THE EVOLUTION AND REGULATION OF THE CHORDATE PARAHOX CLUSTER

Myles Grant Garstang

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

2016

The Evolution and Regulation of the Chordate ParaHox Cluster

Myles Grant Garstang

University of
St Andrews

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

23rd October 2015

1. Candidate's declarations:

I, Myles Grant Garstang, hereby certify that this thesis, which is approximately 73,500 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 a higher degree.

I was admitted as a research student in October, 2011 and as a candidate for the degree of PhD in October, 2012; the higher study for which this is a record was carried out in the University of St Andrews between 2011 and 2015.

Date 22/10/2015 signature of candidate

Myles G. Garstang

2. 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 22/10/2015 signature of supervisor

Dr. David E.K. Ferrier

3. Permission for publication:

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

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

PRINTED COPY

b) Embargo on all or part of print copy for a period of 1 years on the following ground(s):

Supporting statement for printed embargo request:

Two of the thesis chapters are in preparation for publication and the embargo is requested in order for these manuscripts to be 'in press' prior to release of the thesis.

ELECTRONIC COPY

b) Embargo on all or part of electronic copy for a period of 1 years on the following ground(s):

Supporting statement for electronic embargo request:

Two of the thesis chapters are in preparation for publication and the embargo is requested in order for these manuscripts to be 'in press' prior to release of the thesis.

Date 22/10/2015 signature of candidate

Myles G. Garstang

signature of supervisor

Dr. David E.K. Ferrier

Abstract

The ParaHox cluster is the evolutionary sister of the Hox cluster. Like the Hox cluster, the ParaHox cluster is subject to complex regulatory phenomena such as collinearity. Despite the breakup of the ParaHox cluster within many animals, intact and collinear clusters have now been discovered within the chordate phyla in amphioxus and the vertebrates, and more recently within the hemichordates and echinoderms. The archetypal ParaHox cluster of amphioxus places it in a unique position in which to examine the regulatory mechanisms controlling ParaHox gene expression within the last common ancestor of chordates, and perhaps even the wider Deuterostomia. In this thesis, the genomic and regulatory landscape of the amphioxus ParaHox cluster is characterised in detail. New genomic and trascriptomic resources are used to better characterise the *B.floridae* ParaHox cluster and surrounding genomic region, and conserved non-coding regions and regulatory motifs are identified across the ParaHox cluster of three species of amphioxus. In conjunction with this, the impact of retrotransposition upon the ParaHox cluster is examined and analyses of transposable elements and the *AmphiSCP1* retrogene reveal that the ParaHox cluster may be more insulated from outside influence than previously thought. Finally, the detailed analyses of a regulatory element upstream of *AmphiGsx* reveals conserved mechanisms regulating Gsx CNS expression within the chordates, and TCF/Lef is likely a direct regulator of *AmphiGsx* within the CNS. The work in this thesis makes use of new genomic and transcriptomic resources available for amphioxus to better characterise the genomic and regulatory landscape of the amphioxus ParaHox cluster, serving as a basis for the improved identification and characterisation of functional regulatory elements and conserved regulatory mechanisms. This work also highlights the potential of *Ciona intestinalis* as a ‘living test tube’ to allow the detailed characterisation of amphioxus ParaHox regulatory elements.

Acknowledgements

First, and foremost, I would like to thank Dr. David Ferrier, who has been an excellent supervisor and has given me excellent Scientific advice. The often not-so-short discussions that arose from just ‘nipping to Dave’s office’ lead to the majority of the work here. Thank you for your seemingly endless patience, for teaching me how to be a Scientist, and for taking a chance on me in the first place.

I would like to thank the University of St Andrews School of Biology, and the BBSRC DTG scheme for providing the funding that allowed me to carry out this project.

To the past and current members of the EDGe and CoRD groups, thanks for the many tea breaks and Scientific (and non-scientific) discussions had over the past four years, and generally being my lab-partners in crime. In particular Olivia, Clara and Reka for making me feel so welcome when I first arrived and to Simon, for always being on-hand for a tea break or much needed trip to the pub. Also thanks to Dr. Ildiko Somorjai of CoRD, for always happily providing a second opinion and lending her expertise in amphioxus WMISH.

I would also like to thank Dr. Matthew Ronshaugen for first encouraging my interest in the Hox cluster and Evo-Devo, and Jerome, Janet and Tom of the Ronshaugen lab at the University of Manchester for helping me out so much during my time with that group.

Thanks to all of my family, who have always been there for me despite the hard last few years, with a special thanks to mum and dad, who have always encouraged me to pursue what I love. In particular I would like to mention my grandfather Mr. Raymond Duckworth who sadly passed away during my time in St Andrews. He took an active interest in my research and always surprised me with his knowledge despite not being an academic, and was always a voice of encouragement to me.

Last, but not at all least, I would like to thank my better half, Jessica, for standing by me and always supporting me when I needed it. Thank you especially for your patience in listening to me endlessly witter on about regulatory elements and electroporations, and for putting up with me whilst I have been writing this thesis.

Table of Contents

List of Figures

List of Tables

Chapter 1. General Introduction

1.1. Overview of ParaHox Expression

The ParaHox genes are thought to have been ancestrally involved in the anterior-posterior patterning of the central nervous system and gut within the last common ancestor of protostomes and deuterostomes (Annunziata et al., 2013; Arnone et al., 2006; Copf et al., 2003; Hui et al., 2009b; Ikuta et al., 2013; Moreno et al., 2011; Osborne et al., 2009; Schulz et al., 1998; Weiss et al., 1998; Wheeler et al., 2005; Wu and Lengyel, 1998) and also likely involved in the patterning of neurectodermal (Finnerty et al., 2003) and endodermal (Fortunato et al., 2014; Leininger et al., 2014) tissues much deeper in metazoan evolution. Much like the Hox genes, the ParaHox genes are expressed in a collinear fashion. This means that both Hox and ParaHox genes are expressed in the same order along the anterior-posterior axis of the embryo as their physical order along the chromosome (Duboule, 1994; Osborne et al., 2009; Wada et al., 1999). This archetypical ParaHox gene expression within the embryo is well studied within the chordates, with Gsx expression within the anterior CNS (Illes et al., 2009; Osborne et al., 2009), Xlox expression in the midgut (Ohlsson et al., 1993; Osborne et al., 2009), and Cdx expressed in the posterior tailbud, CNS and gut (Beck et al., 1995; Gamer and Wright, 1993; Meyer and Gruss, 1993; Osborne et al., 2009). This is perhaps best observed in the cephalochordate amphioxus where ParaHox expression is thought to be representative of that of the last common ancestor of chordates (Brooke et al., 1998; Osborne et al., 2009). In amphioxus, *Cdx*, the posterior ParaHox gene, is expressed first within the gastrula stage through to the larva in the tailbud, posterior neural tube and posterior endoderm. *Xlox*, the 'middle' ParaHox gene, is expressed next in the posterior endoderm (just anterior to *Cdx* expression) within the neurula, and also within a few cells at the level of the presumptive pigment spot within the neural tube. *Gsx*, the anterior ParaHox gene, is expressed last within a few cells within the presumptive hindbrain (at the level of somite 5) in the neurula stage (just adjacent to *Xlox* neural expression), then later within the cerebral vesical in the late-neurula/Premouth stage (midbrain/forebrain). Now, chordate-like colinear ParaHox expression has also been observed in both echinoderms (Annunziata et al., 2013) and hemichordates (Ikuta et al., 2013), raising the possibility of deeply conserved mechanisms regulating the ParaHox genes not only within the chordates, but also within the wider deuterostomes, and possibly Bilateria.

1.2. Evolution of the Hox/ParaHox clusters.

1.2.1. Origins of the Hox and ParaHox clusters.

The ParaHox genes are well known as the 'evolutionary sister' of the Hox genes and both Hox and ParaHox contain a 180bp Antennepedia-class (ANTP-class) homeobox domain, and group phylogenetically with each other (Brooke et al., 1998). There are three ParaHox groups that are found throughout the Metazoa and although the nomenclature used varies amongst different species, they are known as Gsx (also known as intermediate neuroblasts defective/Ind or Gsh), Xlox (or Xlhbox8, IPF1, PDX1, IDX1, STF1 or Lox3) and Cdx (or Caudal/Cad). These three ParaHox genes, Gsx, Xlox and Cdx, were originally believed to be 'orphaned' homeobox genes that had become separated from the Hox cluster (Li et al., 1996; Sharma et al., 1996) (reviewed in Ferrier and Holland (2001b) and Hui et al. (2012)). However, it was realised that these genes represented a separate, but evolutionarily related, sister cluster to the Hox cluster. The ParaHox genes were observed to be clustered together, much like the Hox cluster, within amphioxus, mouse and human (Brooke et al., 1998; Ferrier et al., 2005). Phylogenetic analysis of amphioxus Hox and ParaHox genes revealed that the ParaHox genes did not form a distinct branch when observed on a phylogenetic tree. Instead, the ParaHox genes were found to have greater sequence similarity within the different Hox 'groups' than they did with each other, with *Gsx* nested with the anterior Hox genes, *Xlox* with the Group 3 genes and *Cdx* with the posterior genes (Brooke et al., 1998). In addition to this, it was observed that the ParaHox cluster, like the Hox cluster, also exhibited a regulatory phenomenon known as collinearity. Two types of collinearity are observed within Hox and ParaHox clusters, Spatial and temporal, and both cause the genes within the Hox and ParaHox clusters to be expressed in the same order along the A/P axis of the embryo as their physical order along the chromosome. Spatial collinearity causes the first gene of the cluster to be expressed most anterior within the embryo and the last gene expressed most posterior. Along with spatial collinearity, temporal collinearity can also be observed in the chordate Hox and ParaHox clusters (Duboule, 1994; Osborne et al., 2009; Wada et al., 1999), however ParaHox temporal expression is in the reverse order to that of the Hox cluster, with the posterior gene (Cdx) expressed first and the anterior gene (Gsx) expressed last (Brooke et al., 1998; Osborne et al., 2009; Wada et al., 1999). Collinearity was first observed in *Drosophila* with the identification of the Hox complex and its role in A/P patterning (Lewis, 1978), and subsequently extended to the vertebrate Hox genes (Gaunt, 1988). Both Hox and ParaHox genes also exhibit collinearity in amphioxus (Brooke et al., 1998; Osborne et al., 2009; Wada et al., 1999), and these clusters are thought to be representative of the ancestral state of the chordate Hox/ParaHox clusters. This suggested that the ParaHox genes were not orphaned Hox genes as previously thought, and were instead a sister cluster to the Hox cluster. In fact the name 'ParaHox' implies just this, that they are 'paralogous to Hox'.

The distinct similarities between Hox and ParaHox genes have received much attention, and the two gene clusters are thought to have arisen via duplication of an ancient ProtoHox cluster. Several theories on the exact nature of ParaHox evolution and the composition of the ProtoHox cluster exist. Brooke *et al.* originally proposed the 'four gene' ProtoHox model in 1998, in which trans-duplication of a ProtoHox cluster resulted in the Hox and ParaHox clusters in the last common ancestor of bilaterians, or ‘Urbilaterian’. The ParaHox had then gone on to lose the 'central' group member, resulting in a three gene cluster, whilst the Hox cluster retained all four groups which then expanded. Other theories have proposed three gene (Finnerty and Martindale, 1999) and two gene ProtoHox cluster models (Garcia-Fernàndez, 2005). A single gene model has also been proposed, where tandem duplication of a ProtoHox-like gene formed the individual ParaHox and Hox gene classes, with separation of the Hox and ParaHox clusters occurring after these duplication events (Ryan et al., 2007). Currently the three and four gene models have the most support and have also been backed by statistical analysis comparing the likelihood of each model. This analysis rejected the single and two gene models, favouring a three or four gene ProtoHox cluster (Lanfear and Bromham, 2008).

The debate as to which ProtoHox evolutionary model is correct, is largely based around the Hox and ParaHox complements of different Phyla (see figure 1.1). Studies examining the Acoelomorpha may give hints as to the ancestral gene complement of the Hox and ParaHox clusters, and by inference the ProtoHox, and support a three-gene Hox/ParaHox complement basal to the Bilateria. It has been discovered that this lineage branched before the divergence of the three main bilaterian super-clades (the Deuterostomia, Lophotrochozoa and Ecdysozoa), placing them as the earliest offshoot of the Bilateria in most phylogenetic analyses (Baguna and Riutort, 2004). It should also be noted that one recent study suggests that the Acoelomorpha may instead be an early branching lineage of the deuterostomes, with the Acoelormorpha forming a monophyletic taxon, the Xenaceolomorpha, with the xenoturbellids (Philippe et al., 2011), though most evidence places them at the previously discussed position. The Hox/ParaHox complement of several species of acoel and one species of nermatodermatid from this lineage have been examined (Moreno et al., 2011), showing that three Hox genes; one anterior, one central and one posterior, appear basal to the Acoelomorpha. The ParaHox compliment is slightly less obvious, and only Cdx homologues have been found in acoels (Cook et al., 2004; Hejnol and Martindale, 2008), though both Xlox and Cdx are present in nermatodermatids (Jimenez-Guri et al., 2006). A lack of Xlox in acoels may be due to their derived gut, in which the gut epithelium has transformed into a syncytial tissue mass (Moreno et al., 2011). It is possible that all three ParaHox genes may be discovered in the Acoelomorpha, as there have been no whole genomes sequenced in this lineage as of yet, though efforts are currently being made (Perea-Atienza et al., 2015). Either way, these studies suggest the presence of an anterior, central and posterior Hox and ParaHox compliment basal to the Bilateria at least, perhaps pointing towards a three gene ancestral ProtoHox cluster. This would have consisted of an anterior, central and posterior gene, with the current group3 Hox representing the ancestral 'central' gene along with the ParaHox gene Xlox. The current 'central' Hox genes (non-group 3) instead may represent an ancient Hox-specific duplication creating the four classical bilaterian Hox groups. Alternatively, it is possible that the ParaHox cluster may have lost a fourth 'central' gene, and the gene compliment seen in the Acoelomorpha represents a lineage specific loss of Hox genes.

In order to gain a true understanding of what the ProtoHox complement of Metazoa looked like, it is also important to examine species belonging to the Cnidaria, the sister group to the Bilateria. The ParaHox complement of Cnidarians is less obvious than in the Bilateria, though studies using *Nematostella vectensis* have found that two ParaHox genes are present; Gsx and an interesting Xlox/Cdx hybrid gene (Chourrout et al., 2006). This holds interesting possibilities with regards to the Hox/ParaHox complement of the cnidarian-bilaterian last common ancestor (C-BLCA) as well as that of the Protohox complex, as there are several possible evolutionary models that could account for this. Possibilities include a two-gene C-BLCA ParaHox cluster, in which an ancestral Xlox/Cdx has undergone tandem duplication and subfunctionalisation in the Bilateria, or perhaps that Xlox and Cdx genes have undergone fusion via unequal crossover during recombination in the *Nematostella* lineage. Alternatively, it could be that Xlox has been lost, and other studies place this Xlox/Cdx gene as a Cdx orthologue only, resulting in a lack of Xlox in *Nematostella*. In other Cnidarian species, both Gsx and Cdx homologues have been identified (Chiori et al., 2009; Yanze et al., 2001), and one study also suggests the presence of an Xlox orthologue in hydrozoans (Quiquand et al., 2009). The potential discovery of Xlox in hydrozoans is a very important addition, as it places all three ParaHox genes present in the basal cnidarian, thus supporting the theory that all ParaHox genes were present in the C-BLCA, with gene losses occurring in different cnidarian lineages.

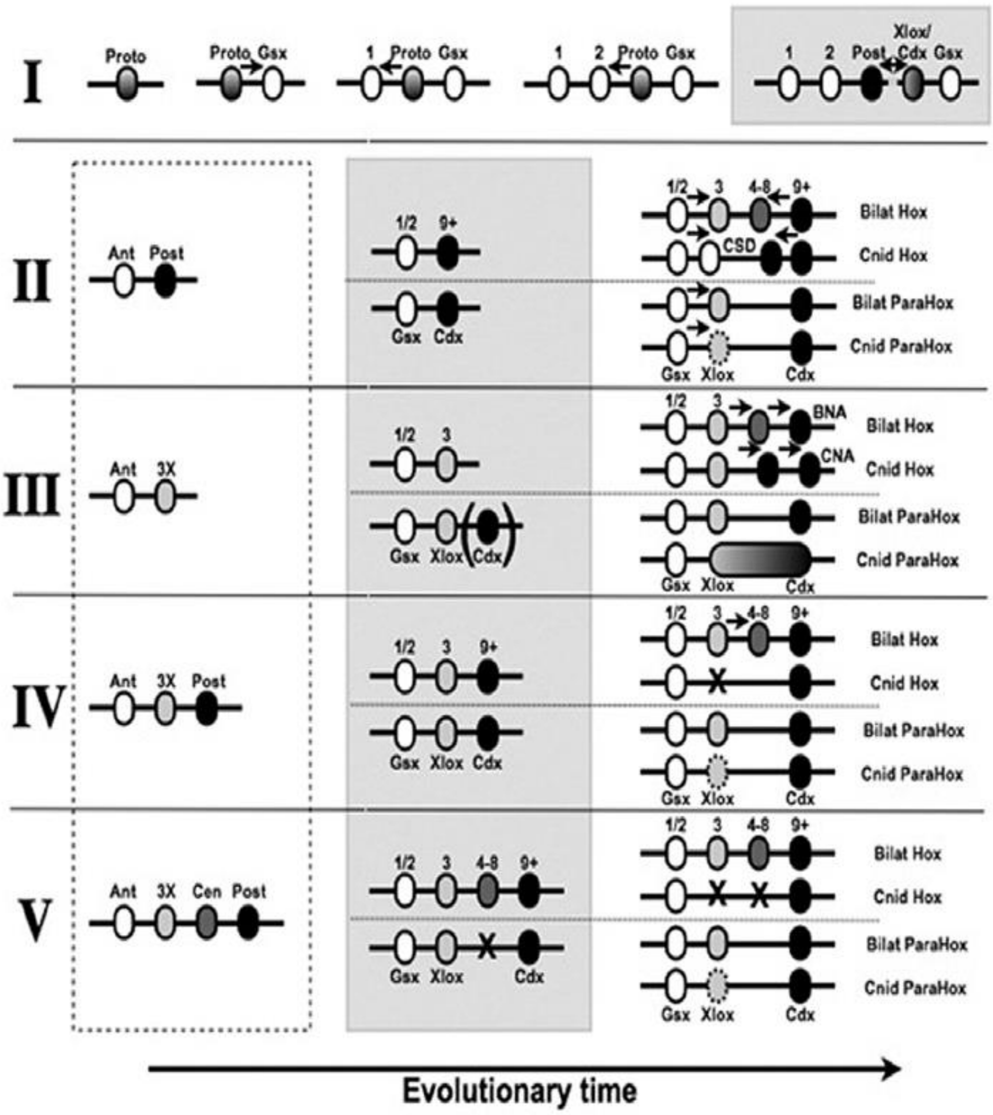


Figure 1.1. Summary of the alternative models proposed for the origin and evolution of the Hox (and ParaHox) clusters.

All except Model I invoke a ProtoHox cluster, the different hypothesized ProtoHox clusters being enclosed in the dashed box. For each model hypothesizing a ProtoHox cluster the evolution of the Hox clusters is given above the dotted line (bilaterian = 'Bilat Hox'; cnidarian = 'Cnid Hox'), whilst the evolution of the ParaHox clusters is below the dotted line (bilaterian = 'Bilat ParaHox'; cnidarian = 'Cnid ParaHox'). Evolutionary time progresses from left to right. Evolutionary time progresses from left to right. Model I: Tandem Duplication is adapted from Ryan et al (2007) and hypothesizes a ProtoHox gene ('Proto') that resides in an expanding gene cluster and repeatedly duplicates to produce the precursors for the different Hox and ParaHox gene families, finally evolving into the precursors for the Posterior Hox and Cdx genes before the Precursor cluster breaks into the Hox and ParaHox clusters (broken horizontal line). Models II and III are alternative versions of a 2-gene ProtoHox. Figure legend continued on next page.

Figure 1, legend continued from previous page. Model II: 2-gene A is adapted from (Garcia-Fernàndez, 2005) and requires extensive independent tandem duplications (denoted by small arrows) within the distinct Hox and ParaHox clusters after they have arisen from a ProtoHox cluster of two genes; one ProtoHox gene is the ancestor of Gsx and Hox1/2 (= 'Ant') whilst the second is the ancestor for Cdx and the Hox9+ genes (= 'Post'). Within the Hox clusters the cnidarian genes other than those orthologous with Hox1/2 and Hox9+ are independent duplications ('CSD' = Cnidarian Specific Duplications). The dotted boundary around the cnidarian Xlox gene in Models II, IV and V represents the fact that Xlox was thought to be absent from cnidarians at the time each model was originally proposed, but has now been shown to be present in some cnidarians (Quiquand et al., 2009). Model III: 2-gene B is adapted from (Chourrout et al., 2006) and hypothesizes a 2-gene ProtoHox cluster containing the ancestor of Hox3 and Xlox (= '3X') instead of the 'Post' ancestor of Model II. This model does not distinguish whether the ParaHox cluster of the Cnidarian-Bilaterian Ancestor (CBA) contained 2 genes (Gsx and Xlox) or 3 genes (Gsx, Xlox and Cdx) (denoted by the brackets around the CBA Cdx gene). In the latter case the present-day cnidarian ParaHox cluster (represented by *Nematostella vectensis*) has been reduced back to a 2-gene cluster, with a gene of indeterminate orthology between Xlox and Cdx (denoted by the stretched gene symbol). Extensive independent duplications are hypothesized for the generation of the Bilaterian Non-Anterior genes ('BNA') and the Cnidarian Non-Anterior genes ('CNA'). Model IV: 3-gene adapted from (Finnerty and Martindale, 1999) and (Ferrier and Holland, 2001a) in which the central Hox genes ('4-8') evolved within the bilaterian Hox cluster and the cnidarian lineage lost a Hox3 orthologue. Model V: 4-gene adapted from (Ferrier and Holland, 2001a) involves loss of Hox3 and Hox4-8 orthologues from the cnidarian Hox cluster and loss of a ParaHox gene paralogous to Hox4-8 in the CBA. The shaded box highlights the hypothesized organization of the Hox/ParaHox genes in the CBA for each model. Small arrows within clusters denote duplication events. Although these are given as arrows the actual direction of the duplication is often unknown (i.e., whether the central Hox might have duplicated from either a Posterior Hox or a Hox3 ancestor). Gene loss events are denoted as 'X' on the horizontal lines, which themselves denote the chromosome. Taken from (Ferrier, 2010)

The Placozoa have proven to be a rather interesting lineage, with regards to both their phylogenetic position within the Metazoa and their Hox/ParaHox compliment. The placozoan *Trichoplax* has only one Hox/ParaHox gene; *Trox-2* (Schierwater and Desalle, 2001). It has been suggested that *Trox-2* represents a putative 'ProtoHox' gene (Jakob et al., 2004; Schierwater et al., 2008), which was accompanied by the theory that the Placozoa represent a basal metazoan lineage (Dellaporta et al., 2006). Others, however, interpreted the phylogenetic analyses as showing *Trox-2* as being a homologue of the ParaHox gene Gsx (Martinelli and Spring, 2004), thus suggesting wide-scale secondary loss of Hox-like genes from *Trichoplax* (Peterson and Sperling, 2007). This prompted Mendivil-Ramos and colleagues to use syntenic analysis of neighbouring genes to show that *Trox-2* in fact lies within a ParaHox locus (Mendivil Ramos et al., 2012). With the sequencing of the *Trichoplax* genome, phylogenies now strongly place the Placozoa as a sister group to the Eumetazoa (cnidarians and bilaterians), with Poriferans basal to the Placozoa (Srivastava et al., 2008).

Current evidence supports an even more ancient origin of the ProtoHox duplication, and has changed the view that distinct Hox and ParaHox clusters must have been a Eumetazoan innovation. In fact, it is now clear that separate Hox and ParaHox clusters present before the divergence of the Porifera even. The identification of 'Ghost' Hox and ParaHox loci in the demosponge *Amphimedon* gave the first hint that these genes may be much more ancient than first thought (Mendivil Ramos et al., 2012), and Syntenic analyses of these regions showed that although both Hox and ParaHox clusters had been lost, genomic loci existed containing genes that are found neighbouring the Hox and ParaHox clusters in many other species. This led to the hypothesis that although the Hox and ParaHox genes themselves were not present in the *Amphimedon* genome, two genomic loci were present that signified a secondary loss of Hox and ParaHox genes rather than their evolution after the divergence of the porifera. Confirmation of this hypothesis has since been realised in the recent analyses of the genomes of *Sycon ciliatum* and *Leucosolenia complicata*, calcisponges that have the ParaHox gene Cdx (Fortunato et al., 2014), firmly placing the origin of the ParaHox genes at prior to the divergence of the Porifera and Eumetazoa. Even more recently, analyses of the first Ctenophoran (Comb Jelly) genome has suggested that *Cdx* may also exist within this enigmatic phylum (Moroz et al., 2014), though further analyses are required. This is particularly intriguing in light of the most recent phylogenetic analysis of the Metazoa. This study, carried out by Whelan and colleagues, represents the most thorough phylogenetic analyses of groupings within the Metazoa yet, and robustly places the Ctenophora basal to the Porifera as the most basal metazoan phylum (Whelan et al., 2015). Hopefully, future analyses of this potential ctenophoran Cdx gene will help to resolve whether the divergence of the Hox and ParaHox genes from ProtoHox occurred after the origin of the Ctenophora, or instead before the last common ancestor of all animals (reviewed in Ferrier (in revision, 2015)).

Finally, when considering the evolution of the ParaHox cluster, it is important to take into account the evolution of the entire ANTP-class of homebox genes, as evidence points towards a common ancestor for the Extended Hox (Hox, Evx, Meox), EHGbox (Gbx, En, Mnx), ParaHox, and NK-like families. The terms 'Hox-linked' and 'NK-linked' have since been adopted as more informative terms representative of chromosome location and linkage patterns and will be used henceforth in this thesis (Hui et al., 2012). The 'mega-cluster' hypothesis suggests that all ANTP-class homeobox genes existed as one huge gene cluster in the metazoan ancestor. The first data supporting this focussed upon the human ANTP-class genes, where ancestral linkage between Hox-linked and NK genes was proposed (Pollard and Holland, 2000). This has since been extended to encompass the ProtoHox hypothesis and the genesis of the ANTP-class homeobox mega-cluster and its evolution from a single ProtoANTP gene into the different ANTP-class homeobox clusters seen today (Garcia-

Fernandez, 2005). A further study went on to examine the ancestral state of all of the ANTP-class genes in the protostome-deuterostome ancestor (PDA) using macrosyntenic analysis, providing an interesting view on the mega-cluster hypothesis (Hui et al., 2012). It was observed that whilst intra-chromosomal rearrangement of a cluster are fairly common, inter-chromosomal rearrangements between the four mega-cluster groups were much rarer, and that only three breaks between the ParaHox, Hox-linked, NK-linked and NK2 clusters seen in *Platynereis* and chordates had occurred since the mega-cluster, if the mega cluster ever did exist. Whether the mega-cluster hypothesis holds true or not, this study does provide strong evidence contrary to the idea that ANTP-class genes may have secondarily come together from scattered locations in an ancestral genome.

This has led to a model where cis-duplication of a ProtoANTP gene resulted in a ProtoHox-like and Proto-NK gene. Further cis-duplication would then have given rise to the ProtoHox, Evx/Meox, EHGbox and ProtoNK genes. Expansion via cis-duplication within each of these gene sub classes would have resulted in the 3 or 4 gene ProtoHox cluster, Evx/Meox, EHGbox and Nk gene families. A further cis-duplication of the ProtoHox+Evx/Meox cluster would then have resulted in a single mega cluster containing ParaHox+Meox, Hox+Evx, EHGbox, and NK genes. Finally, three breakages of this mega-cluster in the metazoan ancestor, or perhaps PDA, resulted in the modern ParaHox, Extended Hox, and NK regions. An alternative model (figure 1.2.) proposed by Hui et al (2012) is more conservative with regards to classification of ANTP-families, highlighting the lack of robust resolution between many of these ANTP-class gene families, in particular the relationship of Dlx to the Hox and NK families (Hui et al., 2012), and represents the current view of the mega-cluster hypothesis.

It is highly likely that the ANTP mega-cluster had broken apart before the last common ancestor of protostomes and deuterostomes, but not much can be said for the integrity of the mega-cluster in the C-BLCA or at the base of the Metazoa. However, it can now be said that the ProtoHox duplication, and NK duplications all occurred prior to the Metazoa, as evidence from the diploblast Cnidarians shows that Hox, Hox-linked, ParaHox and NK-linked genes were all present in the C-BLCA (Finnerty et al., 2004; Gauchat et al., 2000; Ryan et al., 2006), and at least NK and ParaHox genes before the divergence of the Porifera (Fortunato et al., 2014). With the presence of the ParaHox gene Cdx in the Porifera, it can be inferred that the Hox genes must also have been present in the metazoan ancestor, albeit lost in the poriferans examined so far. Any further insights into the ProtoHox and mega-cluster hypotheses will likely come from analysing the genomes of further taxa from basal lineages such as the Porifera and Ctenophora.

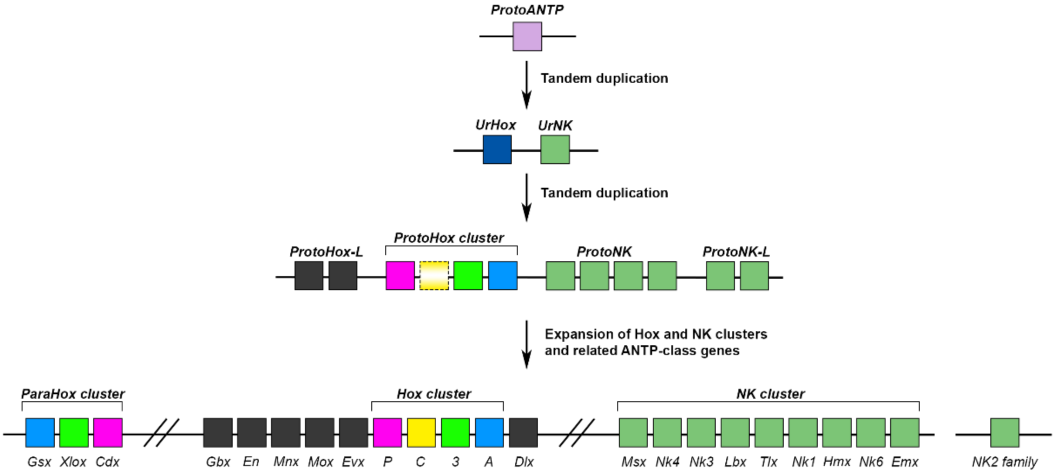


Figure 1.2. ProtoHox and ANTP-class Mega-cluster evolution

A schematic showing the evolution of all Antennepedia-class homeobox genes from a single ProtoANTP gene, and subsequent divergence and duplication of the different ANTP-class genes, resulting in a single 'mega-cluster' in the protostome-deuterostome last common ancestor (though the cluster breakup may well be more ancient). Duplication of a ProtoHox cluster has resulted in separate Hox and ParaHox clusters, which may or may not have been linked in the mega-cluster. NK genes are in green and represent the large scale duplication from a common ancestral UrNK gene. The Hox-linked genes are all coloured black, showing the disputable phylogeny of 'EHG-box', 'Evx/Mox' and Dlx groups, with relative position to other ANTP-class genes providing a better proxy for ancestry. The ProtoHox, Hox and ParaHox genes are coloured according to A/P expression and phylogenetic grouping. Magenta= posterior, Green=group3, Yellow=central and Blue=anterior. The dashed yellow of the ProtoHox represents the lack of resolution between the three and four cluster ProtoHox hypothesis. Double diagonal lines represent the breaks within the chordate ancestor. Arrows represent the expansion of ANTP-class families, though the order of duplications is unknown. Precise order of genes within the mega-cluster are not known except in the Hox, ParaHox and NK clusters. Figure adapted from (Garcia-Fernàndez, 2005) and (Hui et al., 2012).

1.2.2. ParaHox gene clustering: Is temporal collinearity the Key?

The ancestral archetypical structure of the ParaHox cluster is a three gene cluster with Gsx at one end, Xlox in the middle, and Cdx at the other end (Ferrier et al., 2005). This was first observed within the chordates, but with the discovery of intact, three gene archetypal ParaHox clusters within the Ambulacraria this condition was present at least in the deuterstome ancestor. However, within the vast majority of species examined so far, including some chordate lineages, the ParaHox cluster has either broken apart or has lost gene members. This leads us to question why these genes are clustered in some animals, yet not in others, and what sort of mechanisms are involved in cluster maintenance.

It would seem from the Hox and ParaHox complements of vertebrates that there is perhaps a lower selective pressure upon the maintenance of the ParaHox cluster than the Hox cluster. Genome duplications are one event that can cause difficulty in resolving mechanisms surrounding cluster maintenance. Within the vertebrates, two rounds of whole genome duplication have occurred during the early evolution of the vertebrates, resulting in the complement of four intact Hox clusters but only one intact ParaHox cluster, albeit four ParaHox loci, in most modern vertebrates supported by many studies highlighting other genes such as the globins (Hoffmann et al., 2012) and the major histocompatibility complex (MHC) (Abi-Rached et al., 2002). The key supporting evidence though for the 2R hypothesis is the quadruple conserved synteny of vertebrate chromosomal regions to amphioxus scaffolds (Putnam et al., 2008).

Strangely, the teleost fish were found to have seven Hox clusters and it is thought that a further WGD has occurred in this lineage, possibly contributing to the huge success and diversification of the teleosts (Amores et al., 1998; Jaillon et al., 2004; Meyer and Van de Peer, 2005). These whole genome duplication events have been termed the 2R and 3R hypotheses respectively. It is important to note that whilst mammalian genomes contain one intact ParaHox cluster, the further duplication event in the teleost lineage has led to disruption of this cluster, and no intact ParaHox cluster can be found within teleosts (Mulley et al., 2006; Siegel et al., 2007).

One hypothesis is that these genome duplications, particularly the additional 3R duplication of teleosts, may have led to the dispersal of regulatory elements across the different ParaHox loci of vertebrates thus overcoming regulatory-based constraints and facilitating cluster break-up. One key mechanism seen in the Hox complex that may maintain gene clustering is the use of shared regulatory elements, especially enhancers, which regulate the expression of two or more genes in their immediate vicinity. Many shared enhancers have been identified, with examples found to regulate the expression boundaries between Hoxb3-4 and Hoxb4-5 in both murine and zebrafish models (Gould et al., 1997; Hadrys et al., 2006; Sharpe et al., 1998). Such elements may provide a molecular anchor for the genes that they regulate, preventing gene dispersal by tying the regulation of their target genes to one location. Whilst enhancers offer a clear mechanism of shared gene regulation, insulators could also play a role in cluster maintenance. These regions act to block the activity of other enhancers, effectively shielding gene promoters that lie beyond the enhancer (Gaszner and Felsenfeld, 2006). Insulators are found in both protostome and deuterostome Hox clusters, allowing independent gene regulation within the cluster (Belozerov et al., 2003) as well as preventing the effects of enhancers located outside the cluster (Kmita et al., 2002). In light of this, it could be that insulators also act as genetic 'anchors' within a cluster, for if a gene were to disperse from its cluster it may be exposed to foreign enhancer activity producing aberrant expression patterns. In addition, the presence of genomic regulatory blocks, built up of a series of cis-regulatory elements interspersed between and within gene neighbours (Kikuta et al., 2007), provide some evidence that these types of regulatory elements are present within the ParaHox cluster, providing constraints based on regulatory mechanisms. It is possible that the loss of, dispersal, or even absence of such shared regulatory mechanisms within the ParaHox cluster could be a driving factor in the breakup of many ParaHox clusters.

In some cases, the disintegration of Hox/ParaHox clusters can be attributed to a derived development, particularly in species that have related lineages whose gene clusters remain intact. This can be observed in the urochordates, or tunicates, the evolutionary sister lineage to the vertebrates (Delsuc et al., 2006), and is evident in *Ciona*, where the residual collinearity seen in the Hox cluster of *C.intestinalis* is thought to be representative of a cluster undergoing dispersal (Ikuta et al., 2004). Likewise, the ParaHox cluster of *Ciona* has also partially disintegrated along with a loss of collinearity (Ferrier and Holland, 2002). The larvacean *Oikopleura dioica* goes even further in its cluster degradation, having lost all of its central genes and possessing a fully fragmented Hox compliment, with its nine Hox genes located at nine different places in the genome (Seo et al., 2004). Interestingly, these Oikopleura Hox genes appear to have lost temporal, but not spatial collinearity, despite break-up of the cluster. *Oikopleura* exhibits a very fast development even for a tunicate and seems to have evolved to retain its larval form. This process has led to the compaction of its genome, such that it has the smallest chordate genome, being only 60-70Mb in size, yet still contains 15,000 genes (Seo et al., 2001). The ParaHox complement of *Oikopleura* has been degraded, and it contains only Cdx in its genome, though this has undergone tandem duplication and there are three *Oikopleura Cdx* genes (Edvardsen et al., 2005), which coincides with its highly derived development and gene loss. The breakdown of key regulatory mechanisms such as collinearity in the early evolution of the tunicates, may have helped facilitate this further, drastic genome compaction and gene cluster disintegration seen in *Oikopleura*.

The echinoderms are another deuterostome lineage which have evolved a derived development, such that the adults exhibit pentaradial symmetry, though their larvae do retain bilateral symmetry. The sea urchin *Strongylocentrotus purpuratus* has provided a useful developmental model and representative for the echinoderm lineage, and the Hox and ParaHox genes have been studied extensively in this organism. With the loss of a clear A/P axis in adult echinoderms, it might be expected that the Hox genes would be dispersed as in the tunicates, but this is not the case. Whilst we do not see Hox cluster dispersal, there is a large inversion in the cluster, with anterior genes first, then posterior and central in reverse order, with the Hox 11/13 located next to Hox 3. Several other inversions of individual Hox genes in respect to their immediate neighbours are also present, such as in Hox5 and Hox 11/13b (Cameron et al., 2006). The ParaHox genes, however, have not been retained in a cluster and are seen to be dispersed amongst three large scaffolds (each >300kb) and no gene clustering was seen, though it is unknown if these genes reside on the same chromosomes. Though there appears to be no tight linkage between the *S.purpuratus* ParaHox genes, a level of spatial collinearity was still observed in the developing embryo and larval stages (Arnone et al., 2006). Recently, intact collinear ParaHox clusters have been discovered within other members of the Ambulacraria, in both the sea star *Patiria miniata* (Annunziata et al., 2013) and the hemichordate *Ptychodera flava* (Ikuta et al., 2013). Most intriguingly, *P.flava* maintains an intact ParaHox cluster despite the lack of spatial collinearity in the embryo. Temporal collinearity is, however, maintained. Another intact echinoderm ParaHox cluster has also been identified in the sea star *Acanthaster planci*, though the gene expression has not yet been examined. This now makes it highly likely that the last common ancestor of deuterostomes possessed intact, collinear Hox and ParaHox clusters, given the presence of intact collinear ParaHox clusters within amphioxus (Brooke et al., 1998; Osborne et al., 2009), *P.miniata* (Annunziata et al., 2013) and *P.flava* (Ikuta et al., 2013). This improved sampling of phyla within the Ambulacraria has shown how important proper taxon sampling is in resolving evolutionary questions, as previously knowledge of both intact ParaHox clusters, and those displaying collinearity were limited to the chordates.

This clustering of ParaHox genes within both the asteroids and hemichordates, but not the echinoids raises an interesting evolutionary observation regarding larval development and the maintenance of intact ParaHox clusters. As all of the echinoderms possess a derived pentaradial adult form, the general modification of development does not necessarily lead to the break-up of ParaHox clusters. Instead, it may be that the modification of early embryonic development is where selective pressure upon ParaHox clustering may lie. The bipinnaria larvae of asteroids and auricularia larvae of holithuroids both bear similarities in morphology to the tornaria larvae of hemichordates, particularly in the placement of the ciliary bands, suggesting a bipinnaria-tornaria-like, or di-pleurula, larval state at the base of the Ambulacraria. Within the auricularia this is a single ciliary band that loops around the body in a manner identical to the hemichordate tornaria, whilst in bipinnaria larva there are two ciliary bands that loop around the body. In contrast, the pleuteus larvae of echinoids and ophiuroids have ciliary bands that run up the contours of the larval arms, and each arm is supported by a calcitic skeletal rod (echinoderm larval morphology reviewed in Raff and Byrne (2006)). Within the most basal group of echinoderms, the stalked crinoids, the identification of an auricularia-like stage, with a single ciliary band in the same topology as the hemichordate tornaria larvae, supports this ancestral di-pleurula hypothesis (Nakano et al., 2003)(reviewed in Lacalli

(2003)). In addition, the morphological analysis of serotonergic apical organs across the Ambulacraria also suggests that these di-pleurula larvae are more representative of the ancestral di-pleurula than the pleuteus (Byrne et al., 2007). The modification of this ancestral di-pleurula stage to the pleuteus larvae of echinoids could perhaps then have led to the loss of temporal collinearity and breakup of the ParaHox cluster in S.purpuratus.

