism (Huang *et al.*, 2014), and could be sub-optimal for determining the presence of single nucleotide differences in otherwise highly-conserved paralogous regions.

Pax3/7 protein functionality

Although the DNA-binding domains of *Pax3/7a* and *Pax3/7b* are almost identical, it is likely that the differences in the C-terminus and in the linker regions between the conserved domains are sufficient to alter their functionality. Amino-terminal sequence changes have been shown to affect the binding specificity of DNA-binding domains and homeodomains in general (Liu, Matthews, and Bondos 2009; Tzeng and Kalodimos 2012) and Pax genes specifically (reviewed by Mayran, Pelletier, and Drouin 2015).

Small sequence changes have the potential to differentially modify the binding affinity of the paired domain and homeodomain, the binding modality of the paired sub-domains, and subnuclear localisation (Corry *et al.*, 2010). The presence of two DNA-binding domains in Pax4/6 and Pax3/7, which are known to act cooperatively (Corry and Underhill 2005), adds complexity to these modifications: small sequence changes have the potential to differentially modify the binding affinity of the two domains (Vogan, Underhill, and Gros 1996; Vogan and Gros 1997), the binding modality of the paired subdomains (Vogan, Underhill, and Gros 1996), and subnuclear localisation (Corry *et al.*, 2010). The vertebrate Pax3 C-terminus contains a transactivation domain (Chalepakis *et al.*, 1994), and Pax7 and cephalochordate Pax3/7s possess the PHT/OAR/paired tail/C-peptide domain (Vorobyov and Horst 2006), thought to contribute to transcriptional repression (Norris and Kern 2001). The modest differences between Pax3 and Pax7, located mostly in the C-terminus, are enough to produce substantial differences in target activation in myogenesis (Soleimani *et al.*, 2012).

The extent of these substantial functional effects caused by the minor differences in mutants, splice variants and between vertebrate Pax3 and Pax7 is an indication that Pax3/7a and Pax3/7b, which have diverged more than Pax3/Pax7, probably behave differently with regard to target recognition and interaction with cofactors. Such sequence change has been highlighted as an important but under-appreciated mechanism in the evolution of developmental GRNs (Cheatle Jarvela and Hinman 2015).

5.4.2. *Pax3/7 e*xpression in cephalochordate development

Regardless of putative differences in downstream activity, *Pax3/7a* and *Pax3/7b* are expressed differently in gastrulae and neurulae in *B. lanceolatum*, demonstrating that the paralogues have diverged in their *cis*-regulation. *Pax3/7a* and *Pax3/7b* are expressed in partially overlapping but distinct domains in the neural plate (G5 to N0, Figure 5.7a-c, red and blue arrowheads), presumably as the result of modification of an ancestral neural plate domain. *Pax3/7a* is expressed throughout the dorso-posterior mesoderm prior to neurulation (G5 & G6/7, Figure 5.7a & b) while *Pax3/7b* is restricted to smaller, bilaterally symmetrical dorso-posterior regions in both the mesoderm and ectoderm, consistent with a role in the initial specification of the neural plate border. Distinct ectodermal lateral lines of expression do appear in *Pax3/7a* in the late gastrula/early neurula (G7/N0, Figure 5.7c), though diffuse mesodermal expression remains throughout the posterior. By the mid-neurula, the paralogues seem to have switched to a different expression programme, one in which their expression patterns have the least overlap. Particularly notable are the tight, defined neural spots of *Pax3/7b* and the appearance of the asymmetrical, sinistral domain (the left anterior somite, Holland *et al.*, 1999) of expression that first appears in *Pax3/7a* (left of asterisk throughout, N2, Figure 5.7e) and later appears in *Pax3/7b* (N3, Figure 5.7f). As the embryo becomes a larva, the two expression patterns converge until both expression patterns are largely restricted to the asymmetrical domain (L1 & 2, Figure 5.8a-c). Thus, divergence between duplicate expression patterns increases during gastrulation and early neurulation, peaking at mid-neurula stages, consistent with function partitioning.

My results broadly recapitulate previous *Pax3/7* expression data from *B. lanceolatum* (Figure 3H, I & J of Somorjai *et al.*, 2008), considering that the latter used a probe with probable cross-reactivity between the 5' conserved region of *Pax3/7a* and *Pax3/7b*. In contrast, the *B. lanceolatum* expression patterns are not a perfect subset of the *Pax3/7(a)* domains reported for *B. floridae* (Figure 5 of Holland *et al.*, 1999), who used a similarly cross-reactivfe probe. Potentially missing from our patterns are the anterior somitic and mesodermal expression (Figure 5F, G, I & K of Holland *et al.*, 1999), the distinct anterior neural spot (arrow, Figure 5K, M, P & Q of Holland *et al.*, 1999) and the larval axial musculature and notochord expression (Fig. 5M, P, & Q of Holland et al. 1999). Minor discrepancies are not unusual, but significant differences among *Branchiostoma* species are rare (Somorjai *et al.*, 2008). It is possible that these differences are caused by the general variability between probes for the same target, Pax gene probe cross-reactivity, or experimental sensitivity. The probes I used were by necessity relatively short in order to limit possible cross-reaction of highly conserved regions, but the expression patterns I observed are highly specific and reproducible, suggesting they reflect the core domains of *Pax3/7a* and *Pax3/7b*.

In contrast to what is observed in amphioxus, differences between vertebrate *Pax3* and *Pax7* early developmental expression are much less pronounced, to the extent that they have 'swapped' expression profiles during evolution (Monsoro-Burq, Wang, and Harland 2005; Basch, Bronner-Fraser, and García-Castro 2006; Maczkowiak *et al.*, 2010; and see synthesis in Monsoro-Burq 2015). *Pax3/Pax7* appear in the neural plate border during neural induction in the early gastrula, and intensify at the lateral edges to mark the dorsal edge of the closing neural tube, a pattern comparable to late gastrula/early neurula expression in amphioxus. *Pax3* and/or *Pax7* are also expressed throughout the posterior dorsal neuraxis, an approximate analogue of the neural spots in *Pax3/7b* and later *Pax3/7a*, though these spots are more spatiotemporally restricted.

Developmental expression in Asymmetron lucayanum

Although the *in situ* hybridisation images of *A. lucayanum* embryos presented herein (Figure 5.9) suffer from a sub-optimal degree of noise, the expression patterns they show are enough to tentatively suggest that *Pax3/7a* and *Pax3/7b* regulation is not identical between *B. lanceolatum* and *A. lucayanum*.

Although Figure 5.9a is an N1 neurula, it is less developed than the *B. lanceolatum* N1 stage sampled herein (Figure 5.7d), and consequently could represent a transitional point between *B. lanceolatum*-like late gastrula/early neurula (Figure 5.7c-like) and mid neurula (Figure 5.7d-like) patterns rather than a departure from the *B. lanceolatum* patterns. However, the mid neurula patterns (Figure 5.9b) are more obviously divergent. The *A. lucayanum* N2 *Pax3/7b* pattern seems to be a precocious version of the *B. lanceolatum* N3 *Pax3/7b* pattern, having the asymmetrical left anterior somite domain (marked with an asterisk placed just posteriorly throughout) while the medial neural partitions have disappeared. The *A. lucayanum* left anterior somite domain is also noticeably less tightly defined in the anteroposterior axis. However, the *A. lucayanum* N2 *Pax3/7a* expression pattern has no *B. lanceolatum* analogue, being in approximately the same anteroposterior and dorsoventral position as the left anterior somite domain but symmetrical. No other expression domain is reliably discernible within the noise.

The *Pax3/7* loci and products are tightly conserved between *B. lanceolatum* and *A. lucayanum* (*c.f.* Figure 5.3 and Figure 5.6), indicating that differences in deployment are likely due to small changes to the *cis*-regulatory landscape. The cephalochordate *Pax3/7* locus is a promising system with which to understand the evolution of the *cis*-regulation of a homeobox gene paralogue pair.

Myogenic roles of Pax3/7 genes

While *Pax3* and *Pax7* appear to play semi-redundant roles in neural development, they diverge in function in vertebrate myogenesis (reviewed by Buckingham and Relaix 2015). *Pax3* acts broadly from the onset of myogenesis in the presomitic mesoderm to the dermomyotome, while *Pax7* expression is later and restricted to a dermomyotomal subdomain. These PAX3/PAX7 positive cells form a proliferative muscle progenitor population that eventually positions itself underneath the basal lamina on the muscle fibres. In the adult, these cells become a heterogenous population of quiescent satellite cells; all are maintained by *Pax7* expression, but some also express *Pax3*, which is known in this context to be an inadequate substitute, binding 10-fold fewer targets (most of which are also targets of PAX7). During myogenesis, *Pax3* and *Pax7* seem to be responsible for maintaining the cells in a proliferative/quiescent but undifferentiated state. Lack or cessation of *Pax3* or *Pax7* expression in a cell can lead to apoptosis or cell cycle exit and muscle differentiation via MyoD, depending on the precise context.

Although the later myogenic roles of amphioxus *Pax3/7* genes are yet to be thoroughly characterised, at least one of the paralogues is known to be expressed in adult muscle, as *Pax3/7b* has been amplified from adult *B. belcheri* segmental muscle (Chen *et al.*, 2010). Whether both paralogues are involved in adult muscle development redundantly, or rather show temporal or tissue-specific patterns of expression (similar to *Pax3* and *Pax7* in post-embryonic muscle development and regeneration in mice) is still unclear. The initial identification of *Pax3/7b* transcripts in a tail blastema transcriptome clearly identifies a role in the adult regeneration process. However, previous characterisation of *Pax3/7* in a population of satellite-like cells and the nerve cord during tail regeneration (Somorjai *et al.*, 2012) utilised a cross-reactive *in situ* hybridisation probe. Therefore, changes in paralogue function during postembryonic processes in amphioxus cannot currently be ruled out. Future studies are required to determine to what extent divergence has occurred in expression, downstream targets, and interaction with co-factors in both myogenic and neural contexts.

5.4.3. Future work

The cephalochordate *Pax3/7* tandem duplication would be an excellent system in which to perform a detailed study of the dynamics of *cis*-regulatory, sequence, and *trans*-regulatory evolution following a tandem duplication in an otherwise highly-conserved, WGD-free genome. *Pax3/7* duplication is of interest as a homeobox superfamily involved in neurogenesis, which places it at the nexus of several factors known to constrain gene evolvability, as well as being involved in myogenesis and the post-developmental ontogenesis of proliferative regenerative cells.

There are many aspects of *Pax3/7* paralogue evolution which should be investigated, from the completion of assays started herein, to eliminating ambiguities in the currently available data, and beyond to the use of yet-undeveloped techniques to achieve functional insights into *Pax3/7* deployment in cephalochordates. Below are some suggestions for future research directions, divided between investigations into the *cis*-regulatory and differential expression evolution of the *Pax3/7* loci and the *trans*-regulatory effects that the *Pax3/7* paralogues could exert on other genes.

Pax3/7 cis-regulation and differential expression

Several expansions could usefully be made to the *in situ* hybridisation data presented herein. A particular bottleneck in interpreting the expression data from *A. lucayanum* development was a lack of fine temporal sampling between G7/N0 (Figure 5.7c) and N1 (Figure 5.7d), to detail the appearance and sequence of the transition between the expression pattern of the late gastrula/early neurula and the expression pattern of the early/mid neurula. Samples taken from the same fertilization at 20 or 30 minute intervals between 14 and 16 hpf would provide such a time course.

A detailed study of the evolution of the *Pax3/7*s after the radiation of the extant cephalochordates would necessitate more complete and better quality *in situ* data from *A. lucayanum*, which would require the optimisation of the protocol for this species and probably the collection of fresh embryonic material. Additionally, expression data from *B. belcheri* and/or *B. floridae* could be a useful addition given that the *B. lanceolatum Pax3/7* locus is somewhat divergent relative to other *Branchiostoma*, and the imperfect correlation between the domains reported herein and those reported from the cross-reactive *B. floridae AmphiPax3/7* probe by Holland *et al.* (1999), although the observed differences are explainable by differences in embryo developmental progress and experimental conditions.

Another obvious candidate for *in situ* hybridisation would be regenerating tail material. Expression patterns have already been reported using a cross-reactive *Pax3/7* probe and antibody (clone DP312, which targets the middle of the homeodomain). Although paralogue-specific antibodies are unavailable, *in situ* hybridisation using the existing specific probes should be sufficient to reveal important aspects of regenerative *Pax3/7a* and *Pax3/7b* activity. Sectioning *in situ* hybridisation samples would also be invaluable to determining tissue specificity for both developmental and regenerative samples.

A final potential improvement to the expression data would be two-colour *in situ* hybridisation experiments. These would help reveal the precise spatial relationships between *Pax3/7a* and *Pax3/7b* expression where they are not clear when visualised in separate embryos; for example, whether the diffuse dorsal domains of *Pax3/7a* expression in the mid-neurula (white arrows, Figure 5.7d) bookend or overlap the *Pax3/7b* partitioned neural expression, and whether there is an exact overlap between the left anterior somite domains of *Pax3/7a* and *Pax3/7b* (asterisks, Figure 5.7f). Two-colour *in situ* hybridisation or *in situ* hybridisation/immunohistochemistry could also allow comparison with the expression of marker genes (*e.g. SoxB1a-c*), signalling genes (*Wnt*s, *Fgf*s and *BMP*s) other neural plate border specifiers (*i.e. Zic* and *Msx*), and other genes involved in neural crest specification in vertebrates (*e.g. Snail).*

If the miRNA target site analyses were to be completed, it would be necessary to more robustly determine the extent of the 3' UTRs of the *Pax3/7*s, and ideally, to retrieve both 3' UTRs from all four cephalochordate model organisms to reveal commonalities and differences of miRNA targeting. 3' UTRs could be retrieved from total RNA/cDNA samples using 3' RACE (Frohman, Dush, and Martin 1988; Borson, Salo, and Drewes 1992) or predicted *in silico*. Direct sequencing of the paired domains using flanking *Pax3/7a*- and *Pax3/7b*-specific primers would also clarify the potential evidence for recent domain-specific gene conversion.

B. floridae AmphiPax3/7(a) was previously shown to be expressed in multiple splice variants (Short and Holland 2008), including an isoform lacking exon 3, which terminates early in exon 5 and lacks the PHT domain (isoform 3[-]), and isoforms retaining introns 2, 3, and 4 (3, 4, and 5 accounting for the *Pax3/7a* exon 1 described herein). Although these data are almost certainly robust for *Pax3/7a* exons 2-7, it might be worthwhile to repeat these analyses with the information that *Pax3/7a* has 7 exons and on *Pax3/7b*. If efforts to understand the subtleties of *Pax3/7* differential regulation were taken to an extreme, it might be possible to perform *in situ* hybridisation with probes with greater affinity for specific isoforms and miRNA target prediction on the 3' UTRs of these variants.

Other, purely *in silico* analyses of the *Pax3/7* locus could be undertaken. Information about the selective environment of the early evolution of the paralogues might be available from an analysis of the nucleotide sequences using tools and tests like MEME (Murrell *et al.*, 2012), BUSTED (Murrell *et al.*, 2015), RELAX (Wertheim *et al.*, 2015), and PAML (Yang 2007), although the analysis is likely to be limited by the small number of paralogues and species, the heterogenous nucleotide sequence divergence between *Pax3/7a* and *Pax3/7b,* the tight conservation of orthologues between species, and the polymorphism between different sources of sequence data for the same species.

C*is*-regulatory elements including transcription factor binding sites can be predicted with the aid of tools including MULAN/MultiTF (Ovcharenko *et al.*, 2005), PROMO (Farré *et al.*, 2003), jPREdictor (Fiedler and Rehmsmeier 2006), and PreCisIon (Elati *et al.*, 2013). These analyses could be complemented by *B. lanceolatum* ChIP-seq data owned by the laboratory of J.L. Skarmeta.

These *in silico* approaches to analysing the *Pax3/7 cis*-regulatory landscape would only be useful as a source of hypotheses for *in vivo* experimentation. One possibility would be using *Ciona intestinalis* transgenics to test the activity of putative regulatory elements, as was done with the use of regulatory elements of *B. floridae Gsx* to drive expression in the *C. intestinalis* central nervous system (Garstang 2016; Garstang, Osborne, and Ferrier 2016). With the advent of CRISPR-Cas9 genome editing (Ran *et al.*, 2013; reviewed by Doudna and Charpentier 2014; and Hsu, Lander, and Zhang 2014), it may eventually be possible to perform these analyses directly in cephalochordates. Loss-of-function treatments (see below) could also be used on the upstream regulators of *Pax3/7*.

With the experiments suggested above, it may be possible to achieve some insight the relationships between specific regulatory elements and spatiotemporal aspects of the expression patterns, and in so doing gain a detailed understanding of the evolution of differential deployment and function of the *Pax3/7* paralogues in cephalochordates.

Pax3/7 trans-regulation

Two avenues exist for studying the *trans*-regulatory effects of the cephalochordate *Pax3/7* paralogues. One option could be a genome-wide binding site analysis/ChIP-Seq of Pax3/7a and Pax3/7b modelled after Soleimani *et al.*'s (2012) analysis of Pax3/Pax7 binding.

Another option would be the application of loss-of-function techniques to knock down *Pax3/7a* and *Pax3/7b* expression in development and/or regeneration. These could include microinjection (Holland and Onai 2011), electroporation, and/or passive soaking (*e.g.* Luo and Su 2012; Heyland, Hodin, and Bishop 2014) to deliver carefully-controlled (*c.f.* Kok *et al.*, 2015; and Blum *et al.*, 2015), paralogue-specific antisense morpholino oligonucleotide or siRNA treatments to embryos and adult tails. Morpholino treatments could also be used to influence *Pax3/7* splicing (Draper, Morcos, and Kimmel 2001) and miRNA regulation (Choi, Giraldez, and Schier 2007). Over-expression of *Pax3/7a* or *Pax3/7b* could also be induced by ectopic messenger RNA (Holland and Onai 2011). These techniques have been attempted in amphioxus before by researchers and have yielded few published data, indicating that they are not straightforwardly accessible.

The functional manipulations suggested herein could help achieve an understanding of how the differential deployment and C-terminal sequence divergence of *Pax3/7a* and *Pax3/7b* have produced different *trans*-regulatory effects on their downstream targets, and how these have contributed to the evolution of cephalochordate development.

5.5. Conclusions

AmphiPax3/7 was considered a useful proxy for understanding the properties and deployment of the chordate proto-*Pax3/7*. My findings showing independent vertebrate and cephalochordate *Pax3/7* duplications – and the ensuing sequence and regulatory divergence – offer new insight into genomic constraint/plasticity, and evolvability of gene duplicates and GRNs in different duplication contexts. In amphioxus, tandem duplication and divergence of *Pax3/7* has resulted in subfunctionalisation (and possibly neofunctionalisation) of ancestral neural plate border (Li *et al.*, 2017) and muscle related (Liu, Matthews, and Bondos 2009; Konstantinides and Averof 2014) functions, many of which parallel those seen in vertebrate *Pax3* and *Pax7* following WGD. Dissecting the regulatory landscape of *Pax3/7* genes in amphioxus, including the function of the CNEs partitioned between paralogues, should shed further light on genome architecture evolution in chordates.

I have shown that cephalochordates, which are considered to be a significant outgroup to vertebrates in the study of the evolution of the neural crest GRN, have two *Pax3/7* paralogues where it was previously thought that this family was represented by a single-copy gene. This discovery has implications both for previous and future studies of amphioxus development and regeneration and for vertebrate studies in which cephalochordates are used as an outgroup. The amphioxus *Pax3/7* gene pair also offers a tantalising and tractable example of *cis*-regulatory and sequence subfunctionalisation after tandem duplication of a developmental transcription factor involved in the development of key chordate features and in regeneration.

6. General results & discussion

6.1. Homeobox genes in regeneration

An objective of the present study was to compare the regenerative homeobox expression response of *S. lamarcki* and *B. lanceolatum*. The involvement of specific gene families and orthology groups with regard to aspects of *S. lamarcki/B. lanceolatum* regeneration and known developmental/regenerative roles is discussed in sections 6.1.2 & 6.1.4. First, I compare the homeobox content of the mature and regenerative transcriptomes of the two species with a broad quantitative view (section 6.1.1), in an attempt to detect overall differences or similarities in apparent patterns of expression as well as highlight some gene families worthy of candidacy for future investigations.

6.1.1. Quantitative comparison of homeobox deployments in regeneration

A comparison of the genes found in the transcriptomes of *S. lamarcki* and *B. lanceolatum* is given in Table 6.1. Spearman's rho tests of correlation were performed on normalised read counts of genes common between pairs of transcriptomes (Figure 6.1). These indicated that read counts of sequences between the mature, early (2 dpa) and late (6 dpa) regeneration transcriptomes of *S. lamarcki* are strongly correlated, indicating that differential regulation of genes relative to one another is relatively minimal between these transcriptomes, and that regeneration does not entail the transcription of a different, exclusive set of homeobox genes. However, the overall homeobox gene expression is far from static, increasing 426% in early regeneration and then falling from that peak by 31% in late regeneration. Most of the difference between the mature tissue and the regenerating tissue (73.6% of the total 2dpa read count, and 87.1% of the difference) comes from the massive increase in the read counts of just five genes; *Six1/2*, *Dlxa*, *Emx B*, *Emx A*, and *Pax4/6 B* (Figure 6.2). These genes were already among the most transcribed in mature tissue (2nd, 8th, 11th, 10th, and 6th respectively). Read counts of genes outwith these five are also generally increased by 150% overall in early regeneration. Two genes are outliers in their degree of read count change from mature to early regeneration (Figure 6.3); *Dlxb* (13 reads to 746; 5,738%) and *Otx B* (2 reads to 68; 3,400%). In late regeneration, the most prominent outlier is *TALE-I B* (14 reads in mature to 4 reads in early regeneration to 95 reads in late regeneration; 2,375%). Eleven genes (*Hbn*, *Lmx*, *Nk5*, *Pou4 A*, *PRD-VIII*, *Shox*, *En*, *Msxlx*, *Barh*, *Six3/6 B*, & *Nk6*) were expressed *de novo* in the early regeneration transcriptome. All of these except *Lmx, Six3/6 B* and *Nk6* are greatly reduced in late (6dpa) regeneration (Figure 6.4).

Table 6.1. Summary of homeobox genes in the transcriptomes of *S. lamarcki* and *B. lanceolatum*. Gene presence is indicated with an up arrow if the gene is upregulated (based on read count) in the regenerative transcriptomes relative to the mature tissue transcriptome and with a tick if the read count is approximately equal, or a dash if they are absent. For *S. lamarcki*, the arrows are qualified with an 'E' if they are more highly up- or downregulated in the early regenerative (2dpa) transcriptome and 'L' if they are more highly upregulated in the late regenerative (6dpa) transcriptome. The arrows are underlined if they are absent from the mature transcriptome (up arrow) or the regenerative transcriptome (down arrow). Arrows or ticks are placed in parentheses if they are based on a low (≤5) read count. Where gene names are placed in parentheses, the genes in the left and right columns are not detectibly orthologous beyond the class level. HOXL = Hox-linked; NKL = Nk-linked; *S.lam* = *Spirobranchus lamarcki*; *B.lan* = *Branchiostoma lanceolatum*.

S.lam		*B.lan*
	ANTP-HOXL	
▲E ▲E	Dlx	✓
-	Evx	▲
-	Gbx	▲
-	Meox	($\underline{▲}$)
-	Mnx	($\underline{▲}$)
	└ Hox	
-	Hox1	($\underline{▼}$)
-	Hox2	($\underline{▼}$)
-	Hox3	▲
-	Hox4	▲
-	Hox5	▲
▲L	Hox6-8	▼▲▼
-	Hox9-14	▼▲▲ ▲▲▼
	└ ParaHox	
-	Cdx	$\underline{▲}$
-	Gsx	-
-	Xlox	-
	ANTP-NKL	

S.lam		*B.lan*
	PRD	
-	Alx	▲
-	Arx	(▼)
-	Dmbx	▲
-	Drgx	($\underline{▲}$)
▲E	Gsc	(▲)
$\underline{▲}$E	Hbn	-
-	Hopx	▲
-	Isx	-
▲E ▲L	Otp	▲
▲E ▲E	Otx	($\underline{▼}$)
-	Phox	-
-	Pitx	▼
-	Prop	-
▲E	Prrx	▲
-	Rax	▼
-	Repo	(▲)
$\underline{▲}$E	Shox	✓
-	Uncx	(▼)
▼	Vsx	▲

S.lam		*B.lan*
	TALE	
▲L	Irx	($\underline{▲}$) (▲) ▼
▲E ▲L	Meis	▲
▲L	Mkx	✓
▲L	Pbx	▲
▲L	Pknox	✓
▲L	Tgif	▲
▲L ▲L ▼ ▼ ▼ ▼	(Other)	-
	CUT	
x	Acut	($\underline{▼}$)
▲E	Cmp	($\underline{▲}$)
▲L	Cux	✓
▲L	Onecut	✓
	PROS	
-	Prox	▼
	ZF	
-	Azfh	▲

-	Abox	-
-	Barhl	▼
-	Bari	(▼̲)
▲̲E	Barx	-
-	Bsx	-
▲E	Dbx	-
▲E ▲E	Emx	▲▼▼
▲̲E	En	▼
-	Hhex	(▼̲)
-	Hlx	-
-	Hx	▼
-	Lbx	▲
-	Lcx	-
▲L	Msx	▲
▲̲E	Msxlx	-
-	Nedx	▲
-	Noto	-
-	Ro	(✓)
▼	Tlx	(▲̲)
-	Vax	-
-	Ventx	(▲̲) ▲
▲E	(Other)	(▲)
	└ **NK**	
▲L ▲L	Nk1	-
▲E ▲L	Nk2.1	-
▲E	Nk2.2	-
-	Nk3	▲
-	Nk4	▲
▲̲E	Nk5	(✓)
▲̲L	Nk6	▲
-	Nk7	(✓)
▲̲E	(Other)	(▲̲)
	└ **PAX**	
-	Pax2/5/8	-
-	Pax3/7	▲
▲L ▲E	Pax4/6	▼
	LIM	
▲E	Isl	▲
▲L	Lhx1/5	▼
▲L ▲L	Lhx2/9	(▲̲) ▲
-	Lhx3/4	(▼̲)
-	Lhx6/8	-
▲̲L	Lmx	(▲̲)
	POU	
-	Hdx	▼
-	Pou1	(▼̲)
▲L	Pou2	-
▲L	Pou3	▲
▲̲E ▲L	Pou4	(▼)
▼E	Pou6	▲
	HNF	
-	Ahnfx	▲
▲E	Hmbox	▲ (✓)
-	Hnf1	-
	SINE	
▲E	Six1/2	▲
▲̲L	Six3/6	▲
▲L	Six4/5	▼
	CERS	
▲L	Cers	✓
	(OTHERS)	
▲E	(Other)	(▼̲) ▲
-	Tshz	▼
-	Zeb	▼
▲E	Zfhx	✓
-	Zhx	▲

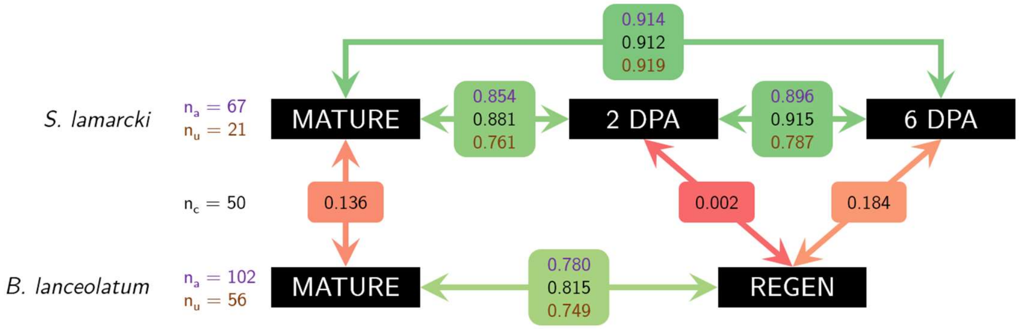


Figure 6.1. Spearman's rho correlations between the read counts of the transcriptomes. Spearman's rho was calculated between the read counts of genes for all sequences (purple numbers), between the read counts of gene families only present in both species' transcriptomes (black numbers), or between the read counts of gene families only present in the transcriptomes of one species (brown numbers). Spearman's rho is given to 3 significant figures. 1.00 = perfect correlation. Colouration is on a red (0.00) to green (1.00) scale. n_a = number of all within-species genes, used to calculate purple numbers; n_c = number of genes common between the two datasets, used to calculate black numbers; n_u = number of within-species genes which are not in n_c. $n_a \neq n_c + n_u$ because of one-to-many and many-to-many orthology relationships between the data sets.

B. lanceolatum also shows a fairly strong correlation between read counts in mature and regenerating tissue (Figure 6.1). Unlike *S. lamarcki*, however, the regenerative response is mild in terms of overall homeobox expression (127% increase). The top five genes with the largest regenerative read count (*Dlx, Hox14, Hox12, Six1/2,* and *Hox11*; Figure 6.2) contribute much less to the change in overall homeobox expression (31% of the early count, and 27% of the difference). The most upregulated genes are *Prrx* (2 reads to 44, 2,200%) and *Hox5* (2 reads to 26, 1,300%). Fourteen genes (*Cdx*, *Vent2*, *Ankx*, *Hmbx-l2* [see section 4.3.2.1], *Lhx2/9-a*, *Lmx*, *IrxA*, *Mox*, *Mnxb*, *Vent1*, *Compass*, *Aprd2*, *Drgx*, *Gsc*) were expressed *de novo* in regeneration, although all but *Cdx* (see section 6.1.2.3) and *Vent2* were represented only by two or fewer reads.

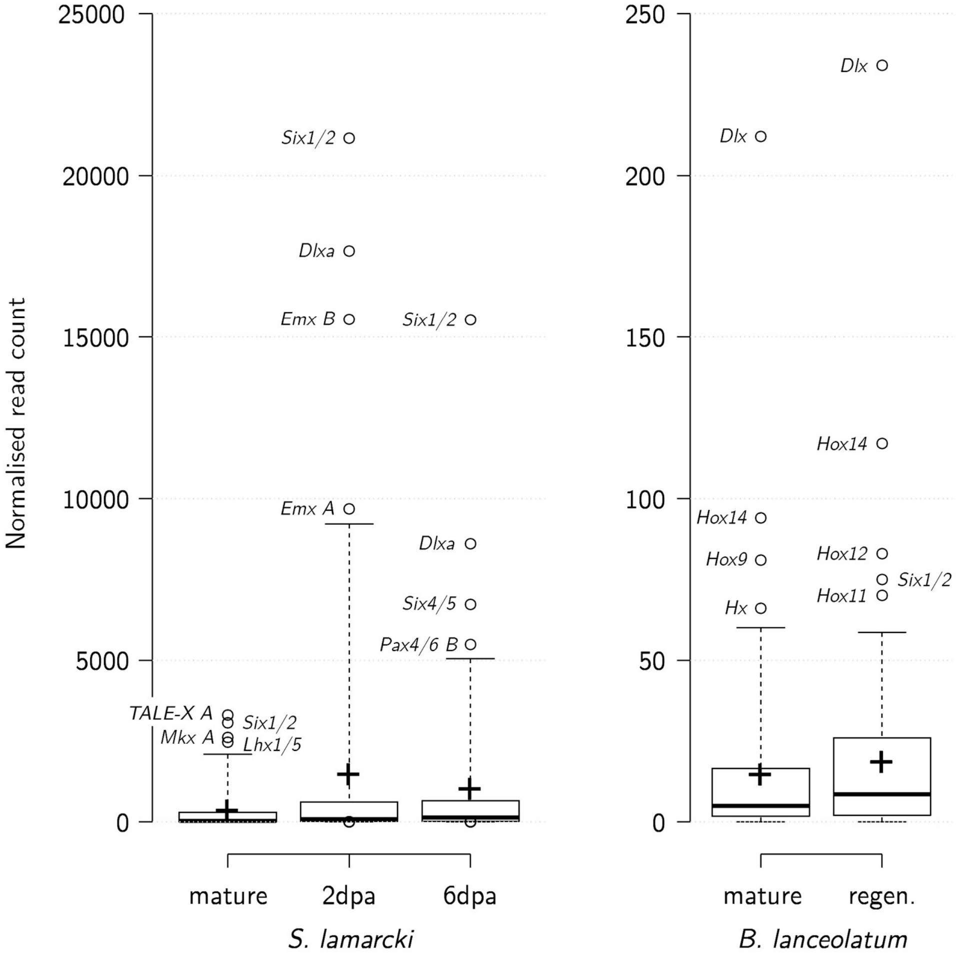


Figure 6.2. Box-and-whisker plots of the homeobox content of the transcriptomes. Center lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend to 5th and 95th percentiles, outliers are represented by dots; crosses represent sample means. *N.b.*, the scale is 0-25,000 reads for *S. lamarcki* and 0-250 reads for *B. lanceolatum*. Adapted from plots produced using BoxPlotR (Spitzer *et al.*, 2014).

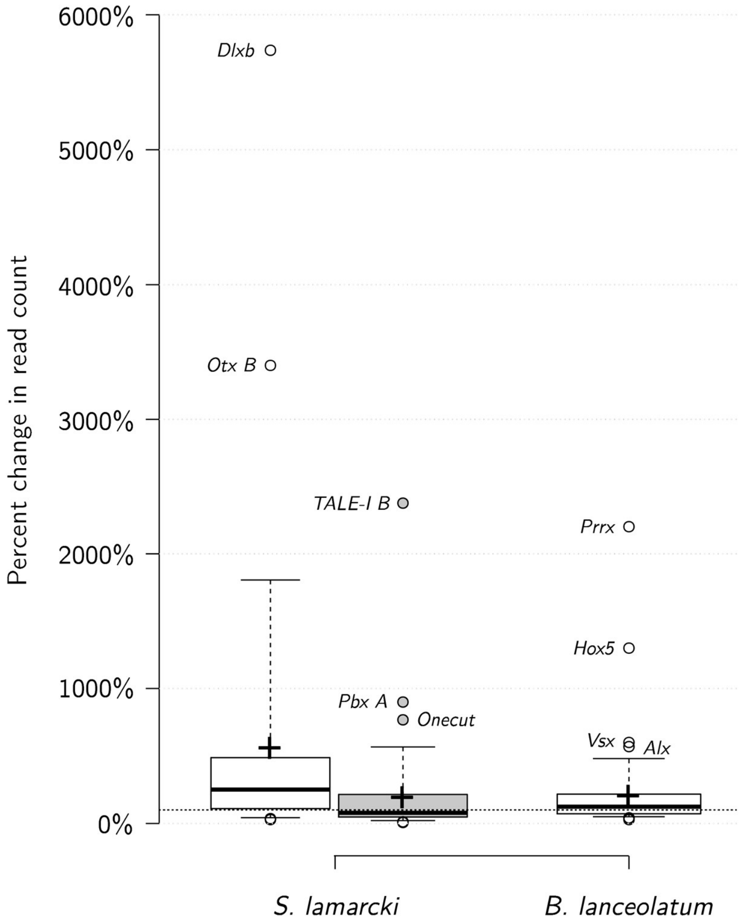


Figure 6.3. Box-and-whisker plot of the percent read count changes between mature and early (2dpa) regeneration (white, *S. lamarcki*); between early (2dpa) and late (6dpa) regeneration (grey, *S. lamarcki*) and between mature and regeneration (white, *B. lanceolatum*). Center lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend to 5th and 95th percentiles, outliers are represented by dots; crosses represent sample means. The dotted black line denotes 100%, *i.e.* parity in read count. Adapted from a plot produced using BoxPlotR (Spitzer *et al.*, 2014).

Normalised read counts represent (to some extent) change in expression as a relative component of the total transcriptome, not absolute change in expression. Read counts are not a reliable indicator of gene expression because they are un-replicated, and therefore no measure of reliability/stochasticity can be made. Gene expression is also usually measured by reads per kilobase of gene length per million reads (or less bias-prone measures, [Zheng, Chung, and Zhao 2011]) to account for the fact that a certain number of transcripts of a long gene will produce more reads than the same number of shorter gene transcripts. Resources for calculating these in *B. lanceolatum* are not particularly robust given the incomplete nature of many predicted transcripts, but are worse for *S. lamarcki*. In the absence of predicted transcripts from the *S. lamarcki* genome, the only measure of transcript length is contig length, which is not independent of read count and is closely correlated (0.878, using Spearman's rho). Accordingly, patterns of expression indicated by contig coverage are not particularly different from those indicated by read count (Figure 6.5).

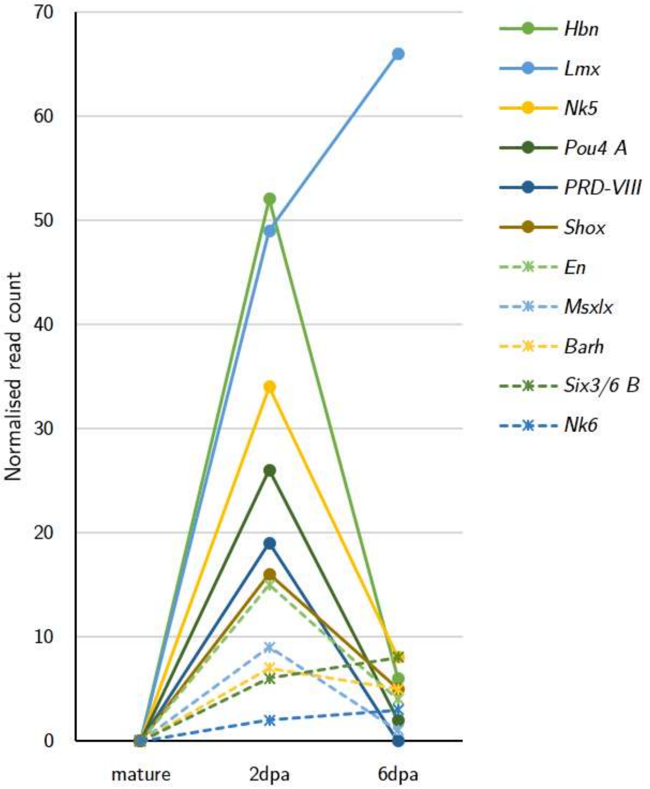


Figure 6.4. Sequences transcribed *de novo* in *S. lamarcki* regeneration, *i.e.* those not found in the mature transcriptome but found in the early (2dpa) and/or late (6dpa) regeneration transcriptomes.

Spearman's rho tests of correlation were calculated for the genes that were found in the transcriptomes of both species ('common,' black numbers, Figure 6.1) and for all the genes found within-species ('all,' purple numbers, Figure 6.1). These numbers were observed to be generally higher than the set of all genes, and correspondingly, when the correlations of read counts for genes found only in one species or the other were calculated (‘unique,’ brown numbers, Figure 6.1), they were found to be lower than the set of all genes. The correlation between read counts in the mature and late (6dpa) regeneration of *S. lamarcki* was an exception to this pattern, and is more closely correlated than the set of all genes. To determine if this could be indicative of a greater degree of differential regulation amongst genes outwith a conserved regenerative, homeostatic, cellular, etc. programme, I calculated the correlations between the read counts of 200 randomly-chosen (using the pseudo-random number generation capability of spreadsheet software) subsets of the data of equivalent size to the unique gene sets (n_u = 21 and 56 for *S. lamarcki* and *B. lanceolatum*, respectively) to determine the probability of this signal occurring by chance. In the *S. lamarcki* random subsets, 10% of the mature-2dpa, 18% of the 2dpa-6dpa, and 76% of the mature-6dpa correlation tests were lower than the unique gene set, and in the *B. lanceolatum* random subsets, 31% of the correlation tests were lower than the unique set. These numbers suggest that the unique gene set is not particularly remarkable and that the observation that it is less strongly correlated between transcriptomes is not likely to be biologically meaningful.

6.1.2. Regenerative axial patterning

A summary of the genes of interest discussed in sections 6.1.2 and 6.1.3 is presented in Table 6.7 (pages 254-255).

6.1.2.1. Homeobox members of a homocratic PD patterning GRN are expressed in operculum regeneration

A highly similar GRN controlling proximodistal axis initiation exists in arthropod and vertebrate appendage development, strongly suggesting some degree of molecular homocracy between the two (Nielsen and Martinez 2003; Svensson 2004; see section 1.2.3.2). However, because arthropod and vertebrate appendages are not considered to be direct morphological homologues, this relationship is referred to as ‘deep homology’ (Shubin, Tabin, and Carroll 1997, 2009). Homocratic, non-homologous relationships are presumed to derive from independent co-option of genes and GRNs; in the case of appendage development, a head patterning GRN is a good candidate (Lemons *et al.*, 2010).

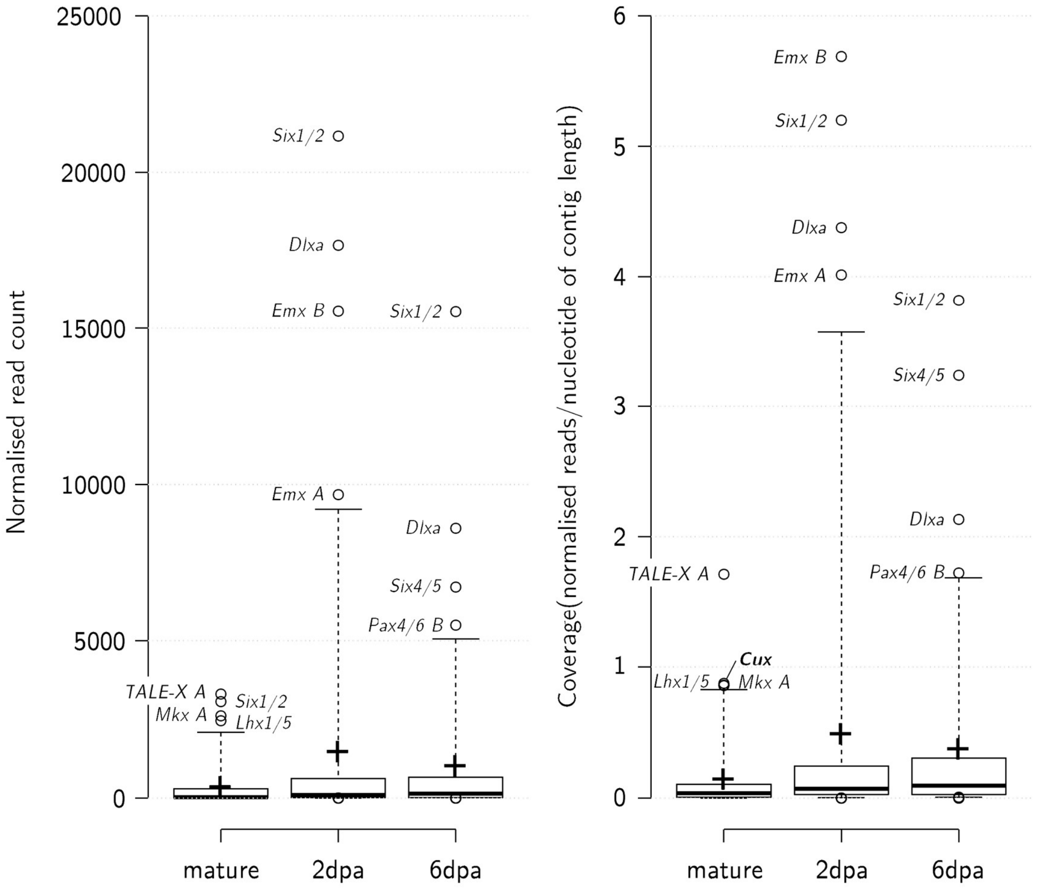


Figure 6.5. Comparison of read count and coverage in *S. lamarcki* transcriptomes; in the graph to the right, the normalised read count (graph on the left) has been divided by the length of the contig to compare the difference in signal. A gene that was not in the outliers indicated in the left graph and in Figure 6.2 (*Cux*) is emboldened. Center lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend to 5th and 95th percentiles, outliers are represented by dots; crosses represent sample means. Adapted from plots produced using BoxPlotR (Spitzer *et al.*, 2014).

The available evidence conflicts on whether a comparable network is operating in the appendage development of errantian polychaetes; components of the network including *Lhx1/5*, *Lhx2/9*, *Dlx*, *Dac* and *omb* orthologues were found to be expressed (but not in regions comparable to arthropods) in *Neanthes arenaceodentata* appendage development (Winchell, Valencia, and Jacobs 2010; Winchell and Jacobs 2013) but Dlx, Meis and Pbx family members and *decapentaplegic* were considered to be expressed in analogous zones in *P. dumerilii,* leading to the suggestion of homology/homocracy (Grimmel, Dorresteijn, and Fröbius 2016). Recently, members of the network including Dlx, Meis, and Pbx family members and *Dac* orthologues have also been found in conserved roles in cephalopod arm development, which is also considered to be an evolutionary novelty (Tarazona *et al.*, 2018 [preprint]).

Orthologues of the homeobox members of this GRN (specifically the Lhx2/9, Meis, Pbx, and Dlx families) are observed to increase in read count during *S. lamarcki* operculum regeneration (summarised in Table 6.2). However, there are some difficulties in interpreting this as the straightforward involvement of a PD patterning GRN (see next subsection). Additionally, orthologues of these gene families are present and increase in read count in *B. lanceolatum* tail regeneration, which suggests it would be wise to be cautious about over-interpreting the available data. The presence of orthologues of non-homeobox commonly-identified members of this GRN, such as *Dac* and the transcription factor *Sp8* (*i.e. buttonhead*) (*e.g.* Lemons *et al.*, 2010; Tarazona *et al.*, 2018), and *in situ* hybridisation of its gene members in development and regeneration would help confirm that this network specifically is at work in *S. lamarcki*.

Table 6.2. Summary of homeobox gene families deployed in appendage patterning. References: [1] Grimmel, Dorresteijn, and Fröbius 2016; [2] Winchell, Valencia, and Jacobs 2010; [3] Winchell and Jacobs 2013; [4] Tarazona *et al.*, 2018 [preprint]; [5] Pueyo and Couso 2005; [6] Williams 2013; [7] Petit, Sears, and Ahituv 2017; [8] Rodriguez-Esteban *et al.*, 1998; [9] Capellini, Zappavigna, and Selleri 2011. Cells include a tick if the gene was, a cross if its absence is specifically reported, or left blank if the study did not report the presence or absence. The numbers given for *S. lamarcki* are the read counts of family members in the mature, 2dpa, and 6dpa transcriptomes; symbols are per Table 6.1. Regen. = regeneration; devo. = development.

Context	Lhx2/9	Dlx	Meis	Pbx
S. lamarcki operculum regen.	✓ 198▲635▲757 31▲92▲144	✓ 978▲17,662▼8,605 13▲746▼434	✓ 1,001▲2,656▼1,240 71≈78▲186	✓ 4≈2▲18
P. dumerilii appendage regen. [1]		✓	✓	✓
N. arenaceodentata appendage devo. [2,3]	X	X		
Cephalopod appendage devo. [4]		✓	✓	✓
Arthropod appendage devo. [5,6,7]	✓	✓	✓	✓
Vertebrate appendage devo. [5,8,9]	✓	✓	✓	✓

Dlx

A previous study of the *S. lamarcki* Dlx paralogues noted the absence of the expression of either paralogue in operculum development, which was considered to be surprising given its conserved roles specifying distal identity (McDougall *et al.*, 2011). Both Dlx paralogues are apparently constitutively expressed in mature opercula, with *Dlxa* being expressed at high levels (978 reads) and *Dlxb* at low levels (13 reads). Both undergo a substantial increase in read count in early regeneration; *Dlxa* increases 18-fold and is the second most abundant homeobox transcript in the early and late regenerating transcriptomes (Figure 6.2) and *Dlxb* increases 57-fold, making it the sequence with the greatest fold change (Figure 6.3). Both see their expression approximately halved in later regeneration. The implication of the absence of Dlx genes in appendage development is that these genes are not recapitulating a developmental role in proximodistal axis specification but have important regeneration-specific functions (see sections 6.1.3.3 & 6.1.4).

An independent co-option of a cephalic patterning GRN?

Tomer *et al.* (2010) reported a GRN responsible for patterning the brain topology of *P. dumerilii* which involved the expression of Lhx2/9, Emx, Pax4/6, Gsx, Nk2.1, Vax and Dlx homeobox family members, a network that they found to share homology with the GRN responsible for patterning the vertebrate brain. The homeobox gene profile of *S. lamarcki* operculum regeneration resembles this network, with very abundant reads reported for multiple paralogues of Emx, Pax4/6 (Figure 6.2), and Nk2.1, as well as the aforementioned Lhx2/9 and Dlx expression. However, neither Gsx nor Vax orthologues were identified.

Although data necessary to support the hypothesis that the evolution of the operculum involved the independent co-option of this brain patterning GRN are not reported herein (for example, the identification in the transcriptomes of the non-homeobox network members identified by Tomer *et al.*), it is tempting to speculate that it could help explain the extremely (*e.g.* Emx, Pax4/6) and relatively high (*e.g.* Lhx2/9, Nk2.1) read counts of members of the homeobox families, which in some cases do not have known roles that help explain their presence.

The evolution of proximodistal axis initiation/polarization and patterning is an important question in evolutionary developmental biology, and spiralian models are taxonomically well-placed to help illuminate these problems. If vertebrates, arthropods, annelids and cephalopods all independently co-opted homologous GRNs for the formation of their appendages, this extreme degree of homocratic convergence would demand either more compelling evidence that appendages are the result of homoplasy and not homology or an explanation for its susceptibility to co-option into appendage development. One such explanation could be that the GRN was involved not just in in cephalic patterning (Lemons *et al.*, 2010) but also in the development of pre-bilaterian anterior sensory appendages (Nielsen and Martinez 2003; Jacobs *et al.*, 2007, 2010; Winchell and Jacobs 2013), from which it was easily co-opted into trunk appendages.

6.1.2.2. Hox genes

Anterior, Medial and Posterior Hox genes are all constitutively expressed in the mature tail, and later in the regenerating tail of *B. lanceolatum*. A comparison of previous reports of Hox gene expression in amphioxus species is given in Table 6.3, indicating that the detection of all Hox gene orthology groups except *Hox10* has not been reported by previous developmental surveys (Oulion *et al.*, 2012; Pascual-Anaya *et al.*, 2012; Yang *et al.*, 2016), although a previous PCR/WMISH survey detected all except *Hox13* (Pascual-Anaya *et al.*, 2012). The present survey has the greatest variety of Hox genes yet reported in transcriptomes of adult tissue (Oulion *et al.*, 2012).

In the case of the Anterior and Medial Hox genes, mature expression seems to be at an extremely low level (a mean of 3.7 reads per gene for *Hox1*-*Hox3* and a mean of 6.2 reads per gene for *Hox4*-*Hox8*). In contrast, for the Posterior Hox genes, many more reads are present (a mean of 57 reads per gene for *Hox9* + *Hox11*-*Hox15*). In both the mature and regenerative transcriptomes, Posterior Hox genes are amongst the most abundant in read counts (Figure 6.2). This finding is consistent with the possibility that Hox genes are expressed in the adult amphioxus tissues in an ongoing role in positional specification, as has been extensively reported in vertebrates (reviewed by Wang, Helms, and Chang 2009; see section 1.2.4.1).

Table 6.3. Comparison of detection of Hox gene expression by previous studies. A plain tick indicates that the study reports the expression pattern of the gene via *in situ* hybridisation; a tick in parentheses indicates that the study reports the expression of a gene via detection in a transcriptome; a tick in square brackets indicates that *in situ* hybridisation did not produce a positive result but PCR did; and 'a' that the gene was found to be absent from transcriptome searches. 8c = 8 cells; gas. = gastrula; 48h = 48 hours post fertilization larvae; mat. = mature tissue; reg. = regenerating tissue; *B. lan.* = *Branchiostoma lanceolatum*; *B. bel.* = *Branchiostoma belcheri.*

Study	*Hox1*	*Hox2*	*Hox3*	*Hox4*	*Hox5*	*Hox6*	*Hox7*	*Hox8*	*Hox9*	*Hox10*	*Hox11*	*Hox12*	*Hox13*	*Hox14*	*Hox15*	Context	Species
Oulion *et al.,* (2012)	(✓)	a	(✓)	(✓)	a	(✓)	a	a	a	a	a	a	a	a	a	8c - adult	*B. lan.*
Pascual-Anaya *et al.,* (2012)	✓	[✓]	✓	✓	[✓]	✓	✓	[✓]	[✓]	✓	[✓]	[✓]	a	✓	[✓]	gas. – 48h	*B. lan.*
Yang *et al.,* (2016)	(✓)	(✓)	(✓)	(✓)	(✓)	(✓)	(✓)	(✓)	a	a	a	a	(✓)	(✓)	a	egg – 48h	*B. bel.*
Present study	(✓)	(✓)	(✓)	(✓)	(✓)	(✓)	(✓)	(✓)	(✓)	a	(✓)	(✓)	(✓)	(✓)	(✓)	mat. + reg.	*B. lan.*

Despite their high read counts, the Posterior Hox gene complement sees a decrease (*Hox9* & *Hox15*) or only a modest increase (*Hox11, Hox12*, *Hox13*, & *Hox14*,) in apparent expression in regeneration (Figure 6.2 & Figure 6.6). This observation differs from a study in the tail regeneration of a newt, which found a 2-to-20-fold increase in *HoxA9, HoxC10, HoxC12,* and *HoxC13* in regeneration (Nicolas *et al.*, 2003). This difference between amphioxus and newt regeneration does not refute the patterning hypothesis, which might not necessarily require strong upregulation. In addition, the early blastemal stage sampled could easily be too early in regeneration to observe Hox genes acting in these roles. As well as their known roles in patterning adult and regenerating tissue, vertebrate *HoxA9*, *HoxC10*, *HoxA13*, and *HoxC13* have been implicated in blastema formation (see section 1.2.4.3), indicating a variety of potential roles could be possible for these genes. The paucity of evidence makes it currently impossible to determine the relationship between developmental, adult, and regenerative expression patterns except to note that the absence of *Hox10* expression, which was detected in posterior tissue in 48 hpf larvae (Pascual-Anaya *et al.*, 2012) could indicate a difference between developmental and adult/regenerative Hox deployment. The absence of evidence for Posterior Hox deployment (other than *Hox10* and *Hox14* in 48 hpf *B. lanceolatum* larvae) probably indicates that these genes are not expressed until later in development or at metamorphosis.

The apparent increase in *Hox3*, *Hox4*, and *Hox5* expression are also a matter of interest. Hox3 semi-orthologue *HoxA3* has been shown to induce endothelial and epithelial (Mace *et al.*, 2005) and stem (Mace *et al.*, 2009) cell migration during wound healing in mice. More evolutionarily distantly, a *Hox3* gene has been found expressed in annelid posterior blastemas (Novikova *et al.*, 2013) and echinoderm arm blastemas (Ben Khadra *et al.*, 2014). *Hox3* and *Hox4* have been reported to be expressed in the posterior of amphioxus 48 hpf larvae (Pascual-Anaya *et al.*, 2012). Hox4 family genes are expressed in the epidermis during development in humans (Kömüves *et al.*, 2002) and in *B. floridae* (Schubert *et al.*, 2004) but surprisingly not in *B. lanceolatum* (Pascual-Anaya *et al.*, 2012), where its expression is restricted to the posterior central nervous system. Hox4 and Hox5 family members are infrequently reported in conjunction with regeneration, although they were also detected in echinoderm blastemas (Ben Khadra *et al.*, 2014). Hox5 family members have even fewer known roles that could be relevant to amphioxus regeneration, and no previously reported amphioxus expression pattern.

Detailed profiling of the expression of Hox genes via qPCR and *in situ* hybridisation – both in late development and in mature and regenerating tails – is undoubtedly a major priority in future studies of amphioxus regeneration.

In contrast to *B. lanceolatum* tails, *S. lamarcki* operculum regeneration does not involve a broad selection of Hox genes. A single Hox gene was found and inferred by phylogenetic analysis and process of elimination to be a highly divergent *Antp* gene (Chapter 3). The *A. virens Antp*-orthologue *Hox7* was reported in bilateral domains underlying the wound epithelium and in the nascent posterior growth zone (Novikova *et al.*, 2013), but *Antp* is unchanging in mature tissue and absent from the blastema of the more closely-related sedentarian *C. teleta* (de Jong and Seaver 2016). Regardless, the protein sequence of *S. lamarcki Antp* is divergent enough to indicate that it may have undergone radical concurrent *cis*-regulatory change to the extent that pre-existing data on annelid expression of *Antp* may be unhelpful in speculating about its roles in *S. lamarcki.* The exciting possibility exists that the divergence of *Antp* could be related to the evolutionary novelty of the development and regeneration of the operculum.

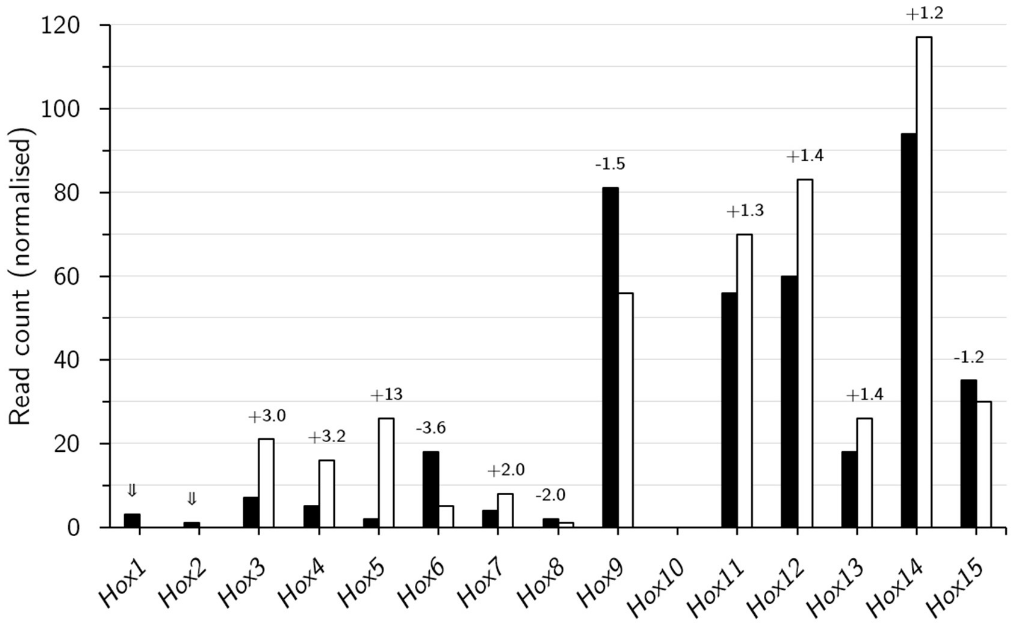


Figure 6.6. Read counts of Hox genes in the *B. lanceolatum* transcriptomes. Normalised read counts from the mature (black) and regenerating (white) tail transcriptomes are shown, with their fold change stated above to two significant figures. *Hox10* was not detected in either transcriptome. Counts were taken from whichever method (*de novo* assembly *vs.* mapping to predicted transcripts) detected more reads.

6.1.2.3. ParaHox genes

The expression of *Cdx* and *Gsx* has previously been reported during the development of the operculum in *S. lamarcki*, in which they are both localised in the pre-adult to the spines and the outward face of the cup (Hui 2008). Surprisingly, no ParaHox genes were detected in the mature or regenerating operculum transcriptomes. It is possible that they are expressed transiently between 2 and 6 dpa and were therefore missed by the two transcriptomes; the pre-adult opercula expressing *Cdx* and *Gsx*, in which the beginnings of mineralization are visible, may be less mature than a 6 dpa regenerating operculum, in which mineralization is well underway and pigmentation is beginning. A PCR survey of a densely-sampled time-course of *S. lamarcki* regeneration would help determine if *Cdx* and *Gsx* are ever expressed.

Cdx was detected in the regeneration of *B. lanceolatum*, where it is apparently expressed strongly *de novo* (0 reads in the mature to 56 reads in the regenerating tail, the joint 8th most abundant homeobox gene in the regenerative transcriptome and the strongest presumptive signal of *de novo* expression). *Cdx* expression has been reported in *B. floridae* in the posterior developing neurectoderm and ectoderm tissue throughout development, including in the posterior neural tube and hindgut as late as 7 dpf (Brooke, Garcia-Fernàndez, and Holland 1998; Osborne *et al.*, 2009). Treatment with exogenous retinoic acid was shown to posteriorly compress its domain of expression (Osborne *et al.*, 2009). In vertebrates, Cdx ohnologues function in specifying posterior cell fates and in axial elongation with various deficiencies being associated with severe posterior truncation (Joly *et al.*, 1992; Subramanian, Meyer, and Gruss 1995; Charite *et al.*, 1998; Isaacs, Pownall, and Slack 1998; van den Akker *et al.*, 2002; Chawengsaksophak *et al.*, 2004; T. Young *et al.*, 2009; Marlétaz *et al.*, 2015, reviewed by Deschamps and Nes 2005; Beck and Stringer 2010). As suggested by the name of the *D. melanogaster* orthologue *caudal* (after which Cdx is named), Cdx has older roots in posterior (*i.e.* caudal) patterning, performing homologous roles in *D. melanogaster* and *C. elegans* as well as less derived protostome models like the red flour beetle *T. castaneum* and the errantean annelids *P. dumerilii* and *A. virens* (Copf, Schröder, and Averof 2004; de Rosa, Prud'homme, and Balavoine 2005; Kulakova, Cook, and Andreeva 2008). Cdx expression has been found to regulate the expression of Hox genes in these posterior contexts (Subramanian, Meyer, and Gruss 1995; Charite *et al.*, 1998; Isaacs, Pownall, and Slack 1998; van den Akker *et al.*, 2002; Chawengsaksophak *et al.*, 2004; Copf, Schröder, and Averof 2004; Deschamps and Nes 2005; T. Young *et al.*, 2009).

The discovery of apparently *de novo Cdx* expression in posterior regeneration in amphioxus is therefore significant because of the possibility that it is recapitulating a developmental programme of induction of posterior elongation. Cdx genes have previously been reported in *P. dumerilii* posterior regenerative blastemas (de Rosa, Prud'homme, and Balavoine 2005) and in nemertean posterior regeneration (Charpignon 2007), but have not been examined by studies of vertebrate posterior regeneration (zebrafish caudal fins or newt tails). Profiling the expression of *Cdx* in regeneration with qPCR and *in situ* hybridisation is another major priority for ongoing study. Further, it might be possible to influence the expression of *Cdx* – and via it, the Hox genes – by treating regenerating animals with retinoic acid, either by microinjection or bead implantation.

6.1.2.4. Other homeobox genes

Evx

Evx family genes have ancient roles in posterior patterning in gastrulation and posterior growth in later development, being found in related roles in ecdysozoans (*e.g.* Copf *et al.*, 2003), annelids (de Rosa, Prud'homme, and Balavoine 2005), vertebrates (Bell *et al.*, 2016), and amphioxus (Ferrier *et al.*, 2001; Yu *et al.*, 2007). Amphioxus possesses two paralogues; *EvxA* is expressed in the tailbud during amphioxus development and is prototypical of the chordates in regards to sequence and expression domain, whereas *EvxB* is highly divergent in sequence and seems to be uniformly and strongly expressed in the ectoderm of post-hatching embryos and larvae (Ferrier *et al.*, 2001). *EvxA* was detected in the mature and regenerating transcriptomes (4 and 6 reads, respectively) whereas *EvxB* was not detected. Although the minimal expression and unconvincing read count change of *EvxA* is rather underwhelming, this gene remains an important candidate for studying developmental patterning mechanisms in amphioxus regeneration, particularly in stages later than the one sampled for the transcriptome.

Emx

The Emx paralogues *Emx A* and *Emx B* prominently increase in read count in the *S. lamarcki* early (2dpa) regeneration transcriptome and continue to be abundant in late (6dpa) regeneration (Figure 6.2 & Figure 6.5). In contrast, the *B. lanceolatum* paralogues *Emxa*, *Emxb* and *Emxc* appear at low levels in the mature transcriptome (5, 23, and 25 reads, respectively) and increase (*Emxa*, to 14 reads) or decrease (*Emxb*, to 12 reads, and *Emxc*, to 13 reads) in regeneration.

Emx genes have been found to have a role in the regeneration of urodele amphibian limbs (Beauchemin *et al.*, 1998; Monaghan *et al.*, 2012) in a proximodistal gradient in the epidermis overlying the regenerative blastema. Given the paucity of reports in other regenerative contexts, it might be imprudent to suppose that the *S. lamarcki* Emx paralogues are engaging in some kind of conserved role in regeneration. Nonetheless, their very high read counts suggest that they could potentially be important. Emx genes in *S. lamarcki* are candidates for future investigation.

6.1.3. Blastemas and stem cells

6.1.3.1. Satellite cells

As well as Pax3/7 family genes (see Chapter 5), several other homeobox families are known to be expressed in and regulate the behaviour of satellite cells, including the Barx (Makarenkova and Meech 2012), Lbx (Watanabe *et al.*, 2007), Pitx (Knopp *et al.*, 2013), Six1/2 (Yajima *et al.*, 2010; Le Grand *et al.*, 2012), and Six4/5 (Yajima *et al.*, 2010) families. The read count data from the *B. lanceolatum* orthologues of these families is presented in Table 6.4.

Table 6.4. Read counts of *B. lanceolatum* genes from families with known roles in satellite cell regulation. Symbols per Table 6.1.

Gene	Transcriptome		
	mature		regen.
Barx	NOT DETECTED		
Lbx	6	≈	8
Pax3/7b	6	▲	21
Pitx	NOT DETECTED		
Six1/2	52	▲	75
Six4/5	16	▼	11

Surprisingly, only one of these genes mirrors the pattern of regenerative upregulation suggested by the read counts of *Pax3/7b*; *Six1/2* is relatively abundant in the mature transcriptome and more so in the regenerative transcriptome; in the latter, it is the 4th most abundant homeobox gene transcript (Figure 6.2). The absence of *Barx* and *Pitx* and apparent *Six4/5* downregulation might be explained by these families having roles in satellite cell differentiation, (Yajima *et al.*, 2010; Makarenkova and Meech 2012; Knopp *et al.*, 2013) in which they would be active after the stage of regeneration sampled in the transcriptome used herein, although Barx also seems to promote proliferation. Vertebrate *Lbx1* is expressed in activated but not quiescent satellite cells, making the minimal response by *B. lanceolatum Lbx* more difficult to explain.

Table 6.5. Read counts of POU-class genes in the transcriptomes of *S. lamarcki* and *B. lanceolatum.* Symbols per Table 6.1.

Family			*S. lamarcki*			*B. lanceolatum*		
	mature		2dpa		6dpa	mature		regen.
Pou1	NOT DETECTED					1	(▼̲)	0
Pou2	185	▲	289	▲	699	NOT DETECTED		
Pou3	90	▼	77	▲	179	30 NOT DETECTED	▲	38
Pou4	0 285	▲̲ ▲	26 585	▼ ≈	2 592	2	(▼)	1
Pou6	14	▼	5	▲	16	9	▲	29

In *S. lamarcki, Pou4 A* is expressed *de novo* in the early (2dpa) regenerative transcriptome and is apparently downregulated in late (6dpa) regeneration, while its paralogue *Pou4 B* is expressed constitutively in mature opercula and is approximately doubled in both regenerative transcriptomes. A putative Pou4 gene (Gold, Gates, and Jacobs 2014), *Smed-POU-P1,* is specifically expressed in planarian neoblasts alongside homologues of up- and down-stream members of the vertebrate pluripotency GRNs (Önal *et al.*, 2012). The other POU-class genes are not detected (*Pou1*), peak in late (6dpa) regeneration (*Pou2*, *Pou3*) or are apparently downregulated in early (2dpa) regeneration (*Pou3*, *Pou6*). The question of the presence of pluripotency or multipotency in the decentralised proliferative cell population observed in opercular regeneration (Szabó and Ferrier 2014) is unresolved, but important.

In the *B. lanceolatum* transcriptome, most POU-class genes are not detected (*Pou2*, *Pou3L*) or very weakly detected and apparently down-regulated (*Pou1*, *Pou4*).

Pou3 and *Pou6* are detected with moderate read counts and apparently upregulated in regeneration (the latter triples in read count in the regenerative transcriptome). Vertebrate adults and adult regenerative processes seem to lack pluripotent cells (Kragl *et al.*, 2009; Slack 2017), particularly the tail regeneration of anuran amphibians, which seems to involve tight lineage restrictions (Gargioli and Slack 2004). Given the overall similarities between cephalochordate and vertebrate regeneration (Somorjai *et al.*, 2012), it would be very surprising to discover the involvement of pluripotent cells in *B. lanceolatum* regeneration, but fascinating given the deleterious tumorigenic properties of induced pluripotent stem cells in vertebrates (*e.g.* Abad *et al.*, 2013) and the contrasting capacity of some tunicates to perform WBR using totipotent blood cells (Rinkevich, Shlemberg, and Fishelson 1995).

6.1.3.2. Muscle dedifferentiation

Msx

Msx genes are expressed in vertebrate regeneration in the blastema (*e.g.* Koshiba *et al.*, 1998; *c.f.* Taghiyar *et al.*, 2017) and in the muscle proximal to the amputation, where it induces dedifferentiation and fragmentation of mature myofibres into multipotent blastema cells (see section 1.2.4.3, reviewed by Frasch 2016). *Msx* has previously been detected in the blastema and overlying mesenchyme of amphioxus (Figure 1.12, Somorjai *et al.*, 2012) and is found in both the mature and regenerating transcriptomes (14 & 18 reads, respectively). The roles of *Msx* in amphioxus regeneration – specifically whether it is involved in the dedifferentiation of muscle fibres proximal to the wound site, which do fragment but in which *Msx* expression was not detected by *in situ* hybridisation (Somorjai *et al.*, 2012) – are an important avenue for future research.

Msx expression was also detected in the *S. lamarcki* transcriptomes (6, 22, and 49 reads in mature, 2dpa and 6dpa transcriptomes respectively). *S. lamarcki* regeneration does not involve a blastema (section 1.3.1) but does involve the fragmentation of muscle fibres (Bubel and Thorp 1985; Bubel *et al.*, 1985), though it is not known if they contribute to a population of proliferative cells, nor anything about the proliferative cells aside from their distribution (Szabó and Ferrier 2014). *Msx* is an important candidate gene for future studies of *S. lamarcki* regeneration.

Barx

Barx gene expression has been found to have a similar effect in causing the dedifferentiation of mature myotubes in a mouse culture model (Meech *et al.*, 2010; Makarenkova and Meech 2012) but unlike *Msx*, also has a role in producing early myotube differentiation. *Barx* is observed to be present in relatively high read counts in the *S. lamarcki* mature and regenerative transcriptomes, appearing to be upregulated in early (2dpa) regeneration and then downregulated to below its mature expression level in late (6dpa) regeneration (847, 1471 and 621 reads respectively). Barx was only relatively recently described in protostomes (Paps *et al.*, 2015). *Barx* was not found in the amphioxus transcriptomes.

6.1.3.3. Blastema markers

Dlx

Like in *S. lamarcki, Dlx* is the most abundant homeobox transcript in the *B. lanceolatum* mature and regenerative transcriptomes (Figure 6.2), showing a slight increase in the latter (212 to 234 reads). Dlx genes have previously been found in mature tissue and tail and limb blastemas of the newts *N. viridescens* and *P. waltl* (Beauchemin and Savard 1992; Nicolas *et al.*, 1996). Expression of *NvDlx-3* was found to be constitutive and constant in the skin, but upregulated two-fold in connection with muscle and central nervous system regeneration (Nicolas *et al.*, 1996). In comparison, *Dlx* reads only increase 1.1-fold in regeneration in *B. lanceolatum*, although the regenerative transcriptome could be drawn from too early a time point to see a substantial increase associated with potential roles in muscle re-differentiation, though ependymal tube elongation is underway (Somorjai *et al.*, 2012). The abundance of *Dlx* reads in amphioxus regeneration suggests that profiling its expression in mature and regenerating tissue is a priority.

Prrx

Prrx is present in the mature *B. lanceolatum* tissue transcriptome (2 reads) and undergoes the strongest apparent upregulation of any gene in the regenerating transcriptome (to 44 reads, Figure 6.3). In contrast, *S. lamarcki Prrx* is apparently expressed moderately in mature tissue (149 reads), is approximately halved in early (2dpa) regeneration (71 reads) and is re-expressed at the same moderate level in late regeneration (153 reads). Prrx genes are expressed constitutively at low levels in urodele skin (Makanae *et al.*, 2013; Lehrberg and Gardiner 2015) and in the distal mesenchyme of limb blastemas, and have been used as blastema markers and sources of blastema-specific enhancers (Satoh *et al.*, 2007; Suzuki *et al.*, 2007; Satoh *et al.*, 2011; Yokoyama *et al.*, 2011; Lehrberg and Gardiner 2015). The read count data suggest that these expression domains could be homologous between cephalochordate posterior blastemas and vertebrate limb blastemas. The strong apparent upregulation of *Prrx* in *B. lanceolatum* regeneration flags it as an important candidate gene for future study.

6.1.4. Biomineralization

Biomineralization is an interesting phenomenon for evolutionary developmental biology. Biomineralized structures are often an evolutionarily important part of the body plans in which they are deployed. They offer a window into the morphology of ancient creatures via their durability, and potential insight into the mechanisms that link genome, GRN, and gene evolution and deployment to the morphology of extant animals. However, our understanding of the evolution of those GRNs and the mechanisms that they govern is still relatively insubstantial (Wilt, Killian, and Livingston 2003; Jackson and Degnan 2016). A conserved deuterostome biomineralization ‘toolkit’ (including the homeobox gene family Alx) was hypothesised by Livingston *et al.* (2006), but the fact that the TFs involved control very different downstream genes indicates these similarities might be coincidental or at best homocratic, rather than homologous. The evidence for a broader nephrozoan, bilaterian or metazoan biomineralization toolkit is even poorer, with few genes found to be conserved between the mineralizing transcriptomes of molluscan classes (reviewed by McDougall and Degnan 2018), let alone between phyla. However, the possibility that a spiralian, protostome or even metazoan biomineralization toolkit might exist has not yet been conclusively discounted.

Although most Evo-Devo studies on biomineralization focus on molluscs (Jackson and Degnan 2016) – which have the advantage of having relevance to ocean ecology and food security – the biomineralizing annelids (the Sabellida, see section 1.3.2.2) are a vital component of robust sampling of the biomineralizing taxa. A previous study profiled the involvement of *msp130*, a cell surface glycoprotein, in *S. lamarcki* regenerative biomineralization; *msp130* homologues are known to be involved in the biomineralizating processes of deuterostomes and other protostomes (Szabó and Ferrier 2015). Cephalochordates do not produce biomineralized structures, so the discussion of biomineralization toolkit genes herein is limited to *S. lamarcki*.

Table 6.6. Read counts of *S. lamarcki* genes from selected families with known roles in spiralian and deuterostome biomineralization. Symbols per Table 6.1.

	Transcriptome				
	mature		2dpa		6dpa
Alx	NOT DETECTED				
Dlxa	978	▲	17662	▼	8605
Dlxb	13	▲	746	▼	434
En	0	▲̲	15	▼	4
Gbx	NOT DETECTED				
Hox1	NOT DETECTED				
Hox4	NOT DETECTED				
Post1	NOT DETECTED				
Post2	NOT DETECTED				
Msx	6	▲	22	▲	49

Several homeobox families have been associated with biomineralization in the molluscs and brachiopods, including Hox1, Hox4, Post1, Post2, Msx, En, Gbx, and Dlx. Some of these (see relevant subsections) also have roles in vertebrate biomineralizing processes. A summary of the read counts of these genes in the *S. lamarcki* transcriptomes is provided in Table 6.6. Mineralization of the distal plate of the regenerating *S. lamarcki* operculum starts at 2-3 dpa (Figure 1.7, Szabó and Ferrier 2014) and is mostly complete by 6dpa.