Though the ParaHox genes have dispersed in many lineages, and collinearity is lost, examples of partial clusters can be found outside of the deuterostomes. Within the lophotrochozoa, *Platynereis dumerilii* shows clustering of Gsx and Xlox, with *Cdx* still linked but at the other end of the chromosome (Hui et al., 2009b), and both *Platynereis* and *Capitella telata* Xlox and Cdx are not clustered, but are linked, albeit with a large distance and many other genes interspersed between the two (Hui et al., 2009b). An example of a partially intact ParaHox cluster can even be found within the cnidarian *Nematostella vectensis*, where Gsx and a single Xlox/Cdx gene are found clustered together (Chourrout et al., 2006; Hui et al., 2008), suggesting that an intact gene cluster was likely the ancestral state in the C-BLCA. The close association of intact Hox and ParaHox clusters with temporal collinearity, but not necessarily spatial collinearity , has led to the hypothesis that temporal collinearity may in fact be key in maintaining an intact ParaHox cluster, and that the breakdown of this regulatory mechanism may facilitate cluster dispersal (Arnone et al., 2006; Ikuta et al., 2013; Seo et al., 2004; Tschopp et al., 2009) (reviewed in Ferrier and Minguillon, (2003) and Garstang and Ferrier, (2013)).

The breakup of Hox and ParaHox clusters in lineages such as the tunicates and echinoids is also associated with the breakdown of collinearity, and could perhaps even be linked to the modified larval development of these lineages. Interestingly, there are no examples so far of disintegrated or partially dispersed ParaHox clusters that still exhibit temporal collinearity (figure 1.3). Conversely, all intact ParaHox clusters so far examined also display temporal collinearity, lending support to the hypothesis that temporal collinearity is key to maintaining an intact ParaHox cluster.

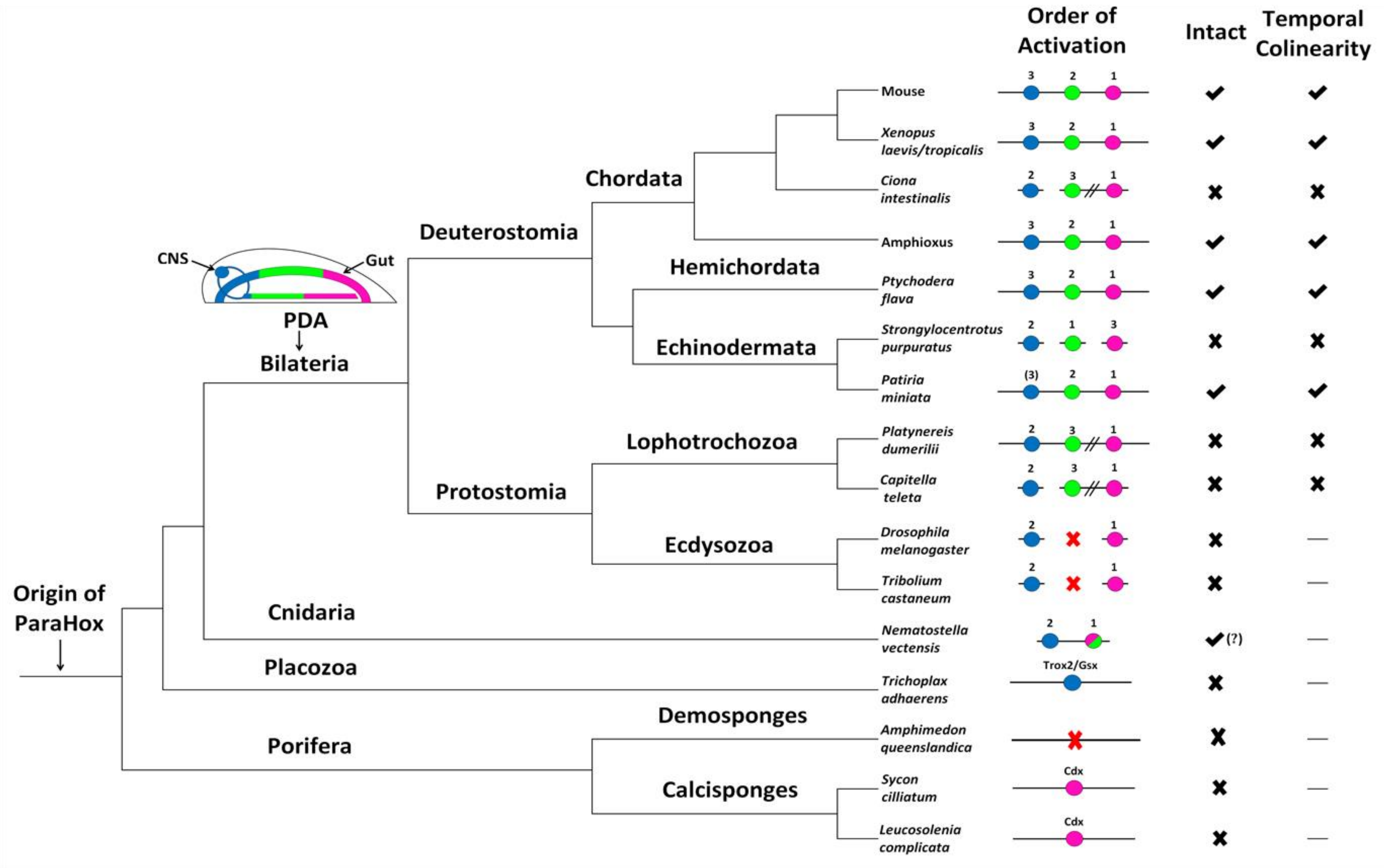

Order of Activation
Intact
Temporal Colinearity
Origin of ParaHox
CNS
Gut
PDA
Bilateria
Deuterostomia
Chordata
Hemichordata
Echinodermata
Protostomia
Lophotrochozoa
Ecdysozoa
Cnidaria
Placozoa
Porifera
Demosponges
Calcisponges
Mouse 3 2 1 ✓ ✓
Xenopus laevis/tropicalis 3 2 1 ✓ ✓
Ciona intestinalis 2 3 1 ✗ ✗
Amphioxus 3 2 1 ✓ ✓
Ptychodera flava 3 2 1 ✓ ✓
Strongylocentrotus purpuratus 2 1 3 ✗ ✗
Patiria miniata (3) 2 1 ✓ ✓
Platynereis dumerilii 2 3 1 ✗ ✗
Capitella teleta 2 3 1 ✗ ✗
Drosophila melanogaster 2 ✗ 1 ✗ —
Tribolium castaneum 2 ✗ 1 ✗ —
Nematostella vectensis 2 1 ✓(?) —
Trichoplax adhaerens Trox2/Gsx ✗ —
Amphimedon queenslandica ✗ ✗ —
Sycon cilliatum Cdx ✗ —
Leucosolenia complicata Cdx ✗ —

Figure 1.3. Animal phylogeny with the correlation between ParaHox cluster integrity and temporal collinearity indicated.

The ParaHox cluster originated before the divergence of the Porifera. The protostome-deuterostome ancestor (PDA) had ParaHox expression in the gut and CNS; with Gsx (blue) anterior, Xlox (green) central and Cdx (pink) posterior (note, this is purely schematic and not intended to illustrate specific morphology or precise expression domain boundaries). Genomic organisation of ParaHox genes for each species is shown, with gene linkage represented by a continuous line connecting individual genes. Double diagonals represent genes located on the same chromosome but separated by large distances, and the inclusion of a red 'X' indicates loss of one or more ParaHox genes. The *Nematostella* cluster has only 2 ParaHox genes, though it is unresolved whether one of these is a Cdx or Xlox homologue and a third gene has been lost relative to other cnidarians (hence the question mark). The order in which ParaHox genes are activated and expressed has been indicated numerically (*Patiria* Gsx activation in parentheses due to presumed later larval expression). The presence of an intact cluster or temporal collinearity is indicated by a check or cross. A horizontal line indicates that temporal collinearity cannot be resolved due to the absence of one or more ParaHox genes. (adapted from (Garstang and Ferrier, 2013)

1.3. The ParaHox genes are ancestrally expressed within the gut and CNS.

The ParaHox genes are known for their role in anterior-posterior (A/P) patterning, and it is thought that the ParaHox genes of Bilateria were ancestrally expressed within the gut and CNS (endoderm and neurectoderm) within the PDA (Annunziata et al., 2013; Arnone et al., 2006; Copf et al., 2003; Hui et al., 2009b; Ikuta et al., 2013; Moreno et al., 2011; Osborne et al., 2009; Schulz et al., 1998; Weiss et al., 1998; Wheeler et al., 2005; Wu and Lengyel, 1998). Here, the expression patterns, and functional studies of ParaHox genes throughout the Metazoa are discussed, highlighting the conserved aspects of expression between Phyla.

1.3.1. Porifera

So far, only one ParaHox gene has been identified within the Porifera, in the calcisponges *Sycon cilliatum* and *Leucosolenia complicata* (Fortunato et al., 2014). This study identified Cdx expressed within the inner cell mass during formation of the choanocyte chamber. This may hold particular relevance to the expression of the ParaHox genes within the endoderm of bilaterians, as recent molecular studies have revived the hypothesis that there is homology between the sponge choanoderm and the bilaterian endoderm (Leininger et al., 2014).

1.3.2. Placozoa

A single ParaHox gene has also been described within the enigmatic Placozoa, within *Trichoplax adherans*, *Trox-2*, which may be a Gsx homologue (Martinelli and Spring, 2004) (though the expression is not reminiscent of Gsx within other phyla). *Trox-2* is expressed in a ring around the periphery of Trichoplax, in small cells located between the upper and lower epithelial cell layers (Jakob et al., 2004). Trichoplax contains a single ParaHox gene *Gsx*, Experiments have been carried out to address the functionality of *Trox-2*, and RNAi or morpholino knockdown of *Trox-2* was shown to cause complete cessation of growth and of binary fission (the *Trichoplax* reproductive mode), suggesting that *Trox-2* function has been greatly modified from ancestral Gsx function. This may, however, be relevant in light of the modern placula hypothesis. Here, *Trichoplax* represents a proxy for a stage in the evolution of animals that has not yet gastrulated/invaginated to form the mouth and archenteron (Osigus et al., 2013; Schierwater et al., 2009). In this case, the entire edge of Trichoplax is potentially homologous to the Eumetazoan structures of the hypostome of cnidarians and the blastopore edge of bilaterians, where Gsx is seen to be expressed (Arnone et al., 2006; Finnerty et al., 2003; Ikuta et al., 2013). It should be noted here that Placozoa are now strongly placed as a sister group to the Cnidaria and Bilateria (Srivastava et al., 2008; Whelan et al., 2015), not a basal metazoan lineage, and the modern placula hypothesis and relation to *Trichoplax Trox-2* expression is not a view widely held.

1.3.3. Cnidaria

Whilst the cnidarians are important for understanding the evolution of ParaHox genes, ParaHox expression in Cnidaria varies greatly between species and the differences in development and morphology make comparisons to the Bilateria difficult. As such, no clear consensus has yet been reached regarding the ancestral role of Hox or ParaHox genes in the Cnidaria, or how this compares to the function of these genes within bilaterians (Chiori et al., 2009; Finnerty et al., 2003; Kamm et al., 2006; Ryan et al., 2007). Despite this, the presence, expression and genomic structure of ParaHox genes within the Cnidaria is useful for piecing together the ancestral state of the ParaHox cluster. When discussing Cnidarian embryology, it is important to remember the anterior/aboral and posterior/oral poles of the embryo, as the two terms for each are used to describe different stages of the embryo, and are used interchangeably across different studies. The anterior/aboral end eventually becomes the ‘foot’ and body, whilst the posterior/oral end becomes the head and tentacles. *Nematostella Anthox2* (Gsx) has been shown to be expressed within the posterior endoderm of the planula larva, before appearing within the presumptive tentacles and tentacle buds within the late larva. In the two-bud larva *Anthox2* is expressed within the oral endoderm, ectoderm and throughout the tentacle buds (Finnerty et al., 2003). This contrasts greatly with *Cnox-2* (Gsx) expression within Hydroids. Both the colonial hydroid *Hydractinia symbiolongicarpus*, and solitary hydroid *Hydra Vulgaris* display *Cnox-2* expression within cell nuclei in the cell body epithelium, at high levels in the aboral body and foot, but at low levels within the head (oral end) (Cartwright et al., 1999; Shenk et al., 1993a; Shenk et al., 1993b). Within the hydrozoan *Podocoryne carnea*, *Gsx* expression begins within the anterior endoderm of 15-34h embryos. Within the 2-day larva, strong anterior endodermal expression is observed, though this extends more weakly to the posterior pole (Yanze et al., 2001). In the coral *Acropora millepora Cnox-2Am* is expressed throughout the ectoderm of early-tentacle-stage embryos, though is excluded from the aboral pole (Hayward et al., 2001). Finally, within the hydrozoan *Clytia hemisphaerica*, *Gsx Ch* expression begins within the posterior/oral endoderm in the late gastrula stage, then extends within aborally within the planula to cover the whole endoderm (Quiquand et al., 2009). Only one gene with clear similarity to Xlox alone has been found within Cnidarians, and *Clytia Pdx Ch* is first expressed one hour after *Gsx Ch* expression, within 1 day old planula in a spotty pattern that extends through both ectoderm and endoderm, but is excluded from the aboral/anterior pole. Within 3 day old planula, expression is restricted to the endoderm but is diffuse (Quiquand et al., 2009). Finally, Cdx expression has also been observed in a few cnidarians. In the Hydrozoan *Clytia CheCdx* expression is expressed maternally, then at the late gastrula stage transcripts are observed throughout the ectoderm of the embryo except within the oral pole. In the 1-day planula, this expression is restricted to the aboral and oral poles within the ectoderm, and then later within the maturing oocytes of the female gonads and within tentacle bulbs (Chiori et al., 2009). Within *Podocoryne*, *Cnox-4* (Cdx) expression is located to the posterior throughout embryonic development (Yanze et al., 2001). Expression of *Cnox-4* within the hydromedusa *Eleutheria dichotoma* is present in the aboral end of the embryo (Kamm et al., 2006). Thus, it is clear from these expression patterns that no clear conserved expression of ParaHox genes is present throughout the Cnidaria.

1.3.4. Bilateria

1.3.4.1. Cdx

Cdx, or *Caudal* was first identified in *Drosophila* and is expressed at the posterior, hence the name *Caudal (Cad)*. Throughout the Ecdysozoa, caudal is involved in the specification of posterior regions and in the formation of the hindgut (Macdonald and Struhl, 1986; Wu and Lengyel, 1998), also showing a similar expression pattern in the short germ arthropods such as the flour beetle *Tribolium castaneum* (Schulz et al., 1998) and the brine shrimp *Artemia franciscana* (Copf et al., 2003). This holds true throughout the Arthropoda and Cdx has been examined in a wide variety of arthropods including; the wasp *Nasonia vitripennis* (Olesnicky et al., 2006), the intermediate germband cricket *Gryllus bimaculata* (Shinmyo et al., 2005), the silkworm *Bombyx mori* (Xu et al., 1994), the grasshopper *Schistocerca gregaria* (Dearden and Akam, 2001), the barnacle *Sacculina* *carcini* (Rabet et al., 2001), the spider *Parasteatoda tepidariorum* (formerly *Achaearanea tepidariorum*)(Akiyama-Oda and Oda, 2003) and the centipede *Strigamia maritima* (Chipman et al., 2004). Perhaps the only example which would need further examination to confirm Cdx expression is in the crayfish *Procambarus clarkia*, where only late embryonic expression has been examined (Abzhanov and Kaufman, 2000). Still, it remains that caudal expression is conserved within posterior segment patterning, particularly in the posterior growth zone, posterior ventral nerve cord, and hindgut. Outside of the arthropods, in the nematode *C.elegans*, the Cdx homologue *pal-1* also appears to play a similar role and is again required for embryonic posterior patterning (Edgar et al., 2001).

In the second super-phylum of the protostomes, the Lophotrochozoa, Cdx expression is much more variable, but conserved roles can still be identified. The annelids are perhaps the most well studied lophotrochozoan lineage, though also the most variable in regards to Cdx expression. In *Capitella teleta*, *Cdx* expression can be seen within the posterior within the presumptive gut as well as in the neurectoderm. However, *Cdx*, is also expressed within the mesoderm, anterior gut and anterior nervous system. It is perhaps only in the late larvae and post-metamorphosis stages where we see more canonical Cdx expression in an expansive region of the hindgut, but also the anterior gut and the ventral nerve cord (Frobius and Seaver, 2006). *Capitella* expression is thus intriguing in that it spans the entirety of the A/P axis, rather than being confined to the posterior. Within polychaetes with a more typical trochophore of stage development (eg, *Platynereis dumerilii)*, Cdx expression is much more reminiscent of that seen in the ecdysozoans, but there are still some divergent aspects to the Cdx expression pattern. *Cdx* expression in *P.dumerilii* begins in the posterior proctodeal ectoderm, as well as in the anterior ectoderm. Some mesendodermal expression can also be seen, with perhaps intermittent staining within the stomadeum (de Rosa et al., 2005; Hui et al., 2009b). Expression becomes restricted to the more canonical hindgut/midgut expression by the 2-day trochophore stage (Hui et al., 2009b). This larval expression is very similar to the expression of Cdx seen in another trochophore-developing polychaete, *Alitta virens* (previously known as *Nereis virens*). In *Alitta Cdx* expression is more typical of the 'canonical' Cdx expression again, and is observed in the proctodeum at 40 hours post fertilisation (hpf), before being restricted later to the hindgut and pygidium. Expression of *Cdx* can also briefly be observed between 100 and 112/115 hpf in the posterior ventral nervous system (Kulakova et al., 2008). The A/P gradient of *Cdx* expression observed in the gut and posterior growth zone of the larvae is also recapitulated in posterior regeneration, and is highly reminiscent of Cdx expression in other phyla during their embryogensis. A final annelid, the oligochaete *Tubifex tubifex* has also been examined with regards to *Cdx* expression, though it is even more highly derived than that of *Capitella*, with expression beginning in the anterior-most part of the mesodermal germ bands, and then expanding posteriorly as development proceeds (Matsuo et al., 2005). The authors here acknowledge the derived expression of *Tubifex* Cdx, and suggest that Cdx expression may have been highly modified within the oligochaete lineage.

In another major lophotrochozoan group, the Mollusca, there are also examples of Cdx expression. In *Patella vulgata Cdx* expression begins in the posterior ectoderm and mesodermal cells, later becoming expressed in the ectoderm of the prototroch, two central, internal clusters of mesodermal cells and two large domains covering the posterior neurectoderm (Le Gouar et al., 2003). Another gastropod, *Gibbula varia* has also been examined. In the trochophore, *Caudal* expression begins at 12 and 18hpf in the posterior neurectoderm, as well as a pair of bilateral cells in the interior of the larvae, much as in *Patella*. In the pretorsional veliger larvae, *Caudal* can be observed within the whole area of the nascent digestive gland, with higher expression in a few cells of the dorsal visceral mass. After torsion, at ~60hpf, strong expression of *Caudal* is observed in the hindgut and rectum, and also weakly throughout the digestive gland (Samadi and Steiner, 2010). The expression within the posterior neurectoderm of the trochophore, as well as within the hindgut and rectum of the post-torsional larvae are reminiscent of the 'canonical' conserved Cdx expression and certainly points to conserved expression of Cdx in all posterior tissues, but particularly the posterior hindgut and nervous system, in the last common ancestor of protostomes.

Moving to the deuterostomes, the Ambulacraria are one group that has only recently begun to be properly characterised with regards to ParaHox expression. Within the echinoderms, Cdx expression can be observed in the hindgut in the gastrula to pleuteus larvae of the sea urchin *Strongylocentrotus purpuratus* (Arnone et al., 2006; Cole et al., 2009). Likewise in the sea star *Patiria miniata* expression is seen within the hindgut throughout the gastrula to late brachiolaria larvae (Annunziata et al., 2013). Differing from the sea urchin, *P.miniata*, also displays expression in the 24hpf blastula stage, within a ring of cells in the vegetal half of the embryo surrounding the blastopore, perhaps representing expression that is more in line with the early recruitment of Cdx to the primitive streak of vertebrates, as discussed later. Within the second ambulacrarian phylum, the hemichordates, *Cdx* expression is much as observed for the echinoderms within the sole species examined so far, *Ptychodera flava*, with expression beginning in the early gastrula and persisting in the hindgut up to the tornaria larva, with anterior expression marking the midgut-hindgut boundary (Ikuta et al., 2013).

Chordate Cdx has been studied in much detail, particularly within the vertebrates, with similar expression throughout the chordate phylum. The cephalochordate amphioxus is thought to have archetypal ParaHox expression, and is thought to represent the ancestral expression patterns of the ParaHox genes within chordates. As such, amphioxus *Cdx* is expressed within the posterior tailbud, extending both dorsally into the posterior neural tube and ventrally into the hindgut (Brooke et al., 1998). Expression begins in the gastrula, in a ring surrounding the blastopore, before extending into the posterior of the forming hindgut and neural tube in the neurula through to larval stages, with both expression domains meeting in the far posterior as a continuous domain of expression expressed in all germ layers (Osborne et al., 2009). *C*dx expression in the tunicates follows a similar pattern, despite them being a derived chordate lineage. *Cdx* expression is more extensive than within other chordates, being expressed throughout the tail within the posterior epidermis, which is unusual for a chordate, the nerve cord and also endodermal strand from the late gastrula onwards in *Ciona intestinalis.* However, expression is notably absent from the posterior tip of the tailbud (Hudson et al., 2007; Ikuta et al., 2010; Kusakabe et al., 2002) (Osborne et al. 2009, Unpublished data). Another ascidian, *Halocynthia roretzi*, displays Cdx expression in a very similar pattern to that of *Ciona*, though it is absent from the epidermis. Experiments injecting dominant negative forms of *Hr-Caudal* mRNA displayed severe posterior truncations and abnormalities, in both *Halocynthia* and also when injected into *Xenopus* embryos, showing the conserved function of Cdx between tunicates and vertebrates (Katsuyama et al., 1999). Cdx has not been studied post-metamorphosis in these species, and as such the canonical hindgut expression has not been observed. There is no functional gut within the larval tunicate, with the undifferentiated endodermal strand going on to form the adult gut during metamorphosis. However, post-metamorphosis Cdx expression has been studied in another tunicate, *Herdmania curvata*. Within the juvenile, *Cdx* expression is localised into two domains, one in the extreme posterior of the gut just anterior to the anus, and a second domain at the stomach-intestine junction (Hinman et al., 2000). This expression is much as expected for canonical *Cdx* expression, with the second more anterior domain representing the midgut-hindgut boundary.

Within the vertebrates, Cdx has been thoroughly examined in its role in posterior patterning, though it is necessary to take into account the multiple homologues when discussing ancestral expression patterns. In the mouse (*Mus musculus*) embryo, expression of all three Cdx genes is activated in the gastrula (Beck et al., 1995; Gamer and Wright, 1993; Meyer and Gruss, 1993) and they go on to be expressed within the posterior of the embryo and the primitive streak within the ectoderm and mesoderm (reviewed in Young and Deschamps (2009)). Interestingly, *Mmu-Cdx2* is also activated within extraembryonic tissues and the placenta (Beck et al., 1995), suggesting a role for Cdx in the development of these mammalian extraembryonic tissues. Expression of mouse Cdx genes in the posterior of the embryo begins with *Mmu-Cdx1*, which has the most rostral anterior limit, with *Mmu-Cdx2* second, followed by *Mmu-Cdx4* with the most caudal anterior limit. *Mmu-*

Cdx1 remains expressed in the posterior neural tube, somatic mesoderm and limb buds (Meyer and Gruss, 1993), whilst both *Mmu-Cdx2* and *MmuCdx-4* remain expressed in the posterior neural tube, presomitic and lateral plate mesoderm (Beck et al., 1995; Gamer and Wright, 1993). Within the posterior endoderm, *Mmu-Cdx2* and *Mmu-Cdx4* are both expressed first at day 8.5, and *Mmu-Cdx1* expression beginning six days later. In the posterior endoderm, expression of *Mmu-Cdx2* is the most anterior whilst *Mmu-Cdx4* is the most posterior. Expression of Cdx genes within the chick, *Gallus gallus,* is very similar to that of mouse, with all three Cdx genes showing expression along the A/P axis within the primitive streak (Marom et al., 1997) and then later in the posterior within all three germ layers, though *CdxB* (Cdx4) is downregulated within the gut (Ehrman and Yutzey, 2001). These patterns of Cdx expression are present throughout the vertebrates, with both the frog *Xenopus tropicalis* (Chalmers et al., 2000; Reece-Hoyes et al., 2002), and the teleost fish *Danio rerio* (zebrafish) again displaying expression of Cdx genes from gastrulation onwards in the posterior of the embryo within all three germ layers, particularly in the posterior neural tube, tailbud, developing gut and the posterior somatic mesoderm. It is noteworthy that the anterior Cdx expression boundary within the *Xenopus* gut terminates at the stomach-intestine boundary in a manner comparable to the invertebrate deuterstomes.

Finally, in vertebrates, several loss-of-function mutant studies and knock-down studies have been carried out to test the function of Cdx genes. The most drastic of these is within *Mmu-Cdx2* homozygous null mutants, which are non-viable as they fail to implant, highlighting the extraembryonic role of *Mmu-Cdx2*. Heterozygous Cdx2 mutants were observed to have shortened tails and intestinal tumours (Chawengsaksophak et al., 1997), in addition to abnormalities of first to third thoracic vertebrae, which are also seen in *Mmu-Cdx1*$^{-/-}$ mutants (Subramanian et al., 1995). Though *Mmu-Cdx1*$^{-/-}$ mutants do not show intestinal defects, there are changes in the expression of *Mmu-Cdx2* (Bonhomme et al., 2008). Interestingly, *Mmu-Cdx4*$^{-/-}$ mutants do not seem to have any effect alone (van Nes et al., 2006). As expected, double Cdx mutants show a much more severe phenotype than single gene mutations, highlighting the redundancy between Cdx paralogues (van den Akker et al., 2002; van Nes et al., 2006). Overexpression studies of the three Cdx genes resulted in anterior to posterior skeletal transformations in all three cases, with *Mmu-Cdx1* producing the most anterior defects, and *Mmu-Cdx2* and *Mmu-Cdx4* producing more posterior defects (Gaunt et al., 2008), perhaps reflecting the different anterior expression boundaries of these genes. In addition, limb defects (*Mmu-Cdx1)* and severe tail defects such as kinks and even splits (*Mmu-Cdx2*) were also observed. Though not examined, it is possible that this effect upon the vertebral segments could be linked to the action of Cdx activating posterior Hox expression (Davidson and Zon, 2006; Deschamps and van de Ven, 2012; Ehrman and Yutzey, 2001; Isaacs et al., 1998) (reviewed in Young and Deschamps (2009)). These studies in mouse have also been backed up by functional studies in *Xenopus*, which showed a similar array of defects caused by Cdx mutants (Chalmers et al., 2000; Isaacs et al., 1998). Finally, morpholino (MO) and knockout experiments in zebrafish have provided further functional data within the teleosts. Though *Dre-Cdx1a* MO knockdowns had no effect, *Dre-Cdx4* null mutants, or MO experiments showed a shortened A/P axis. Again, double knockout/knockdown experiments highlighted the redundancy between these genes, with double *Dre-Cdx4* null mutant/ *Dre-Cdx1a* knockdown zebrafish exhibiting severe posterior truncations (Davidson and Zon, 2006; Shimizu et al., 2005). Knockdown of *Dre-Cdx1b* (Cdx2), which is expressed ubiquitously before becoming restricted to the gut, inhibits intestine differentiation (Flores et al., 2008). Finally, overexpression of *Dre-Cdx4* in the hindbrain induces the expression of spinal cord specific genes, and loss of *Dre-Cdx4 and Dre-Cdx1a* function causes the hindbrain domain of the neural tube to expand. In addition, the loss of Cdx expression causes the caudal neural plate to become responsive to hindbrain signals such as retinoic acid, suggesting that Cdx genes are key to specifying the size of the prospective hindbrain and spinal cord territories. (Skromne et al., 2007).

As an overview of bilaterians, posterior expression of Cdx is conserved throughout the Bilateria with a conserved role in posterior patterning, especially the specification of the posterior endoderm (hindgut). In addition, Cdx expression also appears to be conserved within the neurectoderm, and also the posterior growth zone in all germ layers (as observed between chordates and protostomes). The expression of *Cdx* in the calcisponge choanoderm (Fortunato et al., 2014) suggests that a role for Cdx within the patterning of the endoderm/gut was present in the ancestor of all Metazoa.

1.3.4.2. Xlox

Xlox is the middle ParaHox gene, and has a conserved role in specification of the midgut and neurectoderm. There are currently no studies of Xlox in the Ecdysozoa as the gene has been lost from all ecdysozoan species currently examined, though there is uncertainty about whether Xlox has been lost from *Strigamia* due to the uncertain orthology of the two Strigamia '*Hox3*' genes (Chipman et al., 2014). Thus it remains to be seen whether one of these *Strigamia* Hox3 copies may represent the presence of Xlox within this basal ecdysozoan group. However, multiple studies of lophotrochozoan Xlox (in combination with its presence in deuterostomes) show that Xlox must have been present before divergence of the two protostome lineages. In the leech, *Hirudo medicinalis,* three tandemly duplicated Xlox genes form the Lox3-C cluster, and at least two of these *Lox3* genes are expressed in the dorsal midgut endoderm in 12 transverse stripes at the E10-E15 stage (Wysocka-Diller et al., 1995). These stripes indicate the future crop within the leech midgut, with the first stripe indicating the pharynx-crop boundary. There is also faint expression within the four intestinal segments abutting the crop. In another leech, *Helobdella triserialis*, *Lox3* is also expressed in transverse stripes within the crop in the midgut endoderm, though these are much weaker than the expression within the abutting intestinal segments (Wedeen and Shankland, 1997). Within the polychaetes, *Capitella Xlox* is found only in endodermal cells that form the midgut. The *Capitella* gut forms late in development from endodermal cells that are scattered within the yolk, and the patchy appearance of *Xlox* expression may reflect scattered undifferentiated endodermal cells, because when the midgut epithelium is fully formed *Xlox* expression is no longer detected. This late and derived development of the *Capitella* gut may also explain why *Xlox* is the last *Capitella* ParaHox gene to be transcribed (Fröbius and Seaver, 2006). In *Nereis*, expression begins weakly in the posterior endoderm, but becomes stronger and restricted to the posterior midgut. There is also segmental expression in the ventral nervous system as well as in two lobes of the brain in the nectochaete stage (Kulakova et al., 2008). Finally, *Platynereis Xlox* expression begins in the ventral plate (neurectoderm) by 50hpf, and persists weakly until later stages (Hui et al., 2009b). At 72hpf, expression is initiated within the midgut rudiment, before becoming confined to two distinct cell clusters in the anterior and posterior midgut at 5 days. At this stage, expression is also observed in a bilateral pair of lobes within the brain. In contrast to the diversity of annelid species examined, only a single species of mollusc has been assayed for Xlox expression. In the mollusc *Gibbula varia*. expression of *Xlox* begins at 24hpf in a group of cells in the ventral hyposphere and a pair of symmetrical domains within the medio-ventral episphere in the trochophore. Additionally, *Xlox* is also expressed in 8-9 cells in a semicircle within cells of the ventral neurectoderm around the area where the anus will later open. In the pre-torsional veliger larva, *Xlox* is expressed within the forming digestive gland, as well as five ectodermally derived cells of the ventral nervous system on the right side of the larva. In the post-torsional larva, *Xlox* is expressed in the digestive gland, as well as part of the visceral mass (Samadi and Steiner, 2010). These expression patterns within the lophotrochozoans suggest an ancestral role for Xlox in the patterning of the midgut and neurectoderm at the base of the protostomes. It would be interesting to examine more basal ecdysozoans, for example the further examination of *Strigamia Hox3* genes, to examine if Xlox is retained at all within this group.

Within the Ambulacraria Xlox has remarkably well conserved expression between the echinoderms and hemichordates. In the echinoid *Strongylocentrotus Xlox* is expressed from the late gastrula to pluteus larvae within the developing midgut. Expression begins throughout the posterior gut, but is then restricted to the posterior mid-gut (Annunziata and Arnone, 2014; Arnone et al., 2006; Cole et al., 2009). In addition to this, individual cells within the oral ectoderm, in line with the mid-gut, also begin expressing *Xlox*. These go on to form a subset of neurectodermal cells within the ciliary band of the pluteus (Cole and Arnone, 2009). The expression of *Xlox* within the asteroid *Patiria*, is very similar to that of *Strongylocentrotus*. *Xlox* expression is first observed within the oral ectoderm of the mid-gastrula embryo at 48hpf, followed by an additional domain within the posterior archenteron at 52hpf. This archenteron domain goes on to be restricted to the midgut-hindgut boundary by 72hpf. The ectodermal domain of Xlox expression goes on to form two large bilaterally symmetrical domains of expression within the post-oral ciliary band within the bipinnaria larva (Annunziata et al., 2013). Within the second ambulacrarian phylum, the hemichordates, *Ptychodera* again displays a very similar pattern of *Xlox* expression to the echinoderms (Ikuta et al., 2013).

Within the Chordata, the cephalochordate amphioxus has Xlox expression in a domain anterior to Cdx within the posterior gut (perhaps midgut, see later discussion on the Xlox/Cdx gut boundary) (Brooke et al., 1998). Subsequent work identified that *Xlox* expression begins in the ventral posterior archenteron after gastrulation, before becoming stronger in both the ventral and dorsal posterior endoderm (Osborne et al., 2009). A further domain arises in the posterior, linking these domains during neurulation, and expression is seen within the posterior neural tube and dorsal posterior mesendoderm. The ventral endoderm domain becomes restricted to the midgut-hindgut boundary during the late neurula to larvae stages, whilst the posterior domain becomes downregulated during the late neurula, disappearing by the premouth stage. In addition to these domains, a neural tube specific domain starts at the level of the presumptive pigment spot, arises during the early neurula and persists until the late neurula stage, disappearing by the premouth stage (Osborne et al., 2009). *Ciona Xlox* (*Ci-IPF-1*) expression has been examined in larvae, where it is detected in the sensory vesicle, visceral ganglion and some mesenchymal cells (Corrado et al., 2001). No endodermal expression has been identified for *Ciona* Xlox, though the midgut where Xlox is typically expressed in other taxa does not form until the post-metamorphosis juvenile in *Ciona*, and *Ciona* Xlox has not yet been examined within post-larval stages.

Within the vertebrates, Xlox has been well characterised in its role in patterning the pancreas and the regulation of insulin production (Ohlsson et al., 1993). Mouse *Pdx1* is initially expressed within the dorsal and ventral endoderm of the developing gut, before becoming restricted to the developing buds of the pancreas and the duodenum, which lies between the dorsal and ventral buds of the pancreas. (Ohlsson et al., 1993). Similar expression can be observed in other mammals, within human and rat, though there is also expression within discrete neural cells in the forebrain and hindbrain of the rat (Perez-Villamil et al., 1999; Phan-Hug et al., 2008). This expression within the pancreatic and duodenal endoderm also holds true for the chick, though expression is also seen within the posterior stomach (Kim et al., 1997; Kumar et al., 2003). Much the same can be seen within *Xenopus* and zebrafish and Xlox can be seen within the development of the pancreatic analgan and duodenum of *Xenopus* (Wright et al., 1989), and also within pancreas development in zebrafish (Milewski et al., 1998). Several studies have investigated the function of Xlox in the vertebrates, largely related to its role in pancreatic development and function. In the mouse *Mmu-Pdx1* is required for the transcriptional control of insulin and islet amyloid peptide within pancreatic β and δ cells (Brink, 2003). Mice with mutant *Xlox* also fail to develop a pancreas, and the duodenum does not properly form or differentiate (Jonsson et al., 1994; Offield et al., 1996). Similar results were found from the MO knockdown of Xlox in zebrafish, where the pancreas is specified but no differentiation and outgrowth occurs (Yee et al., 2001).

It would seem from both protostome and deuterostome studies that Xlox has maintained a conserved role within the specification of the midgut, though this has been expanded to include the more specific midgut organs, i.e. the pancreas and duodenum, within the vertebrates. Additionally, the expression of Xlox within restricted domains of neural cells within all the phyla examined suggests that Xlox may also ancestrally be involved in the patterning of the CNS, perhaps in a restricted cell type rather than specifying a large domain.

1.3.4.3. The Xlox-Cdx Midgut-Hindgut boundary

One interesting conserved pattern and potential interaction between the ParaHox genes Cdx and Xlox is the role of Xlox and Cdx in the formation of the midgut/hindgut boundary. Examples of this interaction are most well characterised within the echinoderms, where it has been examined in detail within *S.purpuratus* (Annunziata and Arnone, 2014; Arnone et al., 2006; Cole et al., 2009), but also within *P.miniata* (Annunziata et al., 2013). In these echinoderms, both Xlox and Cdx are expressed within the posterior archenteron and posterior gut in the late gastrula, with Xlox having the more anterior limit. These expression domains then become mutually exclusive within the pluteus so that Xlox forms a restricted band covering the midgut-hindgut boundary and Cdx within the rest of the hindgut. A similar pattern of Xlox-Cdx expression within the midgut and hindgut has been observed within the transforming larvae of the hemichordate *P.flava* (Ikuta et al., 2013) where the domains of Xlox (midgut) and Cdx (hindgut) are adjacent. Within the chordates, amphioxus also exhibits these Xlox and Cdx expression domains present at the midgut-hindgut boundary (Osborne et al., 2009), and they begin overlapping then become restricted to being adjacent, perhaps signifying and Xlox/Cdx midgut/hindgut boundary within the simple gut of amphioxus. Within the mouse, the domains of Xlox and Cdx1/2 display a small overlap(if any) in the midgut-hindgut (Fang et al., 2006), and *Cdx2* restricts *Pdx1* expression within the mouse gut (Grainger et al., 2010). This highly conserved expression, observed throughout the deuterostomes suggests a conserved mechanism of midgut-hindgut patterning, mediated by an Xlox/Cdx expression boundary, at the base of the Deuterostomia. The expression of *Platynereis Xlox* and *Cdx* is also suggestive of a conserved midgut/hindgut boundary controlled by the restriction of *Xlox* and *Cdx* expression within the PDA (Hui et al., 2009b).

1.3.4.4. Gsx

Gsx is the most anteriorly expressed of the ParaHox genes and plays a role in the patterning of neural tissues, though potentially evolutionarily relevant exceptions do exist. This is evident in protostomes, where Gsx is expressed in the anterior neuroectoderm and mediolateral stripes of *Drosophila* (Urbach et al., 2006; Weiss et al., 1998), *Tribolium* (Wheeler et al., 2005), and the anterior neurectoderm of the annelids *Capitella* (Fröbius and Seaver, 2006), *Nereis* (Kulakova et al., 2008) and *Platynereis* (Denes et al., 2007; Hui et al., 2009b), and the mollusc *Gibbula* (Samadi and Steiner, 2010). What is perhaps more intriguing is the expression of Gsx in the stomodeum of the lophotrochozoans *Nereis, Platynereis* and *Gibbula* (Hui et al., 2009b; Kulakova et al., 2008; Samadi and Steiner, 2010). There is debate over whether this structure is homologous to the deuterostome mouth (Christiaen et al., 2007; Hui et al., 2009b), though the lack of Gsx expression in the mouth of deuterostomes supports the theory that the mouth has arisen as a secondary innovation in the Deuterostomia. In addition to the stomodeum expression of *Gibbula Gsx*, the pre-torsional larva expresses Gsx around the mouth opening and within the digestive gland. This digestive gland expression fades within the post-torsional larvae, being replaced by a domain within the foregut (Samadi and Steiner, 2010). This Gsx expression, particularly in the stomodeum and endoderm of *Platynereis* (Hui et al., 2009b), and the stomodeum, mouth opening and foregut of *Gibbula* (Samadi and Steiner, 2010), may represent the ancestral anterior 'gut' expression of Gsx, at least within the protostomes. A wider range of phyla, particularly within the Ecdysozoa would need to be sampled in order to further investigate this.