Dlx

Dlx genes play a part in tooth and bone formation in vertebrate development (Bendall and Abate-Shen 2000; Lézot *et al.*, 2000, 2002; Panganiban and Rubenstein 2002; Morsczeck 2006; Ryoo, Lee, and Kim 2006), and in patterning the shell field in gastropods (Jackson and Degnan 2016). The extremely abundant (*Dlxa,* Figure 6.2) and apparently strongly upregulated (*Dlxb,* Figure 6.3) Dlx paralogue expression in *S. lamarcki* regeneration could be related to a function in biomineralization. A previous study reported the absence of Dlx gene expression in the developing operculum, although whether the opercula of the juvenile animals they used were mineralizing is not reported. However, the animals were de-calcified to remove their nascent habitation tube, and no Dlx expression was reported in the thorax or thoracic collar region (McDougall *et al.*, 2011) from which adult animals produce and deposit mineralized material for their tubes. The necessity of such a step suggests that calcification was actively occurring in this region, indicating that Dlx is not involved in thoracic calcification. Although some other serpulid species produce opercular plates and habitation tubes with similar mineralogical composition (*e.g.* Riedi 2012; referenced by Szabó 2015), *S. lamarcki* opercular plates and habitation tubes are different (Bubel *et al.*, 1983), suggesting the faint possibility of different thoracic and opercular mineralizing processes controlled by distinct genes. A more spatiotemporally rich qPCR time-course or *in situ* hybridisation would help to determine the potential roles of Dlx genes in *S. lamarcki* operculum regeneration.

En

Despite the extreme diversity of mollusc shell transcriptomes, En family genes are quite consistently found as a regulator of shell formation, being found in gastropods (Moshel, Levine, and Collier 1998; Nederbragt, van Loon, and Dictus 2002; Iijima *et al.*, 2008), bivalves (Jacobs *et al.*, 2000; Kin, Kakoi, and Wada 2009), cephalopods (Baratte, Andouche, and Bonnaud 2007), chitons (Jacobs *et al.*, 2000) and scaphopods (Wanninger and Haszprunar 2001), as well as brachiopods (Shimizu *et al.*, 2017) and in bone formation in vertebrates (*e.g.* Deckelbaum *et al.*, 2006). *En* is apparently expressed at a low level *de novo* in early (2dpa) *S. lamarcki* operculum regeneration (15 reads), and then is reduced in late (6dpa) regeneration (4 reads). Despite its low read count, *En* is an important candidate gene for future studies in annelid biomineralization.

Msx

As well as roles in muscle dedifferentiation and in blastemas, Msx genes are expressed in biomineralizing processes in tooth (reviewed by Suryadeva and Khan 2015) and craniofacial development (reviewed by Alappat, Zhang, and Chen 2003) and adult skeletal tissue (Lézot *et al.*, 2000) in the vertebrates, and in oyster shell formation (Zhao *et al.*, 2014). These taxonomically distant examples could be the result of independent co-option, but the biomineralization of the distal opercular plate is another potential role for the *Msx* expression observed in *S. lamarcki* regeneration. Unlike the other genes in Table 6.6, *Msx* is detected most strongly in the transcriptome of late (6dpa) regeneration, well after the onset of biomineralization (Figure 1.7).

Missing genes

Gbx, Pax2/5/8 (Wollesen *et al.*, 2017), Hox1, Hox4 (Hinman *et al.*, 2003; Samadi and Steiner 2009; Wollesen *et al.*, 2017), and Hox9-15 (*i.e.* Post1 and Post2) (Samadi and Steiner 2009) family genes have previously described roles in mollusc shell field formation. Alx family genes were suggested to be master regulators of the deuterostome biomineralization toolkit (Ettensohn *et al.*, 2003; Livingston *et al.*, 2006; Ettensohn 2009). Orthologues of these genes were not detected in *S. lamarcki* operculum regeneration, suggesting that these genes are not involved in *S. lamarcki* biomineralization.

Table 6.7 Summary of the known involvement of various homeobox gene famililes in regenerative processes. The system studied herein to which the families are of potential relevance (Sp. column) are indicated with *B.lan* (*B. lanceolatum*) and *S.lam* (*S. lamarcki*). References for publications are given in the text. Potentially interesting genes (italicized) or gene families which were not represented in the pertinent transcriptome are indicated in parentheses.

Potential relevance	Sp.	Genes/Gene family members	Published systems	Published expression locations & roles
Wound healing, cell migration	*B.lan*	Hox3	Mouse	Endothelial, epithelial, & stem cell migration during healing
Wound healing	*B.lan*	Hox4	Human; *B. floridae* but not *B. lanceolatum*	Epidermal development
Satellite cell control	*B.lan*	(Barx, Pitx), Lbx, Pax3/7, Six1/2, Six4/5	Mouse	Satellite cells proliferation, maintainance and differentiation
Pluripotency	both	POU class	Vertebrates, planarian	Satellite cells & planarian neoblasts
Cell potency & identity control	both	Msx	Planarian, cnidarian	Planarian neoblasts, cnidarian muscle transdifferentiation
Myotube dedifferentiation	both	Msx	Mouse, newt	Myotube dedifferentiation & fragmentation
	S.lam	Barx	Mouse	Early myotube dedifferentiation
Blastema formation/maintainance	*B.lan*	Hox3	*A. virens*	Blastema
		Hox4, Hox5	Echinoderms	Blastema
Blastema marker	*B.lan*	Prrx	Urodele amphibian	Adult skin, distal mesenchyme of limb blastemas
Head appendage patterning, co-option	*S.lam*	(Gsx, Vax), Lhx2/9, Emx, Pax4/6, Nk2.1, Dlx	*P. dumerilii*	Cephalic development
Proximodistal patterning	*S.lam*	Emx	Urodele amphibian	Proximodistal gradient in epidermis over blastema
		(Lhx1/5), Lhx2/9, Meis, Pbx, Dlx	*P. dumerilii* , *N. arenaceodentata* , cephalopod, arthropods, vertebrates	P-D axis patterning in appendage development and regeneration
Posterior patterning	*B.lan*	(*Hox10*)	*B. lanceolatum*	Tail development
		Posterior Hox	Urodele amphibian	Tail regeneration, adult tissue
		Evx	Thoughout the Bilateria	Posterior patterning from gastrulation onwards
		EvxA	*B. floridae*	*Tailbud development*
		(*EvxB*)	*B. floridae*	*Ectoderm of post-hatching embryos and larvae*
Posterior patterning, elongation	*B.lan*	Cdx	*P. dumerilii* , nemertean	Posterior regeneration
			Thoughout the Bilateria	Posterior patterning, Hox gene regulation
Biomineralization	*S.lam*	Dlx	Gastropods	Shell field patterning
			Vertebrates	Tooth and bone development
		En	Throughout the Mollusca	Regulation of shell formation
			Vertebrates	Bone formation
		Msx	*Pinctada fucata*	Shell formation
			Vertebrates	Tooth and craniofacial development; adult skeletal tissue
Muscle and nervous tissue regeneration	both	Dlx	Newt	Muscle and CNS regeneration
Previous detection	*B.lan*	*Msx*	*B. lanceolatum*	*Blastema*
Absence	*S.lam*	(*Cdx, Gsx*)	*S. lamarcki*	Operculum development
		(Gbx, Pax2/5/8, Hox1, Hox4, Hox9-15)	Molluscs	Shell field formation
		(Alx)	Deuterostomes	Master regulators of biomineralization pathway
Unknown	*S.lam*	*Hox7* /*Antp*	*A. virens* but not *C. teleta*	Blastema and Posterior Growth Zone

6.2. Interpreting transcriptomic data

The preceding discussion of the roles of homeobox genes in regeneration has reported the read counts of various *S. lamarcki* and *B. lanceolatum* homeobox genes, and has adopted the implicit premise that a signal of the involvement or not in regenerative processes exists in the presumptive measure of up- or downregulation from read counts. However, this assumption can be entirely misleading. Homeobox genes frequently participate in multiple GRNs, mediating different signals and regulating different sets of targets, in the same tissues. Figure 6.7 depicts a scenario in which a hypothetical homeobox gene (‘Hhx’) appears to be downregulated in regenerating tissue but in fact has important regenerative roles associated with strong expression in a small number of specific cells. Data of this kind are out of reach of transcriptomes produced from amalgamated tissue samples, and consequently care should be taken to avoid over-interpreting read counts.

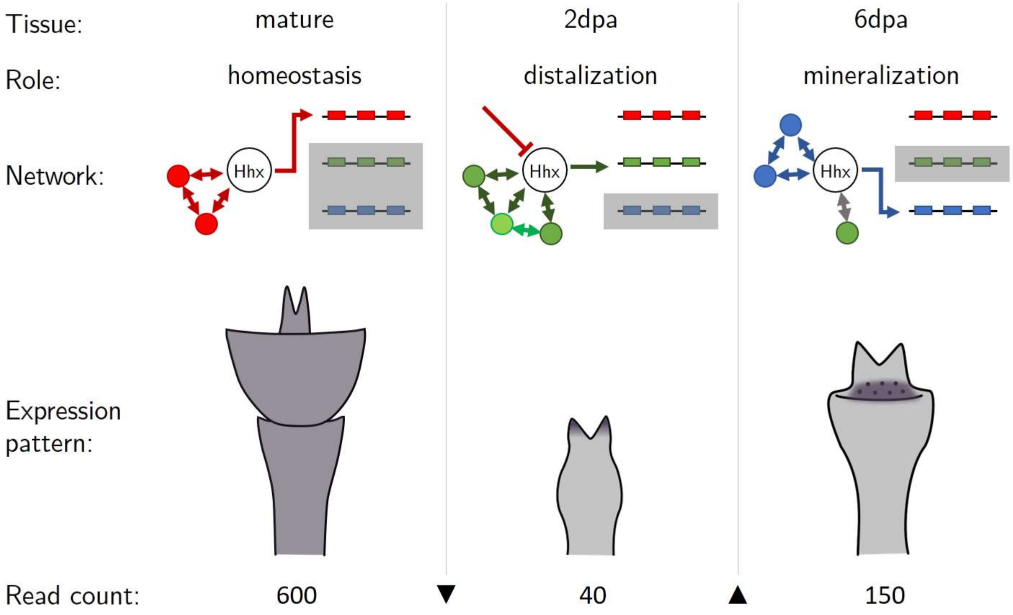


Figure 6.7. How a hypothetical homeobox gene with important roles in regeneration might appear to be downregulated. The hypothetical homeobox gene product (‘Hhx’) is a member of three GRNs (circles, boxes, & arrows), involved in homeostasis (red), distalization (green) and mineralization (blue), and is ubiquitously but weakly expressed in its mature homeostatic role, producing a large read count, but strongly expressed in low numbers of cells in regeneration-specific roles, producing lower read counts and the appearance of downregulation.

The issue of localization and sample amalgamation has broader implications for transcriptomic surveys of genes involved in processes like regeneration. Such surveys often praise the unbiased nature of their own methodology in comparison to a candidate gene approach, which has an inevitable tendency to highlight homology and miss apomorphic GRN reconfiguration or innovation. Beyond the existing criticisms of transcriptomic approaches to identifying important genes (*e.g.* Sarup *et al.*, 2011; Evans 2015), biologically and evolutionarily meaningful spatiotemporal signals enacted by homeobox genes and other transcription factors can be entirely lost by these crude data. Recent developments in single-cell transcriptomics (reviewed by Lee 2017; Svensson, Vento-Tormo, and Teichmann 2018) suggest a future in which it is possible to produce single-cell transcriptomes of an entire regenerative structure in which spatial origin data of cells are preserved, and in the meantime certain types of spatial transcriptomics are already possible (*e.g.* Ståhl *et al.*, 2016).

6.3. The importance of manual curation of genomic data

My analyses of the unusual homeobox genes in *S. lamarcki* regeneration (Chapter 3) and the cephalochordate *Pax3/7* duplication (Chapter 5) are illustrations of the value of manually examining data. The duplication of *Pax3/7* in cephalochordates would not have been discovered if I had relied on an entirely automated process of homeodomain identification, and the searches and classification of the former process relied on manual legwork. Automated pipelines as yet are ill-equipped to notice when they encounter an anomaly that warrants further investigation, let alone to make the qualitative judgements on which are often based our understanding of the orthology relationships between genes.

However, these manual analyses are extremely time-consuming. As genomic data availability increases explosively, as well as hitting storage and skill bottlenecks (Barone, Williams, and Micklos 2017; Papageorgiou *et al.*, 2018), there exists a growing disparity between the sum of genomic data and the data that have been rigorously surveyed. As I suggested in section 3.4.9, some relatively simple tools to integrate, manage and present search data for manual inspection could easily be constructed to minimise the time burden of genomic homeobox survey. Although these obviously don't represent a solution to the growing problem of meaningful big data analysis, it is possible that they would at least be sufficient for the purpose of ongoing robust evolutionary developmental study of homeobox gene diversity.

6.4. Homeobox gene radiation in evolution

It has been proposed that the duplications of ANTP-class genes that produced the Hox, ParaHox and Nk clusters could have contributed to a permissive genomic environment from which emerged the bilaterian bodyplan and with it, the Cambrian explosion (Holland 2015), although the majority of homeobox families were already in place in the planulozoan (Cnidaria + Bilateria) ancestor (see Figure 1.4), including a diverse ANTP class (Thomas-Chollier and Martinez 2016). The idea of (and evidence for) an association between transcription factor diversity and evolutionary novelty or increasing developmental/morphological complexity is an old and enduring one (*e.g.* Holland *et al.*, 1994; Valentine, Collins, and Meyer 1994; Lundin 1999; Miyata and Suga 2001; Levine and Tjian 2003; Wagner, Amemiya, and Ruddle 2003; McCarthy and Enquist 2005; Vogel and Chothia 2006; Degnan *et al.*, 2009; Charoensawan, Wilson, and Teichmann 2010; Pick and Heffer 2012; Mendoza *et al.*, 2013; Schmitz, Zimmer, and Bornberg-Bauer 2016). Increased transcription factor complement allows for an increased diversity and complexity of tissue type and modelling, producing more sophisticated morphology, while duplications allow an escape from the pleiotropy associated with involvement in pre-existing GRNs (Ancliff and Park 2014).

The molecular phylogenies presented in Chapter 3 indicate that an evolutionarily significant radiation in transcription factor diversity might have started and be ongoing in the Lophotrochozoa (specifically, annelids and molluscs). These duplications and divergences should not be seen as analogous to those in the last common bilaterian ancestors; they are occurring in a genomic context of pre-existing regulatory sophistication and unlike canonical homeobox gene complements, vary substantially between different taxa. However, they do offer a stage on which to study the potential integration of new and rapidly-evolving homeobox genes into existing developmental processes (*e.g.* Morino, Hashimoto, and Wada 2017) and the potential creation of new GRNs and novel morphology, as a part of the evolutionary history of two of the most speciose and morphologically diverse phyla in the Metazoa (Giribet 2008). That these sequences are only surfacing now is a testament to the neglect of the Spiralia by evolutionary and developmental biology and the promising insights offered by the recent increased interest and availability of genomic data and techniques for spiralian species.

6.5. General conclusions

Homeobox genes are important, highly conserved master controls of gene regulatory networks that orchestrate complex cellular and organismal processes like development, and as such are potential signals of homology across disparate taxa. Regeneration, a post-ontogenic developmental process, has a convoluted and difficult to understand distribution across the Metazoa. Homeobox genes offer a window into the cryptic relationships between regeneration and ontogenesis and between modes of regeneration between species.

In this thesis, I identified the homeobox gene content of the primarily morphallactic regeneration of the operculum, an evolutionarily novel head appendage of serpulid annelids, in *S. lamarcki* (Chapter 3), and the epimorphic regeneration of the post-anal tail, a chordate synapomorphy, in the highly conserved cephalochordate *B. lanceolatum* (Chapter 4). I found an unexpected diversity of difficult-to-classify homeobox genes in the genome and transcriptome of *S. lamarcki*, leading to an investigation into the unusual evolutionary history of a set of homeobox genes in a variety of spiralian clades. In *B. lanceolatum,* identification of homeobox genes is much simpler but unveiled a previously undiscovered duplication of *Pax3/7* of cephalochordates (Chapter 5), a gene involved in several GRNs which underwent important changes during vertebrate evolution for which cephalochordates are a significant outgroup.

A comparison of broad patterns of homeobox transcription activity (section 6.1.1) yielded potentially interesting quantitative differences between the regenerative reactions of the two systems, contrasting a massive increase in homeobox gene deployment by *S. lamarcki* with a relatively mild rise in expression in *B. lanceolatum*. Finally, the significance of the detection of a selection of candidate genes with regard to their known roles in other systems and contexts was discussed (sections 6.1.2-6.1.4).

This thesis demonstrates that there is value in detailed manual inspection of genetic data, time-consuming though it undoubtedly is. Until artificial intelligences are capable of discerning when they have encountered something biologically meaningfully unexpected and investigating that finding, there will be valuable things to discover by manually performing and carefully examining BLAST searches, protein alignments, and molecular phylogenies. Improvements to automatic processing of these data could dispense with much of the information wrangling necessary to distil signal from noise and vastly improve the productivity of this approach.

The findings of this study reaffirm the importance of the duplication of homeobox genes in evolutionary developmental biology and offer two significant bases in which to study the evolutionary effects of this duplication. One is a tandem duplication of a well-studied gene that resulted in differential paralogue C-terminal sequence and regulatory change and strong conservation between extant cephalochordates, and offers a platform for the study of the tempo of duplication that has typified gene evolution within phyla. The other is an unusual ‘wild west,’ characterised by rapid, unconstrained, and extreme duplication and divergence atypical of previously observed homeobox gene evolution of the last ~500 million years and has the potential to reveal new insights into a more radical pace of homeobox gene evolution.

References

Abad, María, Lluc Mosteiro, Cristina Pantoja, Marta Cañamero, Teresa Rayon, Inmaculada Ors, Osvaldo Graña, *et al.*, 2013. 'Reprogramming *In Vivo* Produces Teratomas and IPS Cells with Totipotency Features'. *Nature* 502 (7471): 340–45.

Abitua, Philip Barron, Eileen Wagner, Ignacio A. Navarrete, and Michael Levine. 2012. 'Identification of a Rudimentary Neural Crest in a Non-Vertebrate Chordate'. *Nature* 492 (7427): 104–7.

Abnave, Prasad, Ellen Aboukhatwa, Nobuyoshi Kosaka, James Thompson, Mark A. Hill, and A. Aziz Aboobaker. 2017. 'Epithelial-Mesenchymal Transition Transcription Factors Control Pluripotent Adult Stem Cell Migration *In Vivo* in Planarians'. *Development* 144 (19): 3440–53.

Aboobaker, A. Aziz, and Mark L. Blaxter. 2003. 'Hox Gene Loss during Dynamic Evolution of the Nematode Cluster'. *Current Biology* 13 (1): 37–40.

Acemel, Rafael D., Juan J. Tena, Ibai Irastorza-Azcarate, Ferdinand Marlétaz, Carlos Gómez-Marín, Elisa de la Calle-Mustienes, Stéphanie Bertrand, *et al.*, 2016. 'A Single Three-Dimensional Chromatin Compartment in Amphioxus Indicates a Stepwise Evolution of Vertebrate Hox Bimodal Regulation'. *Nature Genetics* advance online publication (February).

Ackema, Karin B., and Jeroen Charité. 2008. 'Mesenchymal Stem Cells from Different Organs Are Characterized by Distinct Topographic Hox Codes'. *Stem Cells and Development* 17 (5): 979–92.

Adachi, Kenjiro, and Hans R Schöler. 2012. 'Directing Reprogramming to Pluripotency by Transcription Factors'. *Current Opinion in Genetics & Development*, Cell reprogramming, 22 (5): 416–22.

Agata, K., T. Tanaka, C. Kobayashi, K. Kato, and Y. Saitoh. 2003. 'Intercalary Regeneration in Planarians'. *Developmental Dynamics* 226 (2): 308–16.

Agata, Kiyokazu, Yumi Saito, and Elizabeth Nakajima. 2007. 'Unifying Principles of Regeneration I: Epimorphosis *versus* Morphallaxis'. *Development, Growth & Differentiation* 49 (2): 73–78.

Agoston, Zsuzsa, Naixin Li, Anja Haslinger, Andrea Wizenmann, and Dorothea Schulte. 2012. 'Genetic and Physical Interaction of *Meis2*, *Pax3* and *Pax7* during Dorsal Midbrain Development'. *BMC Developmental Biology* 12 (March): 10.

Aguilera, Felipe, Carmel McDougall, and Bernard M. Degnan. 2017. 'Co-Option and *De No*vo Gene Evolution Underlie Molluscan Shell Diversity'. *Molecular Biology and Evolution* 34 (4): 779–92.

Akam, Michael. 1989. 'Hox and HOM: Homologous Gene Clusters in Insects and Vertebrates'. *Cell* 57 (3): 347–49.

Akimenko, M. A., S. L. Johnson, M. Westerfield, and M. Ekker. 1995. 'Differential Induction of Four Msx Homeobox Genes during Fin Development and Regeneration in Zebrafish'. *Development* 121 (2): 347–57.

Akker, Eric van den, Sylvie Forlani, Kallayanee Chawengsaksophak, Wim de Graaff, Felix Beck, Barbara I. Meyer, and Jacqueline Deschamps. 2002. '*Cdx1* and *Cdx2* Have Overlapping Functions in Anteroposterior Patterning and Posterior Axis Elongation'. *Development* 129 (9): 2181–93.

Alappat, Sylvia, Zun Yi Zhang, and Yi Ping Chen. 2003. 'Msx Homeobox Gene Family and Craniofacial Development'. *Cell Research* 13 (6): 429–42.

Albertin, Caroline B., Oleg Simakov, Therese Mitros, Z. Yan Wang, Judit R. Pungor, Eric Edsinger-Gonzales, Sydney Brenner, Clifton W. Ragsdale, and Daniel S. Rokhsar. 2015. 'The Octopus Genome and the Evolution of Cephalopod Neural and Morphological Novelties'. *Nature* 524 (7564).

Aldea, Daniel, Anthony Leon, Stephanie Bertrand, and Hector Escriva. 2015. 'Expression of Fox Genes in the Cephalochordate *Branchiostoma lanceolatum'*. *Frontiers in Ecology and Evolution* 3.

Altincicek, Boran, and Andreas Vilcinskas. 2007. 'Analysis of the Immune-Related Transcriptome of a Lophotrochozoan Model, the Marine Annelid *Platynereis dumerilii'*. *Frontiers in Zoology* 4 (July): 18.

Alvarado, Alejandro Sánchez, and Phillip A. Newmark. 1999. 'Double-Stranded RNA Specifically Disrupts Gene Expression during Planarian Regeneration'. *Proceedings of the National Academy of Sciences* 96 (9): 5049–54.

Alwes, Frederike, Camille Enjolras, and Michalis Averof. 2016. 'Live Imaging Reveals the Progenitors and Cell Dynamics of Limb Regeneration'. *ELife* 5 (October).

Ancliff, Mark, and Jeong-Man Park. 2014. 'Evolution Dynamics of a Model for Gene Duplication under Adaptive Conflict'. *Physical Review E* 89 (6): 062702.

Andrade, Sónia C. S., Marta Novo, Gisele Y. Kawauchi, Katrine Worsaae, Fredrik Pleijel, Gonzalo Giribet, and Greg W. Rouse. 2015. 'Articulating "Archiannelids": Phylogenomics and Annelid Relationships, with Emphasis on Meiofaunal Taxa'. *Molecular Biology and Evolution* 32 (11): 2860–75.

Andrews, Ethan Allen. 1893. *An Undescribed Acraniate: Asymmetron lucayanum.* Baltimore.

Asakura, Atsushi, Michael A. Rudnicki, and Motohiro Komaki. 2001. 'Muscle Satellite Cells Are Multipotential Stem Cells That Exhibit Myogenic, Osteogenic, and Adipogenic Differentiation'. *Differentiation* 68 (4): 245–53.

Aury, Jean-Marc, Olivier Jaillon, Laurent Duret, Benjamin Noel, Claire Jubin, Betina M. Porcel, Béatrice Ségurens, *et al.*, 2006. 'Global Trends of Whole-Genome Duplications Revealed by the Ciliate *Paramecium tetraurelia'*. *Nature* 444 (7116): 171–78.

Babonis, Leslie S., Mark Q. Martindale, and Joseph F. Ryan. 2016. 'Do Novel Genes Drive Morphological Novelty? An Investigation of the Nematosomes in the Sea Anemone *Nematostella vectensis'*. *BMC Evolutionary Biology* 16 (May): 114.

Backfisch, Benjamin, Vitaly V. Kozin, Stephan Kirchmaier, Kristin Tessmar-Raible, and Florian Raible. 2014. 'Tools for Gene-Regulatory Analyses in the Marine Annelid *Platynereis dumerilii'*. *PLOS ONE* 9 (4): e93076.

Backfisch, Benjamin, Vinoth Babu Veedin Rajan, Ruth M. Fischer, Claudia Lohs, Enrique Arboleda, Kristin Tessmar-Raible, and Florian Raible. 2013. 'Stable Transgenesis in the Marine Annelid *Platynereis dumerilii* Sheds New Light on Photoreceptor Evolution'. *Proceedings of the National Academy of Sciences* 110 (1): 193–98.

Baer, Karl Ernst von. 1828. *Über Entwickelungsgeschichte der Thiere. Beobachtung und Reflexion.* Königsberg, Bei den Gebrüdern Bornträger.

Baguña, Jaume. 2012. 'The Planarian Neoblast: The Rambling History of Its Origin and Some Current Black Boxes'. *The International Journal of Developmental Biology* 56 (1-2–3): 19–37.

Bakalenko, Nadezhda I, Elena L. Novikova, Alexander Y Nesterenko, and Milana A Kulakova. 2013. 'Hox Gene Expression during Postlarval Development of the Polychaete *Alitta virens*'. *EvoDevo* 4 (May): 13.

Baker, Clare V. H, Michael R Stark, and Marianne Bronner-Fraser. 2002. 'Pax3-Expressing Trigeminal Placode Cells Can Localize to Trunk Neural Crest Sites but Are Committed to a Cutaneous Sensory Neuron Fate'. *Developmental Biology* 249 (2): 219–36.

Balavoine, Guillaume, Renaud de Rosa, and André Adoutte. 2002. 'Hox Clusters and Bilaterian Phylogeny'. *Molecular Phylogenetics and Evolution* 24 (3): 366–73.

Bando, Tetsuya, Yoshiyasu Ishimaru, Takuro Kida, Yoshimasa Hamada, Yuji Matsuoka, Taro Nakamura, Hideyo Ohuchi, Sumihare Noji, and Taro Mito. 2013. 'Analysis of RNA-Seq Data Reveals Involvement of JAK/STAT Signalling during Leg Regeneration in the Cricket *Gryllus bimaculatus'*. *Development*, January, dev.084590.

Bannister, Stephanie, Olga Antonova, Alessandra Polo, Claudia Lohs, Natalia Hallay, Agne Valinciute, Florian Raible, and Kristin Tessmar-Raible. 2014. 'TALENs Mediate Efficient and Heritable Mutation of Endogenous Genes in the Marine Annelid *Platynereis dumerilii*'. *Genetics* 197 (1): 77–89.

Baratte, S., A. Andouche, and L. Bonnaud. 2007. 'Engrailed in Cephalopods: A Key Gene Related to the Emergence of Morphological Novelties'. *Development Genes and Evolution* 217 (5): 353–62.

Barone, Lindsay, Jason Williams, and David Micklos. 2017. 'Unmet Needs for Analyzing Biological Big Data: A Survey of 704 NSF Principal Investigators'. *PLOS Computational Biology* 13 (10): e1005755.

Barton-Owen, Thomas B., David E. K. Ferrier, and Ildikó M. L. Somorjai. 2018. '*Pax3/7* Duplicated and Diverged Independently in Amphioxus, the Basal Chordate Lineage'. *Scientific Reports* 8 (1): 9414.

Barton-Owen, Thomas B., Réka Szabó, Ildikó M. L. Somorjai, and David E. K. Ferrier 2018. 'A Revised Spiralian Homeobox Gene Classification Incorporating New Polychaete Transcriptomes Reveals a Diverse TALE Class and a Divergent Hox Gene'. *Genome Biology and Evolution* 10 (9): 2151–67.

Barucca, Marco, Adriana Canapa, and Maria A. Biscotti. 2016. 'An Overview of Hox Genes in Lophotrochozoa: Evolution and Functionality'. *Journal of Developmental Biology* 4 (1): 12.

Basch, Martín L., Marianne Bronner-Fraser, and Martín I. García-Castro. 2006. 'Specification of the Neural Crest Occurs during Gastrulation and Requires *Pax7*'. *Nature* 441 (7090): 218–22.

Beauchemin, Michel, Katia Del Rio-Tsonis, Panagiotis A Tsonis, Monique Tremblay, and Pierre Savard. 1998. 'Graded Expression of *Emx-2* in the Adult Newt Limb and Its Corresponding Regeneration Blastema'. *Journal of Molecular Biology* 279 (3): 501–11.

Beauchemin, Michel, and Pierre Savard. 1992. 'Two Distal-Less Related Homeobox-Containing Genes Expressed in Regeneration Blastemas of the Newt'. *Developmental Biology* 154 (1): 55–65.

Beck, Caroline W., Bea Christen, Donna Barker, and Jonathan M. W. Slack. 2006. 'Temporal Requirement for Bone Morphogenetic Proteins in Regeneration of the Tail and Limb of *Xenopus* Tadpoles'. *Mechanisms of Development* 123 (9): 674–88.

Beck, Caroline W, Bea Christen, and Jonathan M. W. Slack. 2003. 'Molecular Pathways Needed for Regeneration of Spinal Cord and Muscle in a Vertebrate'. *Developmental Cell* 5 (3): 429–39.

Beck, Felix, and Emma J. Stringer. 2010. 'The Role of Cdx Genes in the Gut and in Axial Development'. *Biochemical Society Transactions* 38 (2): 353–57.

Beeman, Richard W. 1987. 'A Homoeotic Gene Cluster in the Red Flour Beetle'. *Nature* 327 (6119): 247–49.

Bell, Charles C., Paulo P. Amaral, Anton Kalsbeek, Graham W. Magor, Kevin R. Gillinder, Pierre Tangermann, Lorena di Lisio, *et al.*, 2016. 'The *Evx1/Evx1as* Gene Locus Regulates Anterior-Posterior Patterning during Gastrulation'. *Scientific Reports* 6 (May): 26657.

Bely, Alexandra E. 2006. 'Distribution of Segment Regeneration Ability in the Annelida'. *Integrative and Comparative Biology* 46 (4): 508–18.

Bely, Alexandra E. 2010. 'Evolutionary Loss of Animal Regeneration: Pattern and Process'. *Integrative and Comparative Biology* 50 (4): 515–27.

Bely, Alexandra E. 2014. 'Early Events in Annelid Regeneration: A Cellular Perspective'. *Integrative and Comparative Biology* 54 (4): 688–99.

Bely, Alexandra E., and Kevin G. Nyberg. 2010. 'Evolution of Animal Regeneration: Re-Emergence of a Field'. *Trends in Ecology & Evolution* 25 (3): 161–70.

Bely, Alexandra E., Eduardo E. Zattara, and James M. Sikes. 2014. 'Regeneration in Spiralians: Evolutionary Patterns and Developmental Processes'. *International Journal of Developmental Biology* 58 (6-7–8): 623–34.

Ben Khadra, Yousra, Khaled Said, Michael Thorndyke, and Pedro Martinez. 2014. ‘Homeobox Genes Expressed during Echinoderm Arm Regeneration’. *Biochemical Genetics* 52 (3–4): 166–80.

Bendall, A.J., and C. Abate-Shen. 2000. ‘Roles for Msx and Dlx Homeoproteins in Vertebrate Development’. *Gene* 247 (1–2): 17–31.

Berger, Michael F., Gwenael Badis, Andrew R. Gehrke, Shaheynoor Talukder, Anthony A. Philippakis, Lourdes Peña-Castillo, Trevis M. Alleyne, *et al.*, 2008. ‘Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences’. *Cell* 133 (7): 1266–76.

Bert, M. P. 1867. ‘On the Anatomy and Physiology of Amphioxus’. *Annals and Magazine of Natural History* 20 (118): 302–4.

Bertolino, Eric, Bernard Reimund, Dunja Wildt-Perinic, and Roger G. Clerc. 1995. ‘A Novel Homeobox Protein Which Recognizes a TGT Core and Functionally Interferes with a Retinoid-Responsive Motif’. *Journal of Biological Chemistry* 270 (52): 31178–88.

Bertrand, Stephanie, and Hector Escriva. 2011. ‘Evolutionary Crossroads in Developmental Biology: Amphioxus’. *Development* 138 (22): 4819–30.

Bharathan, Geeta, Bart-Jan Janssen, Elizabeth A. Kellogg, and Neelima Sinha. 1997. ‘Did Homeodomain Proteins Duplicate before the Origin of Angiosperms, Fungi, and Metazoa?’ *Proceedings of the National Academy of Sciences* 94 (25): 13749–53.

Bhatt, Shachi, Raul Diaz, and Paul A. Trainor. 2013. ‘Signals and Switches in Mammalian Neural Crest Cell Differentiation’. *Cold Spring Harbor Perspectives in Biology* 5 (2): a008326.

Biberhofer, R. 1906. ‘Über Regeneration Bei *Amphioxus lanceolatus’*. *Arch EntwMech Org*, no. 22: 15–17.

Bienz, Mariann. 2005. ‘β-Catenin: A Pivot between Cell Adhesion and Wnt Signalling’. *Current Biology* 15 (2): R64–67.

Bininda-Emonds, Olaf R. P., Jonathan E Jeffery, and Michael K. Richardson. 2003. ‘Inverting the Hourglass: Quantitative Evidence against the Phylotypic Stage in Vertebrate Development’. *Proceedings of the Royal Society of London B: Biological Sciences* 270 (1513): 341–46.

Birchler, James A., and Reiner A. Veitia. 2012. ‘Gene Balance Hypothesis: Connecting Issues of Dosage Sensitivity across Biological Disciplines’. *Proceedings of the National Academy of Sciences* 109 (37): 14746–53.

Blomme, Tine, Klaas Vandepoele, Stefanie De Bodt, Cedric Simillion, Steven Maere, and Yves Van de Peer. 2006. ‘The Gain and Loss of Genes during 600 Million Years of Vertebrate Evolution’. *Genome Biology* 7: R43.

Blum, Martin, Edward M. De Robertis, John B. Wallingford, and Christof Niehrs. 2015. ‘Morpholinos: Antisense and Sensibility’. *Developmental Cell* 35 (2): 145–49.

Bobola, Nicoletta, and Samir Merabet. 2017. ‘Homeodomain Proteins in Action: Similar DNA Binding Preferences, Highly Variable Connectivity’. *Current Opinion in Genetics & Development*, Genome architecture and expression, 43 (April): 1–8.

Bode, Hans R. 2003. ‘Head Regeneration in *Hydra*’. *Developmental Dynamics* 226 (2): 225–36.

Bok, Michael J., Megan L. Porter, Harry A. ten Hove, Richard Smith, and Dan-Eric Nilsson. 2017. ‘Radiolar Eyes of Serpulid Worms (Annelida, Serpulidae): Structures, Function, and Phototransduction’. *The Biological Bulletin* 233 (1): 39–57.

Bone, Q. 1992. ‘Protochordates’. *J Mar Biol Assoc UK* 72: 952–53.

Booker, Tom R., Benjamin C. Jackson, and Peter D. Keightley. 2017. ‘Detecting Positive Selection in the Genome’. *BMC Biology* 15 (October): 98.

Booth, H. Anne F., and Peter W. H. Holland. 2004. ‘Eleven Daughters of NANOG’. *Genomics* 84 (2): 229–38.

Booth, H. Anne F., and Peter W. H. Holland. 2007. ‘Annotation, Nomenclature and Evolution of Four Novel Homeobox Genes Expressed in the Human Germ Line’. *Gene* 387 (1): 7–14.

Bopp, Daniel, Maya Burri, Stefan Baumgartner, Gabriella Frigerio, and Markus Noll. 1986. ‘Conservation of a Large Protein Domain in the Segmentation Gene *paired* and in Functionally Related Genes of Drosophila’. *Cell* 47 (6): 1033–40.

Borson, N. D., W. L. Salo, and L. R. Drewes. 1992. ‘A Lock-Docking Oligo(DT) Primer for 5’ and 3’ RACE PCR.’ *Genome Research* 2 (2): 144–48.

Boyer, Laurie A., Tong Ihn Lee, Megan F. Cole, Sarah E. Johnstone, Stuart S. Levine, Jacob P. Zucker, Matthew G. Guenther, *et al.*, 2005. ‘Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells’. *Cell* 122 (6): 947–56.

Brauchle, Michael, Adem Bilican, Claudia Eyer, Xavier Bailly, Pedro Martínez, Peter Ladurner, Rémy Bruggmann, and Simon G Sprecher. 2018. ‘Xenacoelomorpha Survey Reveals That All 11 Animal Homeobox Gene Classes Were Present in the First Bilaterians’. *Genome Biology and Evolution*, August.

Bray, Nick, Inna Dubchak, and Lior Pachter. 2003. ‘AVID: A Global Alignment Program’. *Genome Research* 13 (1): 97–102.

Breitling, R., and J. K. Gerber. 2000. ‘Origin of the paired Domain’. *Development Genes and Evolution* 210 (12): 644–50.

Brockes, Jeremy P., and Anoop Kumar. 2008. ‘Comparative Aspects of Animal Regeneration’. *Annual Review of Cell and Developmental Biology* 24 (1): 525–49.

Brooke, Nina M., Jordi Garcia-Fernàndez, and Peter W. H. Holland. 1998. ‘The ParaHox Gene Cluster Is an Evolutionary Sister of the Hox Gene Cluster’. *Nature* 392 (6679): 920–22.

Brown, Federico D., Elena L. Keeling, Anna D. Le, and Billie J. Swalla. 2009. ‘Whole Body Regeneration in a Colonial Ascidian, *Botrylloides violaceus’*. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 312B (8): 885–900.

Brunelli, Silvia, and Giulio Cossu. 2005. ‘A Role for Msx2 and Necdin in Smooth Muscle Differentiation of Mesoangioblasts and Other Mesoderm Progenitor Cells’. *Trends in Cardiovascular Medicine* 15 (3): 96–100.

Brunet, Frédéric G., Hugues Roest Crollius, Mathilde Paris, Jean-Marc Aury, Patricia Gibert, Olivier Jaillon, Vincent Laudet, and Marc Robinson-Rechavi. 2006. ‘Gene Loss and Evolutionary Rates Following Whole-Genome Duplication in Teleost Fishes’. *Molecular Biology and Evolution* 23 (9): 1808–16.

Bryant, Susan V., Tetsuya Endo, and David M. Gardiner. 2002. ‘Vertebrate Limb Regeneration and the Origin of Limb Stem Cells.’ *International Journal of Developmental Biology* 46 (7): 887–96.

Bubel, A. 1983a. ‘An Ultrastructural Investigation of Muscle Attachment in the Opercular Filament of a Polychaete Annelid’. *Tissue and Cell* 15 (4): 555–72.

———. 1983b. ‘An Ultrastructural Study of the Opercular Filament Blood Vessel of *Pomatoceros lamarckii* Quatrefages (Polychaeta: Serpulidae)’. *Protoplasma* 115 (2–3): 129–52.

Bubel, A., R.M. Stephens, R.H. Fenn, and P. Fieth. 1983. ‘An Electron Microscope, x-Ray Diffraction and Amino Acid Analysis Study of the Opercular Filament Cuticle, Calcareous Opercular Plate and Habitation Tube of *Pomatoceros lamarckii* Quatrefages (Polychaeta: Serpulidae)’. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry* 74 (4): 837–50.

Bubel, A., and C. H. Thorp. 1985. ‘Tissue Abscission and Wound Healing in the Operculum of *Pomatoceros lamarckii* Quatrefages (Polychaeta: Serpulidae)’. *Journal of Zoology* 1 (1): 95–143.

Bubel, A., C. H. Thorp, Ruth H. Fenn, and D. Livingstone. 1985. ‘Opercular Regeneration in *Pomatoceros lamarckii* Quatrefages (Polychaeta: Serpulidae). Differentiation of the Operculum and Deposition of the Calcareous Opercular Plate’. *Journal of Zoology* 1 (1): 49–94.

Buckingham, Margaret. 2007. ‘Skeletal Muscle Progenitor Cells and the Role of Pax Genes’. *Comptes Rendus Biologies*, Thérapie cellulaire régénérative / Regenerative cell therapy, 330 (6–7): 530–33.

Buckingham, Margaret. 2017. ‘Gene Regulatory Networks and Cell Lineages That Underlie the Formation of Skeletal Muscle’. *Proceedings of the National Academy of Sciences* 114 (23): 5830–37.

Buckingham, Margaret, and Frédéric Relaix. 2007. ‘The Role of Pax Genes in the Development of Tissues and Organs: *Pax3* and *Pax7* Regulate Muscle Progenitor Cell Functions’. *Annual Review of Cell and Developmental Biology* 23 (1): 645–73.

Buckingham, Margaret, and Frédéric Relaix. 2015. ‘PAX3 and PAX7 as Upstream Regulators of Myogenesis’. *Seminars in Cell & Developmental Biology*, Paramutation & Pax Transcription Factors, 44 (August): 115–25.

Buckingham, Margaret, and Stéphane D Vincent. 2009. ‘Distinct and Dynamic Myogenic Populations in the Vertebrate Embryo’. *Current Opinion in Genetics & Development*, Differentation and gene regulation, 19 (5): 444–53.

Burglin, T. R. 1997. ‘Analysis of TALE Superclass Homeobox Genes (MEIS, PBC, KNOX, Iroquois, TGIF) Reveals a Novel Domain Conserved between Plants and Animals’. *Nucleic Acids Research* 25 (21): 4173–80.

Bürglin, Thomas R., and Markus Affolter. 2016. ‘Homeodomain Proteins: An Update’. *Chromosoma* 125: 497–521.

Butts, Thomas, Peter W. H. Holland, and David E. K. Ferrier. 2010. ‘Ancient Homeobox Gene Loss and the Evolution of Chordate Brain and Pharynx Development: Deductions from Amphioxus Gene Expression’. *Proceedings of the Royal Society of London B: Biological Sciences* 277 (1699): 3381–89.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. ‘BLAST+: Architecture and Applications’. *BMC Bioinformatics* 10 (December): 421.

Candiani, Simona, Luca Moronti, Davide De Pietri Tonelli, Greta Garbarino, and Mario Pestarino. 2011. ‘A Study of Neural-Related MicroRNAs in the Developing Amphioxus’. *EvoDevo* 2 (July): 15.

Cannon, Johanna Taylor, Bruno Cossermelli Vellutini, Julian Smith, Fredrik Ronquist, Ulf Jondelius, and Andreas Hejnol. 2016. ‘Xenacoelomorpha Is the Sister Group to Nephrozoa’. *Nature* 530 (7588): 89–93.

Capellini, Terence D., Vincenzo Zappavigna, and Licia Selleri. 2011. ‘Pbx Homeodomain Proteins: TALEnted Regulators of Limb Patterning and Outgrowth’. *Developmental Dynamics* 240 (5): 1063–86.

Carlson, Marc R. J., Susan V. Bryant, and David M. Gardiner. 1998. ‘Expression of *Msx-2* during Development, Regeneration, and Wound Healing in Axolotl Limbs’. *Journal of Experimental Zoology* 282 (6): 715–23.

Carlson, Marc R. J., Y. Komine, S. V. Bryant, and D. M. Gardiner. 2001. ‘Expression of *Hoxb13* and *Hoxc10* in Developing and Regenerating Axolotl Limbs and Tails’. *Developmental Biology* 229 (2): 396–406.

Carroll, Sean B. 1995. ‘Homeotic Genes and the Evolution of Arthropods and Chordates’. *Nature* 376 (6540): 479–85.

Casola, Claudio, and Esther Betrán. 2017. ‘The Genomic Impact of Gene Retrocopies: What Have We Learned from Comparative Genomics, Population Genomics, and Transcriptomic Analyses?’ *Genome Biology and Evolution* 9 (6): 1351–73.

Chain, Frédéric JJ, Jonathan Dushoff, and Ben J. Evans. 2011. ‘The Odds of Duplicate Gene Persistence after Polyploidization’. *BMC Genomics* 12: 599.

Chalepakis, G., F. S. Jones, G. M. Edelman, and P. Gruss. 1994. ‘Pax-3 Contains Domains for Transcription Activation and Transcription Inhibition’. *Proceedings of the National Academy of Sciences* 91 (26): 12745–49.

Chambers, Ian, Douglas Colby, Morag Robertson, Jennifer Nichols, Sonia Lee, Susan Tweedie, and Austin Smith. 2003. ‘Functional Expression Cloning of Nanog, a Pluripotency Sustaining Factor in Embryonic Stem Cells’. *Cell* 113 (5): 643–55.

Chang, Howard Y., Jen-Tsan Chi, Sandrine Dudoit, Chanda Bondre, Matt van de Rijn, David Botstein, and Patrick O. Brown. 2002. ‘Diversity, Topographic Differentiation, and Positional Memory in Human Fibroblasts’. *Proceedings of the National Academy of Sciences* 99 (20): 12877–82.

Charest-Marcotte, Alexis, Catherine R. Dufour, Brian J. Wilson, Annie M. Tremblay, Lillian J. Eichner, Daniel H. Arlow, Vamsi K. Mootha, and Vincent Giguère. 2010. ‘The Homeobox Protein Prox1 Is a Negative Modulator of ERRα/PGC-1α Bioenergetic Functions’. *Genes & Development* 24 (6): 537–42.

Charite, J., W. de Graaff, D. Consten, M. J. Reijnen, J. Korving, and J. Deschamps. 1998. ‘Transducing Positional Information to the Hox Genes: Critical Interaction of cdx Gene Products with Position-Sensitive Regulatory Elements’. *Development* 125 (22): 4349–58.

Charoensawan, Varodom, Derek Wilson, and Sarah A. Teichmann. 2010. ‘Genomic Repertoires of DNA-Binding Transcription Factors across the Tree of Life’. *Nucleic Acids Research* 38 (21): 7364–77.

Charpignon, Véronique. 2007. ‘Homeobox-containing Genes in the Nemertean *Lineus*: Key Players in the Antero-posterior Body Patterning and in the Specification of the Visual Structures’. *Universität Basel & Universität Reims,* 244.

Chawengsaksophak, Kallayanee, Wim de Graaff, Janet Rossant, Jacqueline Deschamps, and Felix Beck. 2004. ‘*Cdx2* Is Essential for Axial Elongation in Mouse Development’. *Proceedings of the National Academy of Sciences* 101 (20): 7641–45.

Cheatle Jarvela, Alys M, and Veronica F Hinman. 2015. ‘Evolution of Transcription Factor Function as a Mechanism for Changing Metazoan Developmental Gene Regulatory Networks’. *EvoDevo* 6 (1).

Chen, Fabian, Hyun Kook, Rita Milewski, Aaron D. Gitler, Min Min Lu, Jun Li, Ronniel Nazarian, *et al.*, 2002. ‘*Hop* Is an Unusual Homeobox Gene That Modulates Cardiac Development’. *Cell* 110 (6): 713–23.

Chen, Jian-Min, David N. Cooper, Nadia Chuzhanova, Claude Férec, and George P. Patrinos. 2007. 'Gene Conversion: Mechanisms, Evolution and Human Disease'. *Nature Reviews Genetics* 8 (10): 762–75.

Chen, K., and N. Rajewsky. 2006. 'Deep Conservation of MicroRNA-Target Relationships and 3'UTR Motifs in Vertebrates, Flies, and Nematodes'. *Cold Spring Harbor Symposia on Quantitative Biology* 71 (January): 149–56.

Chen, Lu, QiuJing Zhang, Wei Wang, and YiQuan Wang. 2010. 'Spatiotemporal Expression of Pax Genes in Amphioxus: Insights into Pax-Related Organogenesis and Evolution'. *Science China Life Sciences* 53 (8): 1031–40.

Chen, Ming, Ming Zou, Beide Fu, Xin Li, Maria D. Vibranovski, Xiaoni Gan, Dengqiang Wang, Wen Wang, Manyuan Long, and Shunping He. 2011. 'Evolutionary Patterns of RNA-Based Duplication in Non-Mammalian Chordates'. *PLoS ONE* 6 (7).

Chen, Sidi, Benjamin H. Krinsky, and Manyuan Long. 2013. 'New Genes as Drivers of Phenotypic Evolution'. *Nature Reviews Genetics* 14 (9): 645–60.

Chen, Sidi, Yong E. Zhang, and Manyuan Long. 2010. 'New Genes in *Drosophila* Quickly Become Essential'. *Science* 330 (6011): 1682–85.

Chen, Wei-Hua, Kalliopi Trachana, Martin J. Lercher, and Peer Bork. 2012. 'Younger Genes Are Less Likely to Be Essential than Older Genes, and Duplicates Are Less Likely to Be Essential than Singletons of the Same Age'. *Molecular Biology and Evolution* 29 (7): 1703–6.

Chen, Ying, Gufa Lin, and Jonathan M. W. Slack. 2006. 'Control of Muscle Regeneration in the *Xenopus* Tadpole Tail by Pax7'. *Development (Cambridge, England)* 133 (12): 2303–13.

Chipman, Ariel D. 2010. 'Parallel Evolution of Segmentation by Co-Option of Ancestral Gene Regulatory Networks'. *BioEssays* 32 (1): 60–70.

Choi, Wen-Yee, Antonio J. Giraldez, and Alexander F. Schier. 2007. 'Target Protectors Reveal Dampening and Balancing of Nodal Agonist and Antagonist by MiR-430'. *Science* 318 (5848): 271–74.

Chou, Hsien-Chao, Natalia Acevedo-Luna, Julie A. Kuhlman, and Stephan Q. Schneider. 2018. 'PdumBase: A Transcriptome Database and Research Tool for *Platynereis dumerilii* and Early Development of Other Metazoans'. *BMC Genomics* 19 (1): 618.

Christen, Bea, Caroline W. Beck, Aurora Lombardo, and Jonathan M. W. Slack. 2003. 'Regeneration-Specific Expression Pattern of Three Posterior Hox Genes'. *Developmental Dynamics* 226 (2): 349–55.

Christen, Bea, Vanesa Robles, Marina Raya, Ida Paramonov, and Juan Carlos Izpisúa Belmonte. 2010. 'Regeneration and Reprogramming Compared'. *BMC Biology* 8 (1): 5.

Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, *et al.*, 2009. ‘Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics’. *Bioinformatics* 25 (11): 1422–23.

Comte, Aurélie, Julien Roux, and Marc Robinson-Rechavi. 2010. ‘Molecular Signaling in Zebrafish Development and the Vertebrate Phylotypic Period’. *Evolution & Development* 12 (2): 144–56.

Conant, Gavin C, James A Birchler, and J Chris Pires. 2014. ‘Dosage, Duplication, and Diploidization: Clarifying the Interplay of Multiple Models for Duplicate Gene Evolution over Time’. *Current Opinion in Plant Biology*, SI: Physiology and metabolism, 19 (June): 91–98.

Conesa, Ana, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, *et al.*, 2016. ‘A Survey of Best Practices for RNA-Seq Data Analysis’. *Genome Biology* 17.

Cook, Charles E., Eva Jiménez, Michael Akam, and Emili Saló. 2004. ‘The Hox Gene Complement of Acoel Flatworms, a Basal Bilaterian Clade’. *Evolution & Development* 6 (3): 154–63.

Copf, Tijana, Nicolas Rabet, Susan E. Celniker, and Michalis Averof. 2003. ‘Posterior Patterning Genes and the Identification of a Unique Body Region in the Brine Shrimp *Artemia franciscana*’. *Development* 130 (24): 5915–27.

Copf, Tijana, Reinhard Schröder, and Michalis Averof. 2004. ‘Ancestral Role of *caudal* Genes in Axis Elongation and Segmentation’. *Proceedings of the National Academy of Sciences* 101 (51): 17711–15.

Copley, Richard R. 2005. ‘The EH1 Motif in Metazoan Transcription Factors’. *BMC Genomics* 6 (November): 169.

Corry, Gareth N., Nikhil Raghuram, Kristal K. Missiaen, Ninghe Hu, Michael J. Hendzel, and D. Alan Underhill. 2010. ‘The PAX3 Paired Domain and Homeodomain Function as a Single Binding Module *In Vivo* to Regulate Subnuclear Localization and Mobility by a Mechanism That Requires Base-Specific Recognition’. *Journal of Molecular Biology* 402 (1): 178–93.

Corry, Gareth N., and D.A. Underhill. 2005. ‘Pax3 Target Gene Recognition Occurs through Distinct Modes That Are Differentially Affected by Disease-Associated Mutations’. *Pigment Cell Research* 18 (6): 427–38.

Costello, Donald P., and Catherine Henley. 1976. ‘Spiralian Development: A Perspective’. *American Zoologist* 16 (3,): 277-.

Crist, Colin G., Didier Montarras, Giorgia Pallafacchina, Didier Rocancourt, Ana Cumano, Simon J. Conway, and Margaret Buckingham. 2009. ‘Muscle Stem Cell Behavior Is Modified by MicroRNA-27 Regulation of *Pax3* Expression’. *Proceedings of the National Academy of Sciences* 106 (32): 13383–87.

Crocker, Justin, Namiko Abe, Lucrezia Rinaldi, Alistair P. McGregor, Nicolás Frankel, Shu Wang, Ahmad Alsawadi, *et al.*, 2015. 'Low Affinity Binding Site Clusters Confer Hox Specificity and Regulatory Robustness'. *Cell* 160 (1): 191–203.

Currie, Ko W., David D. R. Brown, Shujun Zhu, ChangJiang Xu, Veronique Voisin, Gary D. Bader, and Bret J. Pearson. 2016. 'HOX Gene Complement and Expression in the Planarian *Schmidtea mediterranea'*. *EvoDevo* 7 (1).

Dailey, Simon C. 2017. 'Evolutionary Developmental and Genomic Insights from a Tail Regeneration Transcriptome of the Cephalochordate *Branchiostoma lanceolatum*'. Thesis, University of St Andrews.

Dailey, Simon C., Roser Febrero Planas, Ariadna Rossell Espier, Jordi Garcia-Fernàndez, and Ildikó M. L. Somorjai. 2016. 'Asymmetric Distribution of Pl10 and Bruno2, New Members of a Conserved Core of Early Germline Determinants in Cephalochordates'. *Frontiers in Ecology and Evolution* 3.

D'Aniello, Salvatore, Manuel Irimia, Ignacio Maeso, Juan Pascual-Anaya, Senda Jiménez-Delgado, Stephanie Bertrand, and Jordi Garcia-Fernàndez. 2008. 'Gene Expansion and Retention Leads to a Diverse Tyrosine Kinase Superfamily in Amphioxus'. *Molecular Biology and Evolution* 25 (9): 1841–54.

Das, Sunetra, Sharmishtha Shyamal, and David S. Durica. 2016. 'Analysis of Annotation and Differential Expression Methods Used in RNA-Seq Studies in Crustacean Systems'. *Integrative and Comparative Biology* 56 (6): 1067–79.

Davidson, Eric H., and Douglas H. Erwin. 2006. 'Gene Regulatory Networks and the Evolution of Animal Body Plans'. *Science* 311 (5762): 796–800.

Davidson, Eric H., and Michael S. Levine. 2008. 'Properties of Developmental Gene Regulatory Networks'. *Proceedings of the National Academy of Sciences* 105 (51): 20063–66.

Davis, Gregory K., Joseph A. D'Alessio, and Nipam H. Patel. 2005. 'Pax3/7 Genes Reveal Conservation and Divergence in the Arthropod Segmentation Hierarchy'. *Developmental Biology* 285 (1): 169–84.

Davis, Jerel C., and Dmitri A. Petrov. 2004. 'Preferential Duplication of Conserved Proteins in Eukaryotic Genomes'. *PLOS Biology* 2 (3): e55.

Deckelbaum, Ron A., Amit Majithia, Thomas Booker, Janet E. Henderson, and Cynthia A. Loomis. 2006. 'The Homeoprotein Engrailed 1 Has Pleiotropic Functions in Calvarial Intramembranous Bone Formation and Remodeling'. *Development* 133 (1): 63–74.

Degnan, Bernard M, Michel Vervoort, Claire Larroux, and Gemma S Richards. 2009. 'Early Evolution of Metazoan Transcription Factors'. *Current Opinion in Genetics & Development*, Genomes and evolution, 19 (6): 591–99.

Dehal, Paramvir, and Jeffrey L. Boore. 2005. 'Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate'. *PLOS Biology* 3 (10): e314.

Dehal, Paramvir, Yutaka Satou, Robert K. Campbell, Jarrod Chapman, Bernard Degnan, Anthony De Tomaso, Brad Davidson, *et al.*, 2002. ‘The Draft Genome of *Ciona intestinalis*: Insights into Chordate and Vertebrate Origins’. *Science* 298 (5601): 2157–67.

Delgado, Irene, and Miguel Torres. 2017. ‘Coordination of Limb Development by Crosstalk among Axial Patterning Pathways’. *Developmental Biology*, Forming and shaping the field of limb development: A tribute to Dr. John Saunders, 429 (2): 382–86.

Delsuc, Frédéric, Henner Brinkmann, Daniel Chourrout, and Hervé Philippe. 2006. ‘Tunicates and Not Cephalochordates Are the Closest Living Relatives of Vertebrates’. *Nature* 439 (7079): 965–68.

Delsuc, Frédéric, Hervé Philippe, Georgia Tsagkogeorga, Paul Simion, Marie-Ka Tilak, Xavier Turon, Susanna López-Legentil, Jacques Piette, Patrick Lemaire, and Emmanuel J. P. Douzery. 2018. ‘A Phylogenomic Framework and Timescale for Comparative Studies of Tunicates’. *BMC Biology* 16 (April): 39.

Denoeud, France, Simon Henriet, Sutada Mungpakdee, Jean-Marc Aury, Corinne Da Silva, Henner Brinkmann, Jana Mikhaleva, *et al.*, 2010. ‘Plasticity of Animal Genome Architecture Unmasked by Rapid Evolution of a Pelagic Tunicate’. *Science* 330 (6009): 1381–85.

Derelle, Romain, Philippe Lopez, Hervé Le Guyader, and Michaël Manuel. 2007. ‘Homeodomain Proteins Belong to the Ancestral Molecular Toolkit of Eukaryotes’. *Evolution & Development* 9 (3): 212–19.

Deschamps, Jacqueline, and Johan van Nes. 2005. ‘Developmental Regulation of the Hox Genes during Axial Morphogenesis in the Mouse’. *Development* 132 (13): 2931–42.

Dey, Bijan K., Jeffrey Gagan, and Anindya Dutta. 2011. ‘MiR-206 and -486 Induce Myoblast Differentiation by Downregulating *Pax7*’. *Molecular and Cellular Biology* 31 (1): 203–14.

Domazet-Lošo, Tomislav, and Diethard Tautz. 2010. ‘A Phylogenetically Based Transcriptome Age Index Mirrors Ontogenetic Divergence Patterns’. *Nature* 468 (7325): 815–18.

Donoghue, Maria J., Robin Morris-Valero, Yvette R. Johnson, John P. Merlie, and Joshua R. Sanes. 1992. ‘Mammalian Muscle Cells Bear a Cell-Autonomous, Heritable Memory of Their Rostrocaudal Position’. *Cell* 69 (1): 67–77.

Doudna, Jennifer A., and Emmanuelle Charpentier. 2014. ‘The New Frontier of Genome Engineering with CRISPR-Cas9’. *Science* 346 (6213): 1258096.

Draper, Bruce W., Paul A. Morcos, and Charles B. Kimmel. 2001. ‘Inhibition of Zebrafish Fgf8 Pre-MRNA Splicing with Morpholino Oligos: A Quantifiable Method for Gene Knockdown’. *Genesis* 30 (3): 154–56.

Drost, Hajk-Georg, Alexander Gabel, Ivo Grosse, and Marcel Quint. 2015. ‘Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis’. *Molecular Biology and Evolution* 32 (5): 1221–31.

Duboule, Denis. 1994. ‘Temporal Colinearity and the Phylotypic Progression: A Basis for the Stability of a Vertebrate Bauplan and the Evolution of Morphologies through Heterochrony’. *Development*, no. 1994/Supplement (January).

Duboule, Denis. 2007. ‘The Rise and Fall of Hox Gene Clusters’. *Development* 134 (14): 2549–60.

Dunn, Casey W., Andreas Hejnol, David Q. Matus, Kevin Pang, William E. Browne, Stephen A. Smith, Elaine Seaver, *et al.*, 2008. ‘Broad Phylogenomic Sampling Improves Resolution of the Animal Tree of Life’. *Nature* 452 (7188): 745–49.

Dunwell, Thomas L., and Peter W. H. Holland. 2016. ‘Diversity of Human and Mouse Homeobox Gene Expression in Development and Adult Tissues’. *BMC Developmental Biology* 16 (1): 40.

Dupont, S., and M. Thorndyke. 2007. ‘Bridging the Regeneration Gap: Insights from Echinoderm Models’. *Nature Reviews Genetics* 8 (4): 320.

Echeverri, Karen, Jonathan D. W. Clarke, and Elly M. Tanaka. 2001. ‘*In Vivo* Imaging Indicates Muscle Fiber Dedifferentiation Is a Major Contributor to the Regenerating Tail Blastema’. *Developmental Biology* 236 (1): 151–64.

Echeverri, Karen, and Elly M Tanaka. 2002. ‘Mechanisms of Muscle Dedifferentiation during Regeneration’. *Seminars in Cell & Developmental Biology*, Regeneration, 13 (5): 353–60.

Echeverri, Karen, and Elly M. Tanaka. 2005. ‘Proximodistal Patterning during Limb Regeneration’. *Developmental Biology* 279 (2): 391–401.

Edvardsen, Rolf B., Hee-Chan Seo, Marit F. Jensen, Antoine Mialon, Jana Mikhaleva, Marianne Bjordal, Jérome Cartry, *et al.*, 2005. ‘Remodelling of the Homeobox Gene Complement in the Tunicate *Oikopleura dioica*’. *Current Biology* 15 (1): R12–13.

Eickholt, Jesse, and Zheng Wang. 2014. ‘PCP-ML: Protein Characterization Package for Machine Learning’. *BMC Research Notes* 7 (November).

Elati, Mohamed, Rémy Nicolle, Ivan Junier, David Fernández, Rim Fekih, Julio Font, and François Képès. 2013. ‘PreCisIon: PREdiction of CIS-Regulatory Elements Improved by Gene’s PositION’. *Nucleic Acids Research* 41 (3): 1406–15.

Emig, C. C. 1973. ‘Regenerating histogenesis in Phoronida’. *Wilhelm Roux’ Archiv fur Entwicklungsmechanik der Organismen* 173 (3): 235–48.

Endo, Tetsuya, Koji Tamura, and Hiroyuki Ide. 2000. ‘Analysis of Gene Expressions during *Xenopus* Forelimb Regeneration’. *Developmental Biology* 220 (2): 296–306.

Enright, Anton J., Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora S. Marks. 2003. ‘MicroRNA Targets in *Drosophila*’. *Genome Biology* 5 (December): R1.

Ettensohn, Charles A. 2009. ‘Lessons from a Gene Regulatory Network: Echinoderm Skeletogenesis Provides Insights into Evolution, Plasticity and Morphogenesis’. *Development* 136 (1): 11–21.

Ettensohn, Charles A., Michele R. Illies, Paola Oliveri, and Deborah L. De Jong. 2003. 'Alx1, a Member of the Cart1/Alx3/Alx4 Subfamily of Paired-Class Homeodomain Proteins, Is an Essential Component of the Gene Network Controlling Skeletogenic Fate Specification in the Sea Urchin Embryo'. *Development* 130 (13): 2917–2928.

Evans, Tyler G. 2015. 'Considerations for the Use of Transcriptomics in Identifying the "Genes That Matter" for Environmental Adaptation'. *Journal of Experimental Biology* 218 (12): 1925–35.

Fares, Mario A., Orla M. Keane, Christina Toft, Lorenzo Carretero-Paulet, and Gary W. Jones. 2013. 'The Roles of Whole-Genome and Small-Scale Duplications in the Functional Specialization of *Saccharomyces cerevisiae* Genes'. *PLOS Genetics* 9 (1): e1003176.

Farré, Domènec, Romà Roset, Mario Huerta, José E. Adsuara, Llorenç Roselló, M. Mar Albà, and Xavier Messeguer. 2003. 'Identification of Patterns in Biological Sequences at the ALGGEN Server: PROMO and MALGEN'. *Nucleic Acids Research* 31 (13): 3651–53.

Feiner, Nathalie, Rolf Ericsson, Axel Meyer, and Shigehiro Kuraku. 2011. 'Revisiting the Origin of the Vertebrate *Hox14* by Including Its Relict Sarcopterygian Members'. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 316B (7): 515–525.

Felsenstein, Joseph. 1989. 'PHYLIP - Phylogeny Inference Package (Version 3.2)'. *Cladistics* 5: 164–66.

Ferrier, David E. K. 2008. 'When Is a Hox Gene Not a Hox Gene? The Importance of Gene Nomenclature'. *Evolving Pathways: Key Themes in Evolutionary Developmental Biology*, January, 175–93.

Ferrier, David E. K. 2010. 'Evolution of Hox Complexes'. *Advances in Experimental Medicine and Biology* 689: 91–100.

Ferrier, David E. K. 2012. 'Evolutionary Crossroads in Developmental Biology: Annelids'. *Development* 139 (15): 2643–53.

Ferrier, David E. K. 2016. 'Evolution of Homeobox Gene Clusters in Animals: The Giga-Cluster and Primary vs. Secondary Clustering'. *Frontiers in Ecology and Evolution* 4.

Ferrier, David E. K., Carolina Minguillón, Cristina Cebrián, and Jordi Garcia-Fernàndez. 2001. 'Amphioxus Evx Genes: Implications for the Evolution of the Midbrain–Hindbrain Boundary and the Chordate Tailbud'. *Developmental Biology* 237 (2): 270–81.

Ferrier, David E. K., Carolina Minguillón, Peter W. H. Holland, and Jordi Garcia-Fernàndez. 2000. 'The Amphioxus Hox Cluster: Deuterostome Posterior Flexibility and *Hox14'. Evolution & Development* 2 (5): 284–93.

Feuda, Roberto, Martin Dohrmann, Walker Pett, Hervé Philippe, Omar Rota-Stabelli, Nicolas Lartillot, Gert Wörheide, and Davide Pisani. 2017. 'Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals'. *Current Biology* 27 (24): 3864-3870.e4.

Fiedler, Thomas, and Marc Rehmsmeier. 2006. 'JPREdictor: A Versatile Tool for the Prediction of Cis-Regulatory Elements'. *Nucleic Acids Research* 34 (Web Server issue): W546–50.

Finnerty, John R., and Mark Q. Martindale. 1999. 'Ancient Origins of Axial Patterning Genes: Hox Genes and ParaHox Genes in the Cnidaria'. *Evolution & Development* 1 (1): 16–23.

Finotello, Francesca, and Barbara Di Camillo. 2015. 'Measuring Differential Gene Expression with RNA-Seq: Challenges and Strategies for Data Analysis'. *Briefings in Functional Genomics* 14 (2): 130–42.

Flot, J. F., B. Hespeels, X. Li, B. Noel, I. Arkhipova, E. G. Danchin, A. Hejnol, *et al.*, 2013. 'Genomic Evidence for Ameiotic Evolution in the Bdelloid Rotifer *Adineta vaga*.' *Nature* 500 (7463): 453–57.

Force, Allan, Michael Lynch, F. Bryan Pickett, Angel Amores, Yi-lin Yan, and John Postlethwait. 1999. 'Preservation of Duplicate Genes by Complementary, Degenerative Mutations'. *Genetics* 151 (4): 1531–45.

Frank, Uri, Günter Plickert, and Werner A. Müller. 2009. 'Cnidarian Interstitial Cells: The Dawn of Stem Cell Research'. In *Stem Cells in Marine Organisms*, edited by Baruch Rinkevich and Valeria Matranga, 33–59. Dordrecht: Springer Netherlands.

Frasch, Manfred. 2016. 'Chapter Eighteen - Dedifferentiation, Redifferentiation, and Transdifferentiation of Striated Muscles During Regeneration and Development'. In *Current Topics in Developmental Biology*, edited by Paul M. Wassarman, 116:331–55. Essays on Developmental Biology, Part A. Academic Press.

Frazer, Kelly A., Lior Pachter, Alexander Poliakov, Edward M. Rubin, and Inna Dubchak. 2004. 'VISTA: Computational Tools for Comparative Genomics'. *Nucleic Acids Research* 32 (Web Server issue): W273-279.

Fritzsch, Guido, Manja U. Böhme, Mike Thorndyke, Hiroaki Nakano, Olle Israelsson, Thomas Stach, Martin Schlegel, Thomas Hankeln, and Peter F. Stadler. 2008. 'PCR Survey of *Xenoturbella bocki* Hox Genes'. *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution* 310 (3): 278–84.

Fröbius, Andreas C., and Peter Funch. 2017. 'Rotiferan Hox Genes Give New Insights into the Evolution of Metazoan Bodyplans'. *Nature Communications* 8 (1): 9.

Fröbius, Andreas C., David Q. Matus, and Elaine C. Seaver. 2008. 'Genomic Organization and Expression Demonstrate Spatial and Temporal Hox Gene Colinearity in the Lophotrochozoan *Capitella* sp. I'. *PLOS ONE* 3 (12): e4004.

Frohman, M. A., M. K. Dush, and G. R. Martin. 1988. 'Rapid Production of Full-Length CDNAs from Rare Transcripts: Amplification Using a Single Gene-Specific Oligonucleotide Primer'. *Proceedings of the National Academy of Sciences* 85 (23): 8998–9002.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. 'CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data'. *Bioinformatics (Oxford, England)* 28 (23): 3150–52.

Fuentes, Michael, Elia Benito, Stephanie Bertrand, Mathilde Paris, Aurelie Mignardot, Laura Godoy, Senda Jimenez-Delgado, *et al.*, 2007. 'Insights into Spawning Behavior and Development of the European Amphioxus (*Branchiostoma lanceolatum*)'. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 308B (4): 484–93.

Galis, Frietson, Tom J. M. van Dooren, and Johan A. J. Metz. 2002. 'Conservation of the Segmented Germband Stage: Robustness or Pleiotropy?' *Trends in Genetics* 18 (10): 504–9.

Galis, Frietson, and Johan A.J. Metz. 2002. 'Testing the Vulnerability of the Phylotypic Stage: On Modularity and Evolutionary Conservation'. *Journal of Experimental Zoology* 291 (2): 195–204.

Galle, Sabina, Nathalie Yanze, and Katja Seipel. 2005. 'The Homeobox Gene *Msx* in Development and Transdifferentiation of Jellyfish Striated Muscle'. *The International Journal of Developmental Biology* 49 (8): 961–67.

Gans, Carl, and R. Glenn Northcutt. 1983. 'Neural Crest and the Origin of Vertebrates: A New Head'. *Science* 220 (4594): 268–73.

Garcia-Bellido, A., and E. B. Lewis. 1976. 'Autonomous Cellular Differentiation of Homoeotic bithorax Mutants of *Drosophila melanogaster*'. *Developmental Biology* 48 (2): 400–410.

Garcia-Fernàndez, Jordi, and Peter W. H. Holland. 1994. 'Archetypal Organization of the Amphioxus Hox Gene Cluster'. *Nature* 370 (6490): 563–66.

Gardiner, D. M., B. Blumberg, Y. Komine, and S. V. Bryant. 1995. 'Regulation of HoxA Expression in Developing and Regenerating Axolotl Limbs'. *Development* 121 (6): 1731–41.

Gardiner, D. M., and S. V. Bryant. 1996. 'Molecular Mechanisms in the Control of Limb Regeneration: The Role of Homeobox Genes.' *International Journal of Developmental Biology* 40 (4): 797–805.

Gargioli, Cesare, and Jonathan M. W. Slack. 2004. 'Cell Lineage Tracing during *Xenopus* Tail Regeneration'. *Development* 131 (11): 2669–79.

Garstang, Myles G. 2016. 'The Evolution and Regulation of the Chordate ParaHox Cluster'. Thesis, University of St Andrews.

Garstang, Myles G., Peter W. Osborne, and David E. K. Ferrier. 2016. 'TCF/Lef Regulates the *Gsx* ParaHox Gene in Central Nervous System Development in Chordates'. *BMC Evolutionary Biology* 16 (March).

Gaunt, Stephen J. 2015. 'The Significance of Hox Gene Collinearity'. *International Journal of Developmental Biology* 59 (4-5–6): 159–70.

Gehring, W. J., and K. Ikeo. 1999. ‘*Pax 6*: Mastering Eye Morphogenesis and Eye Evolution’. *Trends in Genetics: TIG* 15 (9): 371–77.

Gehring, Walter J. 1985. ‘The Homeo Box: A Key to the Understanding of Development?’ *Cell* 40 (1): 3–5.

Gehring, Walter J., Markus Affolter, and Thomas R. Bürglin. 1994. ‘Homeodomain Proteins’. *Annual Review of Biochemistry* 63: 487–526.

Gehrke, Andrew R., and Neil H. Shubin. 2016. ‘Cis-Regulatory Programs in the Development and Evolution of Vertebrate Paired Appendages’. *Seminars in Cell & Developmental Biology* 57 (September): 31–39.

Gehrke, Andrew R, and Mansi Srivastava. 2016. ‘Neoblasts and the Evolution of Whole-Body Regeneration’. *Current Opinion in Genetics & Development*, Cell reprogramming, regeneration and repair, 40 (October): 131–37.

Gerstein, Mark B., Joel Rozowsky, Koon-Kiu Yan, Daifeng Wang, Chao Cheng, James B. Brown, Carrie A. Davis, *et al.*, 2014. ‘Comparative Analysis of the Transcriptome across Distant Species’. *Nature* 512 (7515): 445–48.

Gharbaran, Rajendra, Gabriel O. Aisemberg, and Susana Alvarado. 2012. ‘Segmental and Regional Differences in Neuronal Expression of the Leech Hox Genes *Lox1* and *Lox2* During Embryogenesis’. *Cellular and Molecular Neurobiology* 32 (8): 1243–53.

Gharbaran, Rajendra, Susana Alvarado, and Gabriel O. Aisemberg. 2014. ‘Regional and Segmental Differences in the Embryonic Expression of a Putative Leech Hox Gene, *Lox2*, by Central Neurons Immunoreactive to FMRFamide-like Neuropeptides’. *Invertebrate Neuroscience: IN* 14 (1): 51–58.

Ghosh, Sukla, Stéphane Roy, Carl Séguin, Susan V. Bryant, and David M. Gardiner. 2008. ‘Analysis of the Expression and Function of *Wnt-5a* and *Wnt-5b* in Developing and Regenerating Axolotl (*Ambystoma mexicanum*) Limbs’. *Development, Growth & Differentiation* 50 (4): 289–97.

Giribet, Gonzalo. 2008. ‘Assembling the Lophotrochozoan (=Spiralian) Tree of Life’. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363 (1496): 1513–22.

Giribet, Gonzalo, and Gregory D. Edgecombe. 2017. ‘Current Understanding of Ecdysozoa and Its Internal Phylogenetic Relationships’. *Integrative and Comparative Biology* 57 (3): 455–66.

Godwin, James W., and Nadia Rosenthal. 2014. ‘Scar-Free Wound Healing and Regeneration in Amphibians: Immunological Influences on Regenerative Success’. *Differentiation*, Exotic Animals in Development, 87 (1): 66–75.

Goecks, Jeremy, Anton Nekrutenko, and James Taylor. 2010. ‘Galaxy: A Comprehensive Approach for Supporting Accessible, Reproducible, and Transparent Computational Research in the Life Sciences’. *Genome Biology* 11 (August): R86.

Gold, David A., Ruth D. Gates, and David K. Jacobs. 2014. ‘The Early Expansion and Evolutionary Dynamics of POU Class Genes’. *Molecular Biology and Evolution*, September, msu243.

Goldstein, Bob, Linda M. Frisse, and W. Kelley Thomas. 1998. ‘Embryonic Axis Specification in Nematodes: Evolution of the First Step in Development’. *Current Biology* 8 (3): 157–60.

Goljanek-Whysall, Katarzyna, Dylan Sweetman, Muhammad Abu-Elmagd, Elik Chapnik, Tamas Dalmay, Eran Hornstein, and Andrea Münsterberg. 2011. ‘MicroRNA Regulation of the Paired-Box Transcription Factor *Pax3* Confers Robustness to Developmental Timing of Myogenesis’. *Proceedings of the National Academy of Sciences* 108 (29): 11936–41.

Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, *et al.*, 2011. ‘Trinity: Reconstructing a Full-Length Transcriptome without a Genome from RNA-Seq Data’. *Nature Biotechnology* 29 (7): 644–52.

Grenier, Jennifer K., Theodore L. Garber, Robert Warren, Paul M. Whitington, and Sean Carroll. 1997. ‘Evolution of the Entire Arthropod Hox Gene Set Predated the Origin and Radiation of the Onychophoran/Arthropod Clade’. *Current Biology* 7 (8): 547–53.

Grieshammer, Uta, David Sassoon, and Nadia Rosenthal. 1992. ‘A Transgene Target for Positional Regulators Marks Early Rostrocaudal Specification of Myogenic Lineages’. *Cell* 69 (1): 79–93.

Grimmel, Jan, Adriaan W. C. Dorresteijn, and Andreas C. Fröbius. 2016. ‘Formation of Body Appendages during Caudal Regeneration in *Platynereis dumerilii*: Adaptation of Conserved Molecular Toolsets’. *EvoDevo* 7 (1): 10.

Grohme, Markus Alexander, Siegfried Schloissnig, Andrei Rozanski, Martin Pippel, George Robert Young, Sylke Winkler, Holger Brandl, *et al.*, 2018. ‘The Genome of *Schmidtea mediterranea* and the Evolution of Core Cellular Mechanisms’. *Nature* 554 (7690): 56–61.

Groves, Andrew K., and Carole LaBonne. 2014. ‘Setting Appropriate Boundaries: Fate, Patterning and Competence at the Neural Plate Border’. *Developmental Biology*, Placodes, 389 (1): 2–12.

Gurley, Kyle A., Sarah A. Elliott, Oleg Simakov, Heiko A. Schmidt, Thomas W. Holstein, and Alejandro Sánchez Alvarado. 2010. ‘Expression of Secreted Wnt Pathway Components Reveals Unexpected Complexity of the Planarian Amputation Response’. *Developmental Biology* 347 (1): 24–39.

Gurley, Kyle A., Jochen C. Rink, and Alejandro Sánchez Alvarado. 2008. ‘β-Catenin Defines Head *Versus* Tail Identity During Planarian Regeneration and Homeostasis’. *Science* 319 (5861): 323–27.

Haeckel. 1874. *Anthropogenie oder Entwickelungsgeschichte des menschen.* W. Engelmann.

Hakes, Luke, John W. Pinney, Simon C. Lovell, Stephen G. Oliver, and David L. Robertson. 2007. ‘All Duplicates Are Not Equal: The Difference between Small-Scale and Genome Duplication’. *Genome Biology* 8 (10): R209.

Halanych, K. M., J. D. Bacheller, A. M. Aguinaldo, S. M. Liva, D. M. Hillis, and J. A. Lake. 1995. 'Evidence from 18S Ribosomal DNA That the Lophophorates Are Protostome Animals'. *Science* 267 (5204): 1641–43.

Halder, Georg, Patrick Callaerts, and Walter J. Gehring. 1995. 'Induction of Ectopic Eyes by Targeted Expression of the Eyeless Gene in *Drosophila*'. *Science* 267 (5205): 1788–92.

Hall, Brian K. 1997. 'Phylotypic Stage or Phantom: Is There a Highly Conserved Embryonic Stage in Vertebrates?' *Trends in Ecology & Evolution* 12 (12): 461–63.

Hammond, Christina L., Yaniv Hinits, Daniel P.S. Osborn, James E.N. Minchin, Gianluca Tettamanti, and Simon M. Hughes. 2007. 'Signals and Myogenic Regulatory Factors Restrict *Pax3* and *Pax7* Expression to Dermomyotome-like Tissue in Zebrafish'. *Developmental Biology* 302 (2): 504–21.

Han, Jun, Mamoru Ishii, Pablo Bringas, Richard L. Maas, Robert E. Maxson, and Yang Chai. 2007. 'Concerted Action of Msx1 and Msx2 in Regulating Cranial Neural Crest Cell Differentiation during Frontal Bone Development'. *Mechanisms of Development* 124 (9): 729–45.

Han, Manjong, Xiaodong Yang, Jennifer E. Farrington, and Ken Muneoka. 2003. 'Digit Regeneration Is Regulated by Msx1 and BMP4 in Fetal Mice'. *Development* 130 (21): 5123–32.

Hanna, J. A., M. R. Garcia, J. C. Go, D. Finkelstein, K. Kodali, V. Pagala, X. Wang, J. Peng, and M. E. Hatley. 2016. 'PAX7 Is a Required Target for MicroRNA-206-Induced Differentiation of Fusion-Negative Rhabdomyosarcoma'. *Cell Death & Disease* 7 (6): e2256.

Hanna, Jacob, Styliani Markoulaki, Patrick Schorderet, Bryce W. Carey, Caroline Beard, Marius Wernig, Menno P. Creyghton, *et al.*, 2008. 'Direct Reprogramming of Terminally Differentiated Mature B Lymphocytes to Pluripotency'. *Cell* 133 (2): 250–64.

Hanson, Sara J., Claus-Peter Stelzer, David B. Mark Welch, and John M. Logsdon. 2013. 'Comparative Transcriptome Analysis of Obligately Asexual and Cyclically Sexual Rotifers Reveals Genes with Putative Functions in Sexual Reproduction, Dormancy, and Asexual Egg Production'. *BMC Genomics* 14 (June): 412.

Harzsch, Steffen, Carsten H. G. Müller, and Yvan Perez. 2015. 'Chaetognatha'. In *Evolutionary Developmental Biology of Invertebrates 1*, 215–40. Springer, Vienna.

Hausdorf, Bernhard, Martin Helmkampf, Maximilian P. Nesnidal, and Iris Bruchhaus. 2010. 'Phylogenetic Relationships within the Lophophorate Lineages (Ectoprocta, Brachiopoda and Phoronida)'. *Molecular Phylogenetics and Evolution* 55 (3): 1121–27.

Hazkani-Covo, Einat, David Wool, and Dan Graur. 2005. 'In Search of the Vertebrate Phylotypic Stage: A Molecular Examination of the Developmental Hourglass Model and von Baer's Third Law'. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 304B (2): 150–58.

Hehenberger, Elisabeth, Denis V. Tikhonenkov, Martin Kolisko, Javier del Campo, Anton S. Esaulov, Alexander P. Mylnikov, and Patrick J. Keeling. 2017. 'Novel Predators Reshape Holozoan

Phylogeny and Reveal the Presence of a Two-Component Signaling System in the Ancestor of Animals'. *Current Biology* 27 (13): 2043-2050.e6.

Heimberg, Alysha M., Lorenzo F. Sempere, Vanessa N. Moy, Philip C. J. Donoghue, and Kevin J. Peterson. 2008. 'MicroRNAs and the Advent of Vertebrate Morphological Complexity'. *Proceedings of the National Academy of Sciences* 105 (8): 2946–50.

Helfenbein, Kevin G., H. Matthew Fourcade, Rohit G. Vanjani, and Jeffrey L. Boore. 2004. 'The Mitochondrial Genome of *Paraspadella gotoi* Is Highly Reduced and Reveals That Chaetognaths Are a Sister Group to Protostomes'. *Proceedings of the National Academy of Sciences* 101 (29): 10639–43.

Helmkampf, Martin, Iris Bruchhaus, and Bernhard Hausdorf. 2008. 'Multigene Analysis of Lophophorate and Chaetognath Phylogenetic Relationships'. *Molecular Phylogenetics and Evolution* 46 (1): 206–14.

Hench, Jürgen, Johan Henriksson, Akram M. Abou-Zied, Martin Lüppert, Johan Dethlefsen, Krishanu Mukherjee, Yong Guang Tong, *et al.*, 2015. 'The Homeobox Genes of *Caenorhabditis elegans* and Insights into Their Spatio-Temporal Expression Dynamics during Embryogenesis'. *PLOS ONE* 10 (5): e0126947.

Heyland, Andreas, Jason Hodin, and Cory Bishop. 2014. 'Manipulation of Developing Juvenile Structures in Purple Sea Urchins (*Strongylocentrotus purpuratus*) by Morpholino Injection into Late Stage Larvae'. *PLoS ONE* 9 (12).

Heyn, Patricia, Martin Kircher, Andreas Dahl, Janet Kelso, Pavel Tomancak, Alex T. Kalinka, and Karla M. Neugebauer. 2014. 'The Earliest Transcribed Zygotic Genes Are Short, Newly Evolved, and Different Across Species'. *Cell Reports* 6 (2): 285–92.

Hilgers, Leon, Stefanie Hartmann, Michael Hofreiter, Thomas von Rintelen, and Iñaki Ruiz-Trillo. 2018. 'Novel Genes, Ancient Genes, and Gene Co-Option Contributed to the Genetic Basis of the Radula, a Molluscan Innovation'. *Molecular Biology and Evolution*.

Hill, Susan D., and Barbara C. Boyer. 2001. 'Phalloidin Labeling of Developing Muscle in Embryos of the Polychaete *Capitella* sp. I'. *The Biological Bulletin* 201 (2): 257–58.

Hinman, Veronica F., Elizabeth K. O'Brien, Gemma S. Richards, and Bernard M. Degnan. 2003. 'Expression of Anterior Hox Genes during Larval Development of the Gastropod *Haliotis asinina'*. *Evolution & Development* 5 (5): 508–21.

Hirai, Hiroyuki, Mayank Verma, Shuichi Watanabe, Christopher Tastad, Yoko Asakura, and Atsushi Asakura. 2010. 'MyoD Regulates Apoptosis of Myoblasts through MicroRNA-Mediated down-Regulation of *Pax3'*. *The Journal of Cell Biology* 191 (2): 347–65.

Hirakow, R., and N. Kajita. 1991. 'Electron Microscopic Study of the Development of Amphioxus, *Branchiostoma belcheri tsingtauense*: The Gastrula'. *Journal of Morphology* 207 (1): 37–52.

Hirakow, R., and N. Kajita. 1994. ‘Electron Microscopic Study of the Development of Amphioxus, *Branchiostoma belcheri tsingtauense*: The Neurula and Larva’. *Kaibogaku Zasshi. Journal of Anatomy* 69 (1): 1–13.

Holder, Thomas, Claire Basquin, Judith Ebert, Nadine Randel, Didier Jollivet, Elena Conti, Gáspár Jékely, and Fulvia Bono. 2013. ‘Deep Transcriptome-Sequencing and Proteome Analysis of the Hydrothermal Vent Annelid *Alvinella pompejana* Identifies the CvP-Bias as a Robust Measure of Eukaryotic Thermostability’. *Biology Direct* 8 (January): 2.

Holland, Linda Z. 2009. ‘Chordate Roots of the Vertebrate Nervous System: Expanding the Molecular Toolkit’. *Nature Reviews Neuroscience* 10 (10): 736–46.

Holland, Linda Z., Ricard Albalat, Kaoru Azumi, Èlia Benito-Gutiérrez, Matthew J. Blow, Marianne Bronner-Fraser, Frederic Brunet, *et al.*, 2008. ‘The Amphioxus Genome Illuminates Vertebrate Origins and Cephalochordate Biology’. *Genome Research* 18 (7): 1100–1111.

Holland, Linda Z., and Jeremy J. Gibson-Brown. 2003. ‘The *Ciona intestinalis* Genome: When the Constraints Are Off’. *BioEssays* 25 (6): 529–32.

Holland, Linda Z., and Takayuki Onai. 2011. ‘Analyses of Gene Function in Amphioxus Embryos by Microinjection of MRNAs and Morpholino Oligonucleotides’. In *Vertebrate Embryogenesis*, 423–38. Methods in Molecular Biology. Humana Press, Totowa, NJ.

Holland, Linda Z., Michael Schubert, Zbynek Kozmik, and Nicholas D. Holland. 1999. ‘*AmphiPax3/7*, an Amphioxus Paired Box Gene: Insights into Chordate Myogenesis, Neurogenesis, and the Possible Evolutionary Precursor of Definitive Vertebrate Neural Crest’. *Evolution & Development* 1 (3): 153–65.

Holland, Nicholas D., Linda Z. Holland, and Alysha Heimberg. 2015. ‘Hybrids Between the Florida Amphioxus (*Branchiostoma floridae*) and the Bahamas Lancelet (*Asymmetron lucayanum*): Developmental Morphology and Chromosome Counts’. *The Biological Bulletin* 228 (1): 13–24.

Holland, Peter W. H. 2012. ‘Evolution of Homeobox Genes’. *Wiley Interdisciplinary Reviews: Developmental Biology* 2 (1): 31–45.

Holland, Peter W. H. 2013. ‘Evolution of Homeobox Genes’. *Wiley Interdisciplinary Reviews: Developmental Biology* 2 (1): 31–45.

Holland, Peter W. H. 2015. ‘Did Homeobox Gene Duplications Contribute to the Cambrian Explosion?’ *Zoological Letters* 1 (1).

Holland, Peter W. H., Jordi Garcia-Fernàndez, Nic A. Williams, and Arend Sidow. 1994. ‘Gene Duplications and the Origins of Vertebrate Development’. *Development* 1994 (Supplement): 125–33.

Holland, Peter W. H., Ferdinand Marlétaz, Ignacio Maeso, Thomas L. Dunwell, and Jordi Paps. 2017. ‘New Genes from Old: Asymmetric Divergence of Gene Duplicates and the Evolution of Development’. *Philosophical Transactions of the Royal Society B: Biological Sciences* 372 (1713).

Holland, Peter WH, H. Anne F. Booth, and Elspeth A. Bruford. 2007. ‘Classification and Nomenclature of All Human Homeobox Genes’. *BMC Biology* 5 (October): 47.

Holstein, Thomas W., E. Hobmayer, and U. Technau. 2003. ‘Cnidarians: An Evolutionarily Conserved Model System for Regeneration?’ *Developmental Dynamics* 226 (2): 257–67.

Holstein, Thomas W., Engelbert Hobmayer, and Charles N. David. 1991. ‘Pattern of Epithelial Cell Cycling in *Hydra*’. *Developmental Biology* 148 (2): 602–11.

Hong, Chang-Soo, and Jean-Pierre Saint-Jeannet. 2007. ‘The Activity of Pax3 and Zic1 Regulates Three Distinct Cell Fates at the Neural Plate Border’. *Molecular Biology of the Cell* 18 (6): 2192–2202.

Hove, H. A. ten, and E. K. Kupriyanova. 2009. ‘Taxonomy of Serpulidae (Annelida, Polychaeta): The State of Affairs’. *Zootaxa* 2036.

Howard-Ashby, Meredith, Stefan C. Materna, C. Titus Brown, Lili Chen, R. Andrew Cameron, and Eric H. Davidson. 2006. ‘Identification and Characterization of Homeobox Transcription Factor Genes in *Strongylocentrotus purpuratus*, and Their Expression in Embryonic Development’. *Developmental Biology*, Sea Urchin Genome: Implications and Insights, 300 (1): 74–89.

Hrycaj, Steven M., and Deneen M. Wellik. 2016. ‘Hox Genes and Evolution’. *F1000Research* 5 (May).

Hsu, Patrick D., Eric S. Lander, and Feng Zhang. 2014. ‘Development and Applications of CRISPR-Cas9 for Genome Engineering’. *Cell* 157 (6): 1262–78.

Hu, Gezhi, Hansol Lee, Sandy M. Price, Michael M. Shen, and Cory Abate-Shen. 2001. ‘Msx Homeobox Genes Inhibit Differentiation through Upregulation of Cyclin D1’. *Development* 128 (12): 2373–84.

Hu, Haiyang, Masahiro Uesaka, Song Guo, Kotaro Shimai, Tsai-Ming Lu, Fang Li, Satoko Fujimoto, *et al.*, 2017. ‘Constrained Vertebrate Evolution by Pleiotropic Genes’. *Nature Ecology & Evolution*, September, 1.

Huang, Shengfeng, Zelin Chen, Xinyu Yan, Ting Yu, Guangrui Huang, Qingyu Yan, Pierre Antoine Pontarotti, *et al.*, 2014. ‘Decelerated Genome Evolution in Modern Vertebrates Revealed by Analysis of Multiple Lancelet Genomes’. *Nature Communications* 5 (1).

Huangfu, Danwei, Kenji Osafune, René Maehr, Wenjun Guo, Astrid Eijkelenboom, Shuibing Chen, Whitney Muhlestein, and Douglas A. Melton. 2008. ‘Induction of Pluripotent Stem Cells from Primary Human Fibroblasts with Only *Oct4* and *Sox2*’. *Nature Biotechnology* 26 (11): 1269–75.

Hueber, Stefanie D., Georg F. Weiller, Michael A. Djordjevic, and Tancred Frickey. 2010. ‘Improving Hox Protein Classification across the Major Model Organisms’. *PLOS ONE* 5 (5): e10820.

Hughes, Austin L. 1994. ‘The Evolution of Functionally Novel Proteins after Gene Duplication’. *Proceedings of the Royal Society of London B: Biological Sciences* 256 (1346): 119–24.

Hui, Jerome H. L. 2008. 'The Evolution of Clustered Homeobox Genes'. Ph.D., Oxford.

Hui, Jerome H. L., Carmel McDougall, Ana S. Monteiro, Peter W. H. Holland, Detlev Arendt, Guillaume Balavoine, and David E. K. Ferrier. 2012. 'Extensive Chordate and Annelid Macrosynteny Reveals Ancestral Homeobox Gene Organization'. *Molecular Biology and Evolution* 29 (1): 157–65.

Hui, Jerome H. L., Florian Raible, Natalia Korchagina, Nicolas Dray, Sylvie Samain, Ghislaine Magdelenat, Claire Jubin, *et al.*, 2009. 'Features of the Ancestral Bilaterian Inferred from *Platynereis dumerilii* ParaHox Genes'. *BMC Biology* 7 (July): 43.

Huminiecki, Lukasz, and Carl Henrik Heldin. 2010. '2R and Remodeling of Vertebrate Signal Transduction Engine'. *BMC Biology* 8: 146.

Huvet, Arnaud, Elodie Fleury, Charlotte Corporeau, Virgile Quillien, Jean Yves Daniel, Guillaume Riviere, Pierre Boudry, and Caroline Fabioux. 2012. '*In Vivo* RNA Interference of a Gonad-Specific Transforming Growth Factor-β in the Pacific Oyster *Crassostrea gigas*'. *Marine Biotechnology* 14 (4): 402–10.

Igawa, Takeshi, Masafumi Nozawa, Daichi G. Suzuki, James D. Reimer, Arseniy R. Morov, Yiquan Wang, Yasuhisa Henmi, and Kinya Yasui. 2017. 'Evolutionary History of the Extant Amphioxus Lineage with Shallow-Branching Diversification'. *Scientific Reports* 7 (1): 1157.

Iijima, Minoru, Takeshi Takeuchi, Isao Sarashina, and Kazuyoshi Endo. 2008. 'Expression Patterns of *Engrailed* and *Dpp* in the Gastropod *Lymnaea stagnalis'*. *Development Genes and Evolution* 218 (5): 237–51.

Ikuta, Tetsuro. 2011. 'Evolution of Invertebrate Deuterostomes and Hox/ParaHox Genes'. *Genomics, Proteomics & Bioinformatics* 9 (3): 77–96.

Ikuta, Tetsuro, Natsue Yoshida, Nori Satoh, and Hidetoshi Saiga. 2004. '*Ciona intestinalis* Hox Gene Cluster: Its Dispersed Structure and Residual Colinear Expression in Development'. *Proceedings of the National Academy of Sciences* 101 (42): 15118–23.

Innan, Hideki, and Fyodor Kondrashov. 2010. 'The Evolution of Gene Duplications: Classifying and Distinguishing between Models'. *Nature Reviews Genetics* 11 (2): 97–108.

Irie, Naoki. 2017. 'Remaining Questions Related to the Hourglass Model in Vertebrate Evolution'. *Current Opinion in Genetics & Development*, Developmental mechanisms, patterning and evolution, 45 (August): 103–7.

Irie, Naoki, and Shigeru Kuratani. 2011. 'Comparative Transcriptome Analysis Reveals Vertebrate Phylotypic Period during Organogenesis'. *Nature Communications* 2 (March): 248.

Irie, Naoki, and Shigeru Kuratani. 2014. 'The Developmental Hourglass Model: A Predictor of the Basic Body Plan?' *Development* 141 (24): 4649–55.

Irie, Naoki, and Atsuko Sehara-Fujisawa. 2007. ‘The Vertebrate Phylotypic Stage and an Early Bilaterian-Related Stage in Mouse Embryogenesis Defined by Genomic Information’. *BMC Biology* 5 (January): 1.

Irvine, Steven Q, and Mark Q Martindale. 2000. ‘Expression Patterns of Anterior Hox Genes in the Polychaete *Chaetopterus*: Correlation with Morphological Boundaries’. *Developmental Biology* 217 (2): 333–51.

Isaacs, Harry V., Mary Elizabeth Pownall, and Jonathan M. W. Slack. 1998. ‘Regulation of Hox Gene Expression and Posterior Development by the *Xenopus* caudal Homologue Xcad3’. *The EMBO Journal* 17 (12): 3413–27.

Jackson, Daniel J., and Bernard M. Degnan. 2016. ‘The Importance of Evo-Devo to an Integrated Understanding of Molluscan Biomineralisation’. *Journal of Structural Biology*, SI:Biomineralization, 196 (2): 67–74.

Jacobs, David K., D Gold, N Nakanishi, D Yuan, A Camara, S Nichols, and V Hartenstein. 2010. ‘Basal Metazoan Sensory Evolution’. In *Key Transitions in Animal Evolution*, edited by Rob DeSalle and Bernd Schierwater, 175–96. Science Publishers.

Jacobs, David K., Nagayasu Nakanishi, David Yuan, Anthony Camara, Scott A. Nichols, and Volker Hartenstein. 2007. ‘Evolution of Sensory Structures in Basal Metazoa’. *Integrative and Comparative Biology* 47 (5): 712–23.

Jacobs, David K., Charles G. Wray, Cathy J. Wedeen, Richard Kostriken, Rob DeSalle, Joseph L. Staton, Ruth D. Gates, and David R. Lindberg. 2000. ‘Molluscan *engrailed* Expression, Serial Organization, and Shell Evolution’. *Evolution & Development* 2 (6): 340–47.

Jain, Kunoor, Virginia Sykes, Tomasz Kordula, and David Lanning. 2008. ‘Homeobox Genes *Hoxd3* and *Hoxd8* Are Differentially Expressed in Fetal Mouse Excisional Wounds’. *Journal of Surgical Research* 148 (1): 45–48.

Jiang, Wen-kai, Yun-long Liu, En-hua Xia, and Li-zhi Gao. 2013. ‘Prevalent Role of Gene Features in Determining Evolutionary Fates of Whole-Genome Duplication Duplicated Genes in Flowering Plants’. *Plant Physiology* 161 (4): 1844–61.

Johnson, Mark, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott McGinnis, and Thomas L. Madden. 2008. ‘NCBI BLAST: A Better Web Interface’. *Nucleic Acids Research* 36 (suppl_2): W5–9.

Jolma, Arttu, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, *et al.*, 2013. ‘DNA-Binding Specificities of Human Transcription Factors’. *Cell* 152 (1): 327–39.

Joly, Jean-Stéphane, Martine Maury, Claire Joly, Philippe Duprey, Habib Boulekbache, and Hubert Condamine. 1992. ‘Expression of a Zebrafish *caudal* Homeobox Gene Correlates with the Establishment of Posterior Cell Lineages at Gastrulation’. *Differentiation* 50 (2): 75–87.

Jong, Danielle M. de, and Elaine C. Seaver. 2016. ‘A Stable Thoracic Hox Code and Epimorphosis Characterize Posterior Regeneration in *Capitella teleta*’. *PLOS ONE* 11 (2): e0149724.

Kachgal, Suraj, Kimberly A. Mace, and Nancy J. Boudreau. 2012. ‘The Dual Roles of Homeobox Genes in Vascularization and Wound Healing’. *Cell Adhesion & Migration* 6 (6): 457–70.

Kaji, Takao, James D. Reimer, Arseniy R. Morov, Shigeru Kuratani, and Kinya Yasui. 2016. ‘Amphioxus Mouth after Dorso-Ventral Inversion’. *Zoological Letters* 2: 2.

Kalinka, Alex T., and Pavel Tomancak. 2012. ‘The Evolution of Early Animal Embryos: Conservation or Divergence?’ *Trends in Ecology & Evolution* 27 (7): 385–93.

Kalinka, Alex T., Karolina M. Varga, Dave T. Gerrard, Stephan Preibisch, David L. Corcoran, Julia Jarrells, Uwe Ohler, Casey M. Bergman, and Pavel Tomancak. 2010. ‘Gene Expression Divergence Recapitulates the Developmental Hourglass Model’. *Nature* 468 (7325): 811–14.

Kaneto, Satoshi, and Hiroshi Wada. 2011. ‘Regeneration of Amphioxus Oral Cirri and Its Skeletal Rods: Implications for the Origin of the Vertebrate Skeleton’. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 316B (6): 409–417.

Kao, Damian, Alvina G Lai, Evangelia Stamataki, Silvana Rosic, Nikolaos Konstantinides, Erin Jarvis, Alessia Di Donfrancesco, *et al.*, 2016. ‘The Genome of the Crustacean *Parhyale hawaiensis*, a Model for Animal Development, Regeneration, Immunity and Lignocellulose Digestion’. *ELife* 5.

Karl, Stefan, and Thomas Dandekar. 2015. ‘Convergence Behaviour and Control in Non-Linear Biological Networks’. *Scientific Reports* 5 (June).

Kassahn, Karin S., Vinh T. Dang, Simon J. Wilkins, Andrew C. Perkins, and Mark A. Ragan. 2009. ‘Evolution of Gene Function and Regulatory Control after Whole-Genome Duplication: Comparative Analyses in Vertebrates’. *Genome Research* 19 (8): 1404–18.

Katoh, Kazutaka, John Rozewicki, and Kazunori D. Yamada. 2017. ‘MAFFT Online Service: Multiple Sequence Alignment, Interactive Sequence Choice and Visualization’. *Briefings in Bioinformatics.*

Katoh, Kazutaka, and Daron M. Standley. 2013. ‘MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability’. *Molecular Biology and Evolution* 30 (4): 772–80.

Kaufman, Thomas C., Ricki Lewis, and Barbara Wakimoto. 1980. ‘Cytogenetic Analysis of Chromosome 3 in *Drosophila melanogaster*: The Homoeotic Gene Complex in Polytene Chromosome Interval 84a-B’. *Genetics* 94 (1): 115–33.

Kawakami, Yasuhiko, Concepción Rodriguez Esteban, Marina Raya, Hiroko Kawakami, Mercè Martí, Ilir Dubova, and Juan Carlos Izpisúa Belmonte. 2006. ‘Wnt/β-Catenin Signaling Regulates Vertebrate Limb Regeneration’. *Genes & Development* 20 (23): 000–000.

Keane, Thomas M., Christopher J. Creevey, Melissa M. Pentony, Thomas J. Naughton, and James O. Mclnerney. 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 Evolutionary Biology* 6 (March): 29.

Kemkemer, Claus, and Manyuan Long. 2014. 'New Genes Important for Development'. *EMBO Reports* 15 (5): 460–61.

Kenny, Nathan J., Erica K. O. Namigai, Ferdinand Marlétaz, Jerome H. L. Hui, and Sebastian M. Shimeld. 2015. 'Draft Genome Assemblies and Predicted MicroRNA Complements of the Intertidal Lophotrochozoans *Patella vulgata* (Mollusca, Patellogastropoda) and *Spirobranchus* (*Pomatoceros*) *lamarcki* (Annelida, Serpulida)'. *Marine Genomics*, Marine genomics for evolution and development, 24, Part 2 (December): 139–46.

Kenny, Nathan J., and Sebastian M. Shimeld. 2012. 'Additive Multiple K-Mer Transcriptome of the Keelworm *Pomatoceros lamarckii* (Annelida; Serpulidae) Reveals Annelid Trochophore Transcription Factor Cassette'. *Development Genes and Evolution* 222 (6): 325–39.

Kertesz, Michael, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. 2007. 'The Role of Site Accessibility in MicroRNA Target Recognition'. *Nature Genetics* 39 (10): 1278.

Kim, Jeong Beom, Vittorio Sebastiano, Guangming Wu, Marcos J. Araúzo-Bravo, Philipp Sasse, Luca Gentile, Kinarm Ko, *et al.*, 2009. 'Oct4-Induced Pluripotency in Adult Neural Stem Cells'. *Cell* 136 (3): 411–19.

Kin, Koryu, Shota Kakoi, and Hiroshi Wada. 2009. 'A Novel Role for Dpp in the Shaping of Bivalve Shells Revealed in a Conserved Molluscan Developmental Program'. *Developmental Biology* 329 (1): 152–66.

Klomp, Jeff, Derek Athy, Chun Wai Kwan, Natasha I. Bloch, Thomas Sandmann, Steffen Lemke, and Urs Schmidt-Ott. 2015. 'A Cysteine-Clamp Gene Drives Embryo Polarity in the Midge *Chironomus*'. *Science* 348 (6238): 1040–42.

Knapp, Dunja, Herbert Schulz, Cynthia Alexander Rascon, Michael Volkmer, Juliane Scholz, Eugen Nacu, Mu Le, *et al.*, 2013. 'Comparative Transcriptional Profiling of the Axolotl Limb Identifies a Tripartite Regeneration-Specific Gene Program'. *PLOS ONE* 8 (5): e61352.

Knopp, Paul, Nicolas Figeac, Mathieu Fortier, Louise Moyle, and Peter S. Zammit. 2013. 'Pitx Genes Are Redeployed in Adult Myogenesis Where They Can Act to Promote Myogenic Differentiation in Muscle Satellite Cells'. *Developmental Biology* 377 (1): 293–304.

Kocot, Kevin M., Torsten H. Struck, Julia Merkel, Damien S. Waits, Christiane Todt, Pamela M. Brannock, David A. Weese, *et al.*, 2017. 'Phylogenomics of Lophotrochozoa with Consideration of Systematic Error'. *Systematic Biology* 66 (2): 256–82.

Kok, Fatma O., Masahiro Shin, Chih-Wen Ni, Ankit Gupta, Ann S. Grosse, Andreas van Impel, Bettina C. Kirchmaier, *et al.*, 2015. 'Reverse Genetic Screening Reveals Poor Correlation between Morpholino-Induced and Mutant Phenotypes in Zebrafish'. *Developmental Cell* 32 (1): 97–108.

Kömüves, László G., Elias Michael, Jeffrey M. Arbeit, Xiao-Kui Ma, Angela Kwong, Eric Stelnicki, Sophia Rozenfeld, M. Morimune, Qian-Chun Yu, and Corey Largman. 2002. 'HOXB4 Homeodomain Protein Is Expressed in Developing Epidermis and Skin Disorders and Modulates Keratinocyte Proliferation'. *Developmental Dynamics* 224 (1): 58–68.

Kon, Takeshi, Masahiro Nohara, Yusuke Yamanoue, Yoshihiro Fujiwara, Mutsumi Nishida, and Teruaki Nishikawa. 2007. 'Phylogenetic Position of a Whale-Fall Lancelet (Cephalochordata) Inferred from Whole Mitochondrial Genome Sequences'. *BMC Evolutionary Biology* 7: 127.

Konstantinides, Nikolaos, and Michalis Averof. 2014. 'A Common Cellular Basis for Muscle Regeneration in Arthropods and Vertebrates'. *Science (New York, N.Y.)* 343 (6172): 788–91.

Koressaar, Triinu, and Maido Remm. 2007. 'Enhancements and Modifications of Primer Design Program Primer3'. *Bioinformatics (Oxford, England)* 23 (10): 1289–91.

Koshiba, Kazuko, Atsushi Kuroiwa, Hiroaki Yamamoto, Koji Tamura, and Hiroyuki Ide. 1998. 'Expression of Msx Genes in Regenerating and Developing Limbs of Axolotl'. *Journal of Experimental Zoology* 282 (6): 703–714.

Kourakis, Matthew J., and Mark Q. Martindale. 2001. 'Hox Gene Duplication and Deployment in the Annelid Leech *Helobdella*'. *Evolution & Development* 3 (3): 145–53.

Kozmik, Zbynek, Nicholas D. Holland, Jana Kreslova, Diana Oliveri, Michael Schubert, Kristyna Jonasova, Linda Z. Holland, Mario Pestarino, Vladimir Benes, and Simona Candiani. 2007. '*Pax–Six–Eya–Dach* Network during Amphioxus Development: Conservation *In Vitro* but Context Specificity *In Vivo*'. *Developmental Biology* 306 (1): 143–59.

Kozmikova, Iryna, Simona Candiani, Peter Fabian, Daniela Gurska, and Zbynek Kozmik. 2013. 'Essential Role of Bmp Signaling and Its Positive Feedback Loop in the Early Cell Fate Evolution of Chordates'. *Developmental Biology* 382 (2): 538–54.

Kozmikova, Iryna, Jana Smolikova, Cestmir Vlcek, and Zbynek Kozmik. 2011. 'Conservation and Diversification of an Ancestral Chordate Gene Regulatory Network for Dorsoventral Patterning'. *PLOS ONE* 6 (2): e14650.

Kozomara, Ana, and Sam Griffiths-Jones. 2011. 'MiRBase: Integrating MicroRNA Annotation and Deep-Sequencing Data'. *Nucleic Acids Research* 39 (suppl_1): D152–57.

Kragl, Martin, Dunja Knapp, Eugen Nacu, Shahryar Khattak, Malcolm Maden, Hans Henning Epperlein, and Elly M. Tanaka. 2009. 'Cells Keep a Memory of Their Tissue Origin during Axolotl Limb Regeneration'. *Nature* 460 (7251): 60.

Kulakova, Milana A., Charles E. Cook, and Tatiana F. Andreeva. 2008. ‘ParaHox Gene Expression in Larval and Postlarval Development of the Polychaete *Nereis virens* (Annelida, Lophotrochozoa)’. *BMC Developmental Biology* 8 (May): 61.

Kulakova, Milana, Nadezhda Bakalenko, Elena L. Novikova, Charles E. Cook, Elena Eliseeva, Patrick R. H. Steinmetz, Roman P. Kostyuchenko, *et al.*, 2007. ‘Hox Gene Expression in Larval Development of the Polychaetes *Nereis virens* and *Platynereis dumerilii* (Annelida, Lophotrochozoa)’. *Development Genes and Evolution* 217 (1): 39–54.

Kumar, Anoop, Cristiana P. Velloso, Yutaka Imokawa, and Jeremy P. Brockes. 2004. ‘The Regenerative Plasticity of Isolated Urodele Myofibers and Its Dependence on *Msx1’*. *PLOS Biology* 2 (8): e218.

Kumar, Sudhir, Glen Stecher, Daniel Peterson, and Koichiro Tamura. 2012. ‘MEGA-CC: Computing Core of Molecular Evolutionary Genetics Analysis Program for Automated and Iterative Data Analysis’. *Bioinformatics* 28 (20): 2685–86.

Kumar, Sudhir, Glen Stecher, and Koichiro Tamura. 2016. ‘MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets’. *Molecular Biology and Evolution* 33 (7): 1870–74.

Kuratani, Shigeru, Rie Kusakabe, and Tatsuya Hirasawa. 2018. ‘The Neural Crest and Evolution of the Head/Trunk Interface in Vertebrates’. *Developmental Biology*, February.

Kuratani, Shigeru, James F. Martin, Stefan Wawersik, Brenda Lilly, Gregor Eichele, and Eric N. Olson. 1994. ‘The Expression Pattern of the Chick Homeobox Gene *GMHox* Suggests a Role in Patterning of the Limbs and Face and in Compartmentalization of Somites’. *Developmental Biology* 161 (2): 357–69.

Kuri, Mauricio, Kyle Belek, and Nancy J. Boudreau. 2011. ‘Homeobox Genes and Wound Healing: Evidence Linking Development and Skin Regeneration’. In *Advances in Wound Care: Volume 2*, 56–61. New Rochelle: Mary Ann Liebert, Inc., publishers.

Kuwajima, Takaaki, Hideo Taniura, Isao Nishimura, and Kazuaki Yoshikawa. 2004. ‘Necdin Interacts with the Msx2 Homeodomain Protein via MAGE-D1 to Promote Myogenic Differentiation of C2C12 Cells’. *Journal of Biological Chemistry* 279 (39): 40484–93.

Kvist, Sebastian, Mercer R. Brugler, Thary G. Goh, Gonzalo Giribet, and Mark E. Siddall. 2013. ‘Pyrosequencing the Salivary Transcriptome of *Haemadipsa interrupta* (Annelida: Clitellata: Haemadipsidae): Anticoagulant Diversity and Insight into the Evolution of Anticoagulation Capabilities in Leeches’. *Invertebrate Biology* 133 (1): 74–98.

Łabaj, Paweł P., and David P. Kreil. 2016. ‘Sensitivity, Specificity, and Reproducibility of RNA-Seq Differential Expression Calls’. *Biology Direct* 11 (1): 66.

Lai, Alvina G., and A. Aziz Aboobaker. 2018. ‘EvoRegen in Animals: Time to Uncover Deep Conservation or Convergence of Adult Stem Cell Evolution and Regenerative Processes’. *Developmental Biology*, Regeneration: from cells to tissues to organisms, 433 (2): 118–31.

Lallemand, Yvan, Marie-Anne Nicola, Casto Ramos, Antoine Bach, Cécile Saint Cloment, and Benoît Robert. 2005. ‘Analysis of *Msx1*;*Msx2* Double Mutants Reveals Multiple Roles for Msx Genes in Limb Development’. *Development* 132 (13): 3003–14.

Lanfear, Robert. 2010. ‘Are the Deuterostome Posterior Hox Genes a Fast-Evolving Class?’ In *Hox Genes*, 111–22. Advances in Experimental Medicine and Biology. Springer, New York, NY.

Langmead, Ben, and Steven L. Salzberg. 2012. ‘Fast Gapped-Read Alignment with Bowtie 2’. *Nature Methods* 9 (4): 357.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, *et al.*, 2007. ‘Clustal W and Clustal X Version 2.0’. *Bioinformatics (Oxford, England)* 23 (21): 2947–48.