Within the Ambulacraria, Gsx is solely restricted to neural expression. In the echinoid *Strongylocentrotus, Gsx* is expressed from the gastrula (first detectable at 24hpf) through to pluteus larvae within two bilaterally symmetrical patches of neurectodermal cells at the level of the midgut (Arnone et al., 2006). No embryonic Gsx expression was observed in the asteroid *Patiria* from 24h to 5-day bipinnaria larva, though expression may occur within later larvae or the adult CNS (Annunziata et al., 2013). Strangely, within the hemichordate *Ptychodera*, *Gsx* expression was detected within a few cells surrounding the blastopore within the late gastrula, but had disappeared by the tornaria larval stage and could not be detected within the transforming larva either (Ikuta et al., 2013). The similar expression of these cells to the Gsx positive cells of *S.purpuratus* (Cole and Arnone, 2009), and position within the ciliary band at later stages make it likely that these cells are also neurectodermal, though these Ptychodera Gsx positive cells have not yet been examined as thoroughly as in *S.purpuratus* to confirm this. This leaves an interesting situation within the Ambulacraria, where the most divergent species morphologically (the echinoids), may in fact display the most ancestral-like Gsx spatial expression, though further studies are clearly needed within a wider range of Ambulacraria.

Within the chordates, amphioxus *Gsx* is first seen within a four cells of the neural tube at the level of somite five, or the presumptive hindbrain, within the mid-neurula stage, overlapping with *Xlox* expression, though this disappears by the premouth stage (Osborne et al., 2009). A second 'late' domain is seen, arising within the premouth stage within the cerebral vesicle on both the right and left sides and remains until the early larval stages (Brooke et al., 1998; Osborne et al., 2009). Within the tunicate *Ciona, Gsx* is first expressed in the mid-gastrula in cells of the neural plate destined to become the anterior and posterior sensory vesicle, where expression is detected up to the late tailbud stage (Hudson and Lemaire, 2001)(also see section 5.3.11 for a comprehensive description of *Ci-Gsx* embryonic expression).

In the vertebrates, Gsx (Gsh) expression is observed throughout the CNS. In mouse, *Mmu Gsh1* expression begins within the hindbrain, but later expands throughout the entire spinal cord. In later stages, expression domains are also observed within the midbrain and forebrain (Valerius et al., 1995). The second Gsx paralogue, *Mmu-Gsh2* is expressed later than *Mmu-Gsh1*, but can be observed within the spinal cord, hindbrain, midbrain and forebrain, again beginning within the hindbrain (Hsiehli et al., 1995). Within the chick, whole embryo expression has not been examined, but expression of chick *Gsh2* has been observed in the developing forebrain, in a domain homologous to the telencephalic domain observed in mice (Von Frowein et al., 2002). Expression within *Xenopus* is again similar to that observed within mouse, with *Gsh1* and *Gsh2* expression beginning within the hindbrain, with expression soon following within the spinal cord and forebrain. Interestingly, *Xenopus Gsh2* is the only vertebrate Gsx gene to show expression within the endoderm, where it is expressed throughout the lateral endoderm in the middle of the embryo, but not the anterior of posterior poles, in stages 21 through to 25 (Illes et al., 2009). Within the teleosts, both medaka and zebrafish *Gsh1* expression patterns have been published. The two species are almost identical and very similar to the expression seen within the other vertebrates, with *Gsh1* expression beginning within the hindbrain, then later encompassing the hindbrain, spinal cord, forebrain and midbrain (Cheesman and Eisen, 2004; Deschet et al., 1998).

As with the other ParaHox genes, several vertebrate studies have also described Gsx mutants and knockdowns in order to assess the function of Gsx genes. Within the mouse, *Gsh2* homozygous mutants display abnormal development of the lateral ganglionic eminence within the forebrain, as well as a delay in the appearance of interneurons (Corbin et al., 2000; Toresson et al., 2000). Though single *Gsh1* mutants display no obvious phenotypes, double *Gsh1/Gsh2* mutants highlight the partial redundancy between the Gsh paralogues and display a much more severe phenotype than either single gene knockout. Severe defects within the striatum and olfactory bulb, both forebrain derivatives of the lateral ganglionic eminence, can be observed (Toresson and Campbell, 2001).

It would seem that bilaterian Gsx expression is conserved within the anterior nervous system, and also within the endoderm in the anterior gut of a few key species (*Platynerieis, Nereis, Gibbula, Xenopus)* (Hui et al., 2009b; Illes et al., 2009; Kulakova et al., 2008; Samadi and Steiner, 2010). Overall, these studies would suggest an ancestral role of Gsx within the patterning of the anterior central nervous system and perhaps also the anterior gut within the PDA.

1.3.4.5. ParaHox expression within the PDA.

Several commonalities are present in the expression patterns of bilaterians when looking across both protostome and deuterostome ParaHox expression, and these can be used to reconstruct the expression pattern most likely within the protostome-deuterostome last common ancestor. Gsx is clearly the anterior most expressed ParaHox gene, Xlox the middle, and Cdx the most posterior across the Bilateria. In addition to this, all three ParaHox genes were likely ancestrally expressed within both the CNS and gut, with Cdx expression also highly conserved within the extreme posterior within all three germ layers. Studies within the deuterostomes also show that it is likely that the ParaHox genes existed as a cluster that exhibited both spatial and temporal collinearity. This is supported by intact ParaHox clusters within all three deuterostome phyla, within the echinoderms (Annunziata et al., 2013; Baughman et al., 2014), hemichordates (Ikuta et al., 2013), and chordates (Ferrier et al., 2005). In addition, the presence of temporal collinearity within the ParaHox clusters of all three phyla (Annunziata et al., 2013; Brooke et al., 1998; Ikuta et al., 2013; Osborne et al., 2009), and spatial collinearity within both chordates (Brooke et al., 1998; Osborne et al., 2009) and echinoderms (Annunziata et al., 2013) provides strong evidence towards the ParaHox cluster of the PDA also exhibiting collinearity, with Cdx being expressed first and most posterior, Xlox second and more centrally, and Gsx last and most anterior (reviewed in Garstang and Ferrier (2013)). The identification of a common Xlox-Cdx boundary within all of these same deuterostome phyla (Annunziata et al., 2013; Ikuta et al., 2013; Osborne et al., 2009) also suggests a more specific role for Xlox and Cdx within the patterning of the midgut-hindgut boundary within the PDA.

It is possible, however, that the collinearity of ParaHox clusters across deuterostome phyla instead represents consolidation of ParaHox genes into a locus that allows the evolution of regulatory mechanisms that can act to regulate multiple ParaHox genes. This has been proposed for the Hox clusters (reviewed in Duboule (2007)), and warns of a vertebrate-centric approach to the examination of gene clusters. This hypothesis suggests that rather than ancestral clustering being lost, it could be that clustered genes have secondarily come together within more complex animal phyla in order to take advantage of regulatory mechanisms that allow the ordered expression of genes along the A/P axis. This could then be the case for the Hox and ParaHox genes within bilateria. With spatial ordering of Hox paralog expression still present in derived organisms such as Oikopleura, which have atomised Hox genes (Seo et al., 2004), it is possible that Hox/ParaHox genes have become clustered only to secondarily take advantage of temporal collinearity. Though just one hypothesis, it is certainly worth bearing in mind that the majority of work examining Hox and ParaHox clustering and expression is carried out with respect to the similarities and differences to the vertebrate Hox clusters.

1.4. The invertebrate chordates as model organisms.

1.4.1. Amphioxus.

Amphioxus, or the lancelet, is the sole surviving order within the chordate sub-phylum Cephalochordata. These small 'fish-like' animals can be found globally within coastal marine habitats and have been recognised as an interesting focus for biological studies for hundreds of years, and the development of amphioxus and its similarity to the vertebrates has been well-studied since the early 1900's (Conklin, 1932). Amphioxus has more recently, since the early 1990s, been 're-discovered' as a model organism for evolutionary developmental biology (evo-devo) (Holland et al., 2004) with the molecular study of the amphioxus genes (Glardon et al., 1998; Holland and Holland, 1996; Holland et al., 1999; Holland et al., 1994; Holland et al., 1992; Holland and Garcia-Fernàndez, 1996; Jackman et al., 2000; Kozmik et al., 1999; Lin et al., 2006; Panopoulou et al., 1998; Schubert et al., 2000a; Schubert et al., 2000b; Schubert et al., 2001; Yu et al., 2004). The phylogenetic position of amphioxus as a pre-2R duplication archetypal chordate (Holland and Garcia-Fernàndez, 1996;

Putnam et al., 2008), and the cephalochordates now being identified as the most basal chordate lineage (Bourlat et al., 2006; Delsuc et al., 2006; Vienne and Pontarotti, 2006), make it well placed for molecular, phylogenetic and morphological studies into the ancestry of the chordate phylum. The release of the *Branchiostoma floridae* genome confirmed the position of amphioxus as a basal pre-duplication chordate (Putnam et al., 2008). This is particularly intriguing as the Cephalochordate sub-phylum has survived as a lineage since before the Cambrian, 542 million years ago (Gradstein and Ogg, 2004; Janvier, 2003) and several examples of possible Cephalochordates exist from this time (most notably *Pikaia* (Mallatt and Holland, 2013; Morris and Caron, 2012)), which helps to justify the suitability of amphioxus as a proxy for the chordate ancestor. These fossils, from the Burgess shale and Chengjiang deposits, have similar morphology to amphioxus (reviewed in Schubert et al. (2006)). Several species of amphioxus now exist, and all are closely related, with three main genera: *Asymmetron, Epigonicthys* and *Branchiostomadae*, with the majority existing within the *Branchiostoma* clade (Kon et al., 2007). Even the most distantly related species are still able to produce hybrids, though they have not yet survived through metamorphosis (Holland et al., 2015).

Amphioxus, particularly the adult, clearly displays morphological characteristics indicative of, and plesiomorphic to, the chordate phylum; including a dorsal hollow nerve cord, elaborated anterior CNS (cerebral vesicle), notochord, post-anal tail, pharyngeal (gill) slits, lateral and segmented chevron-shaped muscles, and an endostyle (homologous to the thyroid in vertebrates . (Holland and Holland, 1996, 1998; Mallatt and Chen, 2003) (reviewed in Holland et al. (2004)). The adult amphioxus also possesses repeated gonads attached to the inner wall of the atrium, that increase in size dramatically during the spawning season (late spring-summer months), and amphioxus adults display separate sexes (Stokes and Holland, 1996). Amphioxus is a filter feeder and spends the majority of its adult life burrowed within sediment, which is usually coarse sand, with the anterior end protruding. The gill slits, which can number up to 200, are clearly visible and open from the pharynx to the atrium and are used in food collection, along with the numerous oral cirri (Brusca and Brusca, 2003). The development of amphioxus embryos is rapid (see description of amphioxus embryology below) and the larvae spend time within the water column before metamorphosis into adults (Stokes and Holland, 1995). Until recently, in order to collect amphioxus embryos, ripe, gravid adults were collected either just before or during the spawning season and a heat shock technique, or electrical stimulation used to induce spawning in the lab (Fuentes et al., 2007; Stokes and Holland, 1995). Efforts have since been made to keep viable laboratory breeding stocks (Benito-Gutierrez et al., 2013; Holland et al., 2015; Yasui et al., 2007). Still, for the large majority of studies, adult amphioxus are still collected from the wild for use in spawning and embryo collection.

Amphioxus is well suited as a model for the evolution and regulation of the chordate ParaHox cluster, as amphioxus contains single Hox and ParaHox clusters (Brooke et al., 1998; Ferrier et al., 2005; Garcia-Fernandez and Holland, 1994), and represents a pre-duplication chordate genome (Holland and Garcia-Fernàndez, 1996; Putnam et al., 2008). Though the ParaHox clusters (Gsx-Cdx) of vertebrates are slightly larger (mouse = ~125kb, human= ~185kb), than the amphioxus ParaHox cluster (=~60kb) relative gene sizes and spacing are maintained between amphioxus and vertebrates (Ferrier et al., 2005). In addition, the amphioxus ParaHox cluster provides an alternative, simpler model for studying gene/gene cluster regulation and evolution than the Hox cluster. In part this is due to the smaller, less complex nature of the ParaHox cluster, with its three genes compared to the Hox cluster's over eight genes, but also because it displays many of the regulatory phenomenon associated with the Hox cluster such as collinearity and pan-cluster response to retinoic acid (Osborne et al., 2009), despite its less complex composition. This places amphioxus in a unique position in which to draw from vertebrate Hox and ParaHox studies and examine regulatory pathways that may have a more widely conserved role in ParaHox regulation.

Unfortunately, despite the apparently ideal phylogenetic position of amphioxus, as well as its archetypal ParaHox cluster, amphioxus is not well-placed as an experimental model. Though some efforts have been made in the micro-injection of amphioxus embryos (Beaster-Jones et al., 2007; Yu et al., 2004), the technique still remains both challenging and uncommon. Embryos can also be treated with signalling molecules or signalling pathway inhibitors by 'bathing' embryos (Holland and Holland, 1996; Onai et al., 2009; Osborne et al., 2009), though these often have undesirable secondary effects and are very crude compared to targeted approaches such as knockouts, knockdowns and over-expression studies. As such, the tunicate *Ciona intestinalis* offers an alternative experimental model in which to carry out cross-species regulatory studies using amphioxus ParaHox regulatory elements, whilst amphioxus can be used for gene expression studies. Both *Branchiostoma floridae* and *Branchiostoma lanceolatum* are used within the studies described within this thesis.

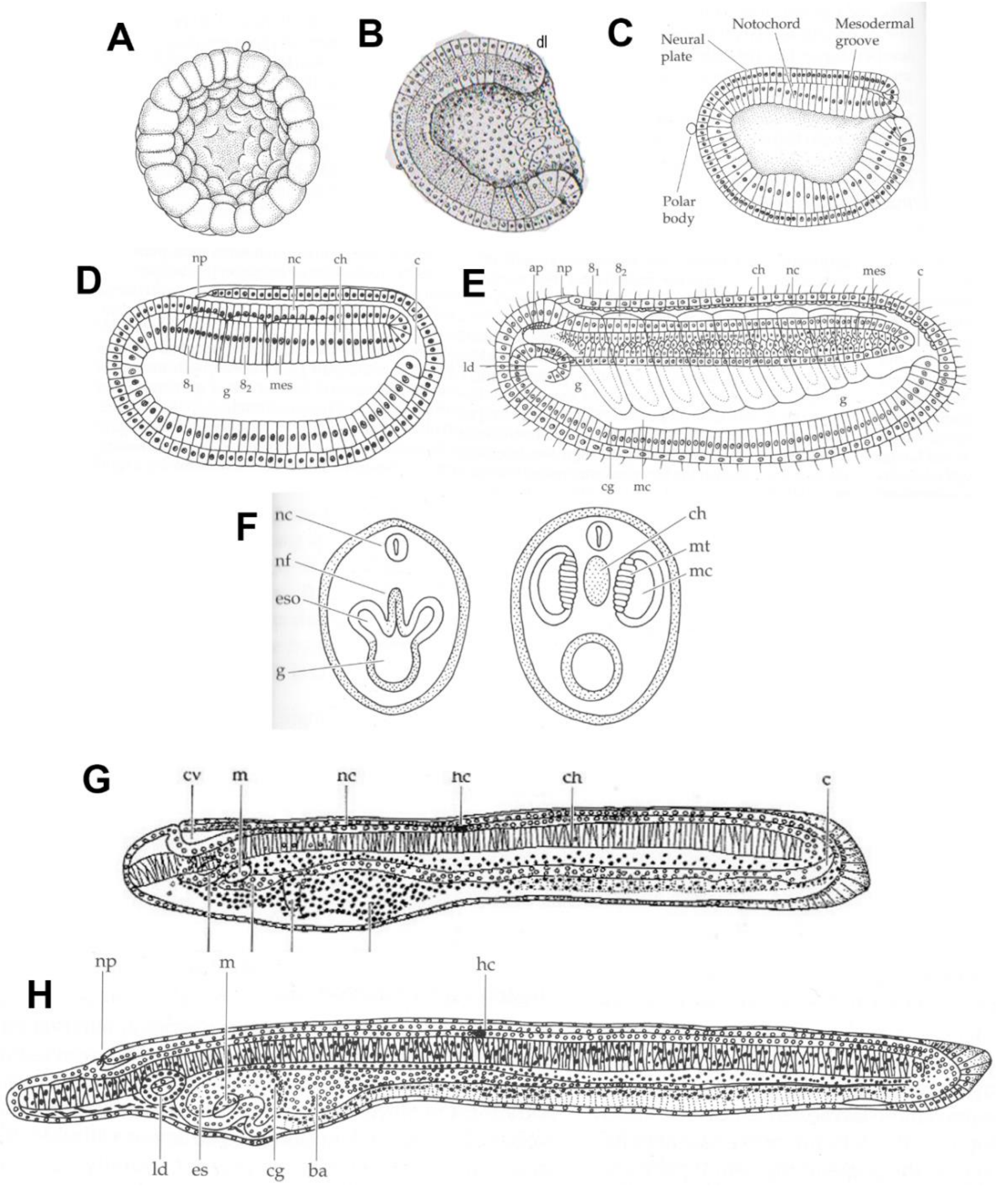
1.4.1.1. Amphioxus development

Perhaps the most descriptive studies of amphioxus embryology still are those of Hatsheck (1893) and Conklin (1905), and the large part of modern knowledge of amphioxus embryology and developmental morphology can be attributed to these works (Conklin, 1932; Hatsheck, 1893). Though development is a continuous process, the development of amphioxus can be roughly divided into several stages; cleavage, gastrulation, neurulation, differentiation and the transition from embryo to larva (Hatsheck, 1893). Development begins in the cleavage stages, which have been well characterised within amphioxus, and the fates of cells within the early embryo mapped out through use of gene expression and functional data, blastomere isolation and dye labelling (Holland and Holland, 2007). Different temperatures have large effects upon the speed of development within amphioxus and timing is thus not an accurate measure of development, and morphological staging should instead be used (Stokes and Holland, 1995). The first cleavage of amphioxus is very rapid, and happens within an hour of fertilisation (though see previous point about timing and developmental staging). Cleavage then proceeds in a roughly synchronous, radial and holoblastic fashion, with blastomeres nearly equal in size that do not tightly adhere to each other. This continues up until the eighth cleavage, or 256-cell stage, when a hollow blastula is formed and marks the end of the cleavage stage of development (Hatsheck, 1893; Holland and Holland, 2007; Whittaker, 1997). During the mid-blastula stage cell adhesion becomes tighter and compaction occurs (figure 1.4.A). Gastrulation of the blastula then begins when the vegetal pole of the embryo flattens and invaginates towards the animal pole and the blastocoel completely disappears resulting in two embryonic layers of cells in a bowl shape. The lips of the blastopore then begin to move together and the embryo begins to elongate along the A/P axis, with the blastopore remaining in the dorsal posterior region of the embryo (Conklin, 1932; Hatsheck, 1893). The dorsal lip of the gastrula stage can be identified by the 'flatter' shape than the more rounded ventral lip of the blastopore (Figure 1.4. B-C) (Hatsheck, 1893; Holland and Holland, 2007).

The next stage of development, neurulation, then proceeds with the flattening of the dorsal side of the embryo as the neural plate extends. The neural plate then begins folding and invaginating, forming the neural tube. This completes as the dorsolateral ectoderm migrates and fuses dorsally over the folded neural tube. Neurulation begins at the posterior, covering the blastopore, and advances to the anterior end of the embryo (figure 1.4. D-E). A small opening in the neural tube, the neuropore, remains until the formation of the anus (Conklin, 1932; Hatsheck, 1893). During neurulation, the somites and notochord also begin to form, starting from the anterior, from the folds of the archenteron (figure 1.4. F). This forms the first eight somites as the embryo elongates posteriorly. The notochord forms following where the somites have pinched off from the archenteron, from dorsal folding of the membranes of the archenteron that extends along the A/P axis as the embryo elongates. At this stage the notochord is still attached to the archenteron by a membrane, but eventually pinches off (figure 1.4 F). The anterior tip of the notochord, an amphioxus specific structure, extends all the way to the anterior of the animal and does not form until after the formation of somite five. At this stage the embryo hatches, and uses ciliated cells located across the entire external surface to move through the water column (figure 1.4. E). Somite position has become asymmetric by the formation of somite seven, and neurulation continues until the closure of the neural tube at the anterior, though the neuropore remains open. The neuropore forms the neural canal, which extends throughout the neural tube round the posterior of the animal and into the archenteron (figure 1.4. D-E). This persists until the formation of the anus (Hatsheck, 1893). The differentiation stage is characterised by the posterior elongation of the embryo, with somites now forming asymmetrically from the tailbud, as well as organ differentiation. Ventral blood vessels begin to form from the somites, and the pigment spot arises within the neural tube at the level of somite 5. Endodermal thickenings mark the beginning of formation of the mouth and the first gill slit, and the anterior/ventral region enlarges to form the future pharynx. In addition the endostyle and club shaped gland begin to form. The neural tube shrinks in diameter throughout most of the embryo, leaving an enlarged swelling within the anterior, forming the cerebral vesicle (figure 1.4. G-H). Finally, the tail caudal fin begins to form from the posterior ectoderm. The rapid formation of the anus, mouth and first gill slit then marks the transformation to the larval stage (Hatsheck, 1893). The larva begins feeding as the mouth opens on the left, whilst the first gill slit opens ventrally but then moves right (figure 1.4. H). Gill slits proceed to form until the larva metamorphoses, with the total number of gill slits reaching well over 100.

Figure 1.4. The embryology and development of amphioxus

(A) The blastula with blastocoel (hollow). (B) The early gastrula has fully invaginated and formed two cell layers; the mesendoderm and the ectoderm, eliminating the blastocoel (C) By the mid/late gastrula the embryo has specified the regions that will form several embryonic structures such as the neurectoderm, notochord and mesodermal groove. (D) In the early/mid-neurula the blastopore has finished closing and the embryo has begun elongating. The neural plate is folding dorsally from the posterior to form the neural tube, with the ectoderm covering it. Somites are beginning to form from the folds of the archenteron. (E) The late neurula has fully neurulated, leaving the neuropore open at the anterior. In preparation for hatching cilia have formed on the exterior of the embryo. The neural canal, neural tube, somites and endoderm are now all visible. (F) Schematic cartoons of transverse views of (D) and (E) showing the formation of the somites and notochord from the folds of the archenteron. (G) Late ‘Pre-mouth’ embryo where the cerebral vesicle, cells of the Hesse organ (develops from the pigment spot), notochord (including anterior notochord). (H) A pre-metamorphosis larva with developing mouth and caudal fin. Abbreviations: dl, dorsal lip of the blastopore; np, neuropore; nc, nerve cord; ch, notochord; c, neuroenteric canal; s1/2, numbered somites; mes, unsegmented mesoderm; ap, anterior process of first somite; ld, left anterior gut diverticulum; mc, myocoel; cv, cerebral vesicle; m, mouth; hc, cells of Hesse organ; es, endostyle; cg, clubshaped gland; ba, branchial anlage. Figures taken and adapted from (Hatsheck, 1893; Holland and Holland, 2007; Whittaker, 1997)


A
B
dl
C
Neural plate
Notochord
Mesodermal groove
Polar body
D
np
nc
ch
c
8_1
g
8_2
mes
E
ap
np
8_1
8_2
ch
nc
mes
c
ld
g
g
cg
mc
F
nc
nf
eso
g
ch
mt
mc
G
cv
m
nc
hc
ch
c
H
np
m
hc
ld
es
cg
ba

1.4.2. Ciona intestinalis

Ciona intestinalis is a member of the tunicates (sometimes known as the urochordates), which are now recognised as the sister group to the vertebrates, having diverged more recently than with amphioxus (Delsuc et al., 2006; Lemaire, 2011). The tunicates are a large and diverse sub-phylum, and containing both sessile and pelagic species, the majority of which belong to the Ascidiidae. The vast majority of tunicate adults, with the exception of the larvaceans, who retain a larval form throughout their life, go through a striking metamorphosis from a chordate-like larval form to a bag-like filter feeding adult, which does not look like any other chordate (Brusca and Brusca, 2003; Lemaire, 2011; Satoh, 1994). The larvae of tunicates display many archetypal chordate features, including a dorsal nerve cord, elaborated anterior CNS (sensory vesicle), notochord and lateral muscle. However, most of these are reabsorbed/degraded during metamorphosis, including the notochord, though the adult ascidian does possess an endostyle. Adult ascidians are generally hermaphroditic, containing separate testes and ovaries, along with oviduct and sperm duct, and solitary ascidians generally reproduce sexually, though colonial ascidians can reproduce both sexually and asexually (Brusca and Brusca, 2003; Lemaire, 2011; Satoh, 1994). Both eggs and sperm can be extracted separately from *Ciona*, allowing for external fertilisation and exact timing of embryonic stages. In addition, the genomes of several tunicate species now exist (both solitary and colonial) allowing comparison between both the tunicates and other phyla. These genomes belong to; *Botryllus schlosseri, Ciona intestinalis, Ciona savignyi, Halocynthia aurantium, Halocynthia roretzi, Molgula occidentalis, Molgula occulata, Phallusia fumigata* and *Phallusia mammilata*.

One advantage over amphioxus is that experimental manipulations are much more established within *Ciona*, and the ability to determine cell fate from early in development aids greatly in both expression and regulatory studies. So far, genetic screens, germline transgenesis, electroporation of plasmid DNA, and micro-injection of morpholinos are all tools available to the *Ciona* geneticist, with hopes for targeted mutagenesis, homologous recombination and RNAi in the near future (Stolfi and Christiaen, 2012). As such, the genetic toolkit available is sufficient for many over-expression, gene knockdown and transgenic studies using the currently available techniques.

In particular, *Ciona* provides a system that is highly amenable to analysis of cis-regulatory elements via embryo electroporation of reporter gene constructs (Corbo et al., 1997), and Cross-species transgenesis between amphioxus and *Ciona* has provided an alternative route to rapidly analysing putative amphioxus regulatory elements (Beaster-Jones et al., 2007; Natale et al., 2011; Wada et al., 2005) in vivo. This allows the use of the substantial *Ciona* toolkit to test the function of amphioxus regulatory elements, when such options are not so readily available in amphioxus itself.

1.4.2.1. *Ciona intestinalis development*

The simple embryo, precisely mapped developmental timing and determinative development of *Ciona* have allowed the production of a precise cell-fate map of early *Ciona* development, and the ANISEED database now contains both comprehensive cell fate maps and 3D reconstructions of many stages of *Ciona* development (Tassy et al., 2010). Ascidian development has again been well characterised since the early 1900's and these studies formed the basis for modern cell fate maps (Conklin, 1905) for Ciona development is extremely rapid, and the free-swimming larval stage is reached by 18hpf at 18˚C (Hotta et al., 2007), and development in the laboratory is normally studied at this temperature (or 16˚C). This development can be grouped into six periods; known as the zygote, cleavage, gastrula, neurula, tailbud and larva. *Ciona* development is even more rapid than that of amphioxus, and the first cell devision begins by one hour at standard temperatures, and the embryo is ready to gastrulate (110-cell stage) within four to five hours (Hotta et al., 2007; Kumano and Nishida, 2007). From the first cleavage, the cells of the *Ciona* embryo are stereotypical, resulting in the AB and <u>AB</u> (or AB*) lineages for the left and right side of the embryo respectively. Cleavage to the four-cell stage results in the specification of anterior (labelled 'A') and posterior (labelled 'B') lineages. The third division establishes the dorso-ventral axis, with dorsal (vegetal) cells labelled A or B (upper case), and ventral (animal) cells labelled a or b (lower case). Thus at the 110 cell stage, a left anterior ventral cell could be a8.32, whilst a right posterior dorsal cell could be <u>B7.1</u> (or B7.1*). Unlike amphioxus, *Ciona* embryos do not have a true blastocoel prior to gastrulation and the blastula is not hollow.

At gastrulation, a single layer of endoderm and mesoderm cells invaginates so that the mesoderm cells lie at the rim of the blastopore, being covered fully by ectodermal cells as the ectoderm migrates towards the vegetal pole (dorsal) to cover the exterior of the embryo (figure 1.5 A). The mesodermal cells are already fated to become the notochord (anterior cells), mesenchyme (posterior cells) and the muscle (lateral cells) even at this early stage (Hotta et al., 2007; Munro et al., 2006). The neural plate has also been specified at the gastrula stage in several rows of 6-8 cells aligned transversely across the dorsal surface, adjacent to the mesodermal cells at the blastopore lip. Neurulation begins at 7hpf as the neural plate sinks medially and the lateral edges curl upward to form the neural tube. Neurulation begins at the posterior, covering the blastopore and forming a neuroenteric canal, then moves anteriorly (Conklin, 1905; Jeffery and Swalla, 1997; Munro et al., 2006). The posterior neural plate will become the spinal cord-like dorsal nerve cord, whilst the anterior becomes the visceral ganglion and brain-like sensory vesicle (figure 1.5).

Next, the posterior of the embryo constricts and begins to extend, forming the classical 'tadpole' tailbud embryo stages, with the tail bending ventrally as it elongates. This bending is driven by the extending notochord, as notochord precursors migrate to the posterior, where they form a single line of cells that secrete a stiff extracellular matrix (figure 1B iii-iv) (Kumano and Nishida, 2007) Meanwhile, the muscle cells form laterally in three distinct rows, ventral, lateral and dorsal as the embryo extends. The muscle in Ciona is not formed within somites like in amphioxus and the vertebrates, and is muscle cells remain as large, single cells with striated actin, but do not form myotubes (Jeffery and Swalla, 1997; Munro et al., 2006) (figure 1.5 Bv). Further mesenchymal cells lie anterior to the tail within the trunk, in two crescent shaped regions, and later become parts of the adult heart and gills and adult muscle and haemolymph (blood) (figure 1C ii).

By the mid-tailbud stage, the central nervous system has changed morphologically, with the dorsal nerve cord running along the length of the tail and consisting of four rows of cells, one dorsal, two lateral and one ventral, whilst the anterior CNS has formed the visceral ganglion in the 'neck' region of the embryo and the sensory vesicle more anteriorly within the head (figure 1.5 B iii-iv, C ii). The *Ciona* larva is non-feeding, and lacks a mouth and gut. The endoderm is instead represented by the endodermal strand, which runs the length of the tail and lies between the dorsal muscle cells and below the notochord, and the head endoderm. This endodermal strand is later reabsorbed during metamorphosis to form parts the adult intestine, along with the head endoderm, (Nakazawa et al., 2013) and the primordial germ cells (Takamura et al., 2002)

In the larva two pigmented sensory organs form within the sensory vesicle; the otolith, a gravity sensing organ; and the ocellus, which develops second and senses light (Jeffery and Swalla, 1997; Satoh, 1994). The tail is motile and the embryos hatch from the chorion at the larval stage into the water column, though they use the tail only transiently before attaching to suitable surfaces using the palps at the extreme anterior of the embryo (Jeffery and Swalla, 1997; Kumano and Nishida, 2007). The tunic has already formed around the hatching larva, whilst the heart, digestive organs, siphons and pharyngeal primordia have also partially formed within the motile larva, and are retained as the larva goes through metamorphosis (Hirano and Nishida, 1997). Metamorphosis takes one or two days, and the body axis rotates to form the juvenile, which is similar to the adult but much smaller, having fewer gill slits and not yet being sexually mature (Chiba et al., 2004) (figure 1.5. E). The sexually mature adult form is reached by approximately three months post metamorphosis (figure 1.5. D)(Jeffery and Swalla, 1997).

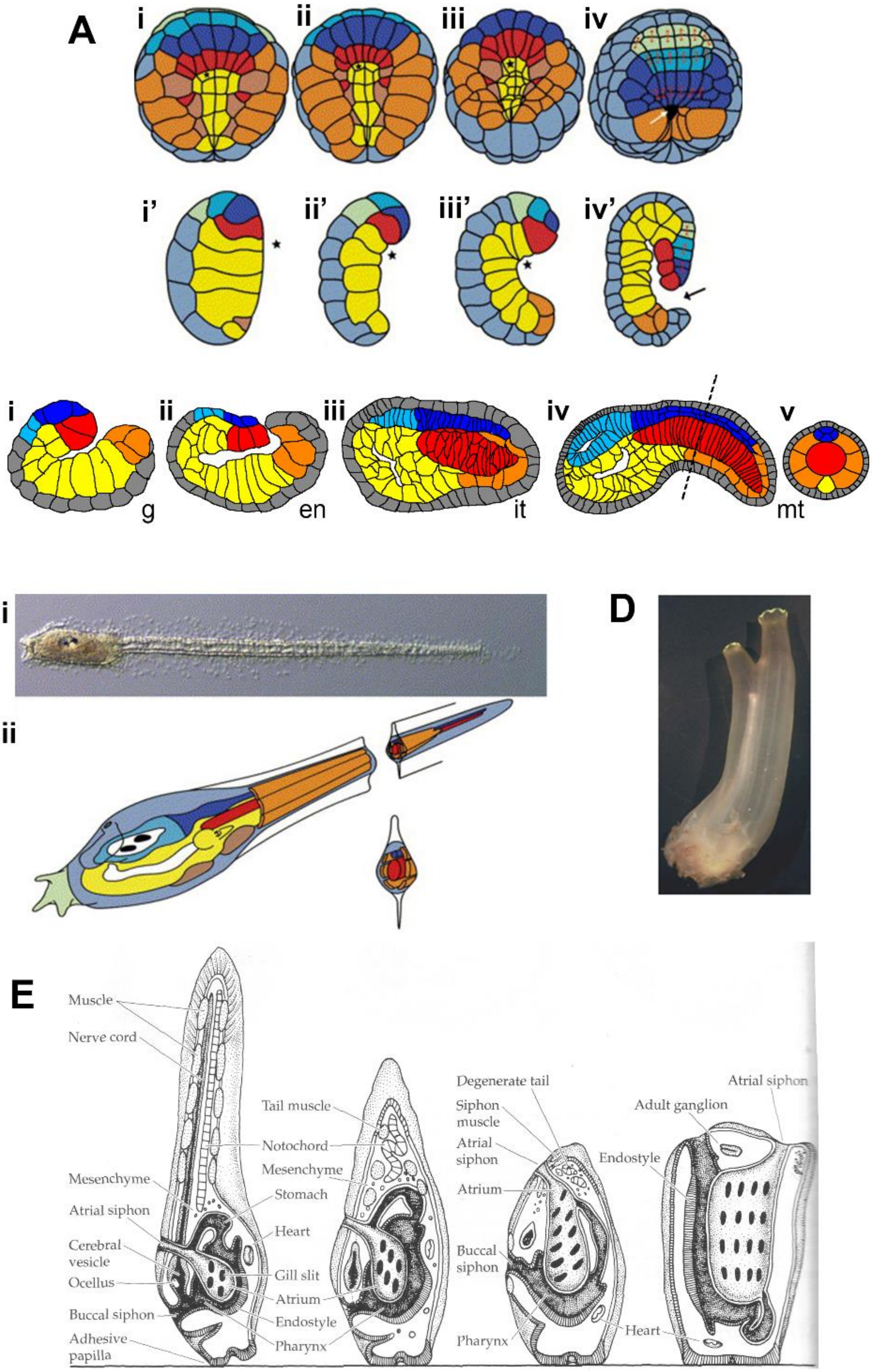

A
i
ii
iii
iv
i'
ii'
iii'
iv'
B
i
ii
iii
iv
v
g
en
it
mt
C
i
ii
D
E
Muscle
Nerve cord
Mesenchyme
Atrial siphon
Cerebral vesicle
Ocellus
Buccal siphon
Adhesive papilla
Tail muscle
Notochord
Mesenchyme
Stomach
Heart
Gill slit
Atrium
Endostyle
Pharynx
Degenerate tail
Siphon muscle
Atrial siphon
Atrium
Buccal siphon
Pharynx
Atrial siphon
Adult ganglion
Endostyle
Heart

Figure 1.5. The embryology and development of *Ciona intesitinalis*

(A) Schematic view of gastrulation of the *Ciona* embryo. The specification of all cell types of the tadpole larva occurs prior to gastrulation. (i-iv) show vegetal/posterior views of gastrulation stage embryps, whilst (I'-iv') display lateral views. (B) (i-iv) Schematic showing cell and tissue fates through developing gastrula to mid tailbud embryos. Schematics are made from traces of the embryos in (figure 5.9 Q E, F, M, O), though numbers of cells displayed may not be absolutely accurate due to cell membranes not being visible in different focal planes. (v) Represents a transverse section through the plane shown by the dotted line in (iv), showing the location of the three rows of lateral muscle cells, dorsal nerve cord, notochord and endodermal strand. (C) (i) Image of a swimming Ciona larva, and schematic (ii) showing the fates of all cell types depicted in (A). (D) Image of a mature *Ciona* adult with visible filled sperm duct. (E) schematic showing ascidian metamorphosis. The tadpole attaches to the substrate via the palps, and reabsorption of the tail occurs, whilst the body axis rotates during metamorphosis to the juvenile stage. Legend: Presumptive notochord cells are shown in red, endoderm yellow, muscle orange, epidermis grey, nerve cord dark blue, sensory vesicle light blue and the palps in green. Lower case lettering refers to the stage of development; g, gastrula; en, early neurula; itb, initial tailbud; mtb, mid tailbud. Figures taken and adapted from (Jeffery and Swalla, 1997; Munro et al., 2006; Sato et al., 2012)

1.5. ParaHox gene regulation

Considering the recent interest in ParaHox genes in the context of gene clustering, evolution and development, very little is known about how they are actually regulated. In order to fully understand the mechanisms and constraints underlying gene clustering and the evolution of the ParaHox genes, it is important to examine the regulatory landscape of these genes. By utilising modern techniques in molecular biology, we are beginning to see how regulatory mechanisms play an important role in the maintenance and regulation of the Hox cluster, but perhaps more importantly the high level of conservation of these mechanisms between lineages as highly divergent as fruit flies and humans. As the evolutionary sister of the ParaHox cluster, we may be able to draw from our understanding of the Hox cluster in our examination of the ParaHox regulation and cluster maintenance. This may potentially work vice versa as well once the ParaHox cluster has been described in more detail, also aiding in our understanding of the Hox cluster.

1.5.1 General gene regulation

Unlike with protein coding regions, which are distinguished by 3bp codons, Methionine start sites and STOP codons, and splice site identifiers, gene regulatory regions have a much looser set of rules. Still, variations in such gene regulatory regions can result in drastic changes in gene expression, and the effect of mutations within promoter regions with regards to the correct expression of genes is now acknowledged to be as potentially harmful as those within coding regions (reviewed in Wray et al. (2003)). Though gene promoters contain a basal promoter region, specified by a 100bp region containing a TATA box as well as various transcription factor binding sites (TFBS) (Lee and Young, 2000; Wray, 2003; Wray et al., 2003), there are no other defining features and nucleotide sequence has proven to be a poor indicator of Promoter function or efficiency beyond these basic features.

Perhaps the most famous example of the importance of cis-regulatory regions to the correct expression of genes is in the case of the *Drosophila* Bithorax complex (*BX-C*) locus and the *even-skipped* (*eve*) gene. Within the *BX-C* locus of *Drosophila*, several non-coding regions, both within introns and intergenic, were found to direct expression of the *Ultrabithorax* (*Ubx*) and *abdominal-a* (*abd-A*) genes (Bender et al., 1983; Karch et al., 1985; Simon et al., 1990; White and Wilcox, 1985). The authors identified the *abx/bx* region (*Ubx* intronic), the *bxd/pbx* region (upstream of *Ubx*), the *iab-2* region (*abd-A intronic*) and the *iab-3* region (*abd-A upstream*). Each of these regulatory regions was able to drive expression of a *LacZ* construct within distinctive parasegments of the *Drosophila* embryo, within parasegments 5 (*abx/bx*), 6 (*bxd/pbx*), 7 (*iab-2*) and 8 (*iab-3*) respectively (Simon et al., 1990). A similar, but much more in depth study, has been carried out specifically for the *eve* gene, where the regulatory region has been thoroughly characterised into discrete ‘modules’ directing the expression of specific *eve* expression bands. Though *eve* is activated in a ubiquitous manner, these modules act as repressors within specific regions to give *eve* its characteristic stripped pattern. One module, directed by *Knirps* defines the third and seventh stripes, with *Knirps* repressing the posterior of stripe three and the anterior of stripe seven. *Hunchback* then sets the anterior limit of stripe three and the posterior limit of stripe seven by causing repression at the anterior and posterior ends of the embryo. Similar, but separate, modules define the expression for all of the stripes of *eve*, resulting in a complex array of repression and activation from relatively simple individual modules, giving the characteristic seven stripes pattern of *Drosophila eve* (Arnosti et al., 1996; Small et al., 1996).

Even small changes and mutations in the regulatory landscape can lead to novel domains of expression and phenotypic consequences, as seen in the pigmented wing patterns of the *Yellow* gene in drosophilids, where subtle changes within cis-regulatory elements, and the co-option of new transcription factors have led to novel and repeated changes in wing pigment patterns (Gompel et al., 2005). It is these regulatory inputs, in the form of transcription factor binding sites, that seem to be the only common factor between promoters, enhancers and repressors, and the suite of binding sites within a cis-regulatory region are key to directing both the temporal and spatial expression of the target gene. This regulatory input results in the transcriptional activation of the target gene if a certain threshold is reached, whether that is an amount of a single factor, or the combination of specific factors binding the cis-regulatory region (Davidson, 2001).