Larroux, Claire, Graham N. Luke, Peter Koopman, Daniel S. Rokhsar, Sebastian M. Shimeld, and Bernard M. Degnan. 2008. ‘Genesis and Expansion of Metazoan Transcription Factor Gene Classes’. *Molecular Biology and Evolution* 25 (5): 980–96.

Laumer, Christopher E., Nicolas Bekkouche, Alexandra Kerbl, Freya Goetz, Ricardo C. Neves, Martin V. Sørensen, Reinhardt M. Kristensen, *et al.*, 2015. ‘Spiralian Phylogeny Informs the Evolution of Microscopic Lineages’. *Current Biology* 25 (15): 2000–2006.

Le Grand, Fabien, Raphaëlle Grifone, Philippos Mourikis, Christophe Houbron, Carine Gigaud, Julien Pujol, Marjorie Maillet, *et al.*, 2012. ‘*Six1* Regulates Stem Cell Repair Potential and Self-Renewal during Skeletal Muscle Regeneration’. *J Cell Biol* 198 (5): 815–32.

Le Grand, Fabien, and Michael A Rudnicki. 2007. ‘Skeletal Muscle Satellite Cells and Adult Myogenesis’. *Current Opinion in Cell Biology*, Cell differentiation / Cell division, growth and death, 19 (6): 628–33.

Lee, Alison K., Christie C. Sze, Elaine R. Kim, and Yuichiro Suzuki. 2013. ‘Developmental Coupling of Larval and Adult Stages in a Complex Life Cycle: Insights from Limb Regeneration in the Flour Beetle, *Tribolium castaneum*’. *EvoDevo* 4 (1): 20.

Lee, Hansol, Raymond Habas, and Cory Abate-Shen. 2004. ‘Msx1 Cooperates with Histone H1b for Inhibition of Transcription and Myogenesis’. *Science* 304 (5677): 1675–78.

Lee, Je H. 2017. ‘Quantitative Approaches for Investigating the Spatial Context of Gene Expression’. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 9 (2): e1369.

Lehrberg, Jeffrey, and David M. Gardiner. 2015. ‘Regulation of Axolotl (*Ambystoma mexicanum*) Limb Blastema Cell Proliferation by Nerves and BMP2 in Organotypic Slice Culture’. *PLOS ONE* 10 (4): e0123186.

Leidenroth, Andreas, and Jane E Hewitt. 2010. ‘A Family History of DUX4: Phylogenetic Analysis of DUXA, B, C and *Duxbl* Reveals the Ancestral DUX Gene’. *BMC Evolutionary Biology* 10 (November): 364.

Leite, Daniel J., Luís Baudouin-Gonzalez, Sawa Iwasaki-Yokozawa, Jesus Lozano-Fernandez, Natascha Turetzek, Yasuko Akiyama-Oda, Nikola-Michael Prpic, *et al.*, 2018. ‘Homeobox Gene Duplication and Divergence in Arachnids’. *Molecular Biology and Evolution*.

Lemke, Steffen, Michael Stauber, Philip J. Shaw, Ab. Matteen Rafiqi, Alexander Prell, and Urs Schmidt-Ott. 2008. ‘Bicoid Occurrence and Bicoid-dependent Hunchback Regulation in Lower Cyclorrhaphan Flies’. *Evolution & Development* 10 (4): 413–20.

Lemons, Derek, Jens H. Fritzenwanker, John Gerhart, Christopher J. Lowe, and William McGinnis. 2010. ‘Co-Option of an Anteroposterior Head Axis Patterning System for Proximodistal Patterning of Appendages in Early Bilaterian Evolution’. *Developmental Biology* 344 (1): 358–62.

Lengner, Christopher J., Fernando D. Camargo, Konrad Hochedlinger, G. Grant Welstead, Samir Zaidi, Sumita Gokhale, Hans R. Scholer, Alexey Tomilin, and Rudolf Jaenisch. 2007. ‘*Oct4* Expression Is Not Required for Mouse Somatic Stem Cell Self-Renewal’. *Cell Stem Cell* 1 (4): 403–15.

Lepper, Christoph, Simon J. Conway, and Chen-Ming Fan. 2009. ‘Adult Satellite Cells and Embryonic Muscle Progenitors Have Distinct Genetic Requirements’. *Nature* 460 (7255): 627–31.

Lepper, Christoph, Terence A. Partridge, and Chen-Ming Fan. 2011. ‘An Absolute Requirement for Pax7-Positive Satellite Cells in Acute Injury-Induced Skeletal Muscle Regeneration’. *Development* 138 (17): 3639–46.

Levin, Michal, Leon Anavy, Alison G. Cole, Eitan Winter, Natalia Mostov, Sally Khair, Naftalie Senderovich, *et al.*, 2016. ‘The Mid-Developmental Transition and the Evolution of Animal Body Plans’. *Nature* 531 (7596): 637–41.

Levin, Michal, Tamar Hashimshony, Florian Wagner, and Itai Yanai. 2012. ‘Developmental Milestones Punctuate Gene Expression in the Caenorhabditis Embryo’. *Developmental Cell* 22 (5): 1101–8.

Levine, Michael, and Robert Tjian. 2003. ‘Transcription Regulation and Animal Diversity’. *Nature* 424 (6945): 147–51.

Lewis, E. B. 1978. ‘A Gene Complex Controlling Segmentation in Drosophila’. *Nature* 276 (5688): 565–70.

Lézot, F., V. Descroix, M. Mesbah, D. Hotton, C. Blin, P. Papagerakis, N. Mauro, *et al.*, 2002. ‘Cross-Talk Between Msx/Dlx Homeobox Genes and Vitamin D During Tooth Mineralization’. *Connective Tissue Research* 43 (2–3): 509–14.

Lézot, F., B. Thomas, D. Hotton, N. Forest, S. Orestes-Cardoso, B. Robert, P. Sharpe, and A. Berdal. 2000. ‘Biomineralization, Life-Time of Odontogenic Cells and Differential Expression of the Two

Homeobox Genes *MSX-1* and *DLX-2* in Transgenic Mice'. *Journal of Bone and Mineral Research* 15 (3): 430–441.

Li, Lihua, Aaron L. Sarver, Setara Alamgir, and Subbaya Subramanian. 2012. 'Downregulation of MicroRNAs MiR-1, -206 and -29 Stabilizes *PAX3* and CCND2 Expression in Rhabdomyosarcoma'. *Laboratory Investigation* 92 (4): 571–83.

Li, Weizhong, and Adam Godzik. 2006. 'Cd-hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences'. *Bioinformatics (Oxford, England)* 22 (13): 1658–59.

Li, Yongbin, Di Zhao, Takeo Horie, Geng Chen, Hongcun Bao, Siyu Chen, Weihong Liu, *et al.*, 2017. 'Conserved Gene Regulatory Module Specifies Lateral Neural Borders across Bilaterians'. *Proceedings of the National Academy of Sciences* 114 (31): E6352–60.

Licciano, Margherita, Joanna Michelle Murray, Gordon James Watson, and Adriana Giangrande. 2012. 'Morphological Comparison of the Regeneration Process in *Sabella spallanzanii* and *Branchiomma luctuosum* (Annelida, Sabellida)'. *Invertebrate Biology* 131 (1): 40–51.

Liu, Ning, Andrew H. Williams, Johanna M. Maxeiner, Svetlana Bezprozvannaya, John M. Shelton, James A. Richardson, Rhonda Bassel-Duby, and Eric N. Olson. 2012. 'MicroRNA-206 Promotes Skeletal Muscle Regeneration and Delays Progression of Duchenne Muscular Dystrophy in Mice'. *The Journal of Clinical Investigation* 122 (6): 2054–65.

Liu, Ying, Kathleen S. Matthews, and Sarah E. Bondos. 2009. 'Internal Regulatory Interactions Determine DNA Binding Specificity by a Hox Transcription Factor'. *Journal of Molecular Biology* 390 (4): 760–74.

Livingston, B. T., C. E. Killian, F. Wilt, A. Cameron, M. J. Landrum, O. Ermolaeva, V. Sapojnikov, D. R. Maglott, A. M. Buchanan, and C. A. Ettensohn. 2006. 'A Genome-Wide Analysis of Biomineralization-Related Proteins in the Sea Urchin *Strongylocentrotus purpuratus*'. *Developmental Biology*, Sea Urchin Genome: Implications and Insights, 300 (1): 335–48.

Lobo, Daniel, and Michael Levin. 2015. 'Inferring Regulatory Networks from Experimental Morphological Phenotypes: A Computational Method Reverse-Engineers Planarian Regeneration'. *PLoS Comput Biol* 11 (6): e1004295.

Lobo, Daniel, Junji Morokuma, and Michael Levin. 2016. 'Computational Discovery and *In Vivo* Validation of Hnf4 as a Regulatory Gene in Planarian Regeneration'. *Bioinformatics* 32 (17): 2681–85.

Longo, Antonella, Gerald P. Guanga, and Robert B. Rose. 2007. 'Structural Basis for Induced Fit Mechanisms in DNA Recognition by the Pdx1 Homeodomain'. *Biochemistry* 46 (11): 2948–57.

Louis, Alexandra, Hugues Roest Crollius, and Marc Robinson-Rechavi. 2012. 'How Much Does the Amphioxus Genome Represent the Ancestor of Chordates?' *Briefings in Functional Genomics* 11 (2): 89–95.

Lozupone, Catherine A., Robin D. Knight, and Laura F. Landweber. 2001. ‘The Molecular Basis of Nuclear Genetic Code Change in Ciliates’. *Current Biology* 11 (2): 65–74.

Luke, Graham N., L. Filipe C. Castro, Kirsten McLay, Christine Bird, Alan Coulson, and Peter W. H. Holland. 2003. ‘Dispersal of NK Homeobox Gene Clusters in Amphioxus and Humans’. *Proceedings of the National Academy of Sciences* 100 (9): 5292–95.

Lundin, Lars- G. 1999. ‘Gene Duplications in Early Metazoan Evolution’. *Seminars in Cell & Developmental Biology* 10 (5): 523–30.

Luo, Yi-Jyun, Miyuki Kanda, Ryo Koyanagi, Kanako Hisata, Tadashi Akiyama, Hirotaka Sakamoto, Tatsuya Sakamoto, and Noriyuki Satoh. 2018. ‘Nemertean and Phoronid Genomes Reveal Lophotrochozoan Evolution and the Origin of Bilaterian Heads’. *Nature Ecology & Evolution* 2: 1.

Luo, Yi-Jyun, and Yi-Hsien Su. 2012. ‘Opposing Nodal and BMP Signals Regulate Left–Right Asymmetry in the Sea Urchin Larva’. *PLOS Biology* 10 (10): e1001402.

Mace, Kimberly A., Scott L. Hansen, Connie Myers, David M. Young, and Nancy Boudreau. 2005. ‘HOXA3 Induces Cell Migration in Endothelial and Epithelial Cells Promoting Angiogenesis and Wound Repair’. *Journal of Cell Science* 118 (12): 2567–77.

Mace, Kimberly A., Terry E. Restivo, John L. Rinn, Agnes C. Paquet, Howard Y. Chang, David M. Young, and Nancy J. Boudreau. 2009. ‘HOXA3 Modulates Injury-Induced Mobilization and Recruitment of Bone Marrow-Derived Cells’. *STEM CELLS* 27 (7): 1654–65.

MacLean, James A., and Miles F. Wilkinson. 2010. ‘The Rhox Genes’. *Reproduction* 140 (2): 195–213.

Maczkowiak, Frédérique, Stéphanie Matéos, Estee Wang, Daniel Roche, Richard Harland, and Anne-Hélène Monsoro-Burq. 2010. ‘The Pax3 and Pax7 Paralogs Cooperate in Neural and Neural Crest Patterning Using Distinct Molecular Mechanisms, in *Xenopus laevis* Embryos’. *Developmental Biology* 340 (2): 381–96.

Magli, Alessandro, Erin Schnettler, Fabrizio Rinaldi, Paul Bremer, and Rita C. R. Perlingeiro. 2013. ‘Functional Dissection of Pax3 in Paraxial Mesoderm Development and Myogenesis’. *Stem Cells (Dayton, Ohio)* 31 (1): 59–70.

Mahdipour, Elahe, and Kimberly Ann Mace. 2011. ‘Hox Transcription Factor Regulation of Adult Bone-Marrow-Derived Cell Behaviour during Tissue Repair and Regeneration’. *Expert Opinion on Biological Therapy* 11 (8): 1079–90.

Makanae, Aki, Ayako Hirata, Yasuko Honjo, Kazumasa Mitogawa, and Akira Satoh. 2013. ‘Nerve Independent Limb Induction in Axolotls’. *Developmental Biology* 381 (1): 213–26.

Makarenkova, Helen P., and Robyn Meech. 2012. ‘Chapter Four - Barx Homeobox Family in Muscle Development and Regeneration’. In *International Review of Cell and Molecular Biology*, edited by Kwang W. Jeon, 297:117–73. Academic Press.

Maki, Nobuyasu, Rinako Suetsugu-Maki, Hiroshi Tarui, Kiyokazu Agata, Katia Del Rio-Tsonis, and Panagiotis A. Tsonis. 2009. ‘Expression of Stem Cell Pluripotency Factors during Regeneration in Newts’. *Developmental Dynamics* 238 (6): 1613–16.

Makino, Takashi, Karsten Hokamp, and Aoife McLysaght. 2009. ‘The Complex Relationship of Gene Duplication and Essentiality’. *Trends in Genetics* 25 (4): 152–55.

Makino, Takashi, and Aoife McLysaght. 2010. ‘Ohnologs in the Human Genome Are Dosage Balanced and Frequently Associated with Disease’. *Proceedings of the National Academy of Sciences* 107 (20): 9270–74.

Mallo, Moisés, and Claudio R. Alonso. 2013. ‘The Regulation of Hox Gene Expression during Animal Development’. *Development* 140 (19): 3951–63.

Mannini, Linda, Paolo Deri, Vittorio Gremigni, Leonardo Rossi, Alessandra Salvetti, and Renata Batistoni. 2008. ‘Two Msh/Msx-Related Genes, *Djmsh1* and *Djmsh2*, Contribute to the Early Blastema Growth during Planarian Head Regeneration’. *The International Journal of Developmental Biology* 52 (7): 943–52.

Mano, Shuhei, and Hideki Innan. 2008. ‘The Evolutionary Rate of Duplicated Genes Under Concerted Evolution’. *Genetics* 180 (1): 493–505.

Manousaki Tereza, Feiner Nathalie, Begemann Gerrit, Meyer Axel, and Kuraku Shigehiro. 2011. ‘Co-orthology of *Pax4* and *Pax6* to the Fly *eyeless* Gene: Molecular Phylogenetic, Comparative Genomic, and Embryological Analyses’. *Evolution & Development* 13 (5): 448–59.

Marchler-Bauer, Aron, Myra K. Derbyshire, Noreen R. Gonzales, Shennan Lu, Farideh Chitsaz, Lewis Y. Geer, Renata C. Geer, *et al.*, 2015. ‘CDD: NCBI’s Conserved Domain Database’. *Nucleic Acids Research* 43 (Database issue): D222-226.

Marchler-Bauer, Aron, Shennan Lu, John B. Anderson, Farideh Chitsaz, Myra K. Derbyshire, Carol DeWeese-Scott, Jessica H. Fong, *et al.*, 2011. ‘CDD: A Conserved Domain Database for the Functional Annotation of Proteins’. *Nucleic Acids Research* 39 (suppl_1): D225–29.

Mariani, Francesca V. 2010. ‘Proximal to Distal Patterning during Limb Development and Regeneration: A Review of Converging Disciplines’. *Regenerative Medicine* 5 (3): 451–62.

Marioni, John C., Christopher E. Mason, Shrikant M. Mane, Matthew Stephens, and Yoav Gilad. 2008. ‘RNA-Seq: An Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays’. *Genome Research*, September.

Marletaz, Ferdinand, Panos Firbas, Ignacio Maeso, Juan J. Tena, Ozren Bogdanovic, Malcolm Perry, Chris DR Wyatt, *et al.*, n.d. ‘Amphioxus Functional Genomics and the Evolution of Vertebrate Regulatory Traits’.

Marlétaz, Ferdinand, André Gilles, Xavier Caubit, Yvan Perez, Carole Dossat, Sylvie Samain, Gabor Gyapay, Patrick Wincker, and Yannick Le Parco. 2008. ‘Chætognath Transcriptome Reveals Ancestral and Unique Features among Bilaterians’. *Genome Biology* 9 (June): R94.

Marlétaz, Ferdinand, Ignacio Maeso, Laura Faas, Harry V. Isaacs, and Peter W. H. Holland. 2015. ‘Cdx ParaHox Genes Acquired Distinct Developmental Roles after Gene Duplication in Vertebrate Evolution’. *BMC Biology* 13 (August).

Marlétaz, Ferdinand, Elise Martin, Yvan Perez, Daniel Papillon, Xavier Caubit, Christopher J. Lowe, Bob Freeman, *et al.*, 2006. ‘Chaetognath Phylogenomics: A Protostome with Deuterostome-like Development’. *Current Biology* 16 (15): R577–78.

Martin, James F., and Eric N. Olson. 2000. ‘Identification of a *Prx1* Limb Enhancer’. *Genesis* 26 (4): 225–29.

Mashanov, Vladimir S., Olga R. Zueva, and José E. García-Arrarás. 2014. ‘Expression of Pluripotency Factors in Echinoderm Regeneration’. *Cell and Tissue Research* 359 (2): 521–36.

Mayor, Chris, Michael Brudno, Jody R. Schwartz, Alexander Poliakov, Edward M. Rubin, Kelly A. Frazer, Lior S. Pachter, and Inna Dubchak. 2000. ‘VISTA: Visualizing Global DNA Sequence Alignments of Arbitrary Length’. *Bioinformatics* 16 (11): 1046–47.

Mayran, Alexandre, Audrey Pelletier, and Jacques Drouin. 2015. ‘Pax Factors in Transcription and Epigenetic Remodelling’. *Seminars in Cell & Developmental Biology*, Paramutation & Pax Transcription Factors, 44 (Supplement C): 135–44.

McCarthy, Megan C, and Brian J Enquist. 2005. ‘Organismal Size, Metabolism and the Evolution of Complexity in Metazoans’, 16.

McCusker, Catherine D., and David M. Gardiner. 2013. ‘Positional Information Is Reprogrammed in Blastema Cells of the Regenerating Limb of the Axolotl (*Ambystoma mexicanum*)’. Edited by Panagiotis A Tsonis. *PLoS ONE* 8 (9): e77064.

McDougall, Carmel. 2008. ‘Comparative Biology of *Pomatoceros lamarckii* and Dlx Evolution in Annelids’. Ph.D., University of Oxford.

McDougall, Carmel, Wei-Chung Chen, Sebastian M. Shimeld, and David E. K. Ferrier. 2006. ‘The Development of the Larval Nervous System, Musculature and Ciliary Bands of *Pomatoceros lamarckii* (Annelida): Heterochrony in Polychaetes’. *Frontiers in Zoology* 3 (1): 16.

McDougall, Carmel, and Bernard M. Degnan. 2018. ‘The Evolution of Mollusc Shells’. *Wiley Interdisciplinary Reviews: Developmental Biology* 7 (3): e313.

McDougall, Carmel, Natalia Korchagina, Jonathan L. Tobin, and David E. K. Ferrier. 2011. ‘Annelid *Distal-Less/Dlx* Duplications Reveal Varied Post-Duplication Fates’. *BMC Evolutionary Biology* 11 (1): 241.

McGrath, Casey L., Jean-Francois Gout, Parul Johri, Thomas G. Doak, and Michael Lynch. 2014. ‘Differential Retention and Divergent Resolution of Duplicate Genes Following Whole-Genome Duplication’. *Genome Research* 24 (10): 1665–75.

McGregor, Alistair P. 2005. ‘How to Get Ahead: The Origin, Evolution and Function of Bicoid’. *BioEssays* 27 (9): 904–13.

McLysaght, Aoife, and Daniele Guerzoni. 2015. ‘New Genes from Non-Coding Sequence: The Role of *De Novo* Protein-Coding Genes in Eukaryotic Evolutionary Innovation’. *Phil. Trans. R. Soc. B* 370 (1678): 20140332.

McLysaght, Aoife, and Laurence D. Hurst. 2016. ‘Open Questions in the Study of *De Novo* Genes: What, How and Why’. *Nature Reviews Genetics* 17 (9): 567–78.

Meech, Robyn, Mariana Gomez, Christopher Woolley, Marietta Barro, Julie-Ann Hulin, Elisabeth C. Walcott, Jary Delgado, and Helen P. Makarenkova. 2010. ‘The Homeobox Transcription Factor *Barx2* Regulates Plasticity of Young Primary Myofibers’. *PLOS ONE* 5 (7): e11612.

Mehr, Shaadi, Aida Verdes, Rob DeSalle, John Sparks, Vincent Pieribone, and David F. Gruber. 2015. ‘Transcriptome Sequencing and Annotation of the Polychaete *Hermodice carunculata* (Annelida, Amphinomidae)’. *BMC Genomics* 16 (1): 445.

Mendivil Ramos, Olivia, Daniel Barker, and David E. K. Ferrier. 2012. ‘Ghost Loci Imply Hox and ParaHox Existence in the Last Common Ancestor of Animals’. *Current Biology* 22 (20): 1951–56.

Mendoza, Alex de, Arnau Sebé-Pedrós, Martin Sebastijan Šestak, Marija Matejčić, Guifré Torruella, Tomislav Domazet-Lošo, and Iñaki Ruiz-Trillo. 2013. ‘Transcription Factor Evolution in Eukaryotes and the Assembly of the Regulatory Toolkit in Multicellular Lineages’. *Proceedings of the National Academy of Sciences* 110 (50): E4858–66.

Mercader, Nadia, Licia Selleri, Luis Miguel Criado, Pilar Pallares, Carlos Parras, Michael L. Cleary, and Miguel Torres. 2009. ‘Ectopic *Meis1* Expression in the Mouse Limb Bud Alters P-D Patterning in a *Pbx1*-Independent Manner’. *International Journal of Developmental Biology* 53 (8-9–10): 1483–94.

Mercader, Nadia, Elly M. Tanaka, and Miguel Torres. 2005. ‘Proximodistal Identity during Vertebrate Limb Regeneration Is Regulated by Meis Homeodomain Proteins’. *Development* 132 (18): 4131–42.

Meulemans, Daniel, and Marianne Bronner-Fraser. 2004. ‘Gene-Regulatory Interactions in Neural Crest Evolution and Development’. *Developmental Cell* 7 (3): 291–99.

Millane, R. Cathriona, Justyna Kanska, David J. Duffy, Cathal Seoighe, Stephen Cunningham, Günter Plickert, and Uri Frank. 2011. ‘Induced Stem Cell Neoplasia in a Cnidarian by Ectopic Expression of a POU Domain Transcription Factor’. *Development* 138 (12): 2429–39.

Miller, M. A., W. Pfeiffer, and T. Schwartz. 2010. ‘Creating the CIPRES Science Gateway for Inference of Large Phylogenetic Trees’. In *2010 Gateway Computing Environments Workshop (GCE)*, 1–8.

Minguillón, Carolina, David E. K. Ferrier, Cristina Cebrián, and Jordi Garcia-Fernàndez. 2002. ‘Gene Duplications in the Prototypical Cephalochordate Amphioxus’. *Gene*, Workshop ‘Comparative Developmental Biology’, Naples, 17- April 2001, 287 (1–2): 121–28.

Mitalipov, Shoukhrat, and Don Wolf. 2009. ‘Totipotency, Pluripotency and Nuclear Reprogramming’. *Advances in Biochemical Engineering/Biotechnology* 114: 185–99.

Mitsui, Kaoru, Yoshimi Tokuzawa, Hiroaki Itoh, Kohichi Segawa, Mirei Murakami, Kazutoshi Takahashi, Masayoshi Maruyama, Mitsuyo Maeda, and Shinya Yamanaka. 2003. ‘The Homeoprotein Nanog Is Required for Maintenance of Pluripotency in Mouse Epiblast and ES Cells’. *Cell* 113 (5): 631–42.

Miyata, Takashi, and Hiroshi Suga. 2001. ‘Divergence Pattern of Animal Gene Families and Relationship with the Cambrian Explosion’. *BioEssays* 23 (11): 1018–27.

Monaghan, James R., Antony Athippozhy, Ashley W. Seifert, Sri Putta, Arnold J. Stromberg, Malcolm Maden, David M. Gardiner, and S. Randal Voss. 2012. ‘Gene Expression Patterns Specific to the Regenerating Limb of the Mexican Axolotl’. *Biology Open* 1 (10): 937–48.

Monsoro-Burq, Anne-Hélène. 2015. ‘PAX Transcription Factors in Neural Crest Development’. *Seminars in Cell & Developmental Biology*, Paramutation & Pax Transcription Factors, 44 (August): 87–96.

Monsoro-Burq, Anne-Hélène, Estee Wang, and Richard Harland. 2005. ‘*Msx1* and *Pax3* Cooperate to Mediate FGF8 and WNT Signals during *Xenopus* Neural Crest Induction’. *Developmental Cell* 8 (2): 167–78.

Monteiro, Ana Sara, and David E. K. Ferrier. 2006. ‘Hox Genes Are Not Always Colinear’. *International Journal of Biological Sciences* 2 (3): 95–103.

Moreno, Eduardo, and Pedro Martínez. 2010. ‘Origin of Bilaterian Hox Patterning System’. In *Encyclopedia of Life Sciences*, edited by John Wiley & Sons, Ltd. Chichester, UK: John Wiley & Sons, Ltd.

Morgan, Thomas Hunt. 1901. *Regeneration*. New York, The Macmillan Company; London, Macmillan & Co., Ltd.

Morino, Yoshiaki, Naoki Hashimoto, and Hiroshi Wada. 2017. ‘Expansion of TALE Homeobox Genes and the Evolution of Spiralian Development’. *Nature Ecology & Evolution* 1 (12): 1942.

Morrison, Jamie I., Paula Borg, and András Simon. 2009. ‘Plasticity and Recovery of Skeletal Muscle Satellite Cells during Limb Regeneration’. *The FASEB Journal* 24 (3): 750–56.

Morrison, Jamie I., Sara Lööf, Pingping He, and András Simon. 2006. ‘Salamander Limb Regeneration Involves the Activation of a Multipotent Skeletal Muscle Satellite Cell Population’. *The Journal of Cell Biology* 172 (3): 433–40.

Morsczeck, C. 2006. ‘Gene Expression of *Runx2*, *Osterix*, *c-Fos*, *DLX-3*, *DLX-5*, and *MSX-2* in Dental Follicle Cells during Osteogenic Differentiation *In Vitro*’. *Calcified Tissue International* 78 (2): 98–102.

Moshel, Sharon M., Michael Levine, and J. R. Collier. 1998. ‘Shell Differentiation and *engrailed* Expression in the *Ilyanassa* Embryo’. *Development Genes and Evolution* 208 (3): 135–41.

Murphy, Malea M., Jennifer A. Lawson, Sam J. Mathew, David A. Hutcheson, and Gabrielle Kardon. 2011. ‘Satellite Cells, Connective Tissue Fibroblasts and Their Interactions Are Crucial for Muscle Regeneration’. *Development* 138 (17): 3625–37.

Murrell, Ben, Steven Weaver, Martin D. Smith, Joel O. Wertheim, Sasha Murrell, Anthony Aylward, Kemal Eren, *et al.*, 2015. ‘Gene-Wide Identification of Episodic Selection’. *Molecular Biology and Evolution* 32 (5): 1365–71.

Murrell, Ben, Joel O. Wertheim, Sasha Moola, Thomas Weighill, Konrad Scheffler, and Sergei L. Kosakovsky Pond. 2012. ‘Detecting Individual Sites Subject to Episodic Diversifying Selection’. *PLOS Genetics* 8 (7): e1002764.

Myohara, Maroko. 2012. ‘What Role Do Annelid Neoblasts Play? A Comparison of the Regeneration Patterns in a Neoblast-Bearing and a Neoblast-Lacking Enchytraeid Oligochaete’. *PLOS ONE* 7 (5): e37319.

Myohara, Maroko, Cintia Carla Niva, and Jae Min Lee. 2006. ‘Molecular Approach to Annelid Regeneration: CDNA Subtraction Cloning Reveals Various Novel Genes That Are Upregulated during the Large-Scale Regeneration of the Oligochaete, *Enchytraeus japonensis’*. *Developmental Dynamics* 235 (8): 2051–70.

Nacu, Eugen, Mareen Glausch, Huy Quang Le, Febriyani Fiain Rochel Damanik, Maritta Schuez, Dunja Knapp, Shahryar Khattak, Tobias Richter, and Elly M. Tanaka. 2013. ‘Connective Tissue Cells, but Not Muscle Cells, Are Involved in Establishing the Proximo-Distal Outcome of Limb Regeneration in the Axolotl’. *Development (Cambridge, England)* 140 (3): 513–18.

Nakano, Tomoyuki, and Tomowo Ozawa. 2007. ‘Worldwide Phylogeography of Limpets of the Order Patellogastropoda: Molecular, Morphological and Palaeontological Evidence’. *Journal of Molluscan Studies* 73 (1): 79–99.

Navet, Sandra, Auxane Buresi, Sébastien Baratte, Aude Andouche, Laure Bonnaud-Ponticelli, and Yann Bassaglia. 2017. ‘The Pax Gene Family: Highlights from Cephalopods’. *PLOS ONE* 12 (3): e0172719.

Neave, Matthew J., Claire Streten-Joyce, Amanda S. Nouwens, Chris J. Glasby, Keith A. McGuinness, David L. Parry, and Karen S. Gibb. 2012. ‘The Transcriptome and Proteome Are Altered in Marine Polychaetes (Annelida) Exposed to Elevated Metal Levels’. *Journal of Proteomics* 75 (9): 2721–35.

Nederbragt, Alexander J., André E. van Loon, and Wim J. A. G. Dictus. 2002. ‘Expression of *Patella vulgata* Orthologs of *engrailed* and *Dpp-BMP2/4* in Adjacent Domains during Molluscan Shell Development Suggests a Conserved Compartment Boundary Mechanism’. *Developmental Biology* 246 (2): 341–55.

Nesnidal, Maximilian P., Martin Helmkampf, Iris Bruchhaus, and Bernhard Hausdorf. 2010. ‘Compositional Heterogeneity and Phylogenomic Inference of Metazoan Relationships’. *Molecular Biology and Evolution* 27 (9): 2095–2104.

Neves, Ricardo, and Sebastian Strempel. 2016. ‘Transcriptome Profiling of *Symbion pandora* (Phylum Cycliophora): Insights from a Differential Gene Expression Analysis’. *Organisms Diversity & Evolution*, November.

Nicolas, Stéphane, Annick Massacrier, Caubit Xavier, Pierre Cau, and Yannick Le Parco. 1996. ‘A Distal-Less-like Gene Is Induced in the Regenerating Central Nervous System of the Urodele *Pleurodeles waltl*’. *Mechanisms of Development* 56 (1): 209–20.

Nicolas, Stéphane, D. Papillon, Y. Perez, X. Caubit, and Y. Le Parco. 2003. ‘The Spatial Restrictions of 5'HoxC Genes Expression Are Maintained in Adult Newt Spinal Cord’. *Biology of the Cell* 95 (9): 389–94.

Nielsen, Claus, and Pedro Martinez. 2003. ‘Patterns of Gene Expression: Homology or Homocracy?’ *Development Genes and Evolution* 213 (3): 149–54.

Nishikawa, Teruaki. 2004. ‘A New Deep-Water Lancelet (Cephalochordata) from off Cape Nomamisaki, SW Japan, with a Proposal of the Revised System Recovering the Genus *Asymmetron*’. *Zoological Science* 21 (11): 1131–36.

Nogi, Taisaku, and Kenji Watanabe. 2001. ‘Position-Specific and Non-Colinear Expression of the Planarian Posterior (Abdominal-B-like) Gene’. *Development, Growth & Differentiation* 43 (2): 177–84.

Nohara, Masahiro, Mutsumi Nishida, Masaki Miya, and Teruaki Nishikawa. 2005. ‘Evolution of the Mitochondrial Genome in Cephalochordata as Inferred from Complete Nucleotide Sequences from Two *Epigonichthys* Species’. *Journal of Molecular Evolution* 60 (4): 526–37.

Norris, Russell A., and Michael J. Kern. 2001. ‘The Identification of *Prx1* Transcription Regulatory Domains Provides a Mechanism for Unequal Compensation by the *Prx1* and *Prx2* Loci’. *Journal of Biological Chemistry* 276 (29): 26829–37.

Novikova, Elena L., N. I. Bakalenko, A. Y. Nesterenko, and M. A. Kulakova. 2016. 'Hox Genes and Animal Regeneration'. *Russian Journal of Developmental Biology* 47 (4): 173–80.

Novikova, Elena L., Nadezhda I. Bakalenko, Alexander Y. Nesterenko, and Milana A. Kulakova. 2013. 'Expression of Hox Genes during Regeneration of Nereid Polychaete *Alitta* (*Nereis*) *virens* (Annelida, Lophotrochozoa)'. *EvoDevo* 4 (1): 14.

Nusbaum, J. 1905. 'Vergleichende Regenerationsstudien. Ueber Die Regeneration Der Polychäten *Amphiglene mediterranea* Leydig Und Nerine Cirratulus Delle Chiaje'. *Z Wiss Zool* 79: 222–307.

Nyberg, Kevin G., Matthew A. Conte, Jamie L. Kostyun, Alison Forde, and Alexandra E. Bely. 2012. 'Transcriptome Characterization via 454 Pyrosequencing of the Annelid *Pristina leidyi*, an Emerging Model for Studying the Evolution of Regeneration'. *BMC Genomics* 13 (June): 287.

Odelberg, Shannon J., Angela Kollhoff, and Mark T. Keating. 2000. 'Dedifferentiation of Mammalian Myotubes Induced by *Msx1*'. *Cell* 103 (7): 1099–1109.

Oinn, Tom, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, *et al.*, 2004. 'Taverna: A Tool for the Composition and Enactment of Bioinformatics Workflows'. *Bioinformatics* 20 (17): 3045–54.

Oliveira, Janaina Lima de, Iderval Silva Sobrinho-Junior, Samira Chahad-Ehlers, and Reinaldo Alves de Brito. 2017. 'Evolutionary Coincidence of Adaptive Changes in *exuperantia* and the Emergence of *bicoid* in Cyclorrhapha (Diptera)'. *Development Genes and Evolution* 227 (5): 355–65.

Önal, Pinar, Dominic Grün, Catherine Adamidi, Agnieszka Rybak, Jordi Solana, Guido Mastrobuoni, Yongbo Wang, *et al.*, 2012. 'Gene Expression of Pluripotency Determinants Is Conserved between Mammalian and Planarian Stem Cells'. *The EMBO Journal* 31 (12): 2755–69.

Onichtchouk, Daria. 2016. 'Evolution and Functions of Oct4 Homologs in Non-Mammalian Vertebrates'. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, The Oct Transcription Factor Family, 1859 (6): 770–79.

Onichtchouk, Daria, and Wolfgang Driever. 2016. 'Chapter Fifteen - Zygotic Genome Activators, Developmental Timing, and Pluripotency'. In *Current Topics in Developmental Biology*, edited by Paul M. Wassarman, 116:273–97. Essays on Developmental Biology, Part A. Academic Press.

Orii, Hidefumi, Kentaro Kato, Yoshihiko Umesono, Takashige Sakurai, Kiyokazu Agata, and Kenji Watanabe. 1999. 'The Planarian HOM/HOX Homeobox Genes (Plox) Expressed along the Anteroposterior Axis'. *Developmental Biology* 210 (2): 456–68.

Osborne, Peter W., Gérard Benoit, Vincent Laudet, Michael Schubert, and David E. K. Ferrier. 2009. 'Differential Regulation of ParaHox Genes by Retinoic Acid in the Invertebrate Chordate Amphioxus (*Branchiostoma floridae*)'. *Developmental Biology* 327 (1): 252–62.

Oshlack, Alicia, Mark D Robinson, and Matthew D Young. 2010. 'From RNA-Seq Reads to Differential Expression Results'. *Genome Biology* 11 (12): 220.

Oulion, Silvan, Stephanie Bertrand, Mohamed R. Belgacem, Yann Le Petillon, and Hector Escriva. 2012. ‘Sequencing and Analysis of the Mediterranean Amphioxus (*Branchiostoma lanceolatum*) Transcriptome’. *PLoS ONE* 7 (5): e36554.

Ovcharenko, Ivan, Gabriela G. Loots, Belinda M. Giardine, Minmei Hou, Jian Ma, Ross C. Hardison, Lisa Stubbs, and Webb Miller. 2005. ‘Mulan: Multiple-Sequence Local Alignment and Visualization for Studying Function and Evolution’. *Genome Research* 15 (1): 184–94.

Oviedo, Néstor J., Cindy L. Nicolas, Dany S. Adams, and Michael Levin. 2008a. ‘Gene Knockdown in Planarians Using RNA Interference’. *Cold Spring Harbor Protocols* 2008 (10): pdb.prot5054.

Oviedo, Néstor J., Cindy L. Nicolas, Dany S. Adams, and Michael Levin. 2008b. ‘Planarians: A Versatile and Powerful Model System for Molecular Studies of Regeneration, Adult Stem Cell Regulation, Aging, and Behavior’. *Cold Spring Harbor Protocols* 2008 (10): pdb.emo101.

Owen, Jennifer, B. Ann Hedley, Claus Svendsen, Jodie Wren, Martijs J. Jonker, Peter K. Hankard, Linsey J. Lister, *et al.*, 2008. ‘Transcriptome Profiling of Developmental and Xenobiotic Responses in a Keystone Soil Animal, the Oligochaete Annelid *Lumbricus rubellus’*. *BMC Genomics* 9 (June): 266.

Özpolat, B Duygu, and Alexandra E Bely. 2016. ‘Developmental and Molecular Biology of Annelid Regeneration: A Comparative Review of Recent Studies’. *Current Opinion in Genetics & Development*, Cell Reprogramming, Regeneration and Repair, 40 (Supplement C): 144–53.

Paixão-Côrtes, Vanessa Rodrigues, Francisco Mauro Salzano, and Maria Cátira Bortolini. 2013. ‘Evolutionary History of Chordate PAX Genes: Dynamics of Change in a Complex Gene Family’. *PLOS ONE* 8 (9): e73560.

Panganiban, Grace, Steven M. Irvine, Chris Lowe, Henry Roehl, Laura S. Corley, Beverley Sherbon, Jennifer K. Grenier, *et al.*, 1997. ‘The Origin and Evolution of Animal Appendages’. *Proceedings of the National Academy of Sciences* 94 (10): 5162–66.

Panganiban, Grace, and John L. R. Rubenstein. 2002. ‘Developmental Functions of the Distal-Less/Dlx Homeobox Genes’. *Development* 129 (19): 4371–86.

Papageorgiou, Louis, Picasi Eleni, Sofia Raftopoulou, Meropi Mantaiou, Vasileios Megalooikonomou, and Dimitrios Vlachakis. 2018. ‘Genomic Big Data Hitting the Storage Bottleneck’. *EMBnet.Journal* 24.

Papillon, Daniel, Yvan Perez, Xavier Caubit, and Yannick Le Parco. 2004. ‘Identification of Chaetognaths as Protostomes Is Supported by the Analysis of Their Mitochondrial Genome’. *Molecular Biology and Evolution* 21 (11): 2122–29.

Paps, Jordi. 2018. ‘What Makes an Animal? The Molecular Quest for the Origin of the Animal Kingdom’. *Integrative and Comparative Biology.*

Paps, Jordi, Jaume Baguñà, and Marta Riutort. 2009a. ‘Lophotrochozoa Internal Phylogeny: New Insights from an up-to-Date Analysis of Nuclear Ribosomal Genes’. *Proceedings of the Royal Society B: Biological Sciences* 276 (1660): 1245–54.

Paps, Jordi, Jaume Baguñà, and Marta Riutort. 2009b. ‘Bilaterian Phylogeny: A Broad Sampling of 13 Nuclear Genes Provides a New Lophotrochozoa Phylogeny and Supports a Paraphyletic Basal Acoelomorpha’. *Molecular Biology and Evolution* 26 (10): 2397–2406.

Paps, Jordi, and Peter W. H. Holland. 2018. ‘Reconstruction of the Ancestral Metazoan Genome Reveals an Increase in Genomic Novelty’. *Nature Communications* 9 (April).

Paps, Jordi, Peter W. H. Holland, and Sebastian M. Shimeld. 2012. ‘A Genome-Wide View of Transcription Factor Gene Diversity in Chordate Evolution: Less Gene Loss in Amphioxus?’ *Briefings in Functional Genomics* 11 (2): 177–86.

Paps, Jordi, Fei Xu, Guofan Zhang, and Peter W. H. Holland. 2015. ‘Reinforcing the Egg-Timer: Recruitment of Novel Lophotrochozoa Homeobox Genes to Early and Late Development in the Pacific Oyster’. *Genome Biology and Evolution* 7 (3): 677–88.

Park, Sook Kyung, Bong-Gun Ju, and Won-Sun Kim. 2009. ‘Msx-1 Acts as a Regulator for Blastema Growth’. *Genes & Genomics* 31 (6): 457–66.

Pascual-Anaya, Juan, Noritaka Adachi, Susana Álvarez, Shigeru Kuratani, Salvatore D’Aniello, and Jordi Garcia-Fernàndez. 2012. ‘Broken Colinearity of the Amphioxus Hox Cluster’. *EvoDevo* 3 (1): 28.

Pascual-Anaya, Juan, Salvatore D’Aniello, and Jordi Garcia-Fernàndez. 2008. ‘Unexpectedly Large Number of Conserved Noncoding Regions within the Ancestral Chordate Hox Cluster’. *Development Genes and Evolution* 218 (11–12): 591–97.

Pascual-Anaya, Juan, Salvatore D’Aniello, Shigeru Kuratani, and Jordi Garcia-Fernàndez. 2013. ‘Evolution of Hox Gene Clusters in Deuterostomes’. *BMC Developmental Biology* 13 (1): 26.

Peel, Andrew D., Maximilian J. Telford, and Michael Akam. 2006. ‘The Evolution of Hexapod Engrailed-Family Genes: Evidence for Conservation and Concerted Evolution’. *Proceedings of the Royal Society of London B: Biological Sciences* 273 (1595): 1733–42.

Pegeta, V. P. 1992. ‘The Regenerative Capacity of the Tail Section of the Cephalochordate’. *Vestn Zool [Zoological Herald] (Kiev)* 1: 74–76.

Perez, Yvan, Carsten H.G. Müller, and Steffen Harzsch. 2014. ‘The Chaetognatha: An Anarchistic Taxon between Protostomia and Deuterostomia’. In *Deep Metazoan Phylogeny: The Backbone of the Tree of Life*, edited by J. Wolfgang Wägele and Thomas Bartolomaeus. Berlin, Boston: DE GRUYTER.

Pesce, Maurizio, and Hans R. Schöler. 2001. ‘*Oct-4*: Gatekeeper in the Beginnings of Mammalian Development’. *STEM CELLS* 19 (4): 271–78.

Peterson, Kevin J., Steven Q. Irvine, R. Andrew Cameron, and Eric H. Davidson. 2000. 'Quantitative Assessment of Hox Complex Expression in the Indirect Development of the Polychaete Annelid *Chaetopterus* sp'. *Proceedings of the National Academy of Sciences* 97 (9): 4487–92.

Petit, Florence, Karen E. Sears, and Nadav Ahituv. 2017. 'Limb Development: A Paradigm of Gene Regulation'. *Nature Reviews Genetics* 18 (4): 245–58.

Pfeifer, Kathrin, Adriaan W. C. Dorresteijn, and Andreas C. Fröbius. 2012. 'Activation of Hox Genes during Caudal Regeneration of the Polychaete Annelid *Platynereis dumerilii*'. *Development Genes and Evolution* 222 (3): 165–79.

Piasecka, Barbara, Paweł Lichocki, Sébastien Moretti, Sven Bergmann, and Marc Robinson-Rechavi. 2013. 'The Hourglass and the Early Conservation Models—Co-Existing Patterns of Developmental Constraints in Vertebrates'. *PLOS Genetics* 9 (4): e1003476.

Pick, Leslie, and Alison Heffer. 2012. 'Hox Gene Evolution: Multiple Mechanisms Contributing to Evolutionary Novelties'. *Annals of the New York Academy of Sciences* 1256 (1): 15–32.

Poss, Stuart G., and Herbert T. Boschung. 1996. 'Lancelets (Cephalochordata: Branchiostomattdae): How Many Species Are Valid?' *Israel Journal of Zoology* 42 (sup1): S13–66.

Probst, G. 1930. 'Regenerationsstudien an Anneliden Und *Branchiostoma lanceolatum* (Pallas)'. *Rev Suisse Zool* 37: 343–52.

Pueyo, Jose Ignacio, and Juan Pablo Couso. 2005. 'Parallels between the Proximal–Distal Development of Vertebrate and Arthropod Appendages: Homology without an Ancestor?' *Current Opinion in Genetics & Development*, Pattern formation and developmental mechanisms, 15 (4): 439–46.

Purschke, Günter. 2002. 'On the Ground Pattern of Annelida'. *Organisms Diversity & Evolution* 2 (3): 181–96.

Putnam, Nicholas H., Thomas Butts, David E. K. Ferrier, Rebecca F. Furlong, Uffe Hellsten, Takeshi Kawashima, Marc Robinson-Rechavi, *et al.*, 2008. 'The Amphioxus Genome and the Evolution of the Chordate Karyotype'. *Nature* 453 (7198): 1064–71.

Putnam, Nicholas H., Mansi Srivastava, Uffe Hellsten, Bill Dirks, Jarrod Chapman, Asaf Salamov, Astrid Terry, *et al.*, 2007. 'Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization'. *Science* 317 (5834): 86–94.

Raff, Rudolf A. 1996. *The Shape of Life: Genes, Development, and the Evolution of Animal Form*. University of Chicago Press.

Raible, Florian, Kristin Tessmar-Raible, Kazutoyo Osoegawa, Patrick Wincker, Claire Jubin, Guillaume Balavoine, David E. K. Ferrier, *et al.*, 2005. 'Vertebrate-Type Intron-Rich Genes in the Marine Annelid *Platynereis dumerilii*'. *Science* 310 (5752): 1325–26.

Rajkovic, Aleksandar, Changning Yan, Wei Yan, Michal Klysik, and Martin M Matzuk. 2002. 'Obox, a Family of Homeobox Genes Preferentially Expressed in Germ Cells'. *Genomics* 79 (5): 711–17.

Rambaut. 2007. ‘FigTree’. 2007.

Ran, F. Ann, Patrick D. Hsu, Jason Wright, Vineeta Agarwala, David A. Scott, and Feng Zhang. 2013. ‘Genome Engineering Using the CRISPR-Cas9 System’. *Nature Protocols* 8 (11): 2281–2308.

Reddien, Peter W. 2011. ‘Constitutive Gene Expression and the Specification of Tissue Identity in Adult Planarian Biology’. *Trends in Genetics* 27 (7): 277–85.

Reddien, Peter W., and Alejandro Sánchez Alvarado. 2004. ‘Fundamentals of Planarian Regeneration’. *Annual Review of Cell and Developmental Biology* 20 (1): 725–57.

Reginelli, A. D., Y. Q. Wang, D. Sassoon, and K. Muneoka. 1995. ‘Digit Tip Regeneration Correlates with Regions of *Msx1* (*Hox 7*) Expression in Fetal and Newborn Mice’. *Development* 121 (4): 1065–76.

Rehmsmeier, Marc, Peter Steffen, Matthias Höchsmann, and Robert Giegerich. 2004. ‘Fast and Effective Prediction of MicroRNA/Target Duplexes’. *RNA* 10 (10): 1507–17.

Relaix, Frédéric, Didier Rocancourt, Ahmed Mansouri, and Margaret Buckingham. 2004. ‘Divergent Functions of Murine *Pax3* and *Pax7* in Limb Muscle Development’. *Genes & Development* 18 (9): 1088–1105.

Relaix, Frédéric, Didier Rocancourt, Ahmed Mansouri, and Margaret Buckingham. 2005. ‘A Pax3/Pax7-Dependent Population of Skeletal Muscle Progenitor Cells’. *Nature* 435 (7044): 948–53.

Relaix, Frederic, and Peter S. Zammit. 2012. ‘Satellite Cells Are Essential for Skeletal Muscle Regeneration: The Cell on the Edge Returns Centre Stage’. *Development* 139 (16): 2845–56.

Richardson, Michael K. 1995. ‘Heterochrony and the Phylotypic Period’. *Developmental Biology* 172 (2): 412–21.

Richardson, Michael K., J. Hanken, M. L. Gooneratne, C. Pieau, A. Raynaud, L. Selwood, and G. M. Wright. 1997. ‘There Is No Highly Conserved Embryonic Stage in the Vertebrates: Implications for Current Theories of Evolution and Development’. *Anatomy and Embryology* 196 (2): 91–106.

Riedi, Marc Andri. 2012. ‘Carbonate Production by Two New Zealand Serpulids : Skeletal Allometry, Mineralogy, Growth and Calcification of *Galeolaria hystrix* and *Spirobranchus cariniferus* (Polychaeta: Serpulidae), Southern New Zealand’. Thesis, University of Otago.

Riesgo, Ana, Sónia C. S. Andrade, Prashant P. Sharma, Marta Novo, Alicia R. Pérez-Porro, Varpu Vahtera, Vanessa L. González, Gisele Y. Kawauchi, and Gonzalo Giribet. 2012. ‘Comparative Description of Ten Transcriptomes of Newly Sequenced Invertebrates and Efficiency Estimation of Genomic Sampling in Non-Model Taxa’. *Frontiers in Zoology* 9 (November): 33.

Rink, Jochen C. 2013. ‘Stem Cell Systems and Regeneration in Planaria’. *Development Genes and Evolution* 223 (1–2): 67–84.

Rinkevich, B., Z. Shlemberg, and L. Fishelson. 1995. ‘Whole-Body Protochordate Regeneration from Totipotent Blood Cells’. *Proceedings of the National Academy of Sciences* 92 (17): 7695–99.

Rinkevich, Yuval, Guy Paz, Baruch Rinkevich, and Ram Reshef. 2007. 'Systemic Bud Induction and Retinoic Acid Signaling Underlie Whole Body Regeneration in the Urochordate *Botrylloides leachi*'. *PLOS Biology* 5 (4): e71.

Rinn, John L., Chanda Bondre, Hayes B. Gladstone, Patrick O. Brown, and Howard Y. Chang. 2006. 'Anatomic Demarcation by Positional Variation in Fibroblast Gene Expression Programs'. *PLOS Genetics* 2 (7): e119.

Rodgers-Melnick, Eli, Shrinivasrao P. Mane, Palitha Dharmawardhana, Gancho T. Slavov, Oswald R. Crasta, Steven H. Strauss, Amy M. Brunner, and Stephen P. DiFazio. 2011. 'Contrasting Patterns of Evolution Following Whole Genome *versus* Tandem Duplication Events in *Populus'*. *Genome Research*, October.

Rodriguez-Esteban, C., J. W. Schwabe, J. D. Pena, D. E. Rincon-Limas, J. Magallon, J. Botas, and J. C. Belmonte. 1998. '*Lhx2*, a Vertebrate Homologue of *apterous*, Regulates Vertebrate Limb Outgrowth'. *Development* 125 (20): 3925–34.

Roensch, Kathleen, Akira Tazaki, Osvaldo Chara, and Elly M. Tanaka. 2013. 'Progressive Specification Rather than Intercalation of Segments During Limb Regeneration'. *Science* 342 (6164): 1375–79.

Ronquist, Fredrik, and John P. Huelsenbeck. 2003. 'MrBayes 3: Bayesian Phylogenetic Inference under Mixed Models'. *Bioinformatics* 19 (12): 1572–74.

Rosa, Renaud de, Benjamin Prud'homme, and Guillaume Balavoine. 2005. '*caudal* and *even-skipped* in the Annelid *Platynereis dumerilii* and the Ancestry of Posterior Growth'. *Evolution & Development* 7 (6): 574–87.

Rose, Alexander S., Anthony R. Bradley, Yana Valasatava, Jose M. Duarte, Andreas Prlić, Peter W. Rose, and Alfonso Valencia. 2018. 'NGL Viewer: Web-Based Molecular Graphics for Large Complexes'. *Bioinformatics*.

Roselló-Díez, Alberto, Carlos G. Arques, Irene Delgado, Giovanna Giovinazzo, and Miguel Torres. 2014. 'Diffusible Signals and Epigenetic Timing Cooperate in Late Proximo-Distal Limb Patterning'. *Development*, January, dev.106831.

Rouhana, Labib, Jennifer A. Weiss, David J. Forsthoefel, Hayoung Lee, Ryan S. King, Takeshi Inoue, Norito Shibata, Kiyokazu Agata, and Phillip A. Newmark. 2013. 'RNA Interference by Feeding *In Vitro*-Synthesized Double-Stranded RNA to Planarians: Methodology and Dynamics: Planarian RNAi by Feeding *in vitro*-Synthesized DsRNA'. *Developmental Dynamics* 242 (6): 718–30.

Rouse, Greg W., Nerida G. Wilson, Jose I. Carvajal, and Robert C. Vrijenhoek. 2016. 'New Deep-Sea Species of *Xenoturbella* and the Position of Xenacoelomorpha'. *Nature* 530 (7588): 94–97.

Roux, Julien, Jialin Liu, and Marc Robinson-Rechavi. 2017. 'Selective Constraints on Coding Sequences of Nervous System Genes Are a Major Determinant of Duplicate Gene Retention in Vertebrates'. *Molecular Biology and Evolution* 34 (11): 2773–91.

Roux, Julien, and Marc Robinson-Rechavi. 2008. ‘Developmental Constraints on Vertebrate Genome Evolution’. *PLOS Genetics* 4 (12): e1000311.

Ruths, Justin, and Derek Ruths. 2014. ‘Control Profiles of Complex Networks’. *Science* 343 (6177): 1373–76.

Ruvkun, Gary, and Oliver Hobert. 1998. ‘The Taxonomy of Developmental Control in *Caenorhabditis elegans*’. *Science* 282 (5396): 2033–41.

Ryan, Joseph F., and Andreas D. Baxevanis. 2007. ‘Hox, Wnt, and the Evolution of the Primary Body Axis: Insights from the Early-Divergent Phyla’. *Biology Direct* 2: 37.

Ryan, Joseph F, Patrick M Burton, Maureen E Mazza, Grace K Kwong, James C Mullikin, and John R Finnerty. 2006. ‘The Cnidarian-Bilaterian Ancestor Possessed at Least 56 Homeoboxes: Evidence from the Starlet Sea Anemone, *Nematostella vectensis*’. *Genome Biology* 7 (7): R64.

Ryan, Joseph F., Kevin Pang, James C. Mullikin, Mark Q. Martindale, Andreas D. Baxevanis, and NISC Comparative Sequencing Program. 2010. ‘The Homeodomain Complement of the Ctenophore *Mnemiopsis leidyi* Suggests That Ctenophora and Porifera Diverged Prior to the ParaHoxozoa’. *EvoDevo* 1 (1): 9.

Rychel, Amanda L., Shannon E. Smith, Heather T. Shimamoto, and Billie J. Swalla. 2006. ‘Evolution and Development of the Chordates: Collagen and Pharyngeal Cartilage’. *Molecular Biology and Evolution* 23 (3): 541–49.

Rychel, Amanda L., and Billie J. Swalla. 2009. ‘Regeneration in Hemichordates and Echinoderms’. In *Stem Cells in Marine Organisms*, edited by Baruch Rinkevich and Valeria Matranga, 245–65. Springer Netherlands.

Ryoo, Hyun-Mo, Mi-Hye Lee, and Youn-Jeong Kim. 2006. ‘Critical Molecular Switches Involved in BMP-2-Induced Osteogenic Differentiation of Mesenchymal Cells’. *Gene* 366 (1): 51–57.

Sadedin, Simon P., Bernard Pope, and Alicia Oshlack. 2012. ‘Bpipe: A Tool for Running and Managing Bioinformatics Pipelines’. *Bioinformatics* 28 (11): 1525–26.

Saló, Emili, Josep F. Abril, Teresa Adell, Francesc Cebriá, Kay Eckelt, Enrique Fernández-Taboada, Mette Handberg-Thorsager, Marta Iglesias, M. Dolores Molina, and Gustavo Rodríguez-Esteban. 2009. ‘Planarian Regeneration: Achievements and Future Directions after 20 Years of Research’. *International Journal of Developmental Biology* 53 (8-9–10): 1317–27.

Samadi, Leyli, and Gerhard Steiner. 2009. ‘Involvement of Hox Genes in Shell Morphogenesis in the Encapsulated Development of a Top Shell Gastropod (*Gibbula varia* L.)’. *Development Genes and Evolution* 219 (9–10): 523–30.

Sambasivan, Ramkumar, Roseline Yao, Adrien Kissenpfennig, Laetitia Van Wittenberghe, Andràs Paldi, Barbara Gayraud-Morel, Hind Guenou, Bernard Malissen, Shahragim Tajbakhsh, and Anne

Galy. 2011. ‘*Pax7*-Expressing Satellite Cells Are Indispensable for Adult Skeletal Muscle Regeneration’. *Development* 138 (17): 3647–56.

Sánchez-Herrero, E., I. Vernós, R. Marco, and G. Morata. 1985. ‘Genetic Organization of *Drosophila* bithorax Complex’. *Nature* 313 (5998): 108–13.

Sandberg, Magnus, Magdalena Källström, and Jonas Muhr. 2005. ‘*Sox21* Promotes the Progression of Vertebrate Neurogenesis’. *Nature Neuroscience* 8 (8): 995–1001.

Sander, K. 1975. ‘Pattern Specification in the Insect Embryo’. *Ciba Foundation Symposium* 0 (29): 241–63.

Sander, K. 1983. ‘The Evolution of Patterning Mechanisms: Gleanings from Insect Embryogenesis and Spermatogenesis.’ In *Development and Evolution, (Eds. B.C. Goodwin, N. Holder & C.C. Wylie)*, pp.124-137. Cambridge University Press.

Sandoval-Guzmán, Tatiana, Heng Wang, Shahryar Khattak, Maritta Schuez, Kathleen Roensch, Eugeniu Nacu, Akira Tazaki, Alberto Joven, Elly M. Tanaka, and András Simon. 2014. ‘Fundamental Differences in Dedifferentiation and Stem Cell Recruitment during Skeletal Muscle Regeneration in Two Salamander Species’. *Cell Stem Cell* 14 (2): 174–87.

Sarup, Pernille, Jesper G. Sørensen, Torsten N. Kristensen, Ary A. Hoffmann, Volker Loeschcke, Ken N. Paige, and Peter Sørensen. 2011. ‘Candidate Genes Detected in Transcriptome Studies Are Strongly Dependent on Genetic Background’. *PLOS ONE* 6 (1): e15644.

Satake, Masanobu, Masakado Kawata, Aoife McLysaght, and Takashi Makino. 2012. ‘Evolution of Vertebrate Tissues Driven by Differential Modes of Gene Duplication’. *DNA Research* 19 (4): 305–16.

Satoh, Akira, David M. Gardiner, Susan V. Bryant, and Tetsuya Endo. 2007. ‘Nerve-Induced Ectopic Limb Blastemas in the Axolotl Are Equivalent to Amputation-Induced Blastemas’. *Developmental Biology* 312 (1): 231–44.

Satoh, Akira, and Aki Makanae. 2014. ‘Conservation of Position-Specific Gene Expression in Axolotl Limb Skin’. *Zoological Science* 31 (1): 6–13.

Satoh, Akira, Aki Makanae, Ayako Hirata, and Yutaka Satou. 2011. ‘Blastema Induction in Aneurogenic State and *Prrx-1* Regulation by MMPs and FGFs in *Ambystoma mexicanum* Limb Regeneration’. *Developmental Biology* 355 (2): 263–74.

Sauka-Spengler, Tatjana, and Marianne Bronner-Fraser. 2006. ‘Development and Evolution of the Migratory Neural Crest: A Gene Regulatory Perspective’. *Current Opinion in Genetics & Development*, Pattern formation and developmental mechanisms, 16 (4): 360–66.

Sauka-Spengler, Tatjana, and Marianne Bronner-Fraser. 2008. ‘A Gene Regulatory Network Orchestrates Neural Crest Formation’. *Nature Reviews Molecular Cell Biology* 9 (7): 557–68.

Schep, Alicia N., and Boris Adryan. 2013. ‘A Comparative Analysis of Transcription Factor Expression during Metazoan Embryonic Development’. *PLOS ONE* 8 (6): e66826.

Schleip, Waldemar. 1929. *Die Determination der Primitiventwicklung: eine Zusammenfassende Darstellung der Ergebnisse über das Determinationsgeschehen in den ersten Entwicklungsstadien der Tiere*. Leipzig: Akademische Verlagsgesellschaft m.b.h.

Schlosser, Gerhard. 2008. 'Do Vertebrate Neural Crest and Cranial Placodes Have a Common Evolutionary Origin?' *BioEssays* 30 (7): 659–72.

Schmerer, Matthew, Robert M. Savage, and Marty Shankland. 2009. 'Paxβ: A Novel Family of Lophotrochozoan Pax Genes'. *Evolution & Development* 11 (6): 689–96.

Schmidt, Kai, and Matthias J. Starck. 2011. 'Testing Evolutionary Hypotheses about the Phylotypic Period of Zebrafish'. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 316B (5): 319–29.

Schmitz, Jonathan F., Fabian Zimmer, and Erich Bornberg-Bauer. 2016. 'Mechanisms of Transcription Factor Evolution in Metazoa'. *Nucleic Acids Research* 44 (13): 6287–97.

Schubert, Michael, Nicholas D. Holland, Hector Escriva, Linda Z. Holland, and Vincent Laudet. 2004. 'Retinoic Acid Influences Anteroposterior Positioning of Epidermal Sensory Neurons and Their Gene Expression in a Developing Chordate (Amphioxus)'. *Proceedings of the National Academy of Sciences* 101 (28): 10320–25.

Seaver, Elaine C, Emi Yamaguchi, Gemma S Richards, and Néva P Meyer. 2012. 'Expression of the Pair-Rule Gene Homologs *runt*, *Pax3/7*, *even-skipped-1* and *even-skipped-2* during Larval and Juvenile Development of the Polychaete Annelid *Capitella teleta* Does Not Support a Role in Segmentation'. *EvoDevo* 3 (April): 8.

Sebé-Pedrós, Arnau, Alex de Mendoza, B. Franz Lang, Bernard M. Degnan, and Iñaki Ruiz-Trillo. 2011. 'Unexpected Repertoire of Metazoan Transcription Factors in the Unicellular Holozoan *Capsaspora owczarzaki*'. *Molecular Biology and Evolution* 28 (3): 1241–54.

Seifert, Anne, David F Werheid, Silvana M Knapp, and Edda Tobiasch. 2015. 'Role of Hox Genes in Stem Cell Differentiation'. *World Journal of Stem Cells* 7 (3): 583–95.

Seixas, Victor Corrêa, Claudia Augusta de Moraes Russo, and Paulo Cesar Paiva. 2017. 'Mitochondrial Genome of the Christmas Tree Worm *Spirobranchus giganteus* (Annelida: Serpulidae) Reveals a High Substitution Rate among Annelids'. *Gene* 605 (Supplement C): 43–53.

Seo, Hee-Chan, Rolf Brudvik Edvardsen, Anne Dorthea Maeland, Marianne Bjordal, Marit Flo Jensen, Anette Hansen, Mette Flaat, *et al.*, 2004. 'Hox Cluster Disintegration with Persistent Anteroposterior Order of Expression in *Oikopleura dioica*'. *Nature* 431 (7004): 67–71.

Shah, Mita V., Erica K. O. Namigai, and Yuichiro Suzuki. 2011. 'The Role of Canonical Wnt Signaling in Leg Regeneration and Metamorphosis in the Red Flour Beetle *Tribolium castaneum*'. *Mechanisms of Development* 128 (7): 342–58.

Shen, Xin, Song Sun, Fang Qing Zhao, Guang Tao Zhang, Mei Tian, Ling Ming Tsang, Jin Feng Wang, and Ka Hou Chu. 2015. 'Phylomitogenomic Analyses Strongly Support the Sister Relationship of the Chaetognatha and Protostomia'. *Zoologica Scripta* 45 (2): 187–99.

Shimeld, Sebastian M. 1997. 'A Transcriptional Modification Motif Encoded by Homeobox and Fork Head Genes'. *FEBS Letters* 410 (2): 124–25.

Shimizu, Keisuke, Yi-Jyun Luo, Noriyuki Satoh, and Kazuyoshi Endo. 2017. 'Possible Co-Option of *engrailed* during Brachiopod and Mollusc Shell Development'. *Biology Letters* 13 (8).

Short, Stephen, and Linda Z. Holland. 2008. 'The Evolution of Alternative Splicing in the Pax Family: The View from the Basal Chordate Amphioxus'. *Journal of Molecular Evolution* 66 (6): 605.

Shubin, Neil, Cliff Tabin, and Sean Carroll. 1997. 'Fossils, Genes and the Evolution of Animal Limbs'. *Nature* 388 (6643): 639–48.

Shubin, Neil, Cliff Tabin, and Sean Carroll. 2009. 'Deep Homology and the Origins of Evolutionary Novelty'. *Nature* 457 (7231): 818–23.

Silva, J. R. M. C., E. G. Mendes, and M. Mariano. 1998. 'Regeneration in the Amphioxus (*Branchiostoma platae*)'. *Zoologischer Anzeiger* 237: 107–12.

Simakov, Oleg, Ferdinand Marletaz, Sung-Jin Cho, Eric Edsinger-Gonzales, Paul Havlak, Uffe Hellsten, Dian-Han Kuo, *et al.*, 2013. 'Insights into Bilaterian Evolution from Three Spiralian Genomes'. *Nature* 493 (7433): 526.

Simion, Paul, Hervé Philippe, Denis Baurain, Muriel Jager, Daniel J. Richter, Arnaud Di Franco, Béatrice Roure, *et al.*, 2017. 'A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals'. *Current Biology* 27 (7): 958–67.

Slack, Jonathan M. W. 2003. 'Phylotype and Zootype'. In *Keywords And Concepts In Evolutionary Developmental Biology (Eds Hall, B. K. & Olson, W. M.)*, 309–318. Harvard Univ. Press.

Slack, Jonathan M. W. 2017. 'Animal Regeneration: Ancestral Character or Evolutionary Novelty?' *EMBO Reports*, July, e201643795.

Slack, Jonathan M. W., P. W. H. Holland, and C. F. Graham. 1993. 'The Zootype and the Phylotypic Stage'. *Nature*, February.

Smith, S. T., and J. B. Jaynes. 1996. 'A Conserved Region of Engrailed, Shared among All En-, Gsc-, Nk1-, Nk2- and Msh-Class Homeoproteins, Mediates Active Transcriptional Repression *In Vivo*'. *Development* 122 (10): 3141–50.

Snow, Peter, and Leo W. Buss. 1994. 'HOM/Hox-Type Homeoboxes from *Stylaria lacustris* (Annelida: Oligochaeta)'. *Molecular Phylogenetics and Evolution* 3 (4): 360–64.

Soleimani, Vahab D., Vincent G. Punch, Yoh-ichi Kawabe, Andrew E. Jones, Gareth A. Palidwor, Christopher J. Porter, Joe W. Cross, *et al.*, 2012. 'Transcriptional Dominance of *Pax7* in Adult

Myogenesis Is Due to High-Affinity Recognition of Homeodomain Motifs'. *Developmental Cell* 22 (6): 1208–20.

Somorjai, Ildikó M L, Stéphanie Bertrand, Alain Camasses, Anne Haguenauer, and Hector Escriva. 2008. 'Evidence for Stasis and Not Genetic Piracy in Developmental Expression Patterns of *Branchiostoma lanceolatum*'. *Development Genes and Evolution* 218 (11–12): 703–13.

Somorjai, Ildikó M. L., Josep Martí-Solans, Miriam Diaz-Gracia, Hiroki Nishida, Kaoru S. Imai, Hector Escrivà, Cristian Cañestro, and Ricard Albalat. 2018. 'Wnt Evolution and Function Shuffling in Liberal and Conservative Chordate Genomes'. *Genome Biology* 19 (1): 98.

Somorjai, Ildikó M L, Rajmund L Somorjai, Jordi Garcia-Fernàndez, and Hector Escrivà. 2012. 'Vertebrate-like Regeneration in the Invertebrate Chordate Amphioxus'. *Proceedings of the National Academy of Sciences of the United States of America* 109 (2): 517–22.

Somorjai, Ildikó M.L. 2017. 'Amphioxus Regeneration: Evolutionary and Biomedical Implications'. *The International Journal of Developmental Biology* 61 (10-11–12): 689–96.

Somorjai, Ildikó M.L., Hector Escrivà, and Jordi Garcia-Fernàndez. 2012. 'Amphioxus Makes the Cut—Again'. *Communicative & Integrative Biology* 5 (5): 499–502.

Song, Honghua, Lili Man, Yingjie Wang, Xue Bai, Sumei Wei, Yan Liu, Mei Liu, Xiaosong Gu, and Yongjun Wang. 2015. 'The Regenerating Spinal Cord of Gecko Maintains Unaltered Expression of β-Catenin Following Tail Amputation'. *Journal of Molecular Neuroscience* 55 (3): 653–62.

Song, Hye-Won, Anilkumar Bettegowda, Blue B. Lake, Adrienne H. Zhao, David Skarbrevik, Eric Babajanian, Meena Sukhwani, *et al.*, 2016. 'The Homeobox Transcription Factor RHOX10 Drives Mouse Spermatogonial Stem Cell Establishment'. *Cell Reports* 17 (1): 149–64.

Spagnuolo, Antonietta, Filomena Ristoratore, Anna Di Gregorio, Francesco Aniello, Margherita Branno, and Roberto Di Lauro. 2003. 'Unusual Number and Genomic Organization of Hox Genes in the Tunicate *Ciona intestinalis*'. *Gene* 309 (2): 71–79.

Spitzer, Michaela, Jan Wildenhain, Juri Rappsilber, and Mike Tyers. 2014. 'BoxPlotR: A Web Tool for Generation of Box Plots'. *Nature Methods* 11 (2): 121–22.

Srivastava, Mansi, Kathleen L. Mazza-Curll, Josien C. van Wolfswinkel, and Peter W. Reddien. 2014. 'Whole-Body Acoel Regeneration Is Controlled by Wnt and Bmp-Admp Signaling'. *Current Biology: CB* 24 (10): 1107–13.

Srivastava, Mansi, Oleg Simakov, Jarrod Chapman, Bryony Fahey, Marie E. A. Gauthier, Therese Mitros, Gemma S. Richards, *et al.*, 2010. 'The *Amphimedon queenslandica* Genome and the Evolution of Animal Complexity'. *Nature* 466 (7307): 720–26.

Ståhl, Patrik L., Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, *et al.*, 2016. 'Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics'. *Science* 353 (6294): 78–82.

Stauber, Michael, Herbert Jäckle, and Urs Schmidt-Ott. 1999. 'The Anterior Determinant *bicoid* of *Drosophila* Is a Derived Hox Class 3 Gene'. *Proceedings of the National Academy of Sciences* 96 (7): 3786–89.

Steinmetz, Patrick R. H., Roman P. Kostyuchenko, Antje Fischer, and Detlev Arendt. 2011. 'The Segmental Pattern of *Otx*, *Gbx*, and *Hox* Genes in the Annelid *Platynereis dumerilii*'. *Evolution & Development* 13 (1): 72–79.

Stelnicki, Eric J., Jeff Arbeit, Darrell L. Cass, Catherine Saner, Michael Harrison, and Corey Largman. 1998. 'Modulation of the Human Homeobox Genes *PRX-2* and *HOXB13* in Scarless Fetal Wounds'. *Journal of Investigative Dermatology* 111 (1): 57–63.

Stocum, David L. 2017. 'Mechanisms of Urodele Limb Regeneration'. *Regeneration* 4 (4): 159–200.

Stolfi, Alberto, Kerrianne Ryan, Ian A. Meinertzhagen, and Lionel Christiaen. 2015. 'Migratory Neuronal Progenitors Arise from the Neural Plate Borders in Tunicates'. *Nature* 527 (7578): 371–74.

Strand, Nicholas. 2016. 'Wnt/β-Catenin Signaling Regulates Regeneration in Diverse Tissues of the Zebrafish'. Thesis, University of Washington.

Struck, Torsten H., Christiane Paul, Natascha Hill, Stefanie Hartmann, Christoph Hösel, Michael Kube, Bernhard Lieb, *et al.*, 2011. 'Phylogenomic Analyses Unravel Annelid Evolution'. *Nature* 471 (7336): 95–98.

Struck, Torsten H., Nancy Schult, Tiffany Kusen, Emily Hickman, Christoph Bleidorn, Damhnait McHugh, and Kenneth M. Halanych. 2007. 'Annelid Phylogeny and the Status of Sipuncula and Echiura'. *BMC Evolutionary Biology* 7 (April): 57.

Struck, Torsten H., Alexandra R. Wey-Fabrizius, Anja Golombek, Lars Hering, Anne Weigert, Christoph Bleidorn, Sabrina Klebow, *et al.*, 2014. 'Platyzoan Paraphyly Based on Phylogenomic Data Supports a Noncoelomate Ancestry of Spiralia'. *Molecular Biology and Evolution* 31 (7): 1833–49.

Struck, Torsten H., Anja Golombek, Anne Weigert, Franziska Anni Franke, Wilfried Westheide, Günter Purschke, Christoph Bleidorn, and Kenneth Michael Halanych. 2015. 'The Evolution of Annelids Reveals Two Adaptive Routes to the Interstitial Realm'. *Current Biology* 25 (15): 1993–99.

Subramanian, Vasanta, Barbara I. Meyer, and Peter Gruss. 1995. 'Disruption of the Murine Homeobox Gene *Cdx1* Affects Axial Skeletal Identities by Altering the Mesodermal Expression Domains of Hox Genes'. *Cell* 83 (4): 641–53.

Sugiura, Takuji, Akira Tazaki, Naoto Ueno, Kenji Watanabe, and Makoto Mochii. 2009. '*Xenopus* Wnt-5a Induces an Ectopic Larval Tail at Injured Site, Suggesting a Crucial Role for Non-canonical Wnt Signal in Tail Regeneration'. *Mechanisms of Development* 126 (1): 56–67.

Sulston, J. E., and H. R. Horvitz. 1977. 'Post-Embryonic Cell Lineages of the Nematode, *Caenorhabditis elegans*'. *Developmental Biology* 56 (1): 110–56.

Sulston, J. E., E. Schierenberg, J. G. White, and J. N. Thomson. 1983. 'The Embryonic Cell Lineage of the Nematode *Caenorhabditis elegans*'. *Developmental Biology* 100 (1): 64–119.

Suryadeva, Sreevalli, and Mohammadi Begum Khan. 2015. 'Role of Homeobox Genes in Tooth Morphogenesis: A Review'. *Journal of Clinical and Diagnostic Research : JCDR* 9 (2): ZE09-ZE12.

Suzuki, Makoto, Akira Satoh, Hiroyuki Ide, and Koji Tamura. 2005. 'Nerve-Dependent and -Independent Events in Blastema Formation during *Xenopus* Froglet Limb Regeneration'. *Developmental Biology* 286 (1): 361–75.

Suzuki, Makoto, Akira Satoh, Hiroyuki Ide, and Koji Tamura. 2007. 'Transgenic *Xenopus* with *Prx1* Limb Enhancer Reveals Crucial Contribution of MEK/ERK and PI3K/AKT Pathways in Blastema Formation during Limb Regeneration'. *Developmental Biology* 304 (2): 675–86.

Suzuki, Makoto, Nayuta Yakushiji, Yasuaki Nakada, Akira Satoh, Hiroyuki Ide, and Koji Tamura. 2006. 'Limb Regeneration in *Xenopus laevis* Froglet'. *The Scientific World Journal* 6: 26–37.

Svensson, Mats E. 2004. 'Homology and Homocracy Revisited: Gene Expression Patterns and Hypotheses of Homology'. *Development Genes and Evolution* 214 (8): 418–21.