The cascade of signalling molecules and transcription factors that result in gene activation or repression results in a complex web of gene regulatory networks (GRNs) that interact and lead to the repression and activation of genes in a variety of cell types and developmental contexts (Davidson, 2010; Davidson and Erwin, 2006). For example, the identification of conserved factors within the endomesoderm GRNs of both sea urchin and *Xenopus* (Hinman et al., 2003; Loose and Patient, 2004) suggests that signalling molecules and their downstream transcription factors could be used to deduce ancestral regulatory interactions. Such GRNs have also been described for other tissues such as the neural crest (Sauka-Spengler et al., 2007) and have even been described for specific developmental events, such as the specification of the midgut-hindgut boundary in the sea urchin (Annunziata and Arnone, 2014). This last study holds particular relevance due to the involvement of the ParaHox genes *Xlox* and *Cdx*, and the presence of the Xlox/Cdx midgut-hindgut boundary throughout the deuterostomes as discussed earlier. It is therefore possible to envisage how such gene regulatory networks might be used to identify the cis-regulatory elements involved in the network, through identification of relevant TFBS, as well as inform the identification of such regulatory interactions in other species. This approach could prove more fruitful than traditional cross-species sequence comparisons, as such approaches have proved to be unreliable for the identification of cis-regulatory elements. Even within the vertebrates there are relatively few cis-regulatory elements that display conserved sequence (Woolfe et al., 2005), and this becomes much harder, though possible in a few cases, as comparisons between larger evolutionary distances are made (Makunin et al., 2013; Pascual-Anaya et al., 2008; Vavouri et al., 2007; Woolfe et al., 2005).

1.5.2. Transcriptional Regulation within the Hox cluster

As the ParaHox cluster is the evolutionary sister of the Hox cluster, and regulatory phenomena such as collinearity have already been observed in both the regulatory studies of the Hox cluster can be looked towards for inferences as to how the ParaHox cluster may be regulated. Many enhancers have been identified within Hox clusters, regulating individual genes such as Hoxb8 (Charite et al., 1995) but also several directing the expression of multiple genes such as the shared neural mesoderm enhancers of Hoxb4 and Hoxb5 (Sharpe et al., 1998), or the shared promoter of zebrafish Hox3a and Hox4a (Hadrys et al., 2006). However, more interesting still, and perhaps more relevant for understanding Hox and ParaHox pan-cluster regulation, are the global enhancers found controlling vertebrate Hox clusters. Such 'global' cis-regulatory regions have been located at both the 3’ and 5’ ends of the HoxD and HoxA clusters (Kmita et al., 2000; Lehoczky et al., 2004; Spitz et al., 2003; Spitz et al., 2005; Tarchini and Duboule, 2006). Similarities between these global enhancers of the HoxD and HoxA clusters, including a 131bp conserved ‘core’ sequence, suggest that these pan-cluster mechanisms were present in the Hox cluster before the vertebrate 2R duplications (Lehoczky et al., 2004). Experiments using inversions of the HoxD cluster have suggested that several global elements exist, one at the 3’ of the cluster and another at the 5’. The first of these, the 3’ Early limb control region, or ELCR, causes temporal activation of the Hox genes dependant on their distance from the ELCR. In this case, *Hoxd1*, the nearest gene, is activated first, with genes being activated later the more 5’ and distant from the ELCR they are located. The second of these global elements is the 5’ POST element, which activates the cluster in a spatial manner along the A/P axis of the developing limb bud. This time, each gene, from *Hoxd10-Hoxd13* is activated in a more posterior domain than its 3’ neighbour, creating spatial collinearity across the developing limb bud (Tarchini and Duboule, 2006). A final element lies further upstream to the 5’ of the cluster than the POST element, and independently activates a second wave of Hoxd expression within the later limb bud. This GCR element activates concomitant expression of *Evx* and the 5’ Hoxd genes, becoming less efficient as distance from the GCR increases. This results in expression of Evx through to Hoxd9, with *Evx* having the strongest expression and *Hoxd9* the weakest (Spitz et al., 2003; Tarchini and Duboule, 2006). This combination of global, shared and individual regulatory regions work in concert to produce the complex expression patterns observed in the Hox cluster.

The regulatory studies detailed above, though detailed, do not tell us much of the regulatory transcription factor input that might be directing such pan-cluster regulatory phenomena. Several studies have sought to examine the regulatory inputs that may be directing expression of the Hox cluster however. The most well studied of these by far is Retinoic acid signalling (RA), and RA has been shown to alter the expression of individual Hox genes and application of exogenous RA produces altered Hox gene expression, as well as the vertebral abnormalities and rhombomeric abnormalities associated with altered Hox gene expression (Conlon and Rossant, 1992; Kessel, 1992). Analysis within human carcinoma cells identified that not only are vertebrate Hox genes activated sequentially, from 3’ to 5’, by application of exogenous RA (Mavilio et al., 1988; Simeone et al., 1990), but also differentially activated by RA according to their position within the cluster (Simeone et al., 1991; Stornaiuolo et al., 1990). In light of this, several studies have identified retinoic acid response elements (RAREs) within the Hox clusters of vertebrates. These RAREs allow the direct regulation of Hox genes by RA through the binding of Retinoic acid receptor (RAR) and Retinoid X receptor (RXR) dimers (Dupe et al., 1997; Langston et al., 1997; Marshall et al., 1996; Marshall et al., 1994; Popperl and Featherstone, 1993). Whilst this mechanism was once thought to be vertebrate-specific, RAREs have since been found within the amphioxus Hox cluster (Wada et al., 2006), and the amphioxus Hox genes also respond to RA signalling (Holland and Holland, 1996; Koop et al., 2010; Onai et al., 2009; Schubert et al., 2005), as do some *Ciona* Hox genes (Kanda et al., 2013)

1.5.3. Regulation of the ParaHox genes

As RA is highly conserved as a regulator of Hox genes, it is an excellent candidate for examining ParaHox regulation in the chordates, and several studies have indeed shown the regulation of ParaHox genes by RA. In mice, *Cdx1* is directly regulated by RA, with excess RA causing posterior expansion of *Cdx1* expression (Houle et al., 2000). A further study identified a DR5 RARE that partially mediated this response (Houle et al., 2003). In addition to the DR5 RARE, a DR2 RARE has been identified in both mouse and chicken *Cdx1/CdxA* introns and deletion of this DR2 element resulted in a shift in LacZ expression of a reporter construct containing this intron (Gaunt et al., 2003). Whilst there are currently no reports of Gsx responding to RA in vertebrates, there are reports of pancreatic Xlox expression being abolished in RA-depleted mouse and zebrafish embryos (Martin et al., 2005; Molotkov et al., 2005; Stafford and Prince, 2002), as well as recovery of Xlox expression when RA is maternally replaced in RA-deficient mice (Martin et al., 2005; Molotkov et al., 2005).

The regulation of ParaHox genes by RA has recently been shown to extend out to the chordates in studies using amphioxus (*B.floridae)*. All three ParaHox genes show response to RA in amphioxus, with the endodermal boundary between *AmphiXlox* and *AmphiCdx* shifting in response to both exogenenous RA and BMS009, an RA antagonist (Osborne et al., 2009). *AmphiGsx* is not expressed in the endoderm, but does shift its expression anteriorly in the neurectoderm, with BMS009 repressing *AmphiGsx* expression altogether (Osborne et al., 2009). Considering the regulation of Hox and ParaHox genes by RA is present in vertebrates, amphioxus and *Ciona*, it is likely that RA regulated both the Hox and ParaHox clusters in the last common ancestor of chordates and possibly much deeper in animal evolution.

Whilst RA signalling is most often associated with axial patterning in chordates, it has come to light that the molecular machinery involved in RA signalling is far more ancient, with retinoic acid receptor (RAR), Retinoid X receptor (RXR) as well as major enzymes involved in RA production and degredation being present throughout the Bilateria (Albalat and Canestro, 2009). It is clear that there have been many independant losses of RA signalling amongst the bilaterian lineages, but it is likely that the C-BLCA possessed a rudimentary RA signalling pathway. It has been observed that classical RAR/RXR heterodimers are not needed for RA signalling to take place, and RXR/RXR homodimers are able to bind to RAREs and mediate signalling (Nowickyj et al., 2008; Vivat-Hannah et al., 2003). This may be interesting in light of the identification of RXR conservation deeper in animal evolution, perhaps allowing RA signalling to function without the need for RAR. The cnidarian *Trypedalia cystophora* possesses RXR with extraordinary similarity to vertebrate RXRs, and it is also involved in pathways regulated by retinoid signalling in vertebrates (Kostrouch et al., 1998), suggesting that this could be the case. RXR is even present in sponges and is upregulated in response to treatment with RA (Wiens et al., 2003), presenting the possibility of RA as a developmentally important signalling molecule in the last common ancestor of all Eumetazoa.

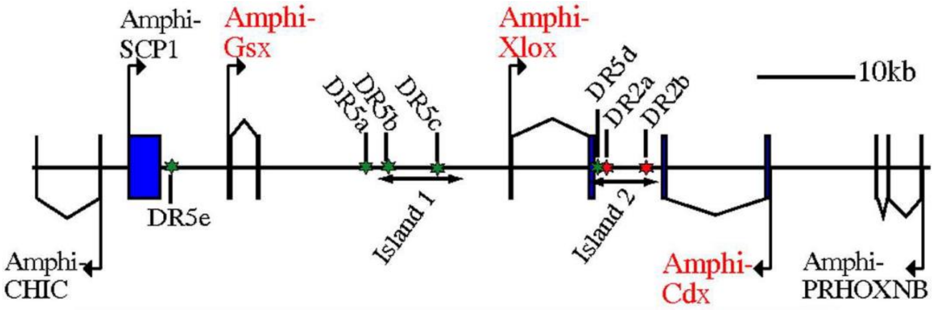


Figure 1.6. The amphioxus ParaHox cluster and mapped positions of Putative RAREs.

Several DR5 and DR2 RAREs have been identified in ParaHox cluster of *B.floridae*. The majority of these putative RAREs are found clustered together in regions termed 'island 1' and 'island 2'. These RAREs may regulate the expression of amphioxus ParaHox genes in response to RA. Figure adapted from (Osborne et al., 2009).

Another pathway that may play a role in the regulation of the ParaHox cluster is that of the lymphoid T-cell-specific transcription factors (TCF/LEF). TCF/LEF forms a heterodimer with β-catenin upon the nuclear localisation of β-catenin in response to Wnt signalling (Behrens et al., 1996; Huber et al., 1996). Wnt signalling is known to play a role in axial patterning, acting alongside molecules such as RA and FGFs in vertebrates (Ikeya and Takada, 2001), and has been noted as an upstream regulator of Cdx genes in vertebrates (Lickert et al., 2000; Pilon et al., 2006; Shimizu et al., 2005). Mouse and chicken studies have even identified TCF/LEF binding sited in Cdx1/Cdxa enhancers, and characterised a response to Wnt signalling via these enhancers (Gaunt et al., 2003; Lickert and Kemler, 2002).

TCF/LEF has been identified in all three chordate sub-phyla, though both amphioxus and *Ciona* only possess one TCF/LEF gene instead of the multiple paralagous genes found in vertebrates, (Lin et al., 2006; Rothbacher et al., 2007) stemming from the 2R WGD in the vertebrate lineage. Upregulation of Wnt signalling using Lithium was shown to have an effect on *Cdx* expression in amphioxus (Onai et al., 2009). However, Lithium is a toxic substance and treatment of embryos with Lithium produces monstrous embryos with abnormal development, so it is not entirely certain whether the changes seen are actually due to direct upregulation of *Cdx* by Wnt, or due to other mechanisms such as posteriorisation, as Wnts are posterior markers, or unknown mechanisms mediated via lithium toxicity. The authors did, however, identify three TCF/LEF binding sites in the *AmphiCdx* upstream regulatory region (Onai et al., 2009). There is also evidence that Wnt signalling and TCF/Lef are involved in regulating Xlox, as Pdx1 expression is lost in *TCF2* knockout mice (Haumaitre et al., 2005), and *TCF2* also induces the expression of *Pdx1* within pancreatic islet cell culture (Quan et al., 2014). In addition, within *Xenopus*, excess Wnt signalling also decreases *Pdx1* expression (McLin et al., 2007). So far, no studies have linked Wnt signalling or TCF/Lef to the regulation of Gsx. Further discussion of TCF/Lef and Wnt with regards to the regulation of ParaHox genes can be found in chapter 5.

Both Wnt and RA have been shown to act in concert to activate Cdx expression within both mouse and chick, and both TCF/Lef binding sites and a RARE have been shown to interact within upstream and intronic Cdx1 enhancers (Gaunt et al., 2003; Gaunt and Paul, 2014). Interestingly, an interaction between the Wnt and RA signalling pathways may also occur in the regulation of Xlox. RA signalling causes activation of the *Ndrg1a* gene, which then goes on to repress Wnt signalling and promote the specification of *Xenopus* pancreas and duodenum (Zhang et al., 2013). Several studies describe this effect of RA activating, and Wnt repressing Pdx1 expression within development of the pancreas (Martin et al., 2005; McLin et al., 2007; Molotkov et al., 2005).

Similarly, several other signalling pathways have been shown to interact and activate Cdx, which then goes on to affect the posterior Hox genes. It may be that Cdx is in fact mediating the effect of many signalling pathways on the posterior Hox genes during posterior development. Examples of this dual-signal>Cdx>Hox mechanism have been shown for the following additional signalling pathways beyond RA+Wnt. Both *Wnt3a/Wnt8* and FGF signals have been shown to activate the expression of *Cdx1a/Cdx4* within zebrafish posterior body formation, and Cdx appears to mediate the transduction of these signals to Hox7a and Hox9a, which are downregulated in the absence of *Wnt3a/Wnt8*, *Cdx1a/Cdx4* and defects in FGF signalling (Shimizu et al., 2005). FGF signalling has been shown to activate the Cdx>Hox pathway within *Xenopus*, with FGF signalling activating *XCad3*, which then activates the posterior Hox genes (Isaacs et al., 1998). The FGF response elements (FREs) of *Xcad3* have been shown to integrate signalling inputs from the FGF, Bmp and Wnt signalling pathways (Haremaki et al., 2003). FGF signalling was also shown to be required for the expression of all three Cdx genes within *Xenopus* during gastrula stages, and interacts with the Wnt signalling pathway (Keenan et al., 2006).

This Wnt+FGF>Cdx>Hox hierarchy has also been shown to interact with RA signalling to clear repressive H3K27me3 histone modifications from the anterior Hox genes in mouse neural progenitors, allowing activation of Hox1-Hox9 (Mazzoni et al., 2013). Similarly, Wnt signalling also acts in concert with BMP signalling, which specifies the dorso-ventral axis in chordates (Panopoulou et al., 1998; Yu et al., 2007a) to again activate the Cdx>Hox pathway. Intriguingly there are two different aspects to this BMP-Wnt-Cdx interaction. The first uses a *BMP4>Wnt3a>Cdx1/4*>Hox pathway to induce ventral-posterior mesoderm, whilst the second acts at a later stage via Wnt>*Lef1>BMP4>Cdx1/4*>Hox to induce hemogenesis. In addition, when BMP signalling is blocked, enforced expression of *Cdx1* or *Cdx4* is able to rescue this latter process (Lengerke et al., 2008). BMP signalling is also involved in the development of the chick hindgut, interacting with *Sonic hedgehog* (*Shh*). Here *Shh* induces the expression of BMP4 and the posterior Hox genes within the hindgut (Roberts et al., 1995). This Shh>BMP4>Hox interaction has also been detailed within the developing limb buds and interacts with the FGF and Wnt signalling inputs to the posterior Hox patterning of limb bud development (Li and Cao, 2003; Sheth et al., 2013). Whilst no link to the ParaHox gene Cdx has yet been made, it is possible that Cdx could also be mediating these pathways involving *Shh* and other signals within both the hindgut and limb bud, particularly considering the Cdx>Hox mediated pathways for the other combinations of axial signaling, the major role of Cdx genes in the patterning of the hindgut and also the expression of *Cdx1* within the limb buds of mouse (Meyer and Gruss, 1993). *Cdx1* overexpression in mice has been shown to cause forelimb defects ranging from no obvious abnormality to severely affected rudimentary limbs (Gaunt et al., 2008). The loss of skeletal elements within the forelimbs of these mice causes in-turning of the limb (from a loss of the radius) and reduction in the number of digits, from one to four. In addition, *Cdx1*$^{-/-}$/*Cdx2*$^{+/-}$ double mutants often show a split digit1 (the big toe) within the hindlimb. Thus Cdx could well be playing a role within the transduction of signals to the Hox genes within the vertebrate limb bud.

Many vertebrate studies have shown the indirect regulation of ParaHox genes, but there are still relatively few looking at the direct regulation of ParaHox genes, or studies examining ParaHox regulatory elements. It is important that this kind of regulatory work is carried out in order to understand the regulatory inputs directing the expression of ParaHox genes in the chordate ancestor. Discovering these ancestral chordate ParaHox regulators holds promise for several human diseases, as Xlox (*Pdx1*) plays a major role in the onset of diabetes (Gannon et al., 2008; Stoffers et al., 1997), and both Xlox (Ma et al., 2008) and Cdx (Colleypriest et al., 2010; Saegusa et al., 2007) have been implicated in cancers. Here, further literature on the direct regulation of ParaHox genes, and the few additional cis-regulatory studies, will be examined. For direct regulation to be established, the authors must have shown binding of the relevant transcription factor, or representative downstream factor of the signalling pathway to a regulatory element of the ParaHox gene.

Cdx is by far the most well studied ParaHox gene in vertebrates, and many regulatory studies additional to those discussed above have been carried out. Human *Cdx2* is able to act in an auto regulatory manner, where *Cdx2* binds to its own TATA-box and upregulates itself. *β-catenin* was also shown to bind to the *Cdx2* promoter region and upregulate *Cdx2* in a manner independent of TCF/Lef binding sites (Saegusa et al., 2007). COUP-TF1 has been shown to competitively bind the RARE in the *Mmu-Cdx1* promoter (Beland and Lohnes, 2005). *Oct1* is a POU homeodomain transcription factor that has been shown to bind OCT sites within the promoters of both mouse *Cdx2* (Jin and Li, 2001) and *Xenopus Cdx4* (Reece-Hoyes et al., 2005), activating expression of Cdx. This latter site was shown to be conserved across all three *Xenopus* Cdx genes, as well as across the vertebrate *Cdx4* genes, making it a very good candidate for a conserved regulator of chordate Cdx (Reece-Hoyes et al., 2005). One region within intron 3 (final intron) of the ParaHox neighbouring gene *PRHOXNB*/*Urad* was found to be necessary for the gut expression of mouse *Cdx2*, binding a combination of the transcription factors *GATA6, HNFα* and *β-catenin/Tcf4* (Benahmed et al., 2008). *NF-κB* is one of the more complex direct regulators of Cdx genes within mice. Through the *PTEN/PI3K-Akt* signalling pathway, *NF-κB* forms a p50 subunit homodimer and directly binds and activates the *Cdx2* promoter. However, *Cdx2* can also be repressed by *NF-κB* binding, where this time a p50/p65 subunit heterodimer binds the *Cdx2* promoter and represses *Cdx2* transcription. Unlike *Cdx2*, *NF-Kb* has no effect on *Cdx1* (Kim et al., 2002). Though no direct binding of Otx to Cdx regulatory elements was shown, the mutually repressive properties, and detailed analysis of transcription of *Xenopus Xcad3* (*Cdx4*) in response to an overexpressed form of *Otx1* with the *Engrailed* repressive domain (OtxEnR) showed that it is likely to be a direct repressor, as overexpression of *Xcad3* would be expected in the case of a secondary intermediate (Isaacs et al., 1999).

Though there are fewer studies on the direct regulation of Xlox, several studies still describe direct binding to regulatory regions. As with *Cdx2*, *Pdx1* has been shown to act in an autoregulatory fashion, binding its own regulatory elements to promote transcription within mouse (Gerrish et al., 2001; Keller et al., 2007; Marshak et al., 2000). Several studies have described the binding of hepatocyte nuclear factor 3 (HNF-3), or FoxA proteins to enhancers of Pdx-1 in mouse. Both Foxa1 and Foxa2 were shown to directly bind regulatory elements (Gao et al., 2008; Marshak et al., 2000; Wu et al., 1997). In addition, mutations of Pax6 and HNF3β (Foxa2) binding sites within the same

Pdx-1 regulatory region were shown to bind *in vitro*, and supported with the in vivo binding of Pax6 and HNF3β (Samaras et al., 2002). Mutagenesis of HNF-1a binding sites within *Pdx-1* regulatory elements has also shown that HNF-1a activates *Pdx-1* expression (Gerrish et al., 2001). HNF-1a, Foxa2, and SP-1 have been shown to cooperatively bind *Pdx-1* regulatory elements and activate *Pdx-1* expression *in vitro*, with cell lysate showing *in vivo* binding in cell culture (Ben-Shushan et al., 2001). Foxa2 has also been shown to interact with another cofactor SIRT1, and both factors cooperatively bind the *Pdx-1* promoter in mice, where SIRT1 deacetylates Foxa2, leading to the activation of the *Pdx-1* promoter (Wang et al., 2013).

One of the NK-families of ANTP-class homeobox proteins has also been shown to regulate *Pdx-1* expression, with Nkx2.2 binding and activating the mouse *Pdx1* promoter both *in vitro* and *in vivo* (Van Velkinburgh et al., 2005). In the mouse, Ptf1a has been shown in several studies to activate and maintain the early expression of *Pdx-1* via binding of an E-box/TC-box motif within the *Pdx-1* promoter (Fukuda et al., 2008; Miyatsuka et al., 2007; Wiebe et al., 2007). PPARα is another factor that has been shown to both lead to increased *Pdx-1* expression when overexpressed, and also bind the *Pdx-1* promoter *in vitro* (Sun et al., 2008). *RIPE3b1* and *MafA/MafB* appear to bind cooperatively and activate the *Pdx1* promoter *in vitro* and *in vivo,* with mutations inhibiting *Pdx1* expression in cell culture (Samaras et al., 2003). Another pair of *Pdx-1* activating factors are HCF-1 and its cofactor E2F1. These factors co-localise to the *Pdx-1* promoter and loss of *E2F1* within pancreatic beta cells causes downregulation of *Pdx1* (Iwata et al., 2013). USF is another factor shown to bind an E-box motif with the promoter of *Pdx-1*, and a reduction in USF greatly decreases Pdx-1 promoter activity within pancreatic beta cells (Qian et al., 1999; Sharma et al., 1996). GATA4 and GATA6 have been shown to be required for the formation of the pancreas, and bind and activate the *Pdx-1* promoter (Carrasco et al., 2012). Finally, though not showing direct regulation of Xlox itself, *Pdx-1* regulatory elements were shown to drive the expression of CRE-recombinase within the brain of rat, within cells of the hypothalamic nuclei, the dorsal raphe and inferior olivary nuclei (Song et al., 2010). This suggests that factors expressed within the brain must also be able to drive neural expression of *Pdx-1* in these cells. Though further investigation is required to identify these and examine neural expression of *Pdx-1*, this is the only study so far that examines regulatory elements of Xlox outside of a pancreatic role.

Currently there are no studies that have examined the direct regulation of *Gsx*, which is surprising given its conserved role in the patterning of the CNS. However, one promising factor is Pax6, which also regulates Xlox (Samaras et al., 2002), and has been shown to mutually repress each other within the forebrain (Corbin et al., 2003; Toresson et al., 2000). In addition, Gsx expression within the telencephalon of *Platynereis dumerilii* is down regulated within Azakenpaullone treated embryos (Tomer et al., 2010), which upregulates Wnt signalling (Schneider and Bowerman, 2007), perhaps indicating an involvement with Wnt signalling in the regulation of Gsx. The regulation of Gsx, both directly and indirectly is further discussed within chapter 5.

Aims

Many studies have characterised vertebrate ParaHox regulation, but currently only one study has looked at regulation of the entire ParaHox cluster (Osborne et al., 2009). This is also the only study to discuss the regulatory state of the ParaHox genes in the chordate ancestor. Understanding the regulation of the ParaHox cluster and of the amphioxus ParaHox genes, is likely to help explain why the ParaHox cluster has remained intact within not only the chordates, but the deuterostomes in general, and also has many applications in understanding vertebrate development and disease. The main aim of this study is to further characterise the regulation of the amphioxus ParaHox cluster, with particular focus on *AmphiGsx*, as this is by far the least characterised of the ParaHox genes in any phylum. As such, the focus of this thesis has been split into three sections.

- Use of new genomic and transcriptomic resources to better characterise the amphioxus ParaHox cluster, and carry out preliminary studies to allow better identification of ParaHox regulatory elements within amphioxus.
- To examine how retrotransposition has impacted the ParaHox cluster and the regulation of ParaHox genes, with particular focus upon the expression and regulation of the *SCP1* retrogene upstream of *AmphiGsx*.
- The in-depth characterisation of an *AmphiGsx* upstream cis-regulatory element to the same depth as many vertebrate and *Drosophila* cis-regulatory studies, using deletion analysis and the mutation of transcription factor binding sites to dissect regulatory function and highlight the power of amphioxus-*Ciona* cross-species transgenesis in the analysis of amphioxus regulatory elements.

Chapter 2. Materials and methods

2.1. Materials

2.1.1. Commercial Kits

Table 2.1. Commercial kits used and their suppliers

Kit	Manufacturer
Expand Long Template PCR System	Roche
Isolate genomic DNA minikit	Bioline
Isolate I PCR and Gel kit	Bioline
Isolate II PCR Gel kit	Bioline
Isolate RNA minikit	Bioline
mini Quick Spin Columns	Roche
Nucleobond Xtra Maxi	Macherey Nagel
pGEM-T-Easy Vector System	Promega
PeqGOLD plasmid minikit I	Peqlab
Tetro cDNA synthesis kit	Bioline

2.1.2. Reagents

Table 2.2. Chemicals and enzymes used and their suppliers

Chemical/Enzyme	Manufacturer
100bp DNA ladder	Bioline
1kb DNA ladder	Fermentas, Thermo Scientific
2-Mercaptoethanol	Sigma
Acetic Anhydride	Sigma
Agarose	Bioline, GibcoBRL
Ampicillin	Fisher Scientific
Anti-dioxigenin-AP, Fab fragments	Roche
Big Dye Terminator + Sequence Buffer	Sequencing facility, Zoology Oxford
Bovine serum albumin (BSA)	Sigma
BM Purple	Roche
$CaCl_2$	Fisher Scientific
Chloroform	Sigma
Denhardt’s solution (50x)	Invitrogen
DIG RNA labelling mix	Roche
Dithiothreitol (DTT)	Fisher Scientific
Dimethyl Formamide	Pierce
D-Mannitol	Acros Organics
DNase I	Fermentas
dNTP set	Fermentas
EDTA	Fisher Scientific
EGTA	Sigma
Ethidium Bromide	Sigma
Ethanol (molecular grade)	Sigma

Formamide (Fluka)	Fisher Scientific
Formamide (deionised)	Sigma
Gentamycin Sulphate	Fisher Scientific
Glycerol	Fisher Scientific
Glycine	BDH
HCl	BDH
Heparin	Acros Organics
Hepes	Acros Organics
IPTG	Fisher Scientific
Isopropanol	Sigma
KCl	Fisher Scientific
LB Agar	Sigma
LB medium	Sigma
$MgCl_2$	Sigma, Fisher Scientific
$MgSO_4$	Fisher Scientific
$MnCl_2$	Acros Organics
Methanol	Fisher Scientific
MOPS	USB, Sigma
Na_2SO_4	Fisher Scientific
NaCl	Fisher Scientific
$NaHCO_3$	Fisher Scientific
NaOH pellets	Fisher Scientific
NBT/BCIP	Roche
Oligonucleotides	Invitrogen. Designed by hand or using Primer3 (Koressaar and Remm, 2007; Untergasser et al., 2012)
Paraformaldehyde	TAAB
PBS, 10x	Fisher Scientific
PIPES	Fisher Scientific
Phenol/Chloroform/Isoamylalcohol	Fisher Scientific
Potassium Ferricyanide	Acros Organics
Potassium Ferrocyanide	Acros Organics
Protease	Sigma
Proteinase K	Sigma
Restriction Enzymes (with 10x Buffers)	Fermentas, Promega, New England Biolabs
RNase A	Ambion, Thermo Scientific
RNase T1	Ambion
RNase ZAP	Ambion
SDS	Fisher Scientific
Sheep Serum	Sigma
Sigmacote	Sigma
Sodium Acetate	Sigma
Sodium Thioglycolate	Sigma
RNA polymerase, T3, T7, SP6	Ambion, Thermo Scientific
RNA polymerase buffer, 10x	Thermo Scientific
RNasein	Promega
T4 DNA ligase (with 10x buffer)	Promega
TAE buffer, 50x	Fisher Scientific
Taq polymerase (with 10x NH_4 buffer, 50mM $MgCl_2$ solution)	Bioline

Total Yeast RNA	Sigma
Trietholamine	Acros Organics
Trireagent	Sigma
Tris Base	Fisher Scientific
Trisodium Citrate dihyrdate	Sigma
Triton X-100	Fisher Scientific
Tween 20	Acros Organics, Sigma
Water, RNase-free	Fisher Scientific
X-GAL	Fermentas

2.1.3. Specialised laboratory equipment

Table 2.3. Specialised laboratory equipment

4 well dishes for IVF	NUNC
Cell/Tissue Homogeniser	Savant Fastprep FP120 cell homogenizer - Thermo Savant
Dissecting microscopes	XTL-3T101 (GX microscopes), Olympus KL300 LED
Electroporation cuvettes (0.4mm)	Biorad
Electroporator	'Home made' (Zeller et al., 2006), BIORAD ShockPod, APELEX PS503 electrophoresis power supply
Hybridisation Ovens	UVP HL-2000 Hybrilinker, HYBAID shake 'n' stack
Imaging upright microscope (with software)	Leica LEITZ DMRB with Qimaging Retiga 2000R Fast1394 with RGB disc (Qcapture suite)
PCR Thermocycler	Techne TC-512
UV Transilluminator	UVP, BioDoc-It^{TM} imaging system- UVP
Spectrophotometers	Nanodrop ND-100, Amersham pharmacia biotech Ultraspec 3300 pro

2.1.4. Solutions

Table 2.4. Solutions made and their components

Solution	**Components**
1:3000 Anti-digoxigenin-AP	15mg amphioxus powder/Ciona embryos, 4ml 0.1% (v/v) Triton-X100 in PBS, heat at 70°C for 30min. Add; 500µl 20mg/ml BSA, 500µl pre-treated sheep serum, 5µl Boehringer anti-DIG Ab, mix and incubate overnight at 4°C with shaking. Add 9.95ml NaPBS, 50µl 20% (v/v) Triton-X100, 500µl pre-treated sheep serum. Make into 1ml aliquots and store frozen at -20°C

Amphioxus hybridisation buffer (HYB)	50% (v/v) formamide (deionised high quality), 100 µg/ml heparin, 5x SSC, 5mM EDTA, 1x Denhardts, 1mg/ml purified yeast RNA, 0.1% (v/v) Tween20.
Amphioxus wash buffer 1	50% (v/v) formamide, 5x SSC, 1% (v/v) SDS
Amphioxus wash buffer 2	50% (v/v) formamide, 2x SSC, 1% (v/v) SDS
Amphioxus wash buffer 3	2x SSC, 0.1% (v/v) Tween 20
Amphioxus wash buffer 4	0.2x SSC, 0.1% (v/v) Tween 20
Amphioxus wash buffer 5	1x PBS, 0.1% (v/v) Tween 20, 2 mg/ml BSA
AP+ buffer	0.1 M Tris-HCl pH 9.6, 0.05 M MgCl2, 0.1 M NaCl. Make fresh and filter through 0.22 µM filter.
AP- buffer (Mg free)	0.1 M Tris-HCl pH 9.6, 0.1 M NaCl. Make fresh and filter through 0.22 µM filter.
Blocking solution	20% heat treated sheep serum in 1x PBS, 0.1% (v/v) Tween 20, 2 mg/ml BSA
Ciona hybridisation buffer (HYB)	50% formamide, 5X SSC, 100µg/ml yeast RNA, 50µg/ml heparin, 0.1% Tween 20
Ciona wash buffer 1	50% (v/v) Formamide, 5x SSC, 1% (v/v) SDS
Ciona wash buffer 2	50% (v/v) Formamide, 2x SSC, 1% (v/v) SDS
Ciona wash buffer 3	2x SSC, 0.1% (v/v) Tween 20
Ciona wash buffer 4	0.2x SSC, 0.1% (v/v) Tween 20
CMF-ASWH (Calcium and magnesium free artificial sea water with Hepes)	463 mM NaCl, 11 mM KCl, 2.15 mM NaHCO3, 25.5 mM Na2SO4. Add H2O and then add Hepes pH 8.0 to 10 mM. pH to 8.0.
Dechorionation Solution	Mix equal volumes of Sodium thioglycolate (2% (w/v)) and Protease (0.1% (w/v)) and pH to 10.5 with 1M NaOH. All solutions need to be made fresh.
FSW (Filtered sea water)	Filtered sea water (50 µM filter, then further 1 µM from SERG tap), further filter sterilised through a 0.22 µM filter.
Glutaraldehyde (25%)	Sigma
Glutaraldehyde (0.2%) in CMF-ASWH	16⍰l Gluteraldehyde (25% (w/v)) + 1984 ⍰L CMF-ASWH
Glycerol 80%	80% (v/v) glycerol in ddH2O.
Glycine 0.2%	0.2% (w/v) glycine in 1x PBT
Glycine 10%	10% (w/v) glycine in ddH2O
LacZ Staining Buffer	1 mM MgCl2, 3 mM K4Fe(CN)6, 3 mM K3Fe(CN)6 (Keep Dark)
LB Agar	LB Agar 4% (w/v) LB Agar in dH2O. Autoclave.
LB Broth	LB Broth 2.5% (w/v) LB in dH2O. Autoclave.
MOPS 5X	0.5 M MOPS, 10 mM MgSO4, 5 mM EGTA, 2.5 M NaCl pH 7.5.
MOPS-PFA, 4%	4% (w/v) PFA in 1x MOPS. Dissolve in NaOH at 60°C and pH to 7.5 with HCl. Filter sterilise through a 0.22 µM filter.
NaPBS	For 50ml: 5ml PBS 10x, 0.45g NaCl, 45ml ddH_2O. (Autoclaved)
NaPBT	250µl 20% (v/v)Tween 20, 49.75ml NaPBS

NaPBS-PFA, 4%	4% (w/v) PFA in NaPBS. Add approximately 5-10 drops 1 M NaOH to help dissolving and rotate in hybridisation oven at 70°C until dissolved. Adjust pH to 7.5 with 1M HCL then add PBS up to 50ml. Make 1ml aliquots and store frozen at -20°C.
PBT	1x PBS, 0.1% (v/v) Tween 20
Protease (0.1%)	0.1% (w/v) Protease in FSW
Proteinase K	10 mg/ml in ddH2O
SDS 20%	20% (w/v) SDS in ddH2O. Heat to 68°C to dissolve.
Sheep Serum (Heat Treated)	100% sheep serum treated at 50°C for 30 min
Sodium Thioglycolate (2%)	2% (w/v) Sodium thioglycolate in FSW
SSC, 20X	3 M NaCl, 0.3 M Trisodium Citrate pH 7. Autoclave.
TE Buffer	10 mM Tris-HCl pH 7.5-8.0, 1 mM EDTA

2.2. Methods

2.2.1. General Laboratory protocols

General laboratory protocols were carried out in Dr David Ferrier's lab in the Scottish Oceans Institute at the University of St Andrews. General laboratory protocols were adapted from (Sambrook et al., 1989). Good chemical and microbiological practice was carried out at all times, and procedures involving genetically modified micro-organisms were followed according to the appropriate guidelines and specifications.

2.2.2. Polymerase Chain Reaction

Polymerase chain reaction (PCR) was set up on ice and carried out under sterile conditions. Reactions were set up with a total volume of 50µl in a 0.2ml PCR tube.

Table 2.5. Quantities used in general PCR reactions.

Component	Stock Concentration	Volume
NH_4 Buffer	10x	5µl
$MgCl_2$	50mM	1.5µl
dNTPs	10mM	2µl
Forward Primer	20µM	1µl
Reverse Primer	20µM	1µl
DNA template	Variable	Variable
Taq Polymerase	5U/µl	0.5µl
ddH_2O		Up to 50µl

Reactions were then carried out in a thermocycler using a PCR program with the following typical steps:

Initial Denaturation:	94-96°C	2 minutes	
Denaturation:	94-96°C	30 seconds	35 cycles
Annealing:	45-65°C	30 seconds	
Extension:	72°C	1-3 minutes	
Final extension:	72°C	7 minutes	
	4°C	Hold	

For the cloning of *Branchiostoma floridae Gsx* upstream regions, a high fidelity Pwo polymerase was used in order to ensure minimal mutations were introduced. This required a different program and PCR mix to that of Taq polymerase.

Table 2.6. Quantities used in High fidelity PCR reactions.

Component	Stock Concentration	Volume (Mix 1)	Volume (Mix 2)
20mM $MgSO_4$ Buffer	10x	-	5µl
dNTPs	10mM	2µl	-
Forward Primer	20µM	1µl	-
Reverse Primer	20µM	1µl	-
DNA template	Variable	Variable	-
Pwo Polymerase	5U/µl	-	0.5µl
ddH_2O		Up to 25µl	Up to 25µl

Mix 1 and Mix 2 were prepared separately on ice and then mixed to a total volume of 50 µl before placing in a thermocycler with the following program:

Initial Denaturation:	94°C	2 minutes	
Denaturation:	94°C	15 seconds	10 cycles
Annealing:	45-65°C	30 seconds	
Extension:	72°C	45 seconds-2 minutes	
Denaturation:	94°C	15 seconds	25 cycles
Annealing:	45-65°C	30 seconds	
Extension:	72°C	45 seconds-2 minutes + 5 seconds/cycle	
Final extension:	72°C	7 minutes	
	4°C	Hold	

For larger fragments, the Expand Long Template PCR System was used with either Buffer 1(<9kb) or Buffer 2 (9-12kb). Long range PCR reactions were set up on ice with a total volume of 50µl in 0.2ml PCR tubes.

Table 2.7. Quantities used in typical Long Template PCR reaction.

Component	Stock Concentration	Volume (Buffer 1)	Volume (Buffer 2)
Buffer	10x	5µl	5µl
dNTPs	10mM	7.5µl	10µl
Forward Primer	20µM	1µl	1µl
Reverse Primer	20µM	1µl	1µl
DNA template	Variable	Variable	Variable
Taq Polymerase	5U/µl	0.75µl	0.75µl
ddH_2O		Up to 50µl	Up to 50µl

Reactions were then carried out in a thermocycler using a PCR program with the following typical steps:

Initial Denaturation:	94°C	2 minutes	
Denaturation:	94°C	30 seconds	10 cycles
Annealing:	45-65°C	45 seconds	
Extension:	72°C	8 minutes +20 seconds/cycle	
Denaturation:	94°C	30 seconds	25 cycles
Annealing:	45-65°C	45 seconds	
Extension:	72°C	8 minutes +20 seconds/cycle	
Final extension:	72°C	7 minutes	
	4°C	Hold	

Annealing temperatures were adjusted according to the predicted melting temperatures for each primer pair, with extension times adjusted dependant on the expected PCR fragment size. Products were run on TAE agarose gels, at concentrations of 0.5-2% dependant on expected fragment size, with 0.5µg/ml of Ethidium bromide added to the gel before pouring. Gels were run in 1x TAE buffer. Band size was compared to an appropriate ladder, either 1Kb or 100bp, and visualised under UV light. For use in cloning, bands were cut using a sterile razor blade upon a UV trans-illuminator. Bands were then extracted using the Bioline Isolate I or Isolate II PCR and Gel extraction kit. Larger fragments arising from Long range PCR were then A-tailed for use in ligation reactions using a further 10 minute reaction at 72°C using the following. A-tailing reactions were set up on ice with a total volume of 10µl in 0.2ml PCR tubes.

Table 2.8. Quantities used in A-tailing reaction.

Component	Stock Concentration	Volume
Buffer	10x	2µl
$MgCl_2$	50mM	0.3µl
dATP	1mM	2µl
DNA template	Variable	1-4µl
Taq Polymerase (Bioline)	5U/µl	0.75µl
ddH_2O		Up to 10µl

Table 2.9. Commonly used Primer sequences.

Primer Name	Primer Sequence
M13F (Forward)	GTA AAA CGA CGG CCA GT
M13R (Reverse)	CAC ACA GGA AAC AGC TAT GAC CAT
T7	AAT ACG ACT CAC TAT AG
SP6	ATT TAG GTG ACA CTA TAG
pCES seq*	GTT TCC GCT TTG CTA CTG AA

*pCES seq is a primer designed to the pCES vector at the 5' end of the basal Cin-Fkh promotor, facing into the multiple cloning site.

2.2.3. Ligation

Purified PCR fragments were ligated into the pGEM-T-easy vector according to the manufacturer's instructions. *Branchiostoma floridae Gsx* upstream regulatory fragments were ligated into the *Ciona* electroporation construct vector pCES (kindly gifted by Dr Clare Hudson, CNRS, Villefranche sur Mer, France) (Harafuji et al., 2002) and primers for these regions contained a 5' PstI site on the forward primer and 3' BamHI site on the reverse primer to facilitate ligation into the appropriate restriction enzyme sites in the pCES multiple cloning site. pCES contains a multiple cloning site (MCS) upstream of a forkhead promoter coupled to a LacZ reporter gene. The following reaction volumes were used for ligation into the pCES vector.

Table 2.10. Quantities used in pCES Ligation reactions

Component	Volume
Linearised SAP* pCES vector	100-200ng
Insert DNA fragment	Variable
10x Ligation buffer	1µl
Ligase (3U/µl)	0.5µl
ddH_2O	Up to 10µl

*Shrimp Alkaline Phosphotase treated to prevent re-annealing of linearised vector.

Ligation reactions were carried out either for 1 hour at room temperature, or overnight at 4°C.

2.2.4. Transformation

5µl of ligation reaction was added to 50µl of competent cells, prepared according to (Inoue et al., 1990) and thawed on ice, and left on ice for 10 minutes. This was then transferred to a heat block and heat shocked at 42°C for 45 seconds, then transferred back to ice for 2 minutes. 200µl of LB-broth was then added and the cells allowed to recover at 37°C for 15 minutes. Cells were then spread onto pre-warmed LB-Agar plates containing 50µg/ml ampicillin, 200mg/ml Xgal and 20mg/ml IPTG and grown overnight at 37°C. In order to check colonies for the correct sized insert, stabs were taken directly from single colonies and placed into a PCR reaction mix with M13F and M13R primers. By analysing these PCR products, promising colonies were then picked and placed into 5ml of LB-broth with 50µl/ml ampicillin and grown overnight at 37°C with shaking. Glycerol stocks were prepared by adding 250µl of sterile 60% glycerol to 750µl transformed cell culture on ice. These were then vortexed briefly and snap frozen in liquid nitrogen and stored at -80°C.

2.2.5. Plasmid Purification

For use in sequencing, in situ probe synthesis and initial cloning of DNA fragments, the Peqlab Peqgold plasmid miniprep kit 1 was used according to the manufacturer's instructions, giving typical yields of between 100 and 200ng/µl. For larger concentrations of plasmid, glycerol stocks of pCES constructs were first streaked out upon LB-agar plates containing 50µl/ml ampicillin, 200mg/ml Xgal and 20mg/ml IPTG and grown overnight at 37°C. A starter culture of 4ml LB-broth containing 50µg/ml ampicillin was then inoculated with a single colony and grown at 37°C with shaking during the day. This was then used to seed 300ml of LB-broth containing 50µg/ml ampicillin and shaken overnight at 37°C. The Nucleobond Xtra maxi plasmid purification kit (Machary Nagel) was then used to purify plasmid from these 300ml cultures according to the manufacturer's instructions, giving typical plasmid concentrations of 1-3µg/µl.

2.2.6. Sequencing

Sequencing reactions were carried out upon purified plasmids using the common primers M13F/R for pGEM-T-easy, or SP6/pCES seq for pCES. They were then set up as follows in a 0.2ml PCR tube:

Table 2.11. Quantities used in sequencing reactions

Component	Volume
Plasmid DNA	2µl
Primer (3.2µl)	1µl
5x Buffer	2µl
Big Dye Terminator v3	1µl
ddH_2O	4µl

Sequencing reactions were next placed in the thermocycler using the following program:

96°C	3 minutes	
96°C	15 seconds	35 cycles
50°C	15 seconds	
60°C	4 minutes	
4°C	Hold	

Samples were then taken out of the thermocycler and transferred to a 1.5ml microfuge tube, with the following added; 1.5µl 3M Sodium Acetate, 31.25µl absolute Ethanol and 7.25µl ddH_2O. This reaction mix was vortexed and left in the dark for 20 minutes before spinning in a microcentrifuge at 4°C, 4,600rpm for 40 minutes. The liquid was then carefully removed, avoiding where the invisible pellet should be and 100µl of 70% Ethanol added. This was then spun for 20 minutes at 4°C, 4,600rpm, and the Ethanol carefully removed again and then air dried in the dark overnight. The finished reaction was then wrapped in foil and submitted for capillary sequencing at the Zoology sequencing service, Department of Zoology, University of Oxford.

2.2.7. Preparation of cDNA

Total RNA was extracted from mixed mid-gastrula to late-tailbud stage embryos (*C.intestinalis*), whole adults (*Branchiostoma lanceolatum*) in RNAlater or from adult gonads, muscle and notochord (*Branchiostoma lanceolatum*) and extracted using the Isolate RNA mini kit (Bioline) with the following modifications. An additional DNase I treatment was included during the protocol during the tissue lysis stage in order to fully remove genomic contamination, with 1µl of DNase I (Fermentas) added to Lysis buffer R and incubated at 37 °C for 30 mins. This was then heat-deactivated at 65°C for 10mins before being processed with the Isolate RNA mini kit (Bioline) according to the manufacturer's instructions.

Alternatively, phenol:chloroform extraction was carried out as follows. Adults in RNA later were dissected and tissue rinsed in RNase-free water several times, before being transferred to 1ml TriReagent on ice. Samples were then homogenised in a D-matrix tube in the Fastprep FP120 cell homogenizer (Thermo Savant) at Speed 6 (6m/s) for 40 seconds, then placed on ice. In the fume hood, the homogenate was then transferred to an RNase free microfuge tube, and 300µl of molecular grade chloroform was added and the sample vortexed for 15 seconds. The sample then separated into 3 phases: a lower pink phase (containing tissue remains, protein and chloroform and DNA), the transparent aqueous phase (containing RNA) and a superficial layer of lipids and denatured protein. This was then centrifuged for 15 minutes at 13000rpm at 4°C. The aqueous phase was then recovered, taking care not to touch the pink phase. This step (adding 300µl of chloroform and centrifuging) was repeated several times until the interface between the Chloroform and aqueous phase was free of white material (denatured protein). 500µl of isopropanol was then added to the aqueous phase and the tube was then inverted 30 times, then vortexed for 15 seconds. The sample was then incubated at RT for 15 minutes. The sample was then centrifuged for 15 minutes at 13000rpm at 4°C. A pellet should form, though is sometimes transparent, and the isopropanol removed by pipetting. 4 washes of EtOH 70% (kept at -20°C) were carried out with the Ethanol removed completely on the final wash. The tube was then left open in the laminar-flow hood and left to dry completely. The RNA pellet was then resuspended in RNAse free water and stored at -80°C for long term storage. An aliquot was also taken to store at -20°C for short-term use. cDNA was then produced from RNA samples using the Tetro cDNA synthesis kit (Bioline) as instructed by the manufacturer, using OligodT's to prime the reaction.

2.2.8. Genomic DNA extraction

gDNA was extracted from either a single whole adult (*B. floridae* & *B. lanceolatum*) or from sperm (C. intestinalis) using the Bioline Isolate genomic DNA Minikit. In the case of whole adults, tissue was first cut into small pieces, and all samples were transferred to Lysis buffer in an MP Biomedicals Lysis Matrix D tube. Samples were then homogenised using a Fastprep FP120 cell homogenizer (Thermo Savant) at 6m/s for 40 seconds. The resulting homogenate was then removed from the D-matrix tube and processed according to the Bioline Isolate genomic DNA Minikit manufacturer's instructions. Alternatively, Phenol:Chloroform:Isoamylalcohol (25:24:1 (v/v)) extraction was carried out. Samples were lysed by adding 200µg proteinase K and 10µl of 10mg/ml RNase A to samples in PBS, or straight to the Ciona sperm sample, and digested for 2 hours at 50°C with occasional gentle swirling. This was inverted gently several times to ensure mixing and centrifuged at 13,000rpm for 5 minutes. The supernatant was transferred to a new microfuge tube and 500µl of phenol:chloroform:isoamylalcohol added and gently inverted to mix. This step was repeated until the protein was fully removed and the interphase was clear of white protein remains. This was then followed by a 500µl chloroform wash and mixed by inversion and centrifuged at

13,000rpm for 5 minutes. This was repeated once and then the top aqueous phase was removed and 2 volumes of absolute ethanol (-20°C) added. This was mixed gently and stored at -20°C for >1 hour to precipitate the gDNA. gDNA was then hooked out with a sterile glass hook and transferred to a new microfuge tube and washed with 300µl 70% Ethanol. The ethanol was then removed and the gDNA allowed to air dry in a laminar flow hood overnight at RT. Finally the dry gDNA was dissolved in ddH_2O.

2.2.9. Antisense RNA Probe synthesis

PCR templates were first synthesised from pGEM-T-easy clones using M13 primers and a general PCR program. These were then run on a 1% agarose gel and purified using the Isolate I or Isolate II PCR and Gel extraction kit. RNA run-off transcription was then carried out, as set up in table (2.12), and incubated at 37°C for 2 hours.

Table 2.12. Quantities used in antisense RNA probe synthesis.

Component	Concentration	Volume
M13 Template	1-2µg	Variable
RNA polymerase buffer	5x	4µl
DTT	0.1M	2µl
Dig-rUTP label mix	10x	2µl
RNasin (RNase inhibitor)	40U/µl	1µl
RNA polymerase (T7 or SP6)	20,000U/ml	1µl

Synthesised probes were then checked by running 1µl on a 1% agarose gel before purification. If probes were made from a particularly long template (approximately 1kb or larger) then the resulting probe was partially hydrolysed for 15 minutes at 60°C using the following mix:

Table 2.13. Quantities used in RNA probe hydrolysis.

Component	Concentration	Volume
RNA	Variable	Variable
Na_2CO_3	200mM	30µl
$NaHCO_3$	200mM	20µl
RNase-free ddH_2O	-	Up to 100µl

The extent of hydrolysis was then checked on a 1% agarose gel before probe purification. Probes were purified using mini Quick spin columns (Roche) according to the manufacturer's instructions and mixed in a 50:50 dilution with deionised formamide.

The *Ci-TCF/Lef* clone used is a 641bp segment covering part of the 5' end of the *Ci-TCF/Lef* mRNA, cloned from gastrula to late tailbud stage *Ciona intestinalis* embryonic cDNA (Genbank accession number KP739765). The *Ci-Gsx* clone was obtained from the cDNA library originally obtained by (Imai et al., 2004), with the location R1CiGC31m18. *Branchiostoma floridae* SCP1 was amplified from the *Branchiostoma floridae* PAC clone -33B4 (Genbank accession number AC129948.3) by PCR, whilst *Branchiostoma lanceolatum* SCP1 was cloned from whole adult cDNA. *Branciostoma lanceolatum* Xlox UTR was cloned from B.la adult gonad cDNA, whilst *Branchiostoma floridae* Xlox UTR was cloned from *Branchiostoma floridae* genomic DNA extracted from a single whole adult.

Table 2.14. Primers used to clone fragments for In situ hybridisation probes

Name	Sequence	Melting temperature (°C)
Ci-TCF F	CAGAGATTCCAGCCACAGAAGT	60.03
Ci-TCF R	TGGTTTCTTCACATATGGCCGA	60.03
PAC7 (F)(B.fl SCP1)	CAGTTTGCTATTGCTTGTGAGTGT	53.1
PAC8 (R)(B.fl SCP1)	GAAGAAGCCAAAAACAGTATC	51.6
Amphi-SCP1 F (B.la SCP1)	GCAGGTGTRTYATCAGCAAGAG	59.90
Amphi-SCP1 R (B.la SCP1)	ACTCRAAGAAGCCAAAAACAGT	59.46
Xlox-ncRNA F (B.fl Xlox UTR)	GAACAAGAGAACGCGCACAG	60.11
Xlox-ncRNA F (B.fl Xlox UTR)	TGTCCTGTTCACGCGTAGTC	60.04
ncRNA1degen F (B.la Xlox UTR)	GATAAAGAGCTCGGTACATCCCTAG	60.11
ncRNA1degen R (B.la Xlox UTR)	TTCTRATACACTTWWACAACAGGCA	58.94

2.2.10. Animal husbandry

Wild *Ciona intestinalis* were collected from two different sites; Croabh Haven, Scotland during May to July and from Arbroath, Scotland in August and September. The differing seawater temperatures at the two sites allowed us to extend the season in which embryos could be collected. Animals were collected from pontoons located in a marina (Croabh Haven) or a small harbour (Arbroath) and then maintained in a flow-through aquarium system with seawater pumped in directly from the North Sea, filtered and pumped into 50L tanks with aeration. Water was allowed to drain out of the tanks and replenished at a constant rate with fresh seawater, keeping a steady flow across the tank. Animals were kept submerged within plastic baskets held up by polystyrene floats, to allow waste to fall to the bottom of the tank rather than it accumulating around the animals. Some food entered as a constant flow of algae provided from the seawater inflow, and this was supplemented once a day with a mixed suspension of *Rhinomonas reticulata* var. *reticulata* (strain number CCAP 995/2), a unicellular red algae, supplied by the Scottish Association for Marine Science (SAMS, Oban, Scotland), and *Tetraselmis sp.*, a unicellular lipid-rich green algae, supplied by Florida Aqua Farms (Dade City, Florida, United States). These algae were grown in culture and concentrated by low-speed centrifugation before 5ml of mixed, concentrated culture was added to the tank by pipette. All in-flow and out-flow of seawater was stopped for 2 hours whilst the animals were fed the algal mix. The presence of a dark green colouration in the gut was used as a visual cue to indicate successful feeding. In order to collect gametes, gravid animals, as distinguished by an abundance of pink/orange eggs visible through the body wall, were selected and gametes liberated by dissection.

2.2.11. Collection and treatment of embryos for *In situ* Hybridisation

2.2.11.1. C. intestinalis

The eggs from 3 to 4 adults were dissected into separate petri-dishes filled with filtered sea water (FSW), whilst sperm was collected from each animal into separate microfuge tubes containing 1ml of FSW and kept on ice. Eggs and sperm were then mixed in a beaker for 10 minutes and then washed through a 100µm mesh filter to wash off excess sperm. Eggs were washed in FSW three times in order to clean excess sperm off, taking care to keep them covered in FSW at all times, and then transferred to a 15ml tube with no more than 2ml of FSW. Fertilised eggs were then dechorionated using 2% sodium thioglycolate and 0.1% protease, prepared separately and then mixed prior to dechorionation. Dechorionation solution was added to the eggs and allowed to sit for 2 minutes in a 15ml tube, then a small sample taken and checked every 30 seconds in a petri dish under a benchtop microscope to check for dechorionation of 50% of fertilised eggs. Dechorionation times varied for animals from different locations, with zygotes from Croabh Haven requiring 3-3.5 minutes and zygotes from Arbroath 6-8mins. These were then washed with FSW several times in a 15ml falcon, gently spinning the zygotes down via a hand centrifuge for no more than 2 minutes in between washes, before transferring to pre-prepared 1% agarose-FSW coated petri-dishes containing approximately 50ml of FSW. 5ml of 100µg/ml Gentamycin in FSW was then added to prevent bacterial growth and embryos allowed to develop at 16°C until the stage required. Embryos were then fixed with 4%-PFA:MOPS solution overnight at 4°C. These were then washed twice in PBT with rocking for 10 minutes, and once in 70% EtOH with rocking for 10 minutes before storing in a 1.5ml microfuge tube in 70% EtOH. All microfuge tubes, 15ml tubes and Pasteur pipettes used in this protocol were silicon-coated using Sigmacote-SL2 (Sigma) to avoid embryos sticking to equipment.

2.2.11.2. B.lanceolatum

Adult *Branchiostoma lanceolatum* were collected by Dr David Ferrier and Clara Coll Lladó at the facilities of Laboratoire Aragó in Banyuls-sur-mèr, France, in 2010. Adult animals were also transferred straight to RNAlater for use in RNA extraction or preserved in absolute ethanol for use in gDNA extraction. Embryos were collected by spawning of ripe amphioxus. These were induced by heat stimulation according to (Fuentes et al., 2007) and embryonic stages (gastrula, early neurula, mid-neurula, late neurula and early larval stage) were collected at regular intervals and fixed in MOPS-PFA, 4% for 1 hour at room temperature or overnight at 4°C. After fixation, embryos for whole-mount *in situ* hybridization (WMISH) were washed three times in 70% ethanol and stored in 70% ethanol at -20°C. Larvae (first gill slit stage) and juvenile amphioxus from previous spawnings were kindly provided by Dr. Héctor Escrivà and Dr. Stéphanie Bertrand. Both developmental stages were fixed and stored in 70% ethanol following the same procedure described for the embryonic stages. Late neurula and early larval stages were kindly donated by Dr. Ildikó Somorjai to complete the amphioxus developmental series.

2.2.11.3. B. floridae

B.floridae adults were collected from Tampa Bay, Florida using a shovel and sieve by Tom Butts and Peter Osborne in 2006. Embryonic stages were collected according to Holland and Yu (2004) and fixed for 1 hour at room temperature or overnight at 4°C in MOPS-PFA, 4%. Embryos were then washed multiple times in 70% Ethanol and stored in 70% ethanol at -20°C. Adults were also preserved in absolute ethanol for use in gDNA extraction.

2.2.12. *In situ* Hybridisation

2.2.12.1. C.intestinalis

In situ hybridisation was carried out as detailed in (Wada et al., 1995) with the following modifications. Embryos were rehydrated through an ethanol series into PBT and then digested for 10 minutes at room temperature in 2µg/ml proteinase K for gastrula to mid-tailbud embryos and 20 minutes for late-tailbud embryos. 4µl of 10% glycine was then added, swirled and the solution removed immediately and replaced with 10% glycine in PBT and washed for 5 minutes. This was then changed for 4% PFA in PBS and fixed for 1 hour at room temperature. After triethanolamine/acetic anhydride washes, embryos were washed three times in PBT before being washed once in 50:50 Hybridisation buffer (HYB) to PBT, then once in Ciona-HYB. This was then changed to fresh Ciona-HYB and embryos were pre-hybridised at 60°C for 3 hours. Approximately 50-100ng of antisense RNA probe in fresh Ciona-HYB was denatured at 70°C for 10 minutes before being added to the embryos. Embryos were then incubated at 70°C for 2 minutes before being moved to an overnight hybridisation at 60°C, rocking gently. Hybridised embryos then underwent 3x

20 minute washes in *Ciona* Wash buffer 1 at hybridisation temperature. This was followed by 2x 20 minute washes at 37°C in *Ciona* wash buffer 2 and then 2x 5 minute washes at room temperature in *Ciona* Wash buffer 3. 3x 5 minute washes at 37°C were followed by 2x 20 minute washes at 50°C, all in *Ciona* Wash buffer 3. A single wash for 20 minutes at 50°C in *Ciona* wash buffer 4 was carried out before 3x 10 minute washes in PBT at room temperature. Embryos were then blocked for 3 hours in blocking solution before adding 1:3000 Anti-digoxigenin-AP Fab fragments in blocking solution and incubating O/N at 4°C. No modifications were made to the staining and post-staining procedures.

2.2.12.2. Amphioxus

In situ hybridisation on *B.floridae* and *B.lanceolatum* embryos was carried out according to (Holland et al., 1996) with the following modifications. Amphioxus embryos were rehydrated through an ethanol series into PBT and then digested for 5 minutes at room temperature in 2µg/ml proteinase K for mid-gastrula to early-neurula embryos, 10 minutes for mid neurula-late neurula embryos and then 15 minutes for premouth embryos and 2-day larvae. After triethanolamine/acetic anhydride washes, embryos were washed once in PBT for 1 minute rotating, then again in PBT for 5 minutes rotating. This was then changed for 100µl of HYB buffer, pre-warmed to 60°C, and rotated for 1 minute. This was then changed for fresh HYB and rocked in the hybridisation buffer for 2 hours. Antisense RNA probe was mixed in 1/200-1/50 dilutions in fresh warm HYB and then denatured at 70°C for 10 minutes, before being added to the embryos. These were then rocked overnight at either 60°C or 62°C in the hybridisation oven. No modifications were made to the day 2 until blocking. RNase steps were carried out with 2µl 10mg/ml RNaseA and 1µl RNaseT1 (10,000U/ml) in 1 ml of Wash solution 3 and 250µl added per well. Wash solution 5 was then removed and 200µl of blocking solution was added to the embryos and rotated for 3 hours at room temperature. Blocking solution was then removed and 1:2000 Anti-digoxigenin-AP Fab fragments in blocking solution were added to the embryos and left incubating overnight at 4°C. On day 3, embryos were washed 4 times in NaPBT for 20 minutes each at room temperature, before 3 washes in AP- followed by 3 washes in AP+. AP+ was then exchanged for staining buffer and embryos were left in the dark at RT for the colour to develop. Signal typically came up overnight, but could take up to 4 days dependant on probe concentration. The final post-staining procedure consisted of 3 washes in AP- for 10 minutes each, rotating and kept dark, followed by 3 washes in NaPBT for 10 minutes each, rotating and kept dark. Embryos were cleaned during the NaPBT washes. They were then fixed in 4%PFA in NaPBS for 1 hour at RT. Finally embryos were washed twice in NaPBT for 10 minutes each before being transferred to 80% glycerol to clear.

2.2.13. Electroporation and *C. intestinalis* transgenes.

Electroporation was carried out using a custom-built electroporator based on the details provided by (Zeller et al., 2006). This was used with settings at 50V, 1000F and either 30 or 40Ω. *C. intestinalis* transgenics were produced by electroporation of fertilised eggs according to (Corbo et al., 1997), with the following modifications. All microfuge tubes, 15ml tubes and Pasteur pipettes used in this protocol were silicon-coated using Sigmacote-SL2 (Sigma) to avoid embryos sticking to equipment. 40-50µg of plasmid DNA was dissolved in 500µl of 0.77M D-Mannitol. Fertilised eggs were dechorionated using 2% sodium thioglycolate and 0.1% protease, prepared separately and then mixed prior to dechorionation. Dechorionation times varied for animals from different locations, with embryos from Croabh Haven requiring 3-3.5 minutes and embryos from Arbroath 6-8mins. These were then washed with filtered seawater several times, gently spinning via a hand centrifuge for no more than 2 minutes in between washes. No more than 200µl of embryos were added to a microfuge tube and the 500µl DNA/Mannitol mix added. This was mixed by gently pipetting and added to a microcuvette (BioRad, electrode width=0.4cm). After the pulse, the DNA/eggs were immediately transferred to a seawater/agarose-coated plate flooded with filtered sea water and 5ml of 100mg/ml Gentamycin. Embryos were then reared at 16°C until the desired time point and fixed for 30 minutes in Glutaraldehyde (0.2%) in CMF-ASWH, then washed 2 times for 10 minutes each in PBT with shaking, then once in LacZ staining buffer for 10 minutes in the dark with shaking. Embryos were then transferred to fresh LacZ staining buffer and allowed to stain overnight at 37°C in the dark. Embryos were then washed three times in PBT and fixed in MOPS-PFA, 4% for 1 hour at room temperature. All constructs were tested in triplicate in separate electroporations, with positive controls (known active constructs) used as well as pCES lacking any amphioxus DNA as a negative control.

2.2.14. Site Directed Mutagenesis

Site directed mutagenesis (SDM) on TCF/Lef sites was carried out using the Phusion SDM kit (Thermo Scientific) to introduce double point mutations. SDM was carried out using RP-HPLC purified 32bp primers with 5’ phosphorylation modifications according to the manufacturer’s instructions. SDM was performed on the pGEM-T-easy copy of each cloned regulatory region, with inserts dropped out via restriction digest and then ligated into the pCES vector. Plasmids underwent a 1:1000 dilution prior to use in the SDM reaction.

Table 2.15. Components used in SDM reaction.

Component	Concentration	Volume
Phusion HF Buffer	5X	10µl
dNTPs	10mM	1µl
Primer A	20µM	1µl
Primer B	20µM	1µl
Plasmid template	Variable	1µl
Phusion Hot start II DNA polymerase	2U/µl	0.5µl
ddH$_2$O	-	Up to 50µl

Table 2.16. Primers used for SDM.

Name	Sequence	Melting Temperature (°C)	Primer Modification
TCF/Lef mut site1F	TGGCAAGAACTGAAAAATTGTTATTCCGTGTT	67	5’ Phosphorylation
TCF/Lef mut site1R	CGCTTTTATACCTGCTCGGACCTTTACTGCTC	72	5’ Phosphorylation
TCF/Lef mut site2F	CGTCGCGTCGAACGCAATTGTGAAGTCCACGT	77	5’ Phosphorylation
TCF/Lef mut site2R	TATCTCTTCATTCGGTGCTGCATACAATTAAC	67	5’ Phosphorylation
TCF/Lef mut site3F	AATCACTAAGGTAGGAATTGATGAAGTCTGCG	68	5’ Phosphorylation
TCF/Lef mut site3R	ATCATTTATACTTGGAAGACATCGTTTCACGG	67	5’ Phosphorylation

The following PCR program was used to carry out the SDM reaction.

Initial Denaturation:	98°C	30 seconds	
Denaturation:	98°C	10 seconds	25 cycles
Annealing:	65-72°C	30 seconds	
Extension:	72°C	1.5-3 minutes	
Final extension:	72°C	7 minutes	
	4°C	Hold	

Ligation and transformation were then carried out according to the standard protocol described earlier.

Chapter 3. Annotation and regulatory analysis of the amphioxus ParaHox cluster.

3.1. Introduction

Genome sequences can form the basis for preliminary understanding of regulatory mechanisms, which in the case of the ParaHox genes have important implications for the evolution of development. Though the *B.floridae* genome has been available for a few years now (Putnam et al., 2008), there remain large gaps within scaffolds as well as poor gene identification via automated methods, with little experimental conformation of gene models. This is particularly evident within the *B.floridae* ParaHox cluster, where the cluster is present upon multiple scaffolds, (namely Scaffolds 24 and 116, http://genome.jgi.doe.gov/Brafl1/Brafl1.home.html), and there are large expanses of poor coverage and assembly artefacts, which have made annotation difficult. This has made the use of these scaffolds difficult for the identification and interpretation of regulatory mechanisms and features. As of yet, detailed genome-wide regulatory analyses such as that available from ENCODE (Thomas et al., 2007), are not available for non-standard model organisms such as amphioxus. However, we can draw from such data available in the vertebrates to aid us in the characterisation of regulatory phenomena at work in amphioxus. These studies are important to elucidate the state of regulatory phenomena in the last common ancestor of chordates and give insights into the evolution of regulatory phenomena within the chordates, as well as those that may have been conserved throughout the evolution of the bilaterians and perhaps even deeper into metazoan evolution.

Within the vertebrates, several studies have identified conserved non-coding elements (CNEs) that not only show deeply conserved sequence, but also show enhancer activity when tested in reporter gene assays (Bhatia et al., 2014; Dermitzakis et al., 2004; Pennacchio et al., 2006; Ray and Capecchi, 2008; Woolfe et al., 2007; Woolfe et al., 2005). Whilst conserved sequence can be identified, even across the vertebrates, only 45% of these putative CNEs show regulatory activity, at least in the assays used (Pennacchio et al., 2006), showing that computational approaches alone do not suffice to predict regulatory function and functional assays must be used to corroborate such data. Despite this, bioinformatic approaches have been useful in highlighting that many of these CNEs are linked to developmental genes (Bejerano et al., 2004; Sandelin et al., 2004; Woolfe et al., 2005), and sequence conservation can be as high as 95% between chicken and mammals over 200bp (Bejerano et al., 2004).

Though most studies are limited to the vertebrates, an increase in genomes has allowed a number of studies to identify CNEs within nematode, fly and tunicate genomes, again showing the same pattern of CNEs clustering around developmental genes (Glazov et al., 2005; Kim et al., 2007a; Siepel et al., 2005), and all animals studied so far have shown this trend, with strong purifying selection occurring in different animal groups (Vavouri and Lehne, 2009). Evidence suggests that CNEs may be in part responsible for the maintenance of gene clusters, such as the Hox genes, in those lineages where intact gene clusters and conserved synteny is observed. In this case, regions spanning sets of enhancers and associated genes, as well as bystander genes whose intergenic regions contain CNEs, are conserved as a genomic regulatory block (GRB) (Kikuta et al., 2007). Both the Hox and ParaHox clusters are also found in the Kikuta lists of GRBs, highlighting the complex suite of genomic regulatory elements present within these two related gene clusters.

With the publication of the first amphioxus genome (*B.floridae*) (Putnam et al., 2008), attempts have been made to draw the comparisons seen in vertebrates out to the invertebrate chordates. The first attempt to compare amphioxus non-coding regions to those of vertebrates focused on the Hox cluster, and found that the 3’ of the Hox clusters, where anterior genes are located, showed more conservation than the rest of the cluster, but non-coding sequences showed little conservation (Amemiya et al., 2008; Manzanares et al., 2000). One difficulty when looking for CNEs across the chordates, including the invertebrate chordates, is that the whole genome duplications of vertebrates mean that regulatory elements have been differentially lost across the vertebrate paralogues. This was identified in a further study when searching for CNEs between the vertebrates and the amphioxus Hox cluster, where a few CNEs were identified in the amphioxus Hox cluster, but their locations were spread out across the four vertebrate Hox loci (Pascual-Anaya et al., 2008). This was also seen in another study, which found two CNEs conserved between the 3’ ends of the vertebrate A & B Hox clusters and amphioxus, again suggesting differential loss (Matsunami et al., 2010). In an attempt to functionally validate some of these amphioxus CNEs, those surrounding *AmphiHox4* were examined in zebrafish, showing expression in the hindbrain, spinal cord, pharyngeal arches and pectoral fins (Punnamoottil et al., 2010). 56 potential CNE’s were also identified within the amphioxus genome paper (Putnam et al., 2008) and several of these tested within reporter constructs in both mouse and amphioxus embryos (Holland et al., 2008). These amphioxus studies are also reviewed in Beaster-Jones (2012). Now that access is available to multiple amphioxus genomes, such approaches examining regions of non-coding conservation to identify regulatory elements may be possible within the Cephalochordata. This would both streamline the identification of regulatory elements in amphioxus and remove the obstacles associated with genomic comparisons across the genomic duplicates of vertebrates, providing an amphioxus regulatory map that can then be experimentally tested. Such an approach has not yet been attempted for the ParaHox cluster.

The identification of amphioxus regulatory elements can also be attempted via looking for conserved transcription factor binding sites and motifs. One promising candidate for regulatory input within the amphioxus ParaHox cluster is CTCF, or CCCTC-binding factor. Indeed, genome-wide mapping of CTCF elements in humans identified that the distribution of CTCF binding sites correlated with genes, but not with transcriptional start sites (Kim et al., 2007b), suggesting an association with regulatory elements. One 20bp CTCF-binding motif, LM2, was found within ~15000 CNEs within the human genome (Xie et al., 2007), implicating these sites within regulatory regions conserved across species.

CTCF has been shown to carry out a wide range of crucial functions in the regulation of genes, and has been shown to have a wide range of functions. It can act directly, as a positive or negative regulator of transcription by binding regulatory elements of target genes (Filippova, 2008; Phillips and Corces, 2009), and also as a mediator of long-range chromatin interactions and insulator protein (Phillips and Corces, 2009; Zlatanova and Caiafa, 2009). Knockout experiments highlighted the crucial role CTCF plays in gene regulation, with dramatic and widespread effects upon gene regulation. Depletion of maternal CTCF caused the widespread misregulation of genes within mouse oocytes (Wan et al., 2008), and knockouts of CTCF in mice were lethal at the pre-implantation stage (Heath et al., 2008; Splinter et al., 2006).

CTCF was first identified as a factor that binds the *Myc* promoter (Lobanenkov et al., 1990), and is best known for its role in insulator elements in vertebrates, having been originally shown to block enhancer function at the chicken *beta-globin* locus (Bell et al., 1999). CTCF is also conserved through to *Drosophila* (Moon et al., 2005). This enhancer-blocking activity of CTCF was subsequently shown to be present in other loci, acting via a chromatin methylation-sensitive interaction within the imprinting region of the *H19/Igf2* locus (Bell and Felsenfeld, 2000; Fedoriw et al., 2004; Hark et al., 2000). This role of CTCF within insulator elements is thought to function via the formation of chromatin loops, where CTCF is able to interact both with DNA and itself, bringing remote CTCF factors bound at different locations together physically, looping the DNA in between (Yusufzai et al., 2004). This looping of DNA by insulator elements, first identified in the *Drosophila gypsy* element (Gerasimova et al., 2000), would prevent enhancer-promoter interactions between elements on different chromatin loops, while allowing or even facilitating those interactions within the same loop (Engel and Bartolomei, 2003; Wallace and Felsenfeld, 2007). CTCF is thought to mediate this looping mechanism through interactions with the co-factor cohesin, which may use its ATP-ase activity to extrude a loop between CTCF binding sites (Alipour and Marko, 2012; Strick et al., 2004). In addition, transcriptional activity may contribute to this cohesin-mediated translocation of DNA (Lengronne et al., 2004) (See figure 3.1 for a model of CTCF function).

In addition to its role in the function of insulator elements, high resolution profiling of histone methylation domains also indicates that CTCF sites function as boundary elements, marking the boundaries between different histone methylation domains (Barski et al., 2007). In keeping with this, recent research has shown the importance of CTCF sites in the formation of topological associating domains (TADs) (Gomez-Marin et al., 2015; Guo et al., 2015; Nora et al., 2012; Vietri Rudan et al., 2015) in a variety of genomic contexts, where TAD borders are designated by CTCF sites with diverging orientations. In these cases, TADs are defined where CTCF sites are able to interact and loop DNA, whereas TAD boundaries represent the regions in-between two looping CTCF-mediated chromatin structures, formed by opposing orientation of CTCF sites. The conserved formation of CTCF-mediated TAD boundaries in the Six locus gives strong evidence that CTCF sites are functioning via identical mechanisms and contexts over vast evolutionary distances, in this case within the same gene cluster in both sea urchins and vertebrates (Gomez-Marin et al., 2015). This conservation of CTCF function may even extend deeper, with a conserved CTCF-Hox 'kernal' identified within the Bilateria (Heger et al., 2012) and a consensus position weight matrix established to identify CTCF binding sites throughout the Bilateria. A summary model of CTCF function within the formation of TADs, insulation and enhancer facilitation is given in figure 3.1.

The deeply conserved association of CTCF-Hox is one that holds particular interest with regards to the regulation of the ParaHox cluster, as we can look to highly conserved Hox mechanisms to begin teasing apart potential shared regulatory phenomena between these two sister clusters. In vertebrates, *in vivo* occupied CTCF binding sites have been shown to be present in all human and murine Hox clusters (Birney et al., 2007; Soshnikova et al., 2010). This is in turn supported by chromosome conformation capture data suggesting that CTCF binding of sites influences chromatin architecture during development (Ferraiuolo et al., 2010). Indeed, recent evidence has shown that CTCF binding sites are involved in the formation of discrete functional domains within the Hox cluster between ESCs and Motor Neurons, and loss of CTCF alters the topological architecture within the *HoxA* locus (Narendra et al., 2015). Here, functional CTCF binding results in the formation of a TAD covering the anterior Hox cluster, whereas mutation of these binding sites results in the caudal spread of this topological domain. It is possible that this 3' topological domain may also extend to amphioxus, as evidence suggests a much higher level of conservation within the 3' of the amphioxus Hox cluster (Matsunami et al., 2010; Pascual-Anaya et al., 2008), as well as between the amphioxus and vertebrate Hox clusters. This pattern of Hox conservation is exemplified in a phenomenon termed 'Deuterostome posterior flexibility' (Ferrier et al., 2000), where deuterostome posterior Hox genes show a higher rate of evolution than their anterior and medial counterparts and protostome orthologs and may represent a functional constraint upon anterior Hox genes. As such, the presence of CTCF-based regulatory mechanisms in the amphioxus ParaHox cluster has the potential to tell us much about the evolution of both Hox and ParaHox clusters.

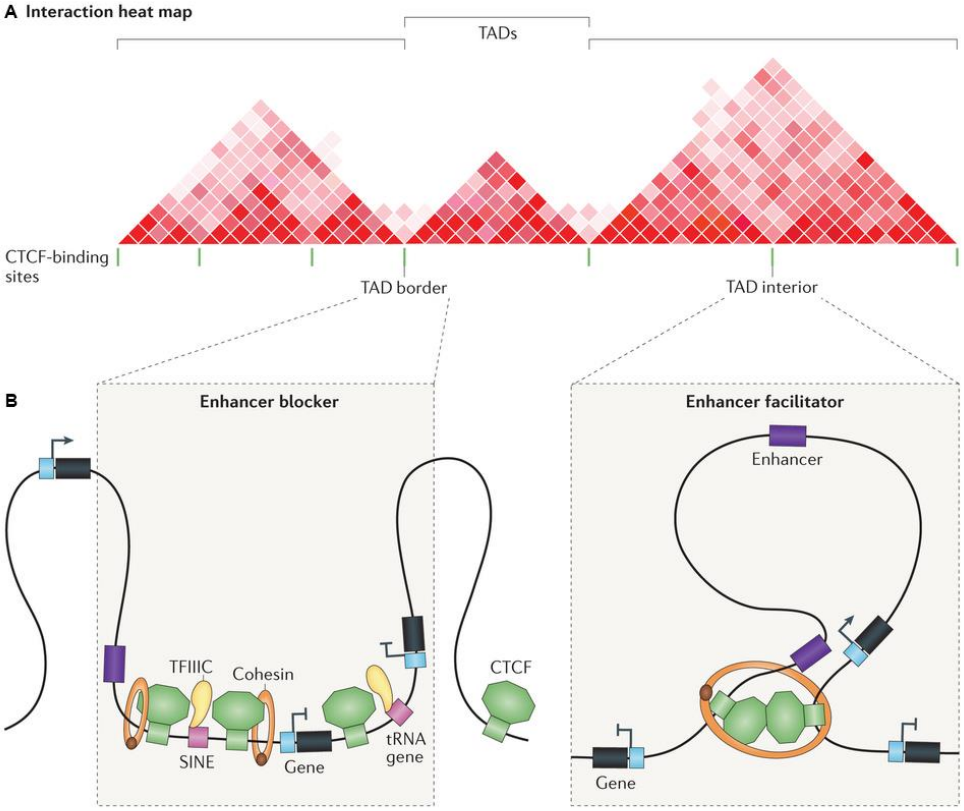


Figure 3.1. Schematic showing the function of CTCF in the formation of TADs, insulators and in facilitating enhancer-promoter interaction.

(A) Schematic of data generated by Chromatin conformation capture techniques used to identify topologically associating domains (TADs). The TADs and their borders are indicated. (B) The presence of multiple binding sites for CCCTC-binding factor (CTCF) and TFIIIC at TAD borders may contribute to the establishment of the border. This arrangement may provide an explanation for the observed function of CTCF as an enhancer blocker. Conversely, CTCF-binding sites within TADs may facilitate enhancer–promoter looping through the recruitment of cohesion and looping of chromatin. The blue box denotes the promoter of the gene. Figure adapted from (Ong and Corces, 2014).

Whilst transcription factors such as CTCF may be involved in the regulation of genes, there are genomic elements which can prove disruptive to such regulation. Transposable elements are one such element that have proven to be important to genomic evolution, and have been shown to have an ability to cause genomic rearrangements (Kidwell, 2002). As such, there has been much interest in the presence of transposable elements in the Hox and ParaHox clusters. The Hox clusters of gnathostomes are relatively devoid of transposable elements (TEs), and TEs that are present are found invading from the ends of Hox clusters (Fried et al., 2004). In concurrence with this gnathostome data, the large intact Hox clusters of beetles, bees and amphioxus also show exclusion of repetitive and transposable elements (Amemiya et al., 2008; Dearden et al., 2006; Shippy et al., 2008). Interestingly, the disrupted protostome Hox clusters of *Drosophila, Anopheles* and *C. elegans* do not exclude TEs (Fried et al., 2004). This suggested a constraint on Hox cluster organisation, hence the exclusion of TEs that could cause rearrangements, at the base of the Bilateria. Unlike the Hox cluster, the intact ParaHox clusters of chordates differ dramatically in their TE content, and studies have determined that TEs are not excluded in a similar manner from the ParaHox cluster and are able to invade this sister cluster (Ferrier et al., 2005; Osborne and Ferrier, 2010; Osborne et al., 2006). TE content, or the lack of TE content, within Hox and ParaHox clusters also has other implications for gene regulation, and exclusion of TEs could be key to avoiding disruption of a high density of regulatory elements (Amemiya et al., 2008). The presence of TEs also gives an indication of the availability/accessibility of these clusters within the germline. The presence of TEs within the ParaHox cluster, but not the Hox, could also be due to the ParaHox cluster being accessible within the germline whilst the Hox is not, and the clustered mouse *Cdx1* ParaHox gene has been shown to be transcriptionally active within germline cells (Kurimoto et al., 2008), where TEs are known to be active and able to invade accessible regions of the genome (Zamudio and Bourc'his, 2010). The potential impact of TEs upon Hox and ParaHox cluster maintenance is perhaps realised in the case of *Ciona intestinalis*, where there may be a causal link between TE invasion and the disruption of the *C.intestinalis* Hox and ParaHox clusters (Ferrier and Holland, 2002).

The ability of the Hox cluster to exclude TEs does not mean that this should be expected of every gene cluster however. In the case of the MHC cluster of mammals, TEs are found within the introns of genes (Doxiadis et al., 2008) and the bovine MHC cluster has undergone an inversion, possibly caused by the presence of TEs at breakpoints (Childers et al., 2006). The widespread presence of TEs across the genomes of eukaryotes could indicate that the Hox cluster is a unique case when it comes to TE exclusion. It has been observed that 45% of the human genome is made up of transposable elements (Lander et al., 2001). This varies widely amongst eukaryotic genomes, with *Tetraodon* having only <10% of its genome made up of TEs, showing remarkable compartmentalisation of TEs (Dasilva et al., 2002), whilst *Lilium* has between 95-99% (reviewed in Biemont and Vieira (2005)). With such high TE content in many genomes, the question has been posed as to whether all of this is truly ‘Junk’ or ‘selfish’ DNA. Indeed evidence from the human genome suggests that a substantial number of promoter and cis-regulatory elements may have evolved from TEs. Jorden et al analysed human promoters and experimentally confirmed cis-regulatory regions and found that 25% of analysed promoter regions contain transposable element derived sequences, as did 2.5% of the experimentally characterised cis-regulatory regions analysed (Jordan et al., 2003). The authors extrapolate this to estimate that roughly 1000 human genes contain cis-regulatory elements derived from TEs, though they stipulate that this is likely an underestimate as many TE-derived sequences will have evolved so that they are no longer traceable. This poses an interesting question as to whether TEs can be 'domesticated' and subverted to be used by Eukaryotic genomes, rather than acting 'selfishly' (Miller et al., 1999). One other possible example of this where TEs are able to influence the position of heterochromatin. Several studies have shown a wide range of effects where the presence of transposable elements directly silences gene expression, from gene silencing in plants to position effect variegation in animals (reviewed in Biemont and Vieira, (2005) and Slotkin and Martienssen (2007)). An interesting case for TE involvement in gene regulation lies in the *Gypsy* insulator elements of *Drosophila*, a transposable element that encodes retroviral proteins (Marlor et al., 1986). However, in addition to being a transposable element, *Gypsy* was the first such insulator element discovered and was able to block enhancer function, much as the CTCF-derived insulators described previously do (Geyer et al., 1986), though through binding of *Suppressor of hairy wing* protein and subsequent recruitment of *CP190* and *modifier of mdg4* (Pai et al., 2004), not CTCF and cohesin. This example shows that TEs may be deeply ingrained in regulatory function, though whether other TEs are able to function as insulators is currently unknown. These studies, highlight how little is truly known about transposable elements and how they affect gene regulation and genome structure. As such the distribution of TEs within and around the ParaHox cluster is interesting in multiple contexts.

Aims: To use bioinformatic approaches to annotate genomic and regulatory features within and surrounding the amphioxus ParaHox cluster. A combination of protein prediction and mapping of EST and transcriptomic data was used to annotate gene models along the *B.floridae* ParaHox reassembly to provide a map of genes surrounding the ParaHox cluster of amphioxus, and overcome the poor assembly quality in genome assembly v1.0, which has thus far hampered such efforts in the *B.floridae* genome. In addition, comparisons with the ParaHox clusters of *B.lanceolatum* and *B.belcheri* are used to confirm the arrangement of genes within the ParaHox cluster across the *Branchiostomidae*. Characterisation of the *Xlox* 3'UTR shows how a combination of bioinformatic and experimental approaches can be used to investigate interesting transcribed non-coding regions. Comparative genomic approaches allow a much more targeted approach to the identification of regulatory elements, highlighting areas of non-coding sequence conserved between amphioxus species to begin mapping the regulatory landscape of the amphioxus ParaHox cluster. The inclusion of transcription factor binding site identification methods, focusing on the transcription factors CTCF and TCF/Lef begin to identify regulatory inputs that may be regulating the ParaHox cluster on both a single gene and pan-cluster scale. Finally, transposable elements are mapped both within and surrounding the ParaHox cluster in order to try and further understand their impact upon ParaHox cluster maintenance and regulation.

3.2. Methods

3.2.1. Bioinformatics approaches

For scaffold annotation, the gene prediction programs fgenesh (Solovyev et al., 2006), with gene prediction settings for *B.floridae*, and AUGUSTUS (Stanke and Morgenstern, 2005) (now located at http://bioinf.uni-greifswald.de/augustus/), with gene prediction settings for *Homo sapiens* and *Petromyzon marinus*, were used to map gene models across the *B.floridae* ParaHox reassembly. Models from these two programs were compared against each other, as well as via BLASTP search to vertebrate proteins, and models where the two programs disagreed drastically were discarded. EST data was collected via BLAST search against the NCBI *B.floridae* EST database, as well as by BLAST search within the *B.floridae* cDNA database (Yu et al., 2008). These were then mapped to the *B.floridae* ParaHox reassembly using gene palette (Rebeiz and Posakony, 2004).

Alignments carried out in this chapter were carried out using VISTA, using the shuffle-LAGAN alignment algorithm. This algorithm was chosen as it takes into account rearrangements that may have occurred and is suited to aligning long genomic segments (Brudno et al., 2003). Translated anchoring was used in conjunction with this to improve the alignment between homologues, particularly in the case of alignments between amphioxus and vertebrate ParaHox clusters.

CTCF sites were identified using FIMO as part of the MEME suite (Bailey et al., 2009; Grant et al., 2011), allowing the input of the (Heger et al., 2012) position weight matrix (PWM) for CTCF. In order to process the *B.floridae* ParaHox reassembly using FIMO, it was first split into regions of 10kb in length using the mEMBOSS splitter function (Rice et al., 2000).

CTCF site location was combined with VISTA analysis to compare CTCF position to that of conserved sequence ‘peaks’ within the ParaHox cluster across amphioxus species. Regions spanning 150kb surrounding the ParaHox cluster, including the immediate neighbouring genes, were obtained from the *B.floridae* ParaHox reassembly, *B.lanceolatum* Sc000038, and *B.belcheri* Sc0000020. These ParaHox regions were then aligned in VISTA, using the shuffle-LAGAN alignment algorithm. The *B.floridae* ParaHox reassembly was used as the base sequence to align against, as annotation data for CTCF binding sites was collected for this scaffold. Though data was collected for the *B.floridae* PAC contig, several sites were not present in this scaffold that appeared to be conserved across the *B.floridae*, *B.lanceolatum* and *B.belcheri* scaffolds, hence the PAC contig was omitted as a base sequence from this analysis to avoid losing potential functionally relevant data. The alignments of *B.floridae/B.lanceolatum* and *B.floridae/B.belcheri* 150kb ParaHox regions were then used to produce sequence identity scores between the ParaHox clusters of *B.floridae*/*B.lanceolatum* and *B.floridae/B.belcheri*. These scores, represented as proportions (with 1 representing 100% identity) were 0.422 and 0.370 respectively for *B.floridae*/*B.lanceolatum* and *B.floridae/B.belcheri.* Bioedit (Hall, 1999) was used to produce all sequence identity scores.

The 500bp CTCF regions *from B.floridae* were processed in the same way, using VISTA and the shuffle-LAGAN alignment program, in order to obtain the corresponding region in each of *B.lanceolatum* and *B.belcheri*. To provide a secondary method of identifying the conserved CTCF 500bp regions in *B.lanceolatum* and *B.belcheri*, BLAST searches were performed for each individual *B.floridae* CTCF 500bp region against both the *B.lanceolatum* and *B.belcheri* ParaHox 150kb regions. These BLAST hits were then observed in the NCBI sequence viewer 3.8 in order to obtain the 500bp region of each alignment. BLAST hits were compared against the VISTA alignment results to confirm that the correct region had been obtained from each species. Each *B.floridae* CTCF 500bp region was then aligned to the corresponding region from each of *B.lanceolatum* and *B.belcheri* using a pairwise alignment with the BLOSUM62 similarity matrix. These pairwise alignments were used to produce sequence identity scores between each CTCF 500bp region *B.floridae*/*B.lanceolatum* and *B.floridae/B.belcheri* comparison. Some CTCF sites produced no alignment, particularly in the case of the *B.belcheri* scaffold. In some cases these were the result of poor assembly quality, or in others, such as with the CTCF8 500bp region, because of a gap within the alignment of the *B.lanceolatum* or *B.belcheri* corresponding scaffold region. Due to the inability to obtain corresponding regions between species in these cases, sequence identity could not be calculated and so was given a score of 0, representing no conservation.

Exact Binomial tests were then carried out to assess whether the conservation between CTCF 500bp regions was significantly higher than the background conservation level for each species, with the null hypothesis that an individual site did not display conservation significantly higher than the background level. In order to carry out Exact Binomial tests, sequence identity scores for each 500bp region were multiplied by the total number of base pairs for each 500bp region (500) to give an observed number of conserved base pairs, rounding to the nearest integer. This number was used along with the number of trials, (500 in 500bp), and the expected proportion of conserved base pairs (i.e. the sequence identity scores calculated for *B.floridae* vs *B.lanceolatum* (0.422) and *B.floridae* vs

B.belcheri (0.370)), to test if the observed level of conservation for each CTCF 500bp region was greater than the expected number under the background level of conservation for the appropriate species comparison. One-tailed Exact Binomial tests were carried out in the statistics software R using these scores to determine whether the observed number of conserved base pairs within a CTCF 500bp region is significantly greater than the background rate of conservation, using the following function.

> binom.test (*nsuccesses*, *ntrials*, p, alternative="greater")

nsuccesses = number of successes observed = observed number of conserved base pairs

ntrails = total number of trails = total number of base pairs = 500

p = hypothesised probability of success = either 0.422 or 0.370

alternative="greater" specifies that the test is one tailed, with the true probability of success is greater than either 0.422 or 0.370. This returned a p-value stating whether the observed number of conserved sites is significantly greater than the background rate.

The numbers of significantly conserved sites and non-conserved sites were then taken and plotted against the expected numbers of conserved and non-conserved sites. Expected numbers of conserved sites were calculated by multiplying the total number of observed CTCF binding sites within the ParaHox cluster against the proportion of observed sites expected under the background rate for each species. The expected number of non-conserved sites were then calculated by subtracting the expected number of conserved sites from the total number of observed sites. These numbers were then used to carry out One-tailed Exact Binomial Tests to determine if the number of sites lying within conserved regions was significantly greater than expected for each species comparison. The R function used previously was also used here, though this time the following variables were specified:

nsuccesses = number of successes observed = observed number of conserved CTCF 500bp regions

ntrails = total number of trails = total number of CTCF 500bp regions = 25

p = hypothesised probability of success = either 0.422 or 0.370

alternative="greater" specifies that the test is one tailed, with the true probability of success is greater than either 0.422 or 0.370. This returned a p-value stating whether the observed number of conserved regions is significantly greater than the number expected under the background rate of conservation.

3.2.2. Experimental procedures.

The Xlox 3’ UTR described in section 3.3.2 was cloned from *B.lanceolatum* adult cDNA according to the methods detailed in section 2.2.2-2.2.6. The primers detailed in table 2.14 were used to clone Xlox 3’ UTR, with the following sequencing primers used to sequence through the centre of the transcript.

Table 3.1. Sequencing Primers for Xlox 3’ UTR

Primer name	Sequence
ncRNA SEQ F (Xlox UTR)	GAAGCTCGCAGGTATTTGTC
ncRNA SEQ R (Xlox UTR)	GCGACGTTTCAACGTGTCCT

An antisense RNA probe was synthesised according to section 2.2.9. and hydrolysed according to section 2.2.10. In situ hybridisation of Xlox 3’ UTR with *B.lanceolatum* embryos was then carried out according to section 2.2.12.

3.3. Results

3.3.1. Annotating the *B.floridae* ParaHox reassembly

To remedy the poor assembly surrounding the ParaHox cluster in the *B.floridae* genome v1.0, a reassembly centred on the *B.floridae* ParaHox cluster was carried out by N.Putman for use by the Ferrier lab. In order to properly characterise the ParaHox cluster, as well as the surrounding genes, multiple protein prediction programs were used to provide corroborating gene models and improve the accuracy of gene prediction. EST data was also used to validate gene models and provide experimental support to intron-exon boundaries and untranslated regions (UTRs), which were not predicted by gene prediction software. This was first carried out for the ParaHox cluster and immediately surrounding genes, including *CHIC, SCP1, Gsx, Xlox, Cdx* and *PRHOXNB*. Exons and UTRs were determined for CHIC, SCP1, Cdx and PRHOXNB. No EST or transcriptomic support was available for Gsx, and Xlox was revealed to have a large 3’UTR (described further in section 3.3.3). In addition, SCP1 was revealed to have both a 3’UTR and multi-exonic 5’UTR (see chapter 4). A schematic of the ParaHox cluster and immediate neighbours is provided in figure 3.2, showing the relative positions and presence of exons and intron boundaries, as well as UTRs.

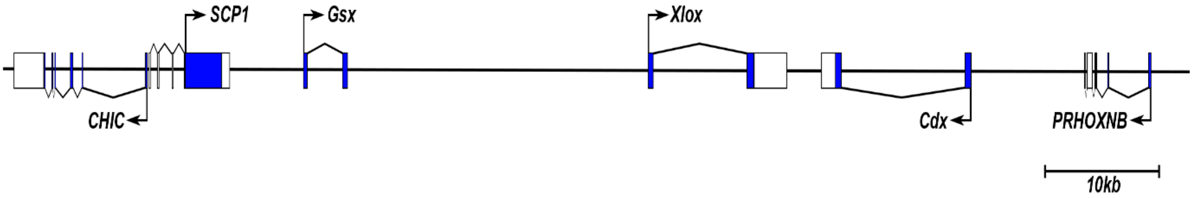


Figure 3.2. A schematic of improved gene models within the ParaHox cluster of *B.floridae*.

EST data has been used to provide improved gene models for the genes within the ParaHox cluster of *B.floridae*. This has provided 3'UTRs for CHIC, SCP1, Xlox, Cdx and PRHOXNB, as well as an SCP1 multi-exonic 5' UTR. Coding sequence is represented in blue, whilst UTR sequence is represented in white. Arrows represent translational start sites. Distances between genes are representative of the actual distances between genes within the *B.floridae* ParaHox reassembly. Arrows at right angles (↲) indicate translational start sites and orientation of transcription. Transcription start sites are unknown.

In addition to the ParaHox cluster and immediate neighbours, the same approach was used to move progressively further out from the ParaHox cluster, mapping gene models and noting if supporting EST models are present. 532Kb out of a total of just over 3.5Mb has been annotated so far, with EST data mapped if possible. These have been mapped using Genepalette (Rebeiz and Posakony, 2004) in order to visualise gene models. See Appendix 7.1 for gene names and positions along the *B.floridae* ParaHox Reassembly. One intriguing feature upstream of Gsx is a large expansion of Alkaline Phosphotase (AP) genes, with a total of 7 AP genes covering a total of ~59Kb from 1.478-1.537Mb along the B.fl ParaHox reassembly. Of these 7 AP genes, two, AP1 and AP2 have ESTs supporting the gene models. These AP genes may represent a series of tandem duplications, but perhaps even more intriguing, a potential reverse transcriptase gene from a Jockey transposable element lies in the centre of this AP cluster between AP2 and AP6, which could perhaps have facilitated such large expansion of AP genes. The close placement of these genes suggests that other Jockey elements may exist, or have ancestrally existed at this location that would have facilitated the movement of the Jockey element. This mechanism could have been hijacked to some extent to facilitate the duplication and insertion of AP genes.

The Genepalette Java-based program, as well as the associated scaffold files used throughout this study can be found on the CD provided with this thesis, and the scaffold sequence files can be requested from Dr David E.K. Ferrier. Within the annotated Genepallete scaffolds, and figure 3.2, genes marked in blue indicate models supported with EST data, whereas genes marked in red indicate either models with no EST or transcriptomic support, or EST data with no supporting gene model. A list of supporting ESTs is provided in the Appendix 7.2.

3.3.2. The organisation of the *B.lanceolatum* and *B.belcheri* ParaHox clusters concur with that of *B.floridae*.

With the release of the *B.belcheri* genome, and access to the preliminary assembly of the *B.lanceolatum* genome, comparisons between these three species have become possible. As such the ParaHox-containing scaffolds were obtained for *B.belcheri* (Sc0000020) and *B.lanceolatum* (Sc0000038) from the *B.belcheri* genome browser v15h11.r2 (Huang et al., 2012; Huang et al., 2014) (http://mosas.sysu.edu.cn/genome/index.php) and *B.lanceolatum* draft genome (public access not yet available) respectively via BLAST searches using the *B.floridae* ParaHox genes Gsx and Cdx as queries.

In order to aid in the comparison of these amphioxus ParaHox clusters, the annotation carried out for the *B.floridae* ParaHox cluster was used as a basis to inform the positions of the ParaHox genes and surrounding *CHIC, SCP1* and *PRHOXNB* genes. BLAST searches were then used to map exonic, intronic and UTR positions of these genes onto both the *B.lanceolatum* and *B.belcheri* ParaHox scaffolds. By comparing the ParaHox clusters of these three amphioxus species, it becomes apparent that the organisation of the ParaHox cluster is conserved between *Branchiostoma* species. However, several assembly errors and artefacts become apparent. Within *B.lanceolatum* Sc0000038 (figure 3.3 B), there is a duplication of *PRHOXNB* exon 2 and surrounding sequence, as well as a duplication of *Cdx* exon1 and surrounding sequence. These are likely to be assembly errors as both the exonic and surrounding non-coding sequence are identical in both cases of the two duplicate exons. Within *B.belcheri* Sc0000020 (figure 3.3 C), both the 5’ and 3’ ends of *SCP1* are missing, with the 3’ containing divergent non-coding sequence at this location and is lacking *SCP1* coding sequence, and the 5’ a long string of N’s. In addition to this, only *PRHOXNB* exon 1 is present, lacking exons 2 and 3 again due to the presence of divergent non-coding sequence at this location. UTRs have been annotated where possible, with EST and Transcriptomic data from *B.floridae* and *B.lanceolatum* used to inform UTR position. UTRs for *Xlox* (Described in section 3.3.3) and *CHIC* shown in *B.lanceolatum* (figure 3.3 B) are supported by *B.lanceolatum* transcriptomic data. Sequence Seq43418.bl (Accession number JT881816.1) was used to provide the mRNA sequence, including 5’ and 3’ UTRs for *B.lanceolatum CHIC*. All other UTRs displayed for both *B.lanceolatum* and *B.belcheri* were identified via BLAST search against *B.floridae* EST data.

It should be noted that the *B.belcheri* genome has undergone updates since v15h11.r2 and additional sequencing since these studies were carried out. In addition, the *B.lanceolatum* genome is still in draft format, and has currently been taken offline whilst additional sequencing and reassembly is carried out. Still, even with the various assembly errors within both *B.lanceolatum*

Sc0000038 and *B.belcheri* Sc0000020 it is clear that the organisation and relative positions of the ParaHox genes and their immediate neighbours remain conserved between *B.floridae* (figure 3.3A), *B.lanceolatum* (figure 3.3 B), and *B.belcheri* (figure 3.3 C).

3.3.3. Amphioxus Xlox has a large 2884bp 3'UTR.

Whilst carrying out annotation of the *B.floridae* reassembly, one peculiarity that arose was that there was no supporting EST or transcriptomic data for the ParaHox gene *Xlox*. However, a large transcript downstream of *Xlox* exon 2 was identified. EST data from both neurula and adult *B.floridae* samples indicated a 2312bp transcript in the same transcriptional orientation as *Xlox*. The ESTs identified as such were bfne153i16 (neurula) (Genbank: **BW890457** and **BW948304**) and bfad018m17 (adult) (Genbank: **BW714759.1**). Amphioxus *Xlox* has previously been found to be difficult to clone from RNA samples, and *B.floridae Xlox* expression was previously examined using an in situ hybridisation probe created from a fused exon template cloned from genomic DNA (Osborne et al., 2009). As such, it was hypothesised that the identified *Xlox* 3' transcript could represent one of two possibilities. In scenario 1, the transcript would represent part of a 3' Xlox UTR that was sequenced despite the difficulties in priming the remainder of the *Xlox* mRNA. The second scenario represents one that is the least likely, given the close proximity and identical transcriptional orientation of *Xlox* and the 3' transcript. In this case, the 3' transcript would represent a novel ncRNA. In order to examine whether this transcript was indeed transcribed with the Xlox coding sequence, RT-PCR was used to attempt to clone both the 3' transcript from whole adult cDNA, as well as a transcript running from *Xlox* exon 2 through to the 3' transcript (Primers for both in table 3.2.).

Looking to the *B.lanceolatum* draft scaffold Sc0000038, the 3' transcript sequence appeared to be well conserved between *B.lanceolatum* and *B.floridae*, suggesting this region is functional. As such, *B.lanceolatum* was used to examine the expression of the *Xlox* 3' transcript, as access to animals for cDNA synthesis was much easier than for *B.floridae*. A transcript covering 2068bp of the Xlox 3' transcript region was amplified, though attempts to clone a transcript linking this region and Xlox exon 2 were unsuccessful (figure 3.4Ai). In addition to RT-PCR, an antisense RNA probe was designed against the cloned *B.lanceolatum* Xlox 3' transcript cDNA in order to examine the expression of this transcript. Due to the position of this transcript between *Xlox* and *Cdx*, embryonic stages covering the expression of both of these genes were chosen. The earliest expression was detected in early neurula stages within the posterior endoderm and neural tube. This expression carries on into the mid neurula stage (figure 3.4 B), where an additional distinct expression domain is gained in the neural tube at the region corresponding to the future pigment spot. This expression domain proves to be transient, persisting through the late neurula but disappearing by the premouth stage. During the premouth and larval stages, the endoderm staining refines to the midgut-hindgut boundary (figure 3.3.B). This expression pattern is identical to that of *B.floridae Xlox.* Attempts were made to examine Xlox expression in *B.lanceolatum* but the probe used (Kindly gifted by Ildiko Somorjai) resulted in considerable background and I was unable to properly compare and image the expression patterns (data not shown).

Subsequent to these analysis, access to the B.lanceolatum draft genome was granted along with additional transcriptomic data. Within this transcriptomic data, a 3359bp read was identified, **comp1023376_c0_seq3**, which displayed transcription going from *Xlox* exon 2 through to the centre of the Xlox 3' transcript (figure 3.4Aii). As such, this transcript, as well as the identical mRNA expression pattern to Xlox, suggest that it is likely that the observed Xlox 3' transcript is an Xlox 3' UTR. An alignment of the *B.floridae Xlox* and 3' genomic sequence (from the ParaHox reassembly), the *B.floridae* Xlox UTR ESTs and *B.lanceolatum* Xlox UTR transcript can be found in Appendix 7.2.

In addition, bioinformatic approaches were used to search for distinctive secondary structures and motifs within the UTR region. In particular conserved RNA secondary structure was examined, using the secondary structure prediction program LocARNA (Smith et al., 2010) with the standard settings. This produced an alignment of the *B.floridae* Xlox downstream region and the *B.lanceolatum* Xlox transcript and identified conserved hairpin forming regions (see Appendix 7.3, figure 7.1). This was then converted to an RNA structure (see Appendix 7.3, figure 7.2). Analysis with both B.lanceolatum and B.floridae using CentroidFold (Hamada et al., 2009) produced entirely different secondary structures again for each species (data not shown). It is clear from these analysis that the central region of the Xlox UTR is not well conserved across species, making it unlikely that the Xlox UTR region is involved in production of secondary structure with regulatory function. Attempts were also made to identify RNA binding protein motifs within this region that might be involved in the function of RNA secondary structure using CatRAPID omics (Agostini et al., 2013), but revealed no obvious candidates (data not shown), showing that it was unlikely that this region was interacting with known RNA-binding proteins such as Polycomb repressive complex (Fatica and Bozzoni, 2014).

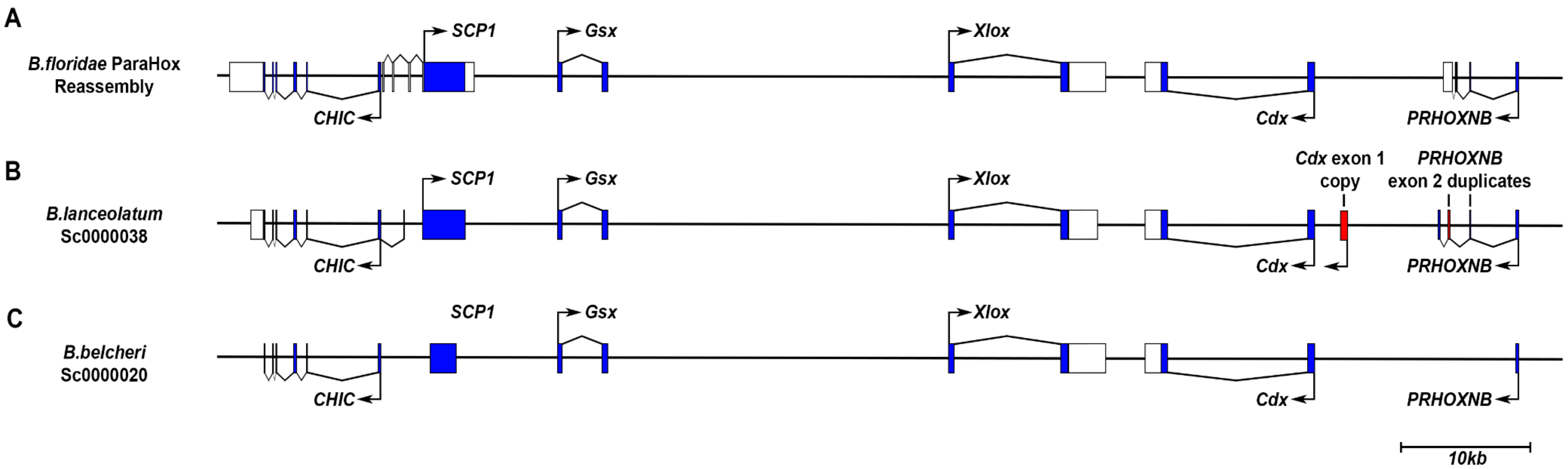


Figure 3.3. Organisation of *B.floridae*, *B.lanceolatum* and *B.belcheri* ParaHox clusters.

(A-C) Schematics showing the relative positions of genes, exons, introns and UTRs within the *B.floridae*, *B.lanceolatum* and *B.belcheri* ParaHox clusters and their immediately neighbouring genes . (A) Schematic representation of the *B.floridae* ParaHox cluster along with surrounding genes *CHIC, SCP1* and *PRHOXNB.* (B) Schematic representation of the *B.lanceolatum* ParaHox cluster along with surrounding genes *CHIC, SCP1* and *PRHOXNB.* Duplications of Cdx exon 1 and PRHOXNB exon 2 are likely assembly artefacts. A 5' UTR of *B.la CHIC* not identified in *B.floridae* (A) is displayed. (C) Schematic representation of the *B.belcheri* ParaHox cluster along with surrounding genes *CHIC, SCP1* and *PRHOXNB. SCP1* is lacking both 5' and 3' coding sequence, whilst PRHOXNB is lacking all 3' exons, with only exon 1 present in the scaffold. Again, these differences are likely due to assembly errors. Arrows at right angles (↲) indicate translational start sites and orientation of transcription. Transcription start sites are unknown.

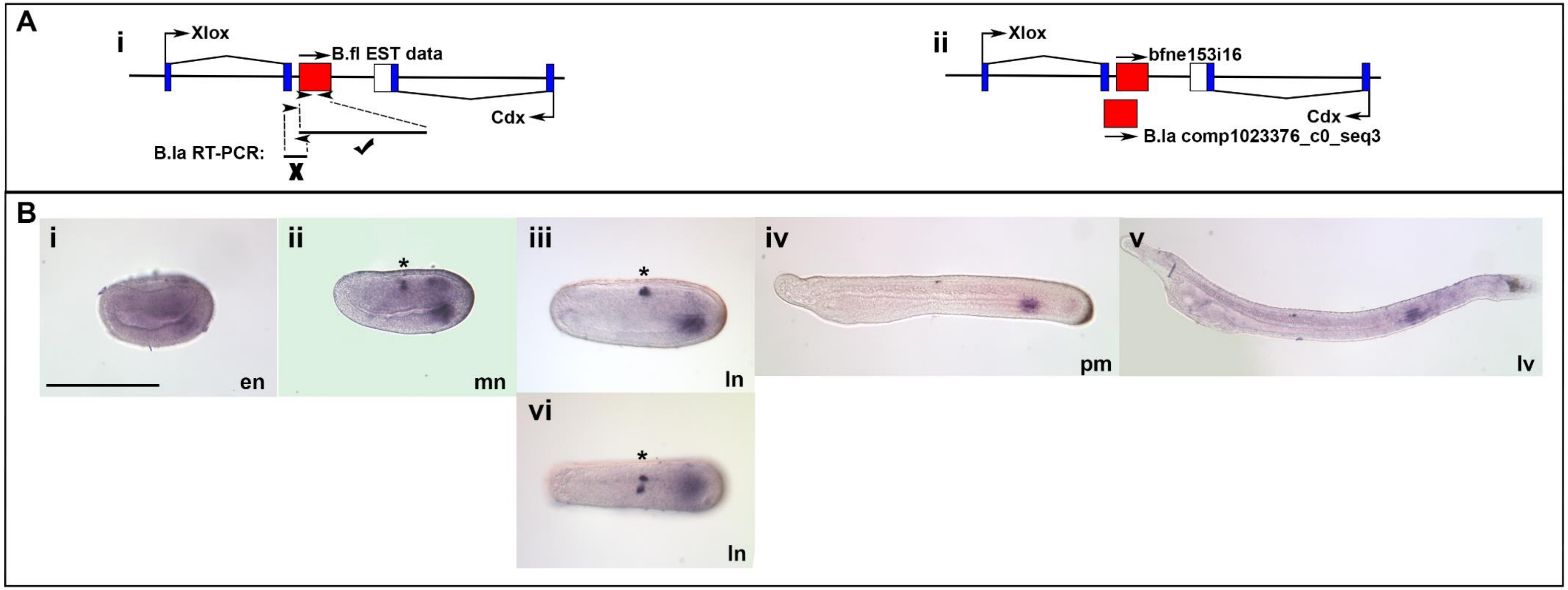


Figure 3.4. The Xlox 3’ transcript is expressed in a pattern identical to Xlox during amphioxus development and represents a 3’ UTR of Xlox.

(A) (i) A schematic showing the inability to clone the Xlox coding region from cDNA, whilst the Xlox 3’ transcript is found in adult cDNA samples. (ii) A *B.lanceolatum* transcript displays continuity of transcription between Xlox exon 2 and the 3’ Xlox transcript. (B) A time-course of *B.lanceolatum* development covering the stages where the Xlox 3’ transcript expression is observed. Expression is seen within the endoderm from the early neurula through to 2-day larvae (i-v). Expression is also seen at the level where the pigment spot will develop (marked by a black aterisk), beginning in the mid neurula (ii) and continuing into the late neurula stage (iii, vi). All embryos have anterior on the left and posterior to the right. (i-v) represent lateral views, whilst (vi) is a dorsal view. en: early neurula, mn: mid neurula, ln: late neurula, pm: premouth, lv: 2-day larvae. Arrows at right angles (↵) indicate translational start sites and orientation of transcription. Transcription start sites are unknown. Scale bar represents 100 µm.

3.3.4. Vista analysis of the ParaHox clusters of three amphioxus species reveals conserved non-coding regions throughout the cluster.

The *B.floridae* ParaHox cluster is archetypal in its genomic structure, with three ParaHox genes located in the order Gsx-Xlox-Cdx, with relatively well conserved exon-intron structure (Ferrier et al., 2005), and it displays regulatory phenomena comparable to and potentially conserved with the Hox cluster, such as a response to retinoic acid (Osborne et al., 2009). As such, the regulatory landscape of the amphioxus ParaHox cluster is of particular interest. Whilst regulatory regions can be identified by screening large intergenic and intronic regions in functional assays such as reporter transgenics, there is a risk of a low success rate to this approach, as much of the non-coding sequence examined may well not show regulatory function, or is perhaps restricted to regulatory contexts not present in the system used to identify these regions. However, if one were able to target regulatory screens to the areas of non-coding sequence conserved between closely-related species, this should increase the chance of a non-coding region displaying regulatory function. In order to achieve this for amphioxus, VISTA analysis has been carried out between the ParaHox clusters of three different amphioxus species, *B.floridae*, *B.lanceolatum* and *B.belcheri*.

The three amphioxus scaffolds, the *B.floridae* reassembly, *B.lanceolatum* Sc0000038, and *B.belcheri* Sc0000020, were aligned against the amphioxus ParaHox PAC contigs (Ferrier et al., 2005) in order to accurately map exon and UTR boundaries and visualise these (see figure 3.5). It is clear that the protein coding regions and UTRs are amongst the most highly conserved regions of each scaffold, though there are some discrepancies in this. These, however, are largely due to sequencing errors in the *B.lanceolatum* and *B.belcheri* scaffolds. These are particularly noticeable within the *B.belcheri* scaffold in the 3' of *SCP1*, the 3' of *Xlox* exon 2 and in the missing PRHOXNB exons (coding exons 2 and 3 and UTR exons). It should be noted that several rounds of resequencing have been carried out for the *B.belcheri* genome since the analysis in this thesis were carried out and subsequent analyses will include this new data.

Upon comparison of the non-coding regions throughout the three amphioxus ParaHox clusters, it became apparent that conservation of sequence varies drastically across the ParaHox cluster. Many 'islands' of high conservation exist that may prove useful in the identification and examination of the amphioxus ParaHox regulatory landscape. One particular region of note is the region immediately upstream of *Gsx*, which actually exhibits higher conservation across the three amphioxus species than the adjacent coding sequence for *Gsx* exon 1. The hypothesis that these highly conserved regions represent functional regulatory elements has been further tested for this *Gsx* upstream region in chapter 5. The regions immediately upstream of all three ParaHox genes appear to be highly conserved across the three amphioxus species, perhaps representing conserved promoter sequences or regulatory elements. Many islands of conservation also exists between genes, as well as within introns. Three of the more notable of these include a large peak downstream of Gsx, located between 25 and 27Kb (figure 3.5), a distinct island of high conservation located between 38 and 40Kb between *Gsx* and *Xlox*, and finally a large conserved stretch surrounding the 74Kb location within the *Cdx* intron. Many more islands of conservation exist (see figure 3.5), and it would be interesting to test some of these in a reporter construct and examine any functional properties of these regions.

3.3.5. The amphioxus ParaHox cluster does not display observable conservation of regulatory elements with the vertebrate ParaHox regions

In order to observe if any conserved non-coding elements (CNEs) are present between the amphioxus and vertebrate ParaHox clusters, the ParaHox clusters of both *H.sapiens* and *M.musculus* were obtained from chromosomes 13 and 5 respectively. Sequence was obtained to the total of 389,053bp from *H.sapiens* (Ensembl v81) and 258,349bp from *M.musculus* (Ensembl v81), covering the ParaHox genes themselves and surrounding intergenic regions upstream of *Gsh1* and *Cdx2*, though including PRHOXNB, in order to include any CNEs that may lie upstream of either Gsx or Cdx. These were then aligned using VISTA and the shuffle-LAGAN alignment algorithm to account for inversions and rearrangements that may have occurred between amphioxus and vertebrates. The relaxed settings of 60 minimum identity and 50bp window used in (Pascual-Anaya et al., 2008) to compare the Hox clusters of vertebrates and amphioxus were also used in the VISTA alignment here, to account for the large evolutionary distances between amphioxus and the vertebrates. When observing the VISTA alignment, there appear to be no conserved non-coding elements between the amphioxus species and the vertebrate ParaHox clusters (figure 3.6), and only some coding regions do show conservation, most notably exon 2 of both Cdx and Xlox.

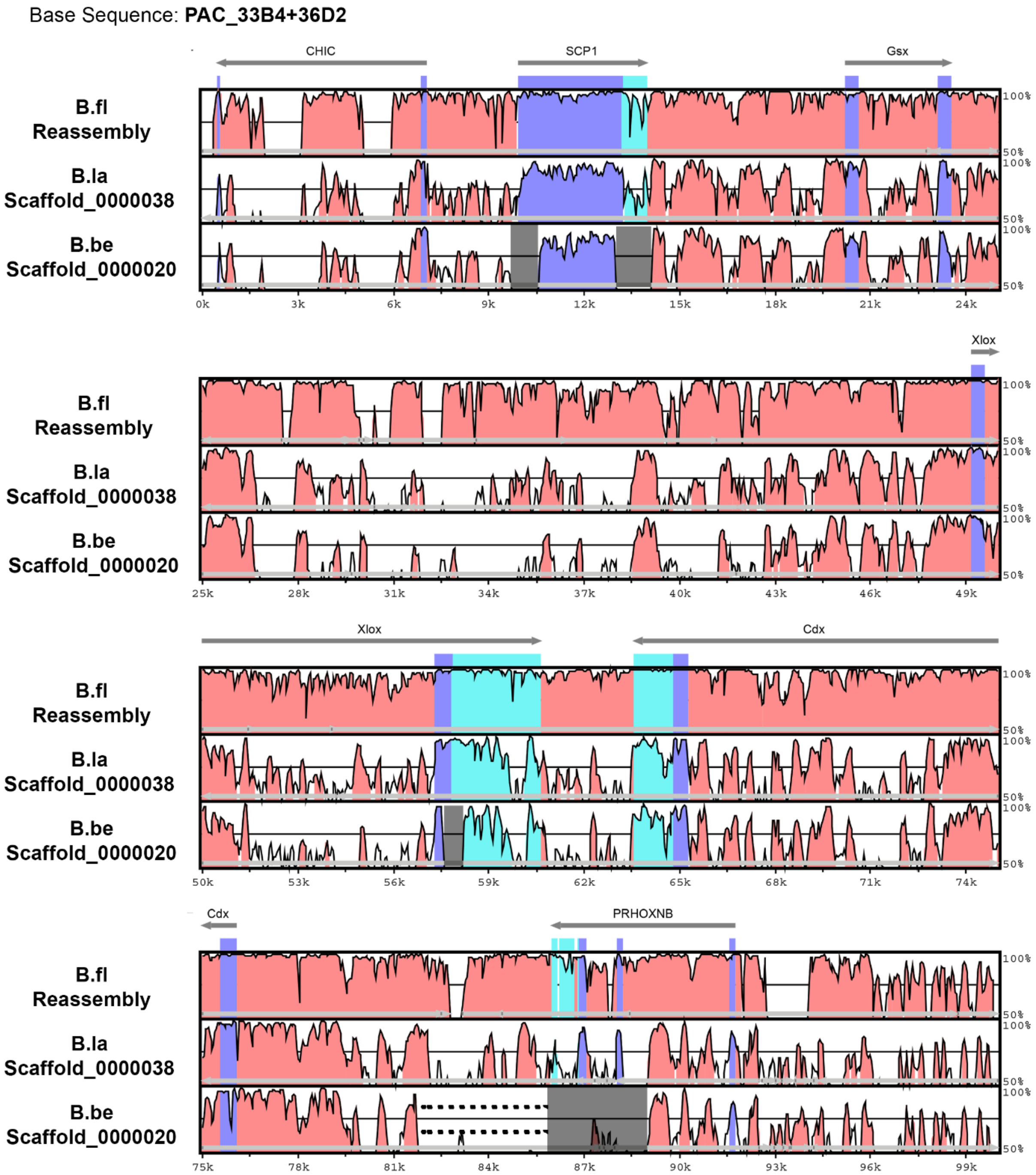


Figure 3.5. VISTA alignment of the *B.floridae*, *B.lanceolatum* and *B.belcheri* ParaHox clusters reveals high levels of conservation between amphioxus ParaHox clusters.

The ParaHox clusters were obtained from the *B.floridae* ParaHox reassembly, *B.lanceolatum* Sc0000038 and *B.belcheri* Sc0000020 and aligned using VISTA against the *B.floridae* PAC ParaHox contig 33B4+36D2 (Ferrier et al., 2005). High levels of conservation of both coding (blue/purple) and non-coding regions (pink) are present between all three species. Light blue regions represent the UTRs identified in section 3.3.1. Black overlays represent regions of assembly error, whilst dotted lines indicate the possible continuation of one of these erroneous regions. Each peak on the VISTA figure represents a conserved region.

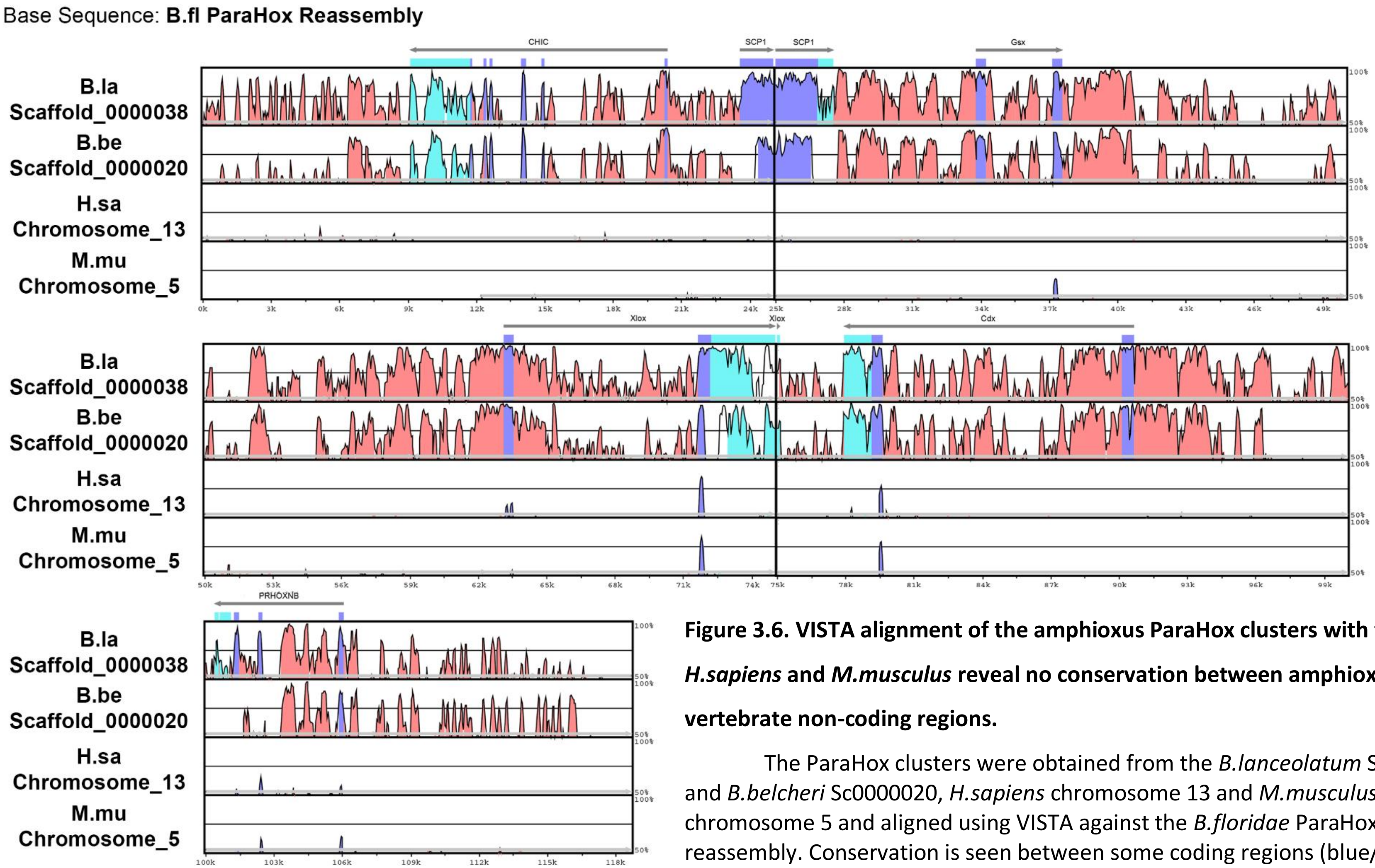


Figure 3.6. VISTA alignment of the amphioxus ParaHox clusters with those of *H.sapiens* and *M.musculus* reveal no conservation between amphioxus and vertebrate non-coding regions.

The ParaHox clusters were obtained from the *B.lanceolatum* Sc0000038 and *B.belcheri* Sc0000020, *H.sapiens* chromosome 13 and *M.musculus* chromosome 5 and aligned using VISTA against the *B.floridae* ParaHox reassembly. Conservation is seen between some coding regions (blue/purple) of *H.sapiens, M.musculus* and the amphioxus ParaHox clusters, but no conservation of non-coding regions (pink) or UTRs (light blue) is observed between vertebrates and amphioxus. Each peak represents a region conserved between the subject and *B.floridae*.

3.3.6. CTCF binding sites show uneven distribution and hint at possible TADs.

In order to identify potential insulating elements, CTCF binding sites were identified according to the consensus sequence and position weight matrix detailed in (Heger et al., 2012). This consensus was compiled from both *Drosophila* and mammalian CTCF binding sites, and represents a bilaterian CTCF consensus sequence (Heger et al., 2012). In order to tally sites across the *B.floridae* ParaHox reassembly, the ~3.5Mb sequence was first split into more manageable 10kb files using the mEMBOSS splitter function (Rice et al., 2000). FIMO, from the MEME suite, (Bailey et al., 2009; Grant et al., 2011) was then used to search for CTCF sites within each 10kb region using the Heger et al. (2012) weight matrix. This approach enabled both the density mapping of CTCF binding sites across the entire *B.floridae* reassembly, providing a general overview of CTCF density in this genomic region, as well as the visualisation and mapping of CTCF site position on a finer scale within the ParaHox cluster itself.

Looking across the whole *B.floridae* ParaHox reassembly (figure 3.7), it becomes apparent that CTCF site density varies considerably across the length of the scaffold, with some 10kb stretches containing no such binding sites (e.g. the 20-30Kb region), whilst others (such as the 100-110kb region) contain 10 sites in the same length of DNA. The mean number of CTCF sites per 10kb is 2.57, and it is clear that there is large variation around this figure across the scaffold. Indeed, the region spanning the ParaHox cluster and immediately flanking genes SCP1 and PRHOXNB looks to have a much more stable number of CTCF sites than the rest of the *B.floridae* ParaHox reassembly, remaining at either 2 or 3 sites/10kb (the average) for the entire 90kb stretch. Looking outside of this region we see that the number of CTCF binding sites varies much more from the average between different 10kb stretches over a similar distance.

Recent experimental data from Chromatin Immunoprecipitation and Chromosome Conformation Capture experiments that have characterised both functional TADs and the scope of CTCF binding within a genomic location also suggest that the directionality of CTCF binding sites across a genomic location contributes to the formation of TADs (Gomez-Marin et al., 2015; Guo et al., 2015). Thus, the directionality of CTCF sites was also characterised, with the intention of using a bioinformatics approach to provide some initial insight into the presence of CTCF-associated TADs in the absence of functional CTCF CHIP-seq data for amphioxus. Though a single 10kb region may contain both sense and antisense oriented CTCF binding sites, there are many regions present within the *B.floridae* ParaHox reassembly that represent a distinct boundary between domains of differing CTCF site orientation. Looking to the ParaHox cluster in particular, there is one such boundary region placed between SCP1 and Gsx (figure 3.7). Here, CTCF site orientation is distinctly antisense from

1.57-1.61Mb before a sense only peak at 1.62Mb, which continues into a domain of roughly equal sense and antisense orientation between 1.63-1.68Mb. Subsequent to this there is another possible boundary element, with another peak in the number of sense CTCF sites and an absence of antisense oriented sites at the 1.69Mb region near PRHOXNB. Extending outwards from the ParaHox cluster and immediate neighbours, a further distinct contrast in CTCF orientation is seen between 1.49 and 1.54Mb, where the solely antisense orientation seen from 1.57-1.61Mb switches to predominantly sense orientated sites at 1.54Mb (figure 3.7).

To further build upon this, and perhaps pinpoint specific boundaries, CTCF sites were individually mapped to the ParaHox cluster on the *B.floridae* ParaHox reassembly, from CHIC to PRHOXNB, using Genepalette (Rebeiz and Posakony, 2004) in order to visualise them in relation to gene introns and exons. When visualised in this way, several boundaries where CTCF binding sites are oriented facing away from each other (antisense<>sense) are revealed.

Once again, one potential boundary is present between SCP1 and Gsx, between the antisense site within the 3' of SCP1 and the intergenic sense site between SCP1 and Gsx (figure 3.8). This correlates well with the wider 10kb count and the domains of CTCF orientation observed on either side of this region (figure 3.8). Several more candidate boundary elements exist where CTCF sites face away from each other (black arrowheads), with particularly striking candidates immediately upstream of both Xlox and Cdx. These candidates immediately upstream of Cdx, or perhaps the one between the triplet of intronic sense PRHOXNB sites and the intergenic antisense sites downstream of PRHOXNB (figure 3.8), also correlates with the switch in CTCF orientation to a domain of sense CTCF sites seen at 1.69Mb in figure 3.7.

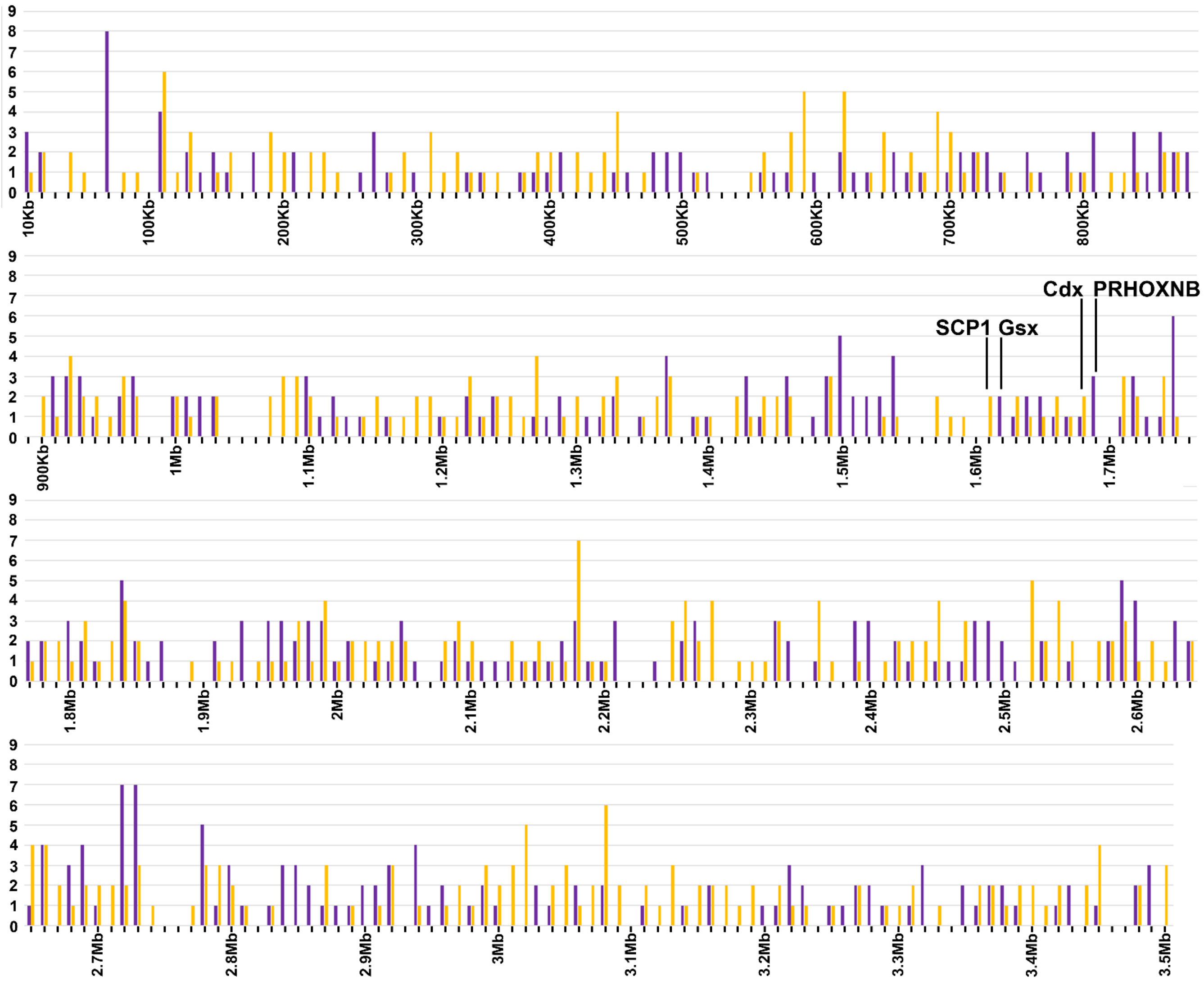


Figure 3.7. CTCF binding sites located throughout the *B.floridae* ParaHox reassembly reveal domains of distinct CTCF site orientation.

CTCF binding sites were identified throughout adjacent 10kb regions of the *B.floridae* ParaHox reassembly using the 15bp position weight matrix taken from (Heger et al., 2012). CTCF site orientation was annotated for both sense (purple) and antisense (yellow) binding sites, with the number of each orientation within each respective 10kb region noted. The positions of both Gsx and Cdx, have been annotated to show the location of the ParaHox cluster. The positions of SCP1 and PRHOXNB have also been annotated to highlight the change in orientation of CTCF binding sites either side of the ParaHox cluster. Dashes on the X axis represent 10kb intervals. The Y axis represent the number of CTCF sites observed.

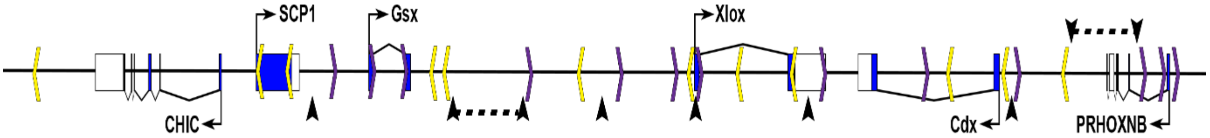


Figure 3.8. Adjacent and Opposing orientation of CTCF binding sites may represent TAD boundaries.

CTCF sites identified in figure 3.7 have been mapped directly to the ParaHox cluster and represented as a schematic. Boundaries between opposing orientation of CTCF binding sites have been indicated by black arrow-heads. Larger stretches between opposing sites have been indicated by two black arrow-heads joined by a dotted line. Sense CTCF binding sites are represented by purple chevrons, whilst antisense CTCF binding sites are indicated by yellow chevrons. Coding regions are indicated in blue, whilst UTRs are indicated in white. Arrows at right angles (↵) indicate translational start sites and orientation of transcription. Transcription start sites are unknown.

3.3.7. The presence of CTCF binding sites within conserved sequence identifies potential insulator elements within the ParaHox.

Whilst the characterisation of CTCF binding site density across the scaffold is informative for a wider genomic context, it is likely that only a portion of the CTCF sites identified in section 3.3.6. are functional. A bioinformatics approach was taken in order to begin addressing whether these sites may lie within potential regulatory regions, with the aim of providing a more directed starting point for future functional analysis. Several CTCF binding sites lying within peaks of high conservation between the three amphioxus species stand out from this analysis (figure 3.9). Of the 25 sites identified, 24 lie within regions conserved between *B.lanceolatum* and *B.floridae*, and 20 between *B.belcheri* and *B.floridae*. Only CTCF site 8 does not occur within a region conserved between at least one of *B.floridae*/ *B.lanceolatum* or *B.floridae/B.belcheri*. Within B.belcheri, CTCF sites 1, 2, 6, 8 and 9 do not lie within conserved peaks, however, site 2 lies within a *B.belcheri* assembly error, and indeed poor assembly quality could also be the cause of a lack of conservation in some of the other CTCF sites as well. Of particular note are CTCF sites 4, 7, 10, 12, 13, 17, 18, 21 and 22-25, which all lie within highly conserved non-coding regions within the ParaHox clusters of all three amphioxus species, as well as CTCF sites 2, 3, 5 and 16, which lie within coding regions. These conserved sites, particularly those lying within highly conserved non-coding regions, may represent potential insulator elements.

From figure 3.9, it appears that CTCF sites are largely localised to regions of conserved sequence between species, with only a small proportion of these sites lacking any conservation. In order to test whether the conservation surrounding CTCF sites is indeed significantly enriched above that of the background level of conservation between the amphioxus ParaHox clusters, 500bp regions were taken surrounding each CTCF site, and used to calculate the level of Identity between the regions surrounding CTCF positions between *B.floridae* and each of *B.lanceolatum* and *B.belcheri*. This was then compared to the 'background' level of identity across the whole alignment between the *B.floridae* ParaHox region (taken from the reassembly) and the *B.lanceolatum* and *B.belcheri* ParaHox regions respectively (figure 3.10).

Exact binomial tests were carried out to assess whether the conservation between CTCF 500bp regions was significantly higher than the background conservation level for each species, with the null hypothesis that an individual site did not display conservation significantly higher than the background level. For *B.lanceolatum*, all CTCF sites, except for site 8, lie within regions showing sequence identity significantly greater (P.values <0.05) than the background identity of 0.422. Likewise in *B.belcheri*, all CTCF 500bp regions identified as being conserved in figures 3.9 show sequence identity to *B.floridae* significantly higher than the background identity of 0.37 (P-values <0.05), though CTCF 500bp regions 1, 2, 6, 8 and 9 could not be aligned. In the case where CTCF sites could not be aligned, conservation scores were given as 0, and P values as 1, representing the null hypothesis that the site does not show conservation significantly higher than the background rate. Figure 3.10, shows the sequence identities between *B.floridae* and *B.lanceolatum* (figure 3.10 A), and *B.floridae* and *B.belcheri* (figure 3.10 B), along with P-values for each sequence identity score.

Finally, an exact binomial test was used to assess whether CTCF sites were enriched within conserved 500bp regions, under the null hypothesis that CTCF sites are distributed between conserved and non-conserved regions at the same rate as the background conservation rate. For this, the significance of the observed number of CTCF sites lying within conserved 500bp regions, and those lying in non-conserved regions, was tested against the expected number under the background rate of conservation between each species. In both cases, the number of CTCF sites present within conserved regions was deemed to be significantly enriched above the background conservation rate (P.values <0.05) (figure 3.10 C, D). This supports what is seen in figure 3.9, that CTCF sites are indeed enriched within regions conserved between amphioxus species.

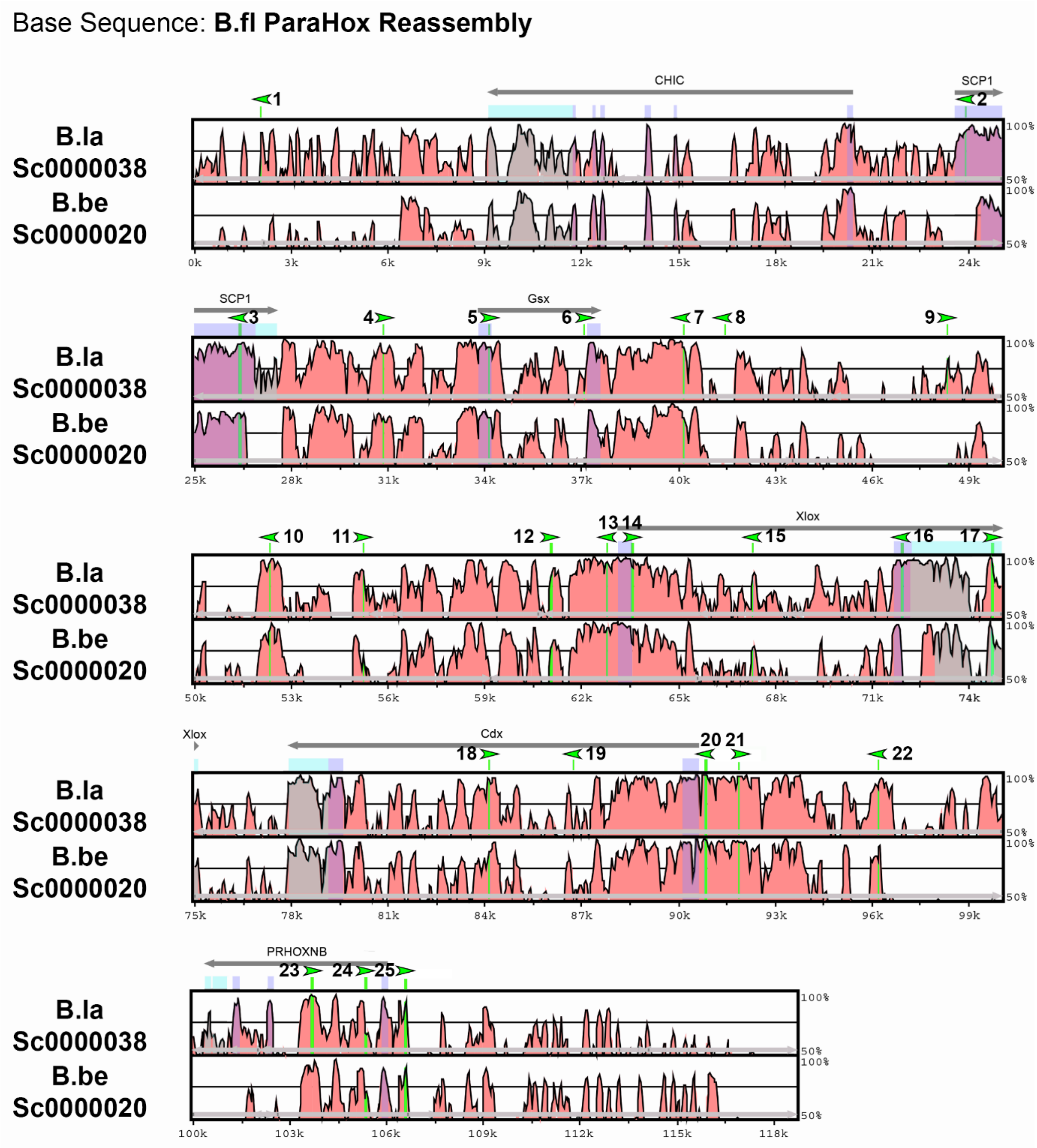


Figure 3.9. CTCF binding sites lie within conserved regions between the ParaHox clusters of *B.floridae*, *B.lanceolatum* and *B.belcheri*.

The 15bp CTCF binding sites located within the ParaHox cluster of *B.floridae* were mapped out and annotated in a VISTA alignment of the ParaHox clusters taken from *B.lanceolatum* Sc0000038 and *B.belcheri* Sc0000020 against the *B.floridae* ParaHox reassembly. All CTCF binding sites (green arrowheads and green lines on peaks) except CTCF site 8 map to conserved regions between *B.lanceolatum* and *B.floridae*, whilst all sites except CTCF sites 1, 2, 6, 8 and 9 map to conserved regions between *B.belcheri* and *B.floridae*. Coding regions are highlighted in blue/purple, UTRs in light blue and non-coding regions are represented in pink. Each peak corresponds to a conserved region.

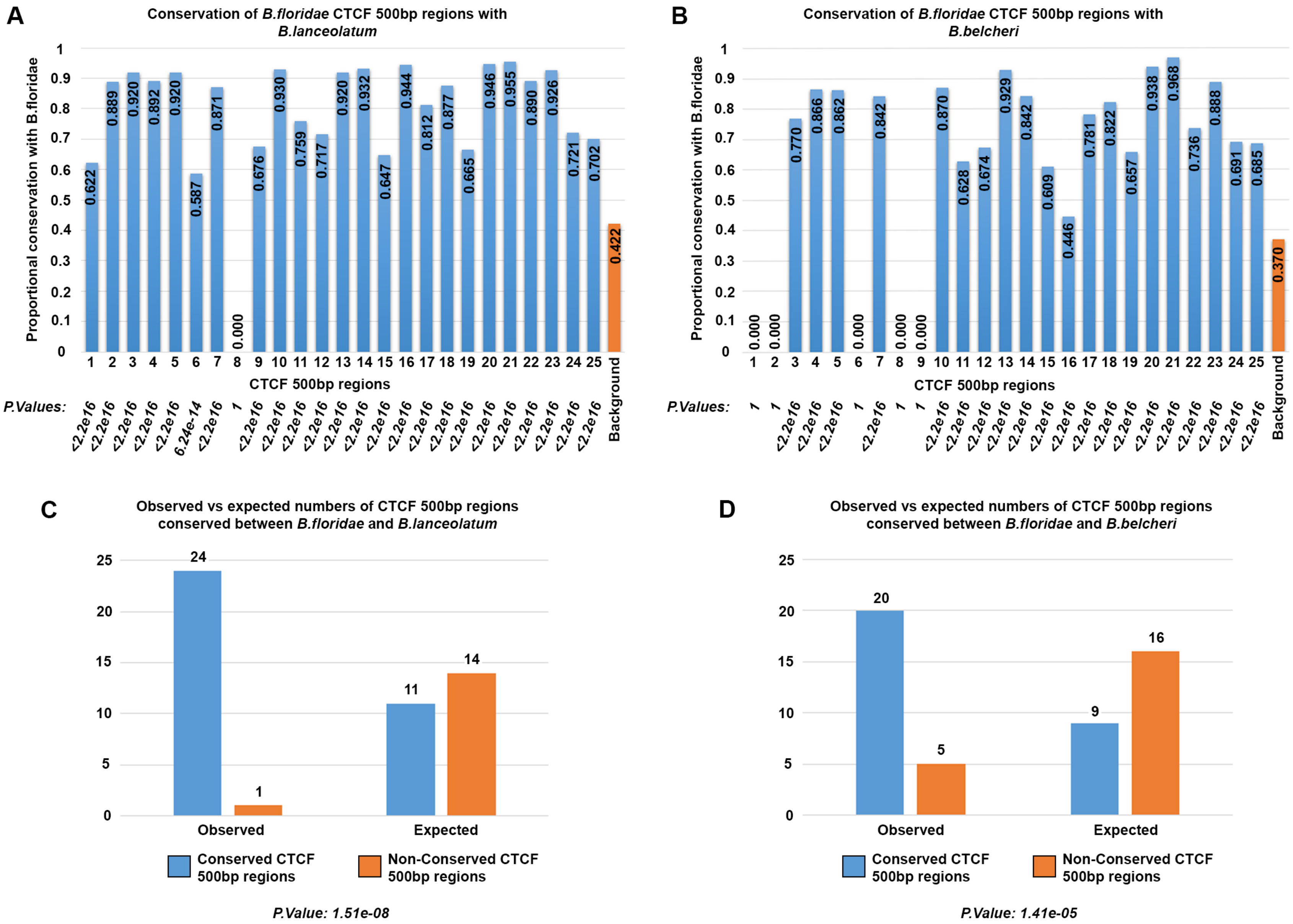

A
Conservation of B.floridae CTCF 500bp regions with B.lanceolatum
Proportional conservation with B.floridae
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
0.622 0.889 0.920 0.892 0.920 0.587 0.871 0.000 0.676 0.930 0.759 0.717 0.920 0.932 0.647 0.944 0.812 0.877 0.665 0.946 0.955 0.890 0.926 0.721 0.702 0.422
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 Background
CTCF 500bp regions
P.Values:
<2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 6.24e-14 <2.2e16 1 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16
B
Conservation of B.floridae CTCF 500bp regions with B.belcheri
Proportional conservation with B.floridae
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
0.000 0.000 0.770 0.866 0.862 0.000 0.842 0.000 0.000 0.870 0.628 0.674 0.929 0.842 0.609 0.446 0.781 0.822 0.657 0.938 0.968 0.736 0.888 0.691 0.685 0.370
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 Background
CTCF 500bp regions
P.Values:
1 1 <2.2e16 <2.2e16 <2.2e16 1 <2.2e16 1 1 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16 <2.2e16
C
Observed vs expected numbers of CTCF 500bp regions conserved between B.floridae and B.lanceolatum
0 5 10 15 20 25
24 1 11 14
Observed
Expected
Conserved CTCF 500bp regions
Non-Conserved CTCF 500bp regions
P.Value: 1.51e-08
D
Observed vs expected numbers of CTCF 500bp regions conserved between B.floridae and B.belcheri
0 5 10 15 20 25
20 5 9 16
Observed
Expected
Conserved CTCF 500bp regions
Non-Conserved CTCF 500bp regions
P.Value: 1.41e-05

Figure 3.10. CTCF sites are significantly enriched within conserved regions of the ParaHox cluster.

(A) A bar chart showing the sequence identity scores of 500bp regions surrounding each CTCF binding site between *B.lanceolatum* and *B.floridae* as proportions. Exact binomial tests were carried out for each CTCF 500bp region and P values given below each site show the significance between the observed conservation between *B.lanceolatum* and *B.floridae* 500bp regions and the background conservation level between *B.lanceolatum* and *B.floridae*. (B) A bar chart showing the sequence identity scores of 500bp regions surrounding each CTCF binding site between *B.belcheri* and *B.floridae* as proportions. Exact binomial tests were carried out for each CTCF 500bp region and P values given below each site show the significance between the observed conservation between *B.belcheri* and *B.floridae* 500bp regions and the background conservation level between *B.belcheri* and *B.floridae*. (C-D) Bar charts showing the observed number of CTCF 500bp regions within conserved and non-conserved regions versus the expected number from the background conservation rate between each of *B.lanceolatum* and *B.floridae* (C) and *B.belcheri* and *B.floridae* (D). Exact Binomial tests were carried out and P-values display the significance of observed number of conserved sites being higher than the expected number of conserved sites for each respective species.

3.3.8. Conserved canonical TCF/Lef sites localise into islands across the amphioxus ParaHox cluster

In order to map the position of TCF/Lef sites conserved across the amphioxus ParaHox cluster, the widely accepted bilaterian TCF/Lef canonical consensus motif 5’-CTTTG[A/T][A/T]-3’ was used to identify sites across the *B.floridae* ParaHox reassembly, *B.lanceolatum* Sc0000038 and *B.belcheri* Sc0000020. The ParaHox regions of *B.lanceolatum* Sc0000038 and *B.belcheri* Sc0000020, from CHIC to PRHOXNB, were aligned using MULAN (Ovcharenko et al., 2005) against the same region taken from the *B.floridae* ParaHox reassembly. This alignment was then submitted to MultiTF (through MULAN), searching specifically for the 5’-CTTTG[A/T][A/T]-3’ motif. Whilst many sites (several hundred) are present within each individual assembly, (see Genepalette annotations for each scaffold on disc), only 24 canonical TCF/Lef binding sites within this region were shown to maintain identical positions across the three species of amphioxus. In order to properly visualise these sites with respect to the regulatory landscape of the amphioxus ParaHox cluster, they were then mapped onto a VISTA alignment of the three amphioxus ParaHox clusters (figure 3.11).

What stands out from this analysis (see figure 3.11) is that the majority of these conserved TCF/Lef binding sites localise into distinct clusters throughout the ParaHox cluster, rather than being evenly spread out throughout conserved regions, as one might expect from a general search of TCF/Lef sites within the ParaHox cluster of one of the amphioxus species (see TCF/Lef consensus sites within genepalette scaffold annotation on disc). Of particular note are two ‘islands’ of high TCF/Lef site density; the first lying downstream of Gsx at the 40kb location and containing TCF/Lef sites 6-9, and the second lying upstream of Cdx, between ~91kb and 94Kb and containing TCF/Lef sites 18-22. In addition to these large TCF/Lef 'islands', TCF/Lef 2-3, 12-13 and 16-17 each form pairs within discreet conserved peaks. Indeed, only TCF/Lef sites 1, 10, 11 and 14 exist as single sites within conserved peaks. Finally, looking across the ParaHox cluster as a whole, it becomes clear that the majority, 16/24, of conserved TCF/Lef binding sites cluster in the areas immediately surrounding *Gsx* and *Cdx* ParaHox genes or within their introns.

3.3.9. Transposable elements localise to non-conserved regions of the ParaHox cluster

Previous work has found that transposable elements (TEs) are not actively excluded from the ParaHox cluster, in stark contrast to chordate Hox clusters, which are largely devoid of TEs (Osborne and Ferrier, 2010). In order to build upon this, further efforts were made to identify TEs both within and surrounding the ParaHox cluster of the *B.floridae* reassembly. Both Censor (Jurka et al., 2005; Jurka et al., 1996) and RepeatMasker (A.F.A. Smit et al., unpublished data) http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) were used for the identification of transposable elements, with BLAST searches used to further confirm and identify the LanceleTn Miniature Inverted-repeat Transposable Elements (MITEs) identified in (Osborne and Ferrier, 2010; Osborne et al., 2006). Both CENSOR and RepeatMasker use the most recent version of RepBase as a catalogue for the identification of TEs (Bao et al., 2015) , which contains TEs and repetitive elements from a variety of eukaryotic organisms, including the published *Branchiostoma floridae* elements. TEs were annotated within the *B.floridae* reassembly, starting within the ParaHox cluster, then moving steadily outwards in both flanking directions up to a total of 480Kb surrounding the ParaHox cluster, building upon the identification of amphioxus TEs within (Osborne and Ferrier, 2010; Osborne et al., 2006). The density of TEs across this region was then mapped, counting the number of TEs present in 20kb intervals across the region annotated (figure 3.12), with the ParaHox cluster at the centre of this region at the 200-280Kb intervals. Though no direct conclusions can yet be drawn from this, it appears from an initial analysis that TE density is considerably higher in the region immediately flanking the ParaHox cluster upstream of *Cdx*, specifically in the region surrounding *PRHOXNB* and the 40kb upstream of *Cdx* at the 260-300kb region (figure 3.12).

Looking at the region immediately within and surrounding the ParaHox cluster itself (figure 3.13), we see that 10 TEs lie within the ParaHox cluster itself and of those, eight lie between Gsx and Xlox, **[LanceleTn2 (196bp), Harbinger-N5 (295bp), 2x BflSINEs (121bp and 239bp), R-TEX-10 (174bp), Crack-24 (313bp), LanceleTn3a (398bp) and Harbinger-N13 (202bp)]** and the other two surround *Cdx* exon 2 **[LanceleTn4 (207bp) and LanceleTn3a (190bp)]**. TEs are not present within the

~4kb regions immediately upstream of the ParaHox genes *Gsx, Xlox* and *Cdx*. Upstream of *SCP1*, a further nine TEs exist in the region upstream and within the introns of *CHIC* **[MER6 (42bp), Mariner-N2 (189bp), Ginger 2-1 (90bp), Mariner-N2 (201bp), Mariner 1N1 (59bp), Gypsy-31-LTR (136bp), LanceleTn1 (173bp), RTE-8 (116bp) and BflSINE1 (248bp)]**. The remaining 27 elements in this region are all found upstream of Cdx, within and surrounding *PRHOXNB* and a gene with similarity to an MFS-type transporter Slc18b1 **[R-TEX-7 (164bp), R-TEX-7 (75bp), LanceleTn3a (129bp), LanceleTn3a (129bp), Harbinger-N13 (113bp), Harbinger-N6 (56bp), Sola3-2 (117bp), LanceleTn2 (213bp), 168bp repeat (168bp), LanceleTn2 (68bp), 168bp repeat (63bp), LanceleTn2 (68bp), LanceleTn3b (170bp), Gypsy-31-LTR (61bp), R-TEX-8 (60bp), LanceleTn1 (433bp), LanceleTn2 (200bp), BflSINE1 (117bp), Harbinger-3 (123bp), LanceleTn3b (189bp), Harbinger-N5 (176bp), Academ-2 (56bp), RTE-13 (55bp), LanceleTn-3a (73bp), RTE-13 (66bp), Sola3-3 (46bp), Sola3-3 (146bp).**

Of the 46 TEs identified within this region, 18 represent TEs specifically identified within amphioxus; 14 LanceleTn MITEs and 4 BflSINEs, with the LanceleTn MITEs being the most abundant TE family. The TEs identified in this study localise to the same regions of the ParaHox cluster seen in (Osborne and Ferrier, 2010; Osborne et al., 2006), despite the sequences being isolated from different animals, though the Osborne and Ferrier 2010 paper has several extra TEs marked. This suggests that there may be a pressure to prevent TEs invading some regions of the amphioxus ParaHox cluster.

Finally, when mapped onto the landscape of conservation between the ParaHox clusters of *B.floridae*, *B.lanceolatum* and *B.belcheri*, it is seen that TEs are largely located outside of conserved regions between the three species (figure 3.14). Only TE 1 (MER6) lies in an area conserved between all three species, within the *CHIC* UTR. TEs 4 (Mariner-N2), 9 (BflSINE1), 11 (Harbinger-N5), 14 (R-TEX-10) and 24 (Harbinger-N6) lie within small peaks of conservation between *B.floridae* and *B.lanceolatum*, though only two of these, 11 (Harbinger-N5) and 14 (R-TEX-10) lie within the ParaHox cluster. TEs 27 (LanceleTn2) and 36 (LanceleTn2) lie within small peaks of conservation between *B.floridae* and *B.lanceolatum*, though these lie outside of the ParaHox cluster within (27) and upstream (36) of *PRHOXNB*. As conserved regions are likely to hold regulatory function, this gives support to the hypothesis that the ParaHox cluster excludes TEs from regions of regulatory function.

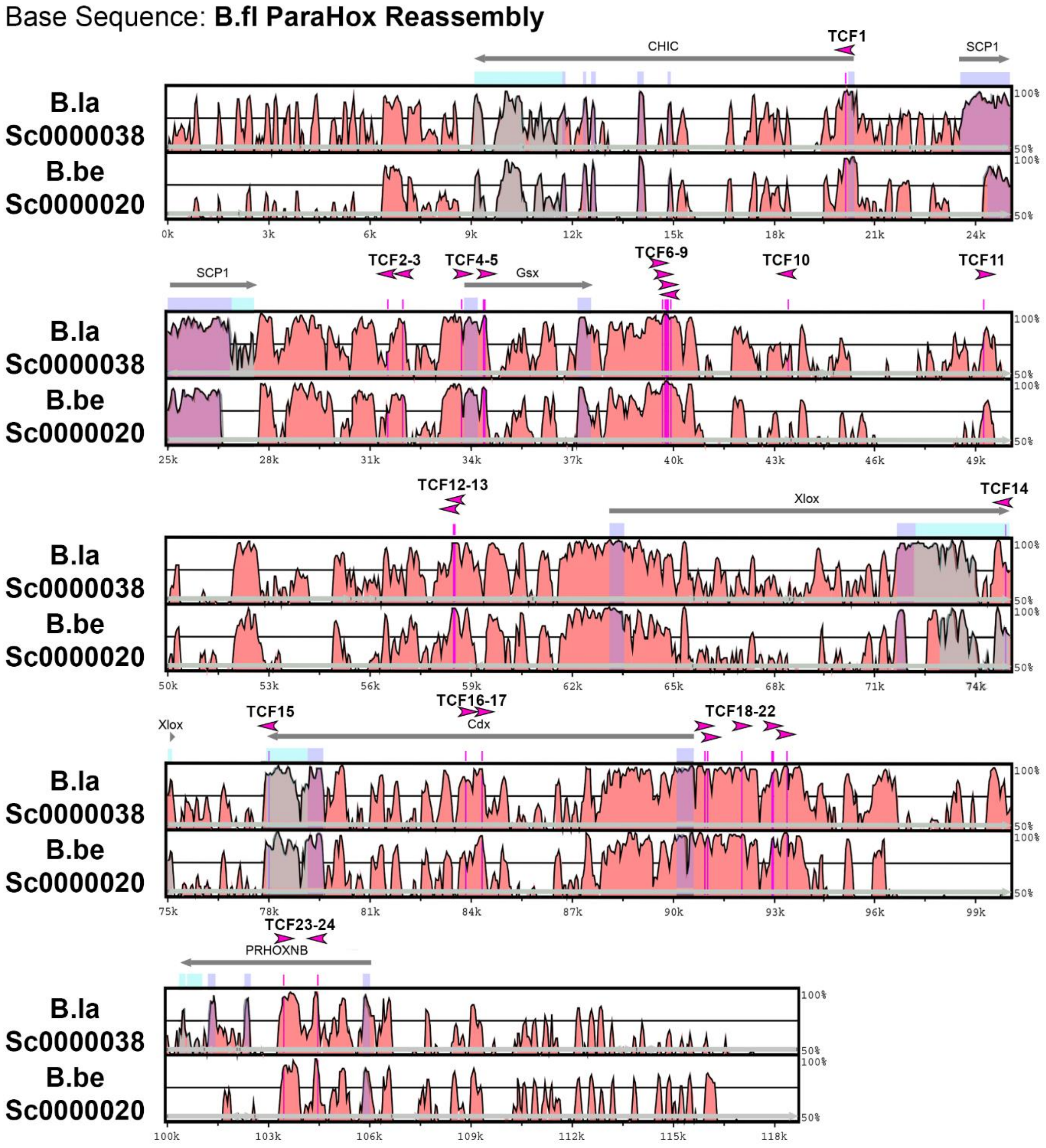


Figure 3.11. TCF/Lef binding sites conserved between the *B.floridae*, *B.lanceolatum* and *B.belcheri* ParaHox clusters group into clusters within conserved peaks.

TCF/Lef binding sites conserved between the ParaHox clusters of *B.lanceolatum* Sc0000038, *B.belcheri* Sc0000020 and the *B.floridae* ParaHox reassembly were identified with MULAN and MultiTF, using the 5'-CTTTG[A/T][A/T]-3' motif (Purple arrowheads and lines within peaks). These were then annotated upon a VISTA alignment of the ParaHox clusters taken from *B.lanceolatum* Sc0000038 and *B.belcheri* Sc0000020 against the *B.floridae* ParaHox reassembly in order to view their position along the ParaHox cluster with respect to conserved sequence. All TCF/Lef sites bar sites 1, 10, 11, 14 and 15 cluster as either pairs or multiple instances within conserved peaks. Coding regions are highlighted in blue/purple, UTRs in light blue and non-coding regions are represented in pink. Each peak corresponds to a conserved region.

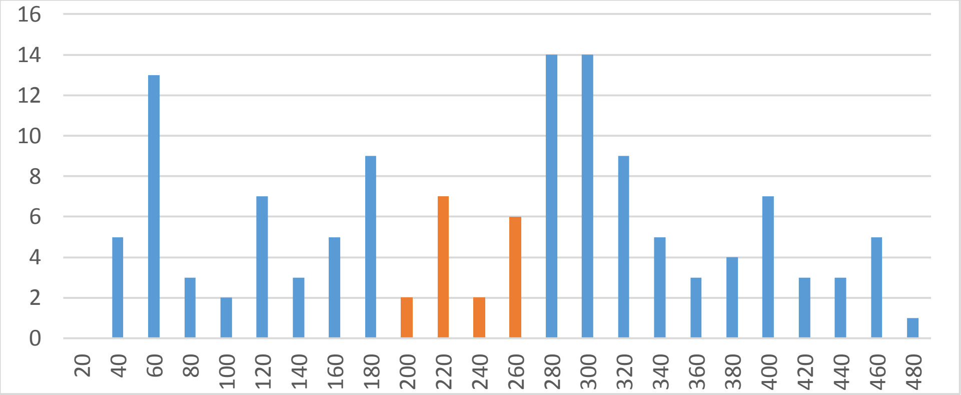


Figure 3.12. Transposable element density across the *B.floridae* ParaHox reassembly.

The density of transposable elements (TEs) was characterised in 20kb windows across a 480kb region surrounding the ParaHox cluster of Gsx-Cdx. A large peak in transposable element density is observed immediately upstream of *Cdx*. The X-axis shows the 20kb region windows, whereas the Y-axis shows the number of TEs observed. The area between *CHIC* and *PRHOXNB* is visible in orange.

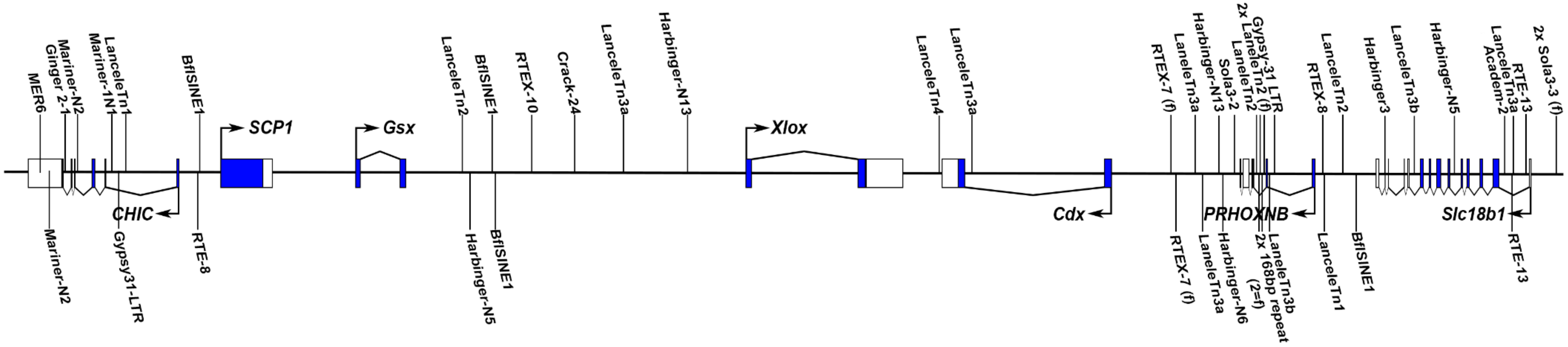


Figure 3.13. Schematic of Transposable elements within and immediately surrounding the B.floridae ParaHox cluster.

A schematic showing the positions of transposable elements identified by Censor and RepeatMasker relative to exon and intron boundaries along the ParaHox cluster and surrounding region of the *B.floridae* ParaHox reassembly. Transposable elements were annotated using Repbase (Bao et al., 2015) TE family identification. 2x represents where two hits to the same TE family lie next to each other, likely due to a splitting of the original element into two fragments. In the case of 2xLanceleTn2 and 2x168bp repeat these elements are present as LancelTn2-168bp-LanceleTn2-168bp. Many of the TEs shown are present only as partial but significantly large copies of the full TE. This analysis includes a larger genomic region than the PAC contig examined within (Osborne and Ferrier, 2010), and includes coverage of the genomic regions flanking the ParaHox cluster.

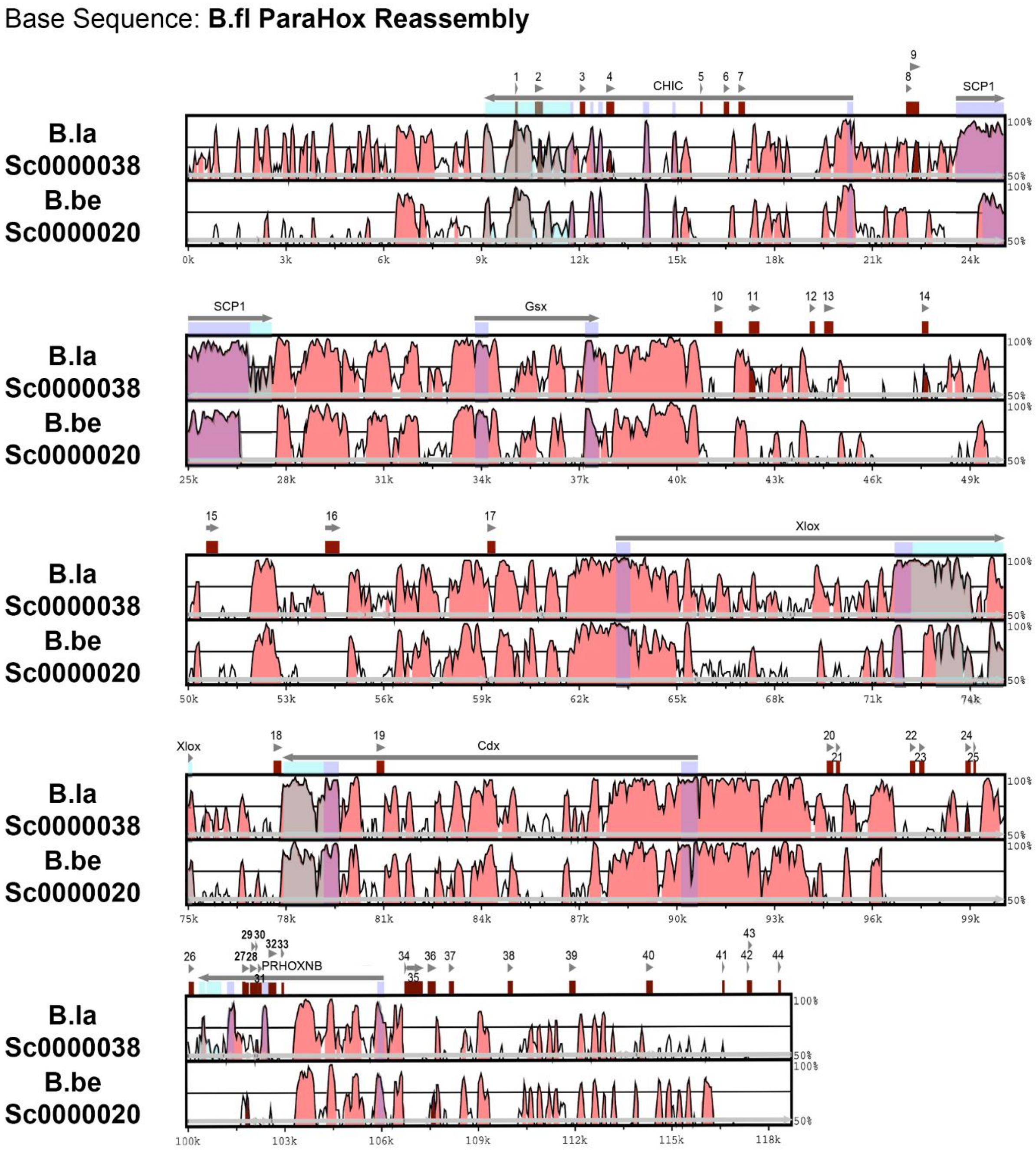


Figure 3.14. *B.floridae* Transposable elements lie within non-conserved regions across the ParaHox cluster.

B.floridae Transposable elements (TEs) (in brown) within and around the ParaHox cluster have been mapped onto a map of conservation between the ParaHox clusters of *B.floridae*, *B.lanceolatum* and *B.belcheri*. TEs locate to areas that are not highly conserved between amphioxus species. Only TE 1 lies in an area conserved between all three species within the *CHIC* UTR. TEs 4, 9, 11, 14 and 24 lie within small peaks (~75% or less) conserved between *B.floridae* and *B.lanceolatum*, though only two of these, 11 (and 14 lie within the ParaHox cluster. TEs 27 and 36 lie within small peaks (~75% or less) conserved between *B.floridae* and *B.lanceolatum*, though these lie outside of the ParaHox cluster within (27) and upstream (36) of *PRHOXNB*. Coding regions are highlighted in blue/purple, UTRs in light blue and non-coding regions are represented in pink. Each peak corresponds to a conserved region.

3.4. Discussion

3.4.1. Improved annotation of the amphioxus ParaHox cluster enables a preliminary analysis of the regulatory landscape.

The poor assembly quality surrounding the ParaHox cluster within the *B.floridae* v1.0 genome (http://genome.jgi.doe.gov/Brafl1/Brafl1.home.html) causes a multitude of issues pertaining to the analysis of the ParaHox cluster. One of these is that gene models are incorrectly predicted and annotated, and EST data is difficult to align and provide accurate gene models for both the ParaHox genes and surrounding genes. In addition to this, the lack of annotation verification of neighbouring genes, beyond automated prediction by fgenesh (Solovyev et al., 2006), makes further attempts at syntenic analysis difficult, as most of these predictions are based on vertebrate models and as such are not always accurate. The improved assembly within the *B.floridae* ParaHox reassembly has allowed the proper annotation of gene models within and surrounding the ParaHox cluster with transcriptomic data providing verification and expansion of gene models. In particular, the genes within and immediately surrounding the ParaHox cluster now have annotated 3' UTRs, including *CHIC*, *SCP1*, *Xlox*, *Cdx* and *PRHOXNB*, though *Gsx* does not have supporting EST data and so no UTR has been annotated for *Gsx* (figure 3.2 and genepalette annotation of the *B.floridae* ParaHox reassembly). In addition, the 3' UTR of Xlox has also been experimentally verified via *in situ* hybridisation (figure 3.4). This was particularly fruitful as this Xlox 3' UTR now provides a means to successfully test the presence of Xlox within RNA pools, which has previously proven to be difficult using the coding sequence (personal communication with several members of the amphioxus community). A 5' UTR of SCP1 has also been identified, and is discussed further in chapter 4. In the annotation of regions outside of the ParaHox cluster proper, several exonic ESTs have been mapped to which there are no BLAST hits to vertebrate or the wider Bilateria. These may represent novel genes that have not yet been characterised, or perhaps multi-exonic long non-coding RNAs.

The main body of this analysis will enable the characterisation and mapping of further regulatory, and non-regulatory features within and surrounding the ParaHox cluster. In addition, the continuation of this annotation may reveal further synteny surrounding the amphioxus and vertebrate ParaHox clusters. Of the genes annotated so far along the 532kb out of 3.5Mb examined, none of the immediately neighbouring genes, other than CHIC and PRHOXNB, show conserved synteny with the ParaHox loci of vertebrates. Since this covers a large region outside of the ParaHox cluster, it is possible that continued annotation will not reveal further genes showing conserved synteny with the ParaHox cluster. It would therefore be more interesting to look at the presence of the genes currently annotated (Appendix 7.1) within species such as *Ptychodera flava* and Patiria miniata, which also exhibit intact ParaHox clusters, in order to examine if any of these genes represent ParaHox neighbours within other phyla. Interestingly, none of the FLT/VEGFR/PDGFR/KIT Superfamily genes appear within the B.floridae ParaHox reassembly. These genes are found adjacent to the mammalian ParaHox clusters (Ferrier et al., 2005), as well as Flt1 adjacent to the *P.flava* ParaHox cluster (Ikuta et al., 2013). Though the B.floridae ParaHox reassembly cannot be used to tell us where these genes are located within amphioxus, the B.floridae v.1.0 genome can be used instead. Within the *B.floridae* genome v1.0, the ParaHox genes, CHIC and PRHOXNB are located adjacent to one another upon scaffolds 24 and 116 (see table 3.2). Though the genomic locations are misleading due to sequencing and assembly errors, *CHIC-Gsx-Xlox-Cdx-PRHOXNB* have also been confirmed to be adjacent and in this order within the amphioxus genome in both the PAC contigs (Ferrier et al., 2005) and within the B.floridae ParaHox reassembly used in this chapter. Members of the FLT/VEGFR/PDGFR/KIT Superfamily, including FLT-1, cannot be found upon these ParaHox scaffolds. Instead, they can be found upon scaffolds 295 and 783 (see table 3.2), with non-ParaHox genes adjacent. Though the FLT/VEGFR/PDGFR/KIT are not linked to the ParaHox cluster within amphioxus, those present upon scaffolds 295 and 783 (which are likely different haplotypes), there is a linkage of up to 3 adjacent FLT/VEGFR/PDGFR/KIT genes. Within the vertebrates, 3 separate families (PDGFRA/B, FLT3/KIT/CSF1R, and FLT1/KDR/ FLT4) make up this superfamily, and lie adjacent to one another within the genome. BLAST searches were unable to resolve the identity of the FLT/VEGFR/PDGFR/KIT superfamily genes located within scaffolds 295 and 783, and this FLT/VEGFR/PDGFR/KIT cluster of genes could represent members from each family, or a tandem duplication of one of these FLT/VEGFR/PDGFR/KIT family members. Having said that, the gene present at scaffold_783:59575-60631 may represent *AmphiFLT-1*, as it bears 3 exons, a number conserved with *P.flava* FLT-1. It should be noted that the vertebrate FLT1, KDR and FLT4 genes contain many more introns, so the presence of 3 exons is by no means characteristic of the FLT1 family. Further phylogenetic analysis with other chordate, and perhaps *Ptychodera*, FLT/VEGFR/PDGFR/KIT superfamily genes would be required to resolve gene identities and family groupings beyond the superfamily level.

It is likely that a genomic rearrangement has occurred within *B.floridae*, resulting in the relocation of either FLT/VEGFR/PDGFR/KIT family members, or perhaps the ParaHox cluster and its immediate neighbourhood (i.e CHIC and PRHOXNB). The maintenance of the *CHIC*-ParaHox-*PRHOXNB* group may then be linked to the regulatory landscape, with a GRB covering this region and ParaHox regulatory elements interdigitated throughout the *CHIC*-ParaHox-*PRHOXNB* region, but not extending to the FLT/VEGFR/PDGFR/KIT genes in amphioxus. This would be in contrast to *P.flava*

(Ikuta et al., 2013) and the vertebrates (Ferrier et al., 2005), where FLT1 has been retained as a ParaHox neighbouring gene. Another possibility is that FLT/VEGFR/PDGFR/KIT regulatory elements are interdigitated amongst the ParaHox clusters of vertebrates and *P.flava*, but not in amphioxus. This could potentially be the scenario if the 2/3 FLT/VEGFR/PDGFR/KIT superfamily genes identified in B.floridae Scaffolds 295 and 783 do represent members of the separate PDGFRA/B, FLT3/KIT/CSF1R, and FLT1/KDR/ FLT4 families, and represent clustered FLT/VEGFR/PDGFR/KIT within both amphioxus and vertebrates. In the case of genomic rearrangement splitting the *CHIC*-ParaHox-*PRHOXNB* and FLT/VEGFR/PDGFR/KIT regions in amphioxus, the FLT/VEGFR/PDGFR/KIT region could represent a further locus with which ancestral ParaHox neighbouring genes could be identified.

Table 3.2 Genomic locations of ParaHox genes and conserved neighbours in *B.floridae* (genome v1.0)

Gene Name	Genomic Coordinates (*B.floridae*)
CHIC	**scaffold_24**:622803-631218 **scaffold_116**:1777081-1785599
Gsx	**scaffold_24**:601597-601803 **scaffold_116**:1759788-1763506
Xlox	**scaffold_24**:487192-497991 **scaffold_116:** missing sequence in scaffold
Cdx	**scaffold_24**:553762-553965 **scaffold_116:** missing sequence in scaffold
PRHOXNB	**scaffold_24**:527677-532480 **scaffold_116**:1694807-1694994
FLT/VEGFR/PDGFR/KIT Superfamily	**scaffold_783**:48227-57595, 59575-60631 (**FLT1?**), 61486-67275 **scaffold_295**:284814-291533, 291776-301611

In addition to the annotation of the *B.floridae* ParaHox Reassembly, the ParaHox clusters of both *B.lanceolatum* and *B.belcheri* (figure 3.3) can now be used in conjunction to identify conserved sequences and motifs that allow improved identification of regulatory features. The ability to compare genomic regions between amphioxus species, both within the ParaHox cluster as well as in other genomic locations, now allows approaches such as phylogenetic footprinting, and provides a much more comprehensive view of the regulatory landscape of the cephalochordate sub-phylum.

3.4.2. High conservation of non-coding sequence within the ParaHox cluster of amphioxus allows the identification of potential regulatory elements.

Previously, the poor assembly quality surrounding the amphioxus ParaHox cluster has made the identification of potential regulatory elements a laborious and difficult process. However, with the release of the *B.belcheri* genome, and upcoming release of the *B.lanceolatum* genome, it is now possible to compare orthologous regions of these three genomes and observe regions of conservation. Indeed the high levels of conservation of non-coding regions observed between *B.floridae*, *B.lanceolatum*, and *B.belcheri* in (figure 3.5) strongly support the presence of conserved ParaHox regulatory elements within the cephalochordates (one of which is more thoroughly examined in chapter 5). Several regions, including the more typical upstream candidate regions as well as intergenic regions, stand out from this analysis as showing high conservation between the three species. This data can be used to further inform the functional analysis of amphioxus ParaHox regulatory elements, providing a strong starting point on which to base the cloning of putative regulatory elements. Previously, in order to identify regulatory elements, identification was targeted to regions expected to show regulatory activity, such as regions immediately upstream of a gene, or perhaps within introns. This was both a time consuming and relatively inefficient method of identifying potential regulatory elements, with many of the regions examined displaying no regulatory function (P. Osborne, unpublished data) at all. In addition, the polymorphism studies required to identify conservation within potential regulatory elements were time consuming, requiring the cloning and sequencing of regulatory elements from multiple individuals. This analysis makes it possible to quickly refine and target potential regulatory elements to regions of conserved sequence. This should decrease the number of regions examined that display no regulatory potential, eliminating trialling those regions that have undergone no positive selective pressure, and also refining regions where excess sequence may interfere with regulatory function in a reporter context.

3.4.3. The ParaHox conservation seen between amphioxus species does not extend out to the vertebrates.

Whilst the ParaHox cluster shows high levels of conservation between *B.floridae*, *B.lanceolatum* and *B.belcheri*, this does not appear to extend out to the vertebrate ParaHox clusters. This is in contrast to the Hox cluster of amphioxus, which does present some CNEs between the amphioxus and vertebrate Hox clusters (Manzanares et al., 2000; Matsunami et al., 2010; Pascual-Anaya et al., 2008; Punnamoottil et al., 2010), particularly between the 3’ regions of Hox clusters.

Even so these are still few, with 75 conserved VISTA hits identified at >60% identity and 50bp length, but only between the amphioxus Hox cluster and individual human Hox clusters (Pascual-Anaya et al., 2008). Of these, only 2 out of 75 amphioxus CNEs aligned with the ‘core regions’ of vertebrate Hox CNES and can be considered conserved across chordates (Matsunami et al., 2010). It is maybe unsurprising that the same conservation between amphioxus and vertebrate ParaHox clusters is not observed, as whilst vertebrates have maintained four intact Hox clusters, only a single intact ParaHox cluster has been maintained. This has resulted in three further vertebrate regions containing a single ParaHox gene each; Gsh2, Cdx1 and Cdx4 respectively, but having lost the other ParaHox paralogues through gene loss (Ferrier et al., 2005). This in turn may have resulted in regulatory elements being spread across these four vertebrate ParaHox regions, much like the Hox though with substantial gene loss. Though there was not time here to do so, further work should carry out similar VISTA analysis between all four vertebrate ParaHox loci and the amphioxus ParaHox cluster, particularly as it has been observed that amphioxus Hox CNEs are only conserved with a single vertebrate cluster (Amemiya et al., 2008; Pascual-Anaya et al., 2008). If anything, such analysis should be easier between the amphioxus and vertebrate ParaHox clusters due to their much smaller size, as specialised software such as Tracker, a Perl-based program had to be used to identify CNEs across the much larger Hox clusters of amphioxus and vertebrates (Amemiya et al., 2008), whilst VISTA could only be used for short sections of the amphioxus Hox cluster (Pascual-Anaya et al., 2008). Even Tracker was run at the limits of its sensitivity within the vertebrate Hox analysis (Amemiya et al., 2008) and the approach using shorter regions within VISTA was much more sensitive (Pascual-Anaya et al., 2008). Thus, it should be entirely possible to identify ParaHox CNEs between the amphioxus ParaHox cluster and other vertebrate ParaHox loci, if they exist, as ParaHox regions are easily aligned within VISTA (see figure 3.6), even though no CNEs are present between the amphioxus ParaHox cluster and the vertebrate ParaHox cluster containing locus.

3.4.4. Binding sites for the ‘insulator associated’ protein CTCF are associated with conserved non-coding regions of the amphioxus ParaHox cluster.

Though Chromatin Immunoprecipitation and chromosome conformation capture analysis are required to fully characterise CTCF binding, the analysis carried out in this chapter show that CTCF sites are not only present (figures 3.7 and 3.8), but also lie within conserved regions of the ParaHox cluster (figure 3.9 and 3.10). This shows similarity to the CTCF sites seen in the vertebrate Hox clusters, where CTCF sites are not only abundant within vertebrate CNEs (Xie et al., 2007), but a CTCF-Hox ‘kernal’ has been identified extending across the Bilateria (Heger et al., 2012). Bearing in mind the importance of CTCF in the regulation of Hox genes (Ferraiuolo et al., 2010; Narendra et al., 2015; Soshnikova et al., 2010), it is intriguing that CTCF sites are also present and conserved within amphioxus ParaHox clusters. The data from ENCODE also shows several sites of experimentally confirmed CTCF binding within the vertebrate ParaHox clusters (Birney et al., 2007; Thomas et al., 2007), providing a further reason to analyse these sites within the amphioxus ParaHox cluster. The presence of large domains of differing CTCF orientation across the *B.floridae* scaffold, as well as immediately surrounding the ParaHox cluster itself (figure 3.7) suggests that CTCF sites could also be directing the formation of TADs within the *B.floridae* genome. CTCF sites have been shown to define the boundaries of TADs in both the Hox cluster (Narendra et al., 2015) of vertebrates and the Six cluster of vertebrates and sea urchins (Gomez-Marin et al., 2015). Indeed several recent studies have added to a growing body of evidence pointing towards the presence of CTCF sites in divergent orientations as an indicator of TAD boundaries, the 'non-looped' structure between TADs, and the importance of CTCF site orientation in gene regulation and TADs (Gomez-Marin et al., 2015; Guo et al., 2015; Vietri Rudan et al., 2015) by forming looped chromatin domains between CTCF sites of a convergent orientation that are 'facing each other' (Rao et al., 2014). The presence of CTCF association with TADs within bilaterian gene clusters, particularly the Hox cluster, provides a strong basis to carry out further work looking into this association within the amphioxus ParaHox clusters. The preliminary bioinformatics analysis shown in figure 3.7, is certainly consistent with CTCF sites forming domains of orientation, particularly within and around the ParaHox cluster, perhaps in a comparable fashion to TADs in the six cluster (Gomez-Marin et al., 2015). In addition to this, we see in figure 3.8 that several pairs of CTCF binding sites with opposing orientation exist within the amphioxus ParaHox cluster, with two such pairs lying upstream of Cdx and another upstream of Gsx. These potential CTCF boundary elements lie either side of the ParaHox cluster and also at the boundary of CTCF orientation domains observed in figure 3.7.

The CTCF sites within the *B.floridae* ParaHox cluster also significantly associate with regions conserved across the amphioxus species (figures 3.9 and 3.10), suggesting an evolutionary pressure and functionality to these sites. These CTCF sites, particularly those within discrete non-coding peaks of conservation, may represent regulatory elements that function in the regulation of the ParaHox cluster and individual ParaHox genes. The role of CTCF sites in insulator elements has been well studied (Bell et al., 1999; Saitoh et al., 2000), and was shown to function within the *Drosophila Abdominal-B* Hox locus via the fab-8 insulator (Moon et al., 2005). These insulators are thought to act by introducing a physical barrier to the interaction of distal enhancers with a promoter by DNA looping in a mechanism termed 'enhancer blocking' (Geyer and Corces, 1992; Kellum and Schedl, 1992) (reviewed in West et al. (2002)). However, evidence suggests that CTCF may play a much more versatile role and also be involved in the function of active enhancers. The presence of CTCF binding elements in promoters (Kim et al., 2007b) and enhancers (Handoko et al., 2011) gives the possibility that CTCF may also be playing a direct role in the function of enhancers, as well as through the more established repression via insulators. A recent study examining the binding sites of CTCF and BORIS, an amniote paralog of CTCF expressed in the germ cells that shares an almost identical DNA-binding domain that recognises the same DNA sequence *in-vivo* and *in-vitro* (Kosaka-Suzuki et al., 2011; Pugacheva et al., 2010; Sleutels et al., 2012; Suzuki et al., 2010), examined the presence of clustered CTCF sites within enhancer and promoter regions. Interestingly, '2XCTS' sites were found to be highly enriched in these regions, confirming a functional significance in transcription. This study also challenges the idea that all CTCF binding sites are equal, with different 1xCTS (single) and 2XCTS (clustered) sites binding single CTCF or BORIS proteins, or BORIS/CTCF hetero- and homodimers under different conditions in cell lines (Pugacheva et al., 2015).

The studies discussed here reveal the breadth and scope of CTCF function, its importance in gene regulation, and also the ongoing discovery of different mechanisms of CTCF action. The versatility and widespread presence of this transcription factor makes it an excellent target for further study as a potential regulator of the amphioxus ParaHox cluster. Further work could test the regions identified in this work within a reporter context, both alone and in front of a strong enhancer to test their regulatory function and screen for insulator function. In addition, chromatin immunoprecipitation and Chromatin Conformation Capture experiments would be key in elucidating the full extent of CTCF binding, and testing the hypothesis that amphioxus ParaHox CTCF binding sites are involved in cluster-wide regulatory mechanisms as well as the formation of TADs. Based on chromatin conformation capture, Hi-C sequencing could also prove valuable in elucidating the presence and function of chromatin loops and CTCF sites in the ParaHox cluster of chordates, and would provide additional spatial context to genomic studies (Belton et al., 2012), allowing the investigation of CTCF involvement in TADs and orientation and 3D-chromatin structure dependent regulatory mechanisms.

3.4.5. Annotation of conserved binding sites demarcates potential regulatory inputs within the ParaHox cluster.

Both CTCF (figure 3.8) and TCF/Lef binding sites (3.10) have been identified across the three amphioxus ParaHox clusters examined, with CTCF sites in *B.floridae* showing a clear correlation with highly conserved sequence between *B.floridae*, *B.lanceolatum* and *B.belcheri*. In addition, TCF/Lef sites conserved between the three amphioxus species have also been identified. The identification of binding sites alone is rather uninformative, particularly for factors such as TCF/Lef, which has a short hexameric binding site, where many hundreds of sites can be identified. However, by identifying sites conserved across the three amphioxus species many of those that are less likely to be functional can be discarded swiftly, leaving those with conserved position. This has proven to be a successful approach for the tunicates, another invertebrate chordate group, where the closely related tunicate species *Ciona intestinalis* and *Ciona savignyi* have been processed in a similar manner to predict transcription factor binding sites. Of those identified in this manner, mutation of these motifs produced observable changes in expression within a reporter background (Chen et al., 2014; Kanda et al., 2013). The use of VISTA analysis in combination also allows the visualisation of where such sites are positioned in respect to conserved sequence, improving both the identification of amphioxus regulatory elements as well as potential regulatory inputs for these elements. This approach could be expanded beyond the CTCF and TCF/Lef factors to other targets that may be of interest, for example retinoic acid response elements (RAREs), as retinoic acid and RAREs have been shown to direct the expression of ParaHox and Hox genes in amphioxus (Manzanares et al., 2000; Osborne et al., 2009; Schubert et al., 2005; Wada et al., 2006).

One functional assay that would greatly improve the identification of regulatory elements, and would complement the conserved binding site and VISTA approaches used in this work would be Chromatin Immunopreciptitation with sequencing, also known as ChIP-seq. This would allow the identification of *in vivo* protein, and protein complex, binding sites that demarcate various regulatory elements in different states. For example, the use of RNA polymerase II (Pol II) or p300 with the histone methylation H3K4me3 to denote active promoters, and histone modifications such as H3K4me1 with H3K27ac to denote active enhancers, H3K4me1 alone for primed enhancers, or H3K4me1 and H3K27me3 to denote closed or poised enhancers (reviewed in Levine et al. (2014) and Shlyueva et al. (2014)). Chip-seq could also be used to examine the binding of other transcription factors, such as TCF/Lef and RAR/RXR (for RA signalling), to examine the state of *in vivo* binding across the entire ParaHox cluster. Currently, Chip-seq experiments are being carried out to determine the state of chromatin modifications across the *B.floridae* genome, with particular focus upon the amphioxus Hox cluster (Personal communication with Jose Luis Skarmeta and Hector Escriva). This data will be extremely useful in determining whether the amphioxus Hox cluster shows similar patterns of chromatin modification to vertebrates, and will strengthen the possibility of these mechanisms occurring more widely throughout the genome, such as in the sister ParaHox cluster. It is unknown, as of writing, which methylations and protein markers this amphioxus Chip-seq will focus upon, and whether it will extend outside of the amphioxus Hox cluster into other regions of the genome. In addition to this, the use of ChIA-PET, or 3C or 4C Chromatin Conformation Capture techniques would also aid in the identification of regulatory elements, as well as providing experimental conformation of their linkage to a target gene or genomic location.

3.4.6. Transposable element content within the ParaHox cluster

With the ParaHox cluster being the sister cluster to the Hox, it was expected that it may also exclude transposable elements. This, however, was not seen to be the case, and three studies have now identified transposable elements throughout the amphioxus ParaHox cluster (Ferrier et al., 2005; Osborne and Ferrier, 2010; Osborne et al., 2006), despite conserved gene spacing and cluster organisation across the Chordata (Ferrier et al., 2005). This places the chordate ParaHox cluster in stark contrast to the Hox cluster of chordates (Osborne and Ferrier, 2010), where TEs seem to be actively excluded and pushed to the 5' and 3' of the chordate Hox clusters (Amemiya et al., 2008). The work here (section 3.3.9 and figures 3.12-3.14) further supports this and an abundance of additional TEs and TE fragments are seen spread throughout the *B.floridae* ParaHox cluster.

It is thought that the ability of TEs to invade the ParaHox cluster, unlike the Hox cluster, may be linked to the open state of the ParaHox cluster within the germline (Osborne and Ferrier, 2010), with the clustered mouse *Cdx1* ParaHox gene shown to be transcriptionally active within germline cells (Kurimoto et al., 2008). This would then suggest additional selective pressures are involved in maintaining an intact ParaHox cluster within the chordates given the propensity of TEs to facilitate genomic rearrangement. One such constraint could be the presence of genomic regulatory blocks (GRBs) and long range enhancers maintaining the relative positions and organisation of genes (Kikuta et al., 2007). The state of TE content has not yet been examined within the recently discovered intact ParaHox clusters of the echinoderms *Patiria miniata* and *Acanthaster planci* (Annunziata et al., 2013; Baughman et al., 2014) and hemichordate *Ptychodera flava* (Ikuta et al., 2013). It would be interesting to examine the TE content of these species to determine whether TEs are invading the ParaHox cluster throughout the Deuterostomia, or if this is a unique feature of the chordate phylum. One interesting aspect that has been made clearer from this analysis is that the density of TEs immediately surrounding the ParaHox cluster is much higher than that of within the ParaHox cluster, particularly surrounding *PRHOXNB*, where 28 TEs exist in the 40kb upstream of *Cdx* (figure 3.10). Of those elements that do exist within the ParaHox cluster, including the regions immediately upstream of *Gsx* and *Cdx*, a total of 10 TEs exist within the ParaHox cluster proper, and all bar two TEs exist in the intergenic region between *Gsx* and *Xlox*, something also seen in (Osborne et al., 2006), though the number of elements described here is less than the 16 described in this intergenic region within the ParaHox PAC sequence (Osborne and Ferrier, 2010). It would thus be interesting to examine whether the Osborne et al TEs similarly localise to non-conserved regions. The remaining two elements, LanceleTn4 and LanceleTn3a, are located either side of Cdx exon2. Taken together, this suggests that there may be some constraint at work preventing TEs from invading some regions of the ParaHox cluster, perhaps instead targeting them to regions less integral to the regulation of the ParaHox genes. Further work would be needed to analyse whether any constraints on the ability of TEs to invade the ParaHox cluster do indeed exist. Several intact ParaHox clusters are now available beyond those analysed in (Osborne and Ferrier, 2010). One could start by analysing the TE content of *B.lanceolatum* and *B.belcheri* to examine if the localisation of TEs across these clusters agree with that of *B.floridae*, as well as those of the aforementioned echinoderm and hemichordate clusters. The presence of TEs within the ParaHox cluster may also serve as a useful tool for the identification of regions with important regulatory function, as TEs are unlikely to invade such regions without harming the development of embryos. When combined with the other analyses carried out in this work, such as the VISTA analysis in section 3.3.4 (figure 3.4), it may help inform the targeted screening for regions of regulatory importance to the ParaHox genes.

The high abundance of TEs within the intergenic region between Gsx and Xlox in amphioxus poses the question as to whether this could serve a functional purpose. The *Gypsy* TE in *Drosophila* functions as an insulator, preventing distal enhancers from interacting with a promoter region, and other TEs have been shown to have epigenetic affects (reviewed in Slotkin and Martienssen (2007)). This raises the question that perhaps the opening of intergenic regions within the ParaHox cluster to TE invasion may be intrinsically involved in their regulation. The presence of TEs between ParaHox genes would place them in regions where insulator elements might be expected. This is particularly notable in as TEs lie within the intergenic region between Gsx and Xlox, but not immediately upstream of ParaHox genes where presumptive promoter regions would lie, nor within intronic regions (bar a single LanceleTn element immediately next to Cdx exon 2). Of course, this is just speculation, but may be worth further investigation. Still, it does appear from this study, and by comparison with previous work by Osborne et al. and Ferrier (Ferrier et al., 2005; Osborne and Ferrier, 2010; Osborne et al., 2006) that TEs are excluded from the regions where promoter and enhancer elements may be expected, in the immediate upstream and intronic regions of the ParaHox genes, suggesting some constraint on TE invasion into the ParaHox cluster.

Chapter 4. SCP1: An example of retrogene replacement adjacent to the amphioxus ParaHox cluster.

4.1. Introduction.

Retroposons are a class of transposable element, and much like those discussed in chapter 3 result from the duplication of one element and insertion of the retrocopy into a distant, otherwise unrelated region of the genome. Unlike retrotransposons, retroposons are non-autonomous and do not encode their own reverse transcriptase, but are created by RNA-based duplication, in which an RNA intermediate is reverse transcribed into cDNA and reintegrated into a new location in the genome (Weiner et al., 1986). These retroposons are represented by the long interspersed elements (SINEs and LINEs), such as the amphioxus *B.floridae* SINE elements (Holland, 2006), and also the retrogenes. Retrogenes are a peculiar class of retroposon in which a multi-exonic parent gene is duplicated, giving rise to a single-exon daughter retrocopy due to the reverse transcription of the RNA tending to take place after splicing has occurred (Mighell et al., 2000). This strange class of gene was first identified in the rodent genome, where the rat *Insulin I* gene was discovered to be a functional retrocopy of the *Insulin II* gene (Soares et al., 1985). Interestingly, this method of gene duplication proved to be widespread, with examples soon found in other mammalian genomes; in both human (McCarrey and Thomas, 1987) and mouse (Ashworth et al., 1990), as well as in *Drosophila* (Betran et al., 2002; Long and Langley, 1993). It has even been hypothesised that a burst of retroduplication within the primates (Ohshima et al., 2003), particularly the retroduplication of genes within the lineage leading to humans may be responsible for a substantial amount of genomic innovation leading up to and within the human lineage (Marques et al., 2005).

Now, almost 4000 retrogene copies are thought to exist in the human genome (Marques et al., 2005), and though it is unknown how many of these are actually transcribed, 4 – 6% of these were found to be abundantly expressed (Harrison et al., 2005). In one case, over 1000 transcribed retrogenes were identified, 120 of which had evolved into fully functional genes in their own right (Vinckenbosch et al., 2006). This would put the figure at 30.1% of (human) retrogenes being transcribed, whilst other studies put the estimate at 2-3% (Yano et al., 2004; Yu et al., 2007b), or even ~1% (Sakai et al., 2007). Still, despite these discrepancies it remains clear that transcribed and even functional retrogenes do exist. The transcriptional analysis of retrogenes narrowed the number of functional, bona fide retrogenes within the human genome down to ~117 (Vinckenbosch et al., 2006). One study has even identified approximately 8000 non-functional, processed copies of genes (retropseudogenes) within the human genome, suggesting a very high rate of retrotransposition in the human lineage (Zhang et al., 2003). In contrast, the number of retropseudogenes within

Drosophila is approximately 20, and the total number of retrogenes is about 100, with a sixth of these as candidate processed pseudogenes (Harrison et al., 2003). The rate of occurrence of functional retrogene recruitment is thought to be 1 every million years for the human lineage (Marques et al., 2005), and 0.51 retrogenes/million years within *Drosophila* (Bai et al., 2007), and the lack of retrogenes within the *Drosophila* genome is thought to be due to the high rate of DNA loss within *Drosophila*. This might reduce the likelihood of retrocopies becoming fixed genes by causing them to become deleted before they are able to recruit upstream regulatory elements, or due to strong negative selection (Harrison et al., 2003; Petrov, 2002; Petrov and Hartl, 1998).

Interestingly, retrogenes are often found to be expressed in the testis and in fact most genes that give rise to retrogenes are those that were originally expressed within the testis (Marques et al., 2005; Vinckenbosch et al., 2006), with 22.9% of intact retrocopies having evolved a function, or simply expression, within the testis. It seems that many retrogenes may even be initially transcribed within the testis, before gaining additional functions, due to the promiscuous transcription in this tissue. This is largely because of the permissive state of chromatin within germ cells that results from extensive repackaging of DNA during spermatogenesis (Soumillon et al., 2013). This initial expression within the testis may prevent pseudogenisation, and allow the acquisition of new regulatory elements and more defined expression within other tissues within older retrogenes. This is highlighted in the disparity in testis expression between young and old retrocopies, where 10.7% of young retrocopy ESTs are found in the testis, compared to the smaller bias towards testis expression, 5.4%, of ESTs for older retrocopies (Vinckenbosch et al., 2006). This has led to the 'Out of the Testis' hypothesis, in which functional retrogenes often emerge from the testis, whether there is function within the testes or not (Kleene et al., 1998).

In a similar manner, the 'Out of the X' hypothesis provides another avenue for the production of functional retrogenes. During the meiosis of spermatogenesis, the sex chromosomes are heterochromatinised and segregated into the 'XY body' (Solari, 1974), and the transcriptional machinery is excluded from this region (Ayoub et al., 1997; Richler et al., 1994). This poses a particular problem for the expression of essential housekeeping genes upon the X chromosome, as transcription is repressed for six days in mice, and 15 days in humans (Goetz et al., 1984; Wang, 2004). This method of silencing is distinct from female X-inactivation and the Xist RNA is not required for silencing within the XY body in male spermatogenesis (McCarrey et al., 2002). In this case, the X chromosome of Humans, mouse and flies have generated many functional retrogenes, which can be found on the autosomes (Betran et al., 2002; Emerson et al., 2004). These retrocopies are then able to carry out essential functions that would otherwise be lost during the silencing of the X chromosome (McCarrey and Thomas, 1987) . Several of these 'Out of X' genes that can be considered essential have been described (Emerson et al., 2004), including the ribosomal protein retrogenes RPL36AL and RP10L, of which the RP10L retrogene is conserved across mouse, rat and dog genomes (Vinckenbosch et al., 2006). These mark particularly