Svensson, Valentine, Roser Vento-Tormo, and Sarah A. Teichmann. 2018. 'Exponential Scaling of Single-Cell RNA-Seq in the Past Decade'. *Nature Protocols* 13 (4): 599–604.

Švorcová, Jana. 2012. 'The Phylotypic Stage as a Boundary of Modular Memory: Non Mechanistic Perspective'. *Theory in Biosciences* 131 (1): 31–42.

Szabó, Réka. 2015. 'Regeneration and Calcification in the *Spirobranchus lamarcki* Operculum: Development and Comparative Genetics of a Novel Appendage'. University of St Andrews.

Szabó, Réka, and David E. K. Ferrier. 2014. 'Cell Proliferation Dynamics in Regeneration of the Operculum Head Appendage in the Annelid *Pomatoceros lamarckii*'. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 322 (5): 257–68.

Szabó, Réka, and David E. K. Ferrier. 2015. 'Another Biomineralising Protostome with an *msp130* Gene and Conservation of *msp130* Gene Structure across Bilateria'. *Evolution & Development* 17 (3): 195–97.

Tabin, Clifford J., Sean B. Carroll, and Grace Panganiban. 1999. 'Out on a Limb Parallels in Vertebrate and Invertebrate Limb Patterning and the Origin of Appendages'. *Integrative and Comparative Biology* 39 (3): 650–63.

Taghiyar, Leila, Mahdi Hesaraki, Forough Azam Sayahpour, Leila Satarian, Samaneh Hosseini, Nasser Aghdami, and Mohamadreza Baghaban Eslaminejad. 2017. 'Msh Homeobox 1 (*Msx1*)− and Msx2−overexpressing Bone Marrow−derived Mesenchymal Stem Cells Resemble Blastema Cells and Enhance Regeneration in Mice'. *Journal of Biological Chemistry*, May, jbc.M116.774265.

Takahashi, Kazutoshi, Koji Tanabe, Mari Ohnuki, Megumi Narita, Tomoko Ichisaka, Kiichiro Tomoda, and Shinya Yamanaka. 2007. 'Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors'. *Cell* 131 (5): 861–72.

Takahashi, Kazutoshi, and Shinya Yamanaka. 2006. 'Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors'. *Cell* 126 (4): 663–76.

Takahashi, Tokiharu, Carmel McDougall, Jolyon Troscianko, Wei-Chung Chen, Ahamarshan Jayaraman-Nagarajan, Sebastian M. Shimeld, and David E. K. Ferrier. 2009. 'An EST Screen from the Annelid *Pomatoceros lamarckii* Reveals Patterns of Gene Loss and Gain in Animals'. *BMC Evolutionary Biology* 9 (1): 240.

Takatori, Naohito, Thomas Butts, Simona Candiani, Mario Pestarino, David E. K. Ferrier, Hidetoshi Saiga, and Peter W. H. Holland. 2008. 'Comprehensive Survey and Classification of Homeobox Genes in the Genome of Amphioxus, *Branchiostoma floridae*'. *Development Genes and Evolution* 218 (11–12): 579–90.

Takeo, Makoto, Chikako Yoshida-Noro, and Shin Tochinai. 2008. 'Morphallactic Regeneration as Revealed by Region-Specific Gene Expression in the Digestive Tract of *Enchytraeus japonensis* (Oligochaeta, Annelida)'. *Developmental Dynamics* 237 (5): 1284–1294.

Takeo, Makoto, Chikako Yoshida-Noro, and Shin Tochinai. 2009. 'Functional Analysis of *grimp*, a Novel Gene Required for Mesodermal Cell Proliferation at an Initial Stage of Regeneration in *Enchytraeus japonensis* (Enchytraeidae, Oligochaete)'. *International Journal of Developmental Biology* 54 (1): 151–60.

Tanaka, Hibiki Vincent, Nathaniel Chuen Yin Ng, Zhan Yang Yu, Martin Miguel Casco-Robles, Fumiaki Maruo, Panagiotis A. Tsonis, and Chikafumi Chiba. 2016. 'A Developmentally Regulated Switch from Stem Cells to Dedifferentiation for Limb Muscle Regeneration in Newts'. *Nature Communications* 7 (March): ncomms11069.

Tantin, Dean. 2013. 'Oct Transcription Factors in Development and Stem Cells: Insights and Mechanisms'. *Development* 140 (14): 2857–66.

Tapia, Natalia, Peter Reinhardt, Annett Duemmler, Guangming Wu, Marcos J. Araúzo-Bravo, Daniel Esch, Boris Greber, *et al.*, 2012. 'Reprogramming to Pluripotency Is an Ancient Trait of Vertebrate Oct4 and Pou2 Proteins'. *Nature Communications* 3 (December): 1279.

Tarazona, Oscar A., Davys H. Lopez, Leslie A. Slota, and Martin J. Cohn. 2018. 'Evolution of Limb Development in Cephalopod Mollusks'. *BioRxiv*, July, 379735.

Tautz, Diethard, and Tomislav Domazet-Lošo. 2011. 'The Evolutionary Origin of Orphan Genes'. *Nature Reviews Genetics* 12 (10): 692.

Telenti, Amalio, Christoph Lippert, Pi-Chuan Chang, and Mark DePristo. 2018. 'Deep Learning of Genomic Variation and Regulatory Network Data'. *Human Molecular Genetics* 27 (R1): R63–71.

Teshima, Kosuke M., and Hideki Innan. 2004. ‘The Effect of Gene Conversion on the Divergence Between Duplicated Genes’. *Genetics* 166 (3): 1553–60.

Thomas, Kelsey, Adam J. Engler, and Gretchen A. Meyer. 2015. ‘Extracellular Matrix Regulation in the Muscle Satellite Cell Niche’. *Connective Tissue Research* 56 (1): 1–8.

Thomas-Chollier, Morgane, Valérie Ledent, Luc Leyns, and Michel Vervoort. 2010. ‘A Non-Tree-Based Comprehensive Study of Metazoan Hox and ParaHox Genes Prompts New Insights into Their Origin and Evolution’. *BMC Evolutionary Biology* 10 (1): 73.

Thomas-Chollier, Morgane, and Pedro Martinez. 2016. ‘Origin of Metazoan Patterning Systems and the Role of ANTP-Class Homeobox Genes’. In *ELS*, 1–10. American Cancer Society.

Thompson, Jennifer A., Andreas Zembrzycki, Ahmed Mansouri, and Mel Ziman. 2008. ‘*Pax7* Is Requisite for Maintenance of a Subpopulation of Superior Collicular Neurons and Shows a Diverging Expression Pattern to *Pax3* during Superior Collicular Development’. *BMC Developmental Biology* 8 (May): 62.

Thompson, Jennifer A., and Mel Ziman. 2011. ‘Pax Genes during Neural Development and Their Potential Role in Neuroregeneration’. *Progress in Neurobiology* 95 (3): 334–51.

Thummel, Ryan, Shan Bai, Michael P. Sarras, Peizhen Song, Jeffrey McDermott, Jeffrey Brewer, Martin Perry, Xiaoming Zhang, David R. Hyde, and Alan R. Godwin. 2006. ‘Inhibition of Zebrafish Fin Regeneration Using *In Vivo* Electroporation of Morpholinos against *Fgfr1* and *Msxb*’. *Developmental Dynamics* 235 (2): 336–46.

Thummel, Ryan, Mila Ju, Michael P. Sarras, and Alan R. Godwin. 2007. ‘Both *Hoxc13* Orthologs Are Functionally Important for Zebrafish Tail Fin Regeneration’. *Development Genes and Evolution* 217 (6): 413–20.

Timm, Tarmo, and Patrick J. Martin. 2015. ‘Chapter 21 - Clitellata: Oligochaeta’. In *Thorp and Covich’s Freshwater Invertebrates (Fourth Edition)*, edited by James H. Thorp and D. Christopher Rogers, 529–49. Boston: Academic Press.

Tiozzo, Stefano, and Richard R. Copley. 2015. ‘Reconsidering Regeneration in Metazoans: An Evo-Devo Approach’. *Frontiers in Ecology and Evolution* 3.

Tiras, Kh P., and K. B. Aslanidi. 2016. ‘Two Populations of Pluripotent Stem Cells in Planarians *Girardia tigrina*’. *Biochemistry (Moscow) Supplement Series A: Membrane and Cell Biology* 10 (1): 46–52.

Tolkunova, Elena N., Miki Fujioka, Masatomo Kobayashi, Deepali Deka, and James B. Jaynes. 1998. ‘Two Distinct Types of Repression Domain in Engrailed: One Interacts with the Groucho Corepressor and Is Preferentially Active on Integrated Target Genes’. *Molecular and Cellular Biology* 18 (5): 2804–14.

Tomer, Raju, Alexandru S. Denes, Kristin Tessmar-Raible, and Detlev Arendt. 2010. ‘Profiling by Image Registration Reveals Common Origin of Annelid Mushroom Bodies and Vertebrate Pallium’. *Cell* 142 (5): 800–809.

Treisman, J., E. Harris, and C. Desplan. 1991. ‘The paired Box Encodes a Second DNA-Binding Domain in the Paired Homeo Domain Protein.’ *Genes & Development* 5 (4): 594–604.

Tweedt, Sarah M. 2017. ‘Gene Regulatory Networks, Homology, and the Early Panarthropod Fossil Record’. *Integrative and Comparative Biology* 57 (3): 477–87.

Tzeng, Shiou-Ru, and Charalampos G. Kalodimos. 2012. ‘Protein Activity Regulation by Conformational Entropy’. *Nature* 488 (7410): 236–40.

Untergasser, Andreas, Ioana Cutcutache, Triinu Koressaar, Jian Ye, Brant C. Faircloth, Maido Remm, and Steven G. Rozen. 2012. ‘Primer3—New Capabilities and Interfaces’. *Nucleic Acids Research* 40 (15): e115.

Uyeno, Lori A., Jennifer A. Newman-Keagle, Irene Cheung, Thomas K. Hunt, David M. Young, and Nancy Boudreau. 2001. ‘*Hox D3* Expression in Normal and Impaired Wound Healing’. *Journal of Surgical Research* 100 (1): 46–56.

Valentine, James W., Allen G. Collins, and C. Porter Meyer. 1994. ‘Morphological Complexity Increase in Metazoans’. *Paleobiology* 20 (2): 131–42.

Vidal, Berta, Anthony Santella, Esther Serrano-Saiz, Zhirong Bao, Chiou-Fen Chuang, and Oliver Hobert. 2015. ‘*C. elegans* SoxB Genes Are Dispensable for Embryonic Neurogenesis but Required for Terminal Differentiation of Specific Neuron Types’. *Development*, July, dev.125740.

Vogan, Kyle J., and Philippe Gros. 1997. ‘The C-Terminal Subdomain Makes an Important Contribution to the DNA Binding Activity of the Pax-3 Paired Domain’. *Journal of Biological Chemistry* 272 (45): 28289–95.

Vogan, Kyle J., D. A. Underhill, and P. Gros. 1996. ‘An Alternative Splicing Event in the Pax-3 Paired Domain Identifies the Linker Region as a Key Determinant of Paired Domain DNA-Binding Activity.’ *Molecular and Cellular Biology* 16 (12): 6677–86.

Vogel, Christine, and Cyrus Chothia. 2006. ‘Protein Family Expansions and Biological Complexity’. *PLoS Computational Biology* 2 (5).

Voordeckers, Karin, Ksenia Pougach, and Kevin J Verstrepen. 2015. ‘How Do Regulatory Networks Evolve and Expand throughout Evolution?’ *Current Opinion in Biotechnology*, Systems biology; Nanobiotechnology, 34 (August): 180–88.

Vorobyov, Eugene, and Jürgen Horst. 2006. ‘Getting the Proto-Pax by the Tail’. *Journal of Molecular Evolution* 63 (2): 153–64.

Voskoboynik, Ayelet, Noa Simon-Blecher, Yoav Soen, Baruch Rinkevich, Anthony W. De Tomaso, Katherine J. Ishizuka, and Irving L. Weissman. 2007. 'Striving for Normality: Whole Body Regeneration through a Series of Abnormal Generations'. *The FASEB Journal* 21 (7): 1335–44.

Wada, Hiroshi, Peter W. H. Holland, Shigeru Sato, Hiroaki Yamamoto, and Noriyuki Satoh. 1997. 'Neural Tube Is Partially Dorsalized by Overexpression of *HrPax-3/7*: The Ascidian Homologue of *Pax-3* and *Pax-7*'. *Developmental Biology* 187 (2): 240–52.

Wada, Shuichi, Miki Tokuoka, Eiichi Shoguchi, Kenji Kobayashi, Anna Di Gregorio, Antonietta Spagnuolo, Margherita Branno, *et al.*, 2003. 'A Genomewide Survey of Developmentally Relevant Genes in *Ciona intestinalis* II. Genes for Homeobox Transcription Factors'. *Development Genes and Evolution* 213 (5–6): 222–34.

Wagner, Gunte P., Chris Amemiya, and Frank Ruddle. 2003. 'Hox Cluster Duplications and the Opportunity for Evolutionary Novelties'. *Proceedings of the National Academy of Sciences* 100 (25): 14603–6.

Wang, Heng, and András Simon. 2016. 'Skeletal Muscle Dedifferentiation during Salamander Limb Regeneration'. *Current Opinion in Genetics & Development*, Cell reprogramming, regeneration and repair, 40 (October): 108–12.

Wang, Kevin C., Jill A. Helms, and Howard Y. Chang. 2009. 'Regeneration, Repair and Remembering Identity: The Three Rs of Hox Gene Expression'. *Trends in Cell Biology* 19 (6): 268–75.

Wang, Shi, Jinbo Zhang, Wenqian Jiao, Ji Li, Xiaogang Xun, Yan Sun, Ximing Guo, *et al.*, 2017. 'Scallop Genome Provides Insights into Evolution of Bilaterian Karyotype and Development'. *Nature Ecology & Evolution* 1 (5): 0120.

Wang, Sue-Hong, Ming-Shiun Tsai, Ming-Fu Chiang, and Hung Li. 2003. 'A Novel NK-Type Homeobox Gene, ENK (Early Embryo Specific NK), Preferentially Expressed in Embryonic Stem Cells'. *Gene Expression Patterns* 3 (1): 99–103.

Wang, Wei, Huai-Liang Xu, Lu-Ping Lin, Bing Su, and Yi-Quan Wang. 2005. 'Construction of a BAC Library for Chinese Amphioxus *Branchiostoma belcheri* and Identification of Clones Containing Amphi-Pax Genes'. *Genes & Genetic Systems* 80 (3): 233–36.

Wang, Wei, Jing Zhong, and Yi-Quan Wang. 2010. 'Comparative Genomic Analysis Reveals the Evolutionary Conservation of Pax Gene Family'. *Genes & Genetic Systems* 85 (3): 193–206.

Wang, Zhuo, Juan Pascual-Anaya, Amonida Zadissa, Wenqi Li, Yoshihito Niimura, Zhiyong Huang, Chunyi Li, *et al.*, 2013. 'The Draft Genomes of Soft-Shell Turtle and Green Sea Turtle Yield Insights into the Development and Evolution of the Turtle-Specific Body Plan'. *Nature Genetics* 45 (6): 701–6.

Wanninger, Andreas, and Gerhard Haszprunar. 2001. ‘The Expression of an engrailed Protein during Embryonic Shell Formation of the Tusk-Shell, *Antalis entalis* (Mollusca, Scaphopoda)’. *Evolution & Development* 3 (5): 312–21.

Wasik, Kaja, James Gurtowski, Xin Zhou, Olivia Mendivil Ramos, M. Joaquina Delás, Giorgia Battistoni, Osama El Demerdash, *et al.*, 2015. ‘Genome and Transcriptome of the Regeneration-Competent Flatworm, *Macrostomum lignano’*. *Proceedings of the National Academy of Sciences* 112 (40): 12462–67.

Watanabe, Shuichi, Shunzo Kondo, Michiko Hayasaka, and Kazunori Hanaoka. 2007. ‘Functional Analysis of Homeodomain-Containing Transcription Factor *Lbx1* in Satellite Cells of Mouse Skeletal Muscle’. *Journal of Cell Science* 120 (23): 4178–87.

Waterhouse, Andrew M., James B. Procter, David M. A. Martin, Michèle Clamp, and Geoffrey J. Barton. 2009. ‘Jalview Version 2—a Multiple Sequence Alignment Editor and Analysis Workbench’. *Bioinformatics* 25 (9): 1189–91.

Webb, Graham C., Paul Q. Thomas, Judith H. Ford, and Peter D. Rathjen. 1993. ‘Hesx1, a Homeobox Gene Expressed by Murine Embryonic Stem Cells, Maps to Mouse Chromosome 14, Bands A3-B’. *Genomics* 18 (2): 464–66.

Wehner, Daniel, and Gilbert Weidinger. 2015. ‘Signaling Networks Organizing Regenerative Growth of the Zebrafish Fin’. *Trends in Genetics* 31 (6): 336–43.

Weigert, Anne, and Christoph Bleidorn. 2016. ‘Current Status of Annelid Phylogeny’. *Organisms Diversity & Evolution* 16 (2): 345–62.

Weigert, Anne, Conrad Helm, Matthias Meyer, Birgit Nickel, Detlev Arendt, Bernhard Hausdorf, Scott R. Santos, *et al.*, 2014. ‘Illuminating the Base of the Annelid Tree Using Transcriptomics’. *Molecular Biology and Evolution* 31 (6): 1391–1401.

Wells, James M., and Fiona M. Watt. 2018. ‘Diverse Mechanisms for Endogenous Regeneration and Repair in Mammalian Organs’. *Nature* 557 (7705): 322–28.

Wertheim, Joel O., Ben Murrell, Martin D. Smith, Kosakovsky Pond, Sergei L, and Konrad Scheffler. 2015. ‘RELAX: Detecting Relaxed Selection in a Phylogenetic Framework’. *Molecular Biology and Evolution* 32 (3): 820–32.

White, Philip, David W. Thomas, Steven Fong, Eric Stelnicki, Fritz Meijlink, Corey Largman, and Phil Stephens. 2003. ‘Deletion of the Homeobox Gene *PRX-2* Affects Fetal but Not Adult Fibroblast Wound Healing Responses’. *Journal of Investigative Dermatology* 120 (1): 135–44.

Whittington, Niteace, Doreen Cunningham, Thien-Kim Le, David De Maria, and Elena M. Silva. 2015. ‘*Sox21* Regulates the Progression of Neuronal Differentiation in a Dose-Dependent Manner’. *Developmental Biology* 397 (2): 237–47.

Wicht, Helmut, and Thurston C Lacalli. 2005. ‘The Nervous System of Amphioxus: Structure, Development, and Evolutionary Significance’. *Canadian Journal of Zoology* 83 (1): 122–50.

Williams, Nic A., and Peter W. H. Holland. 2000. ‘An Amphioxus *Emx* Homeobox Gene Reveals Duplication During Vertebrate Evolution’. *Molecular Biology and Evolution* 17 (10): 1520–28.

Williams, Terri A. 2013. ‘Mechanisms of Limb Patterning in Crustaceans’. In *Functional Morphology and Diversity*, edited by Les Watling and Martin Thiel, 74–102. Oxford University Press.

Wilt, Fred H., Christopher E. Killian, and Brian T. Livingston. 2003. ‘Development of Calcareous Skeletal Elements in Invertebrates’. *Differentiation* 71 (4–5): 237–50.

Winchell, Christopher J., and David K. Jacobs. 2013. ‘Expression of the Lhx Genes *apterous* and *Lim1* in an Errant Polychaete: Implications for Bilaterian Appendage Evolution, Neural Development, and Muscle Diversification’. *EvoDevo* 4 (1): 4.

Winchell, Christopher J., Jonathan E. Valencia, and David K. Jacobs. 2010. ‘Expression of *Distal-less*, *dachshund*, and *optomotor blind* in *Neanthes arenaceodentata* (Annelida, Nereididae) Does Not Support Homology of Appendage-Forming Mechanisms across the Bilateria’. *Development Genes and Evolution* 220 (9–10): 275–95.

Wollesen, Tim, Maik Scherholz, Sonia Victoria Rodríguez Monje, Emanuel Redl, Christiane Todt, and Andreas Wanninger. 2017. ‘Brain Regionalization Genes Are Co-Opted into Shell Field Patterning in Mollusca’. *Scientific Reports* 7 (1): 5486.

Wong, Eunice, Elena K. Kupriyanova, Pat Hutchings, María Capa, Vasily I. Radashevsky, and Harry A. ten Hove. 2014. ‘A Graphically Illustrated Glossary of Polychaete Terminology: Invasive Species of Sabellidae, Serpulidae and Spionidae’. *Memoirs of Museum Victoria* 71: 327–42.

Wong, Yue Him, Taewoo Ryu, Loqmane Seridi, Yanal Ghosheh, Salim Bougouffa, Pei-Yuan Qian, and Timothy Ravasi. 2014. ‘Transcriptome Analysis Elucidates Key Developmental Components of Bryozoan Lophophore Development’. *Scientific Reports* 4 (October): 6534.

Wu, Rimao, Hu Li, Lili Zhai, Xiaoting Zou, Jiao Meng, Ran Zhong, Changyin Li, Haixia Wang, Yong Zhang, and Dahai Zhu. 2015. ‘MicroRNA-431 Accelerates Muscle Regeneration and Ameliorates Muscular Dystrophy by Targeting *Pax7* in Mice’. *Nature Communications* 6 (July): 7713.

Xie, Ting, Qing-Yong Yang, Xiao-Tao Wang, Aoife McLysaght, and Hong-Yu Zhang. 2016. ‘Spatial Colocalization of Human Ohnolog Pairs Acts to Maintain Dosage-Balance’. *Molecular Biology and Evolution* 33 (9): 2368–75.

Xu, Fei, Tomislav Domazet-Lošo, Dingding Fan, Thomas L. Dunwell, Li Li, Xiaodong Fang, and Guofan Zhang. 2016. ‘High Expression of New Genes in Trochophore Enlightening the Ontogeny and Evolution of Trochozoans’. *Scientific Reports* 6 (October): 34664.

Xu, Jin, Rui Zhang, Yang Shen, Guojing Liu, Xuemei Lu, and Chung-I Wu. 2013. ‘The Evolution of Evolvability in MicroRNA Target Sites in Vertebrates’. *Genome Research* 23 (11): 1810–16.

Yajima, Hiroshi, Norio Motohashi, Yusuke Ono, Shigeru Sato, Keiko Ikeda, Satoru Masuda, Erica Yada, *et al.*, 2010. 'Six Family Genes Control the Proliferation and Differentiation of Muscle Satellite Cells'. *Experimental Cell Research* 316 (17): 2932–44.

Yampolsky, Lev Y, and Michael A Bouzinier. 2014. 'Faster Evolving *Drosophila* Paralogs Lose Expression Rate and Ubiquity and Accumulate More Non-Synonymous SNPs'. *Biology Direct* 9 (January): 2.

Yan, Fang, Shaojie Luo, Yu Jiao, Yuewen Deng, Xiaodong Du, Ronglian Huang, Qingheng Wang, and Weiyao Chen. 2014. 'Molecular Characterization of the *BMP7* Gene and Its Potential Role in Shell Formation in *Pinctada martensii'*. *International Journal of Molecular Sciences* 15 (11): 21215–28.

Yang, Kevin Yi, Yuan Chen, Zuming Zhang, Patrick Kwok-Shing Ng, Wayne Junwei Zhou, Yinfeng Zhang, Minghua Liu, Junyuan Chen, Bingyu Mao, and Stephen Kwok-Wing Tsui. 2016. 'Transcriptome Analysis of Different Developmental Stages of Amphioxus Reveals Dynamic Changes of Distinct Classes of Genes during Development'. *Scientific Reports* 6 (March): 23195.

Yang, Qiumei, Jie Yu, Bing Yu, Zhiqing Huang, Keying Zhang, De Wu, Jun He, Xiangbing Mao, Ping Zheng, and Daiwen Chen. 2016. 'PAX3+ Skeletal Muscle Satellite Cells Retain Long-Term Self-Renewal and Proliferation'. *Muscle & Nerve* 54 (5): 943–51.

Yang, Ziheng. 2007. 'PAML 4: Phylogenetic Analysis by Maximum Likelihood'. *Molecular Biology and Evolution* 24 (8): 1586–91.

Ye, Wenbin, Qing Lv, Chung-Kwun Amy Wong, Sean Hu, Chao Fu, Zhong Hua, Guoping Cai, Guoxi Li, Burton B. Yang, and Yaou Zhang. 2008. 'The Effect of Central Loops in MiRNA:MRE Duplexes on the Efficiency of MiRNA-Mediated Gene Regulation'. *PLOS ONE* 3 (3): e1719.

Yeh, Jennifer, Lydia M. Green, Ting-Xin Jiang, Maksim Plikus, Eunice Huang, Richard N. Chang, Michael W. Hughes, Cheng-Ming Chuong, and Tai-Lan Tuan. 2009. 'Accelerated Closure of Skin Wounds in Mice Deficient in the Homeobox Gene *Msx2'*. *Wound Repair and Regeneration* 17 (5): 639–48.

Yi, Buqing, Dan Bumbarger, and Ralf J. Sommer. 2009. 'Genetic Evidence for Pax-3 Function in Myogenesis in the Nematode *Pristionchus pacificus'*. *Evolution & Development* 11 (6): 669–79.

Yin, Hang, Feodor Price, and Michael A. Rudnicki. 2013. 'Satellite Cells and the Muscle Stem Cell Niche'. *Physiological Reviews* 93 (1): 23–67.

Yokoyama, Hitoshi. 2008. 'Initiation of Limb Regeneration: The Critical Steps for Regenerative Capacity'. *Development, Growth & Differentiation* 50 (1): 13–22.

Yokoyama, Hitoshi, Tamae Maruoka, Akio Aruga, Takanori Amano, Shiro Ohgo, Toshihiko Shiroishi, and Koji Tamura. 2011. '*Prx-1* Expression in *Xenopus laevis* Scarless Skin-Wound Healing and Its Resemblance to Epimorphic Regeneration'. *Journal of Investigative Dermatology* 131 (12): 2477–85.

Yokoyama, Hitoshi, Hajime Ogino, Cristi L. Stoick-Cooper, Rob M. Grainger, and Randall T. Moon. 2007. ‘Wnt/β-Catenin Signaling Has an Essential Role in the Initiation of Limb Regeneration’. *Developmental Biology* 306 (1): 170–78.

Yong, Luok Wen, Stéphanie Bertrand, Jr-Kai Yu, Hector Escriva, and Nicholas D. Holland. 2017. ‘Conservation of *BMP2/4* Expression Patterns within the Clade *Branchiostoma* (Amphioxus): Resolving Interspecific Discrepancies’. *Gene Expression Patterns* 25–26 (November): 71–75.

Young, Nathan M., Diane Hu, Alexis J. Lainoff, Francis J. Smith, Raul Diaz, Abigail S. Tucker, Paul A. Trainor, Richard A. Schneider, Benedikt Hallgrímsson, and Ralph S. Marcucio. 2014. ‘Embryonic Bauplans and the Developmental Origins of Facial Diversity and Constraint’. *Development* 141 (5): 1059–63.

Young, Teddy, Jennifer Elizabeth Rowland, Cesca van de Ven, Monika Bialecka, Ana Novoa, Marta Carapuco, Johan van Nes, *et al.*, 2009. ‘Cdx and Hox Genes Differentially Regulate Posterior Axial Growth in Mammalian Embryos’. *Developmental Cell* 17 (4): 516–26.

Yu, Jr-Kai, Daniel Meulemans, Sonja J. McKeown, and Marianne Bronner-Fraser. 2008. ‘Insights from the Amphioxus Genome on the Origin of Vertebrate Neural Crest’. *Genome Research* 18 (7): 1127–32.

Yu, Jr-Kai, Yutaka Satou, Nicholas D. Holland, Tadasu Shin-I, Yuji Kohara, Noriyuki Satoh, Marianne Bronner-Fraser, and Linda Z. Holland. 2007. ‘Axial Patterning in Cephalochordates and the Evolution of the Organizer’. *Nature* 445 (7128): 613–17.

Yue, Jia-Xing, Iryna Kozmikova, Hiroki Ono, Carlos W. Nossa, Zbynek Kozmik, Nicholas H. Putnam, Jr-Kai Yu, and Linda Z. Holland. 2016. ‘Conserved Noncoding Elements in the Most Distant Genera of Cephalochordates: The Goldilocks Principle’. *Genome Biology and Evolution* 8 (8): 2387–2405.

Yue, Jia-Xing, Jr-Kai Yu, Nicholas H. Putnam, and Linda Z. Holland. 2014. ‘The Transcriptome of an Amphioxus, *Asymmetron lucayanum*, from the Bahamas: A Window into Chordate Evolution’. *Genome Biology and Evolution* 6 (10): 2681–96.

Zakany, Jozsef, and Denis Duboule. 2007. ‘The Role of Hox Genes during Vertebrate Limb Development’. *Current Opinion in Genetics & Development*, Pattern formation and developmental mechanisms, 17 (4): 359–66.

Zakas, Christina, Nancy Schult, Damhnait McHugh, Kenneth L. Jones, and John P. Wares. 2012. ‘Transcriptome Analysis and SNP Development Can Resolve Population Differentiation of *Streblospio benedicti*, a Developmentally Dimorphic Marine Annelid’. *PLOS ONE* 7 (2): e31613.

Zalts, Harel, and Itai Yanai. 2017. ‘Developmental Constraints Shape the Evolution of the Nematode Mid-Developmental Transition’. *Nature Ecology & Evolution* 1 (5): 0113.

Zantke, Juliane, Stephanie Bannister, Vinoth Babu Veedin Rajan, Florian Raible, and Kristin Tessmar-Raible. 2014. ‘Genetic and Genomic Tools for the Marine Annelid *Platynereis dumerilii*’. *Genetics* 197 (1): 19–31.

Zapata, Felipe, Nerida G. Wilson, Mark Howison, Sónia C. S. Andrade, Katharina M. Jörger, Michael Schrödl, Freya E. Goetz, Gonzalo Giribet, and Casey W. Dunn. 2014. ‘Phylogenomic Analyses of Deep Gastropod Relationships Reject Orthogastropoda’. *Proc. R. Soc. B* 281 (1794): 20141739.

Zattara, Eduardo E., and Alexandra E. Bely. 2016. ‘Phylogenetic Distribution of Regeneration and Asexual Reproduction in Annelida: Regeneration Is Ancestral and Fission Evolves in Regenerative Clades’. *Invertebrate Biology* 135 (4): 400–414.

Zattara, Eduardo E., Kate W. Turlington, and Alexandra E. Bely. 2016. ‘Long-Term Time-Lapse Live Imaging Reveals Extensive Cell Migration during Annelid Regeneration’. *BMC Developmental Biology* 16 (1): 6.

Zeltser, L., C. Desplan, and N. Heintz. 1996. ‘Hoxb-13: A New Hox Gene in a Distant Region of the HOXB Cluster Maintains Colinearity’. *Development* 122 (8): 2475–84.

Zeng, An, Hua Li, Longhua Guo, Xin Gao, Sean McKinney, Yongfu Wang, Zulin Yu, *et al.*, 2018. ‘Prospectively Isolated Tetraspanin+ Neoblasts Are Adult Pluripotent Stem Cells Underlying Planaria Regeneration’. *Cell* 173 (7): 1593-1608.e20.

Zhang, Qiu-jin, Guang Li, Yi Sun, and Yi-quan Wang. 2009. ‘Chromosome Preparation and Preliminary Observation of Two Amphioxus Species in Xiamen: Chromosome Preparation and Preliminary Observation of Two Amphioxus Species in Xiamen’. *Zoological Research* 30 (2): 131–36.

Zhang, Qiu-Jin, Yi-Jyun Luo, Hui-Ru Wu, Yen-Ta Chen, and Jr-Kai Yu. 2013. ‘Expression of Germline Markers in Three Species of Amphioxus Supports a Preformation Mechanism of Germ Cell Development in Cephalochordates’. *EvoDevo* 4: 17.

Zhao, Mi, Maoxian He, Xiande Huang, and Qi Wang. 2014. ‘A Homeodomain Transcription Factor Gene, *PfMSX*, Activates Expression of Pif Gene in the Pearl Oyster *Pinctada fucata*’. *PLOS ONE* 9 (8): e103830.

Zhao, Mi, Yu Shi, Maoxian He, Xiande Huang, and Qi Wang. 2016. ‘PfSMAD4 Plays a Role in Biomineralization and Can Transduce Bone Morphogenetic Protein-2 Signals in the Pearl Oyster *Pinctada fucata*’. *BMC Developmental Biology* 16 (April): 9.

Zheng, Wei, Lisa M. Chung, and Hongyu Zhao. 2011. ‘Bias Detection and Correction in RNA-Sequencing Data’. *BMC Bioinformatics* 12 (1): 290.

Zhong, Ying-Fu, Thomas Butts, and Peter WH Holland. 2008. ‘HomeoDB: A Database of Homeobox Gene Diversity’. *Evolution & Development* 10 (5): 516–518.

Zhong, Ying-Fu, and Peter WH Holland. 2011. ‘HomeoDB2: Functional Expansion of a Comparative Homeobox Gene Database for Evolutionary Developmental Biology’. *Evolution & Development* 13 (6): 567–568.

Zhong, Ying-fu, and Peter WH Holland. 2011. ‘The Dynamics of Vertebrate Homeobox Gene Evolution: Gain and Loss of Genes in Mouse and Human Lineages’. *BMC Evolutionary Biology* 11 (June): 169.

Zhou, Xue, Ping Jin, Sheng Qin, Liming Chen, and Fei Ma. 2012. ‘Systematic Investigation of Amphioxus (*Branchiostoma floridae*) MicroRNAs’. *Gene* 508 (1): 110–16.

Zielins, Elizabeth R., Ryan C. Ransom, Tripp E. Leavitt, Michael T. Longaker, and Derrick C. Wan. 2016. ‘The Role of Stem Cells in Limb Regeneration’. *Organogenesis* 12 (1): 16–27.

Zwarycz, Allison S., Carlos W. Nossa, Nicholas H. Putnam, and Joseph F. Ryan. 2016. ‘Timing and Scope of Genomic Expansion within Annelida: Evidence from Homeoboxes in the Genome of the Earthworm *Eisenia fetida*’. *Genome Biology and Evolution* 8 (1): 271–81.

Appendices

All appendices associated with this thesis are distributed digitally. Chapters 1 and 6 do not have associated appendices; therefore, there are no appendices numbered 1.x or 6.x.

Appendices in Chapter 2

APPENDIX 2.1 .... List of the genomic, transcriptomic, miRNA, & CNE databases used

APPENDIX 2.2 .... Python and .sh scripts used

LARGE_TOPOLOGY_COLLIDER/ Scripts to map support vals between trees

microVISTA.py Script used to perform the analysis presented in Figure 5.4

PHYLIP/ Scripts used to reformat sequence names for running trees in PHYLIP

rename.sh Example of batch .sh file used to rename sequence names in tree files

APPENDIX 2.3 .... *in situ* protocols

ISH STOCKS.docx reagent recipes

WMISH amphioxus protocol.docx short form protocol

APPENDIX 2.4 .... Information for VISTA analysis (Figure 5.3)

APPENDIX_2.4.pdf scaffold curation details

Sc0000222-complete_mVISTA_annotations.txt B.lan scaffold annotation

Appendices in Chapter 3

APPENDIX 3.1 .... Work as submitted for publication in Genome Biology & Evolution

APPENDIX 3.2 .. *S.lam* transcriptome results + sequences in the trees (3.3) and alignments (3.4)

APPENDIX_3.2a.pdf summary of S.lam transcriptome results

APPENDIX_3.2b.xlsx full details of transcriptome + h'box search results

APPENDIX 3.3 .... NEWICK format consensus tree files of *S.lam* homeobox genes

APPENDIX 3.3a_all.tree all homeodomains (Figure 3.1)

APPENDIX_3.3b_hoxparahox.tree Hox & Parahox homeodomains (Figure 3.2)

APPENDIX_3.3c_TALE.tree TALE homeodomains inc. STE (Figure 3.4)

APPENDIX_3.3d_PRD.tree *PRD homeodomains inc. SPE (Figure 3.6)*

APPENDIX_3.3e_NK.tree *NK, Msx, Lbx, Tlx homeodomains (Figure 3.8)*

APPENDIX 3.4 ... FASTA format homeodomain alignments used to generate trees in (3.3)

APPENDIX_3.4a_all.fa *all homeodomains (Figure 3.2)*

APPENDIX_3.4b_hoxparahox.fa *Hox & Parahox homeodomains (Figure 3.2)*

APPENDIX_3.4c_TALE.fa *TALE homeodomains inc. STE (Figure 3.4)*

APPENDIX_3.4d_PRD.fa *PRD homeodomains inc. PRD-E (Figure 3.6)*

APPENDIX_3.4e_NK.fa *NK, Msx, Lbx, Tlx homeodomains (Figure 3.8)*

APPENDIX_3.4f_TALE-IV.fa *TALE-IV full alignment*

APPENDIX 3.5 .... Sequences in Paps *et al.*, (2015), Morino et al (2017) and present study

APPENDICES IN CHAPTER 4

APPENDIX 4.1 .... Identity of predicted homeobox transcripts in the *B.lan* genome

APPENDIX 4.2 .... Alignments used to establish the identities in (4.1)

APPENDICES IN CHAPTER 5

APPENDIX 5.1 .. NT + AA alignment of the PRD domains of the *Branchiostoma Pax3/7*s

APPENDIX 5.2 .... AA alignment of Pax3/7 sequences from all cephalochordates

APPENDIX 5.3 .... Chordate Pax3/7 exon boundaries

APPENDIX 5.4 .... Pax3/7 phylogeny files

APPENDIX_5.4a.pdf *accession numbers of sequences used*

APPENDIX_5.4b.fa *alignment used to generate phylogeny*

APPENDIX 5.5 .... Work as submitted for publication to Scientific Reports

APPENDIX_5.5.docx *main text*

FIG1_v5.pdf *vector format version of paper Figure 1*

FIG2.pdf *vector format version of paper Figure 2*

FIG3_rearrange_hires.pdf *vector format version of paper Figure 3*

Supplementary Files & Notes4.pdf *supplementary material of the paper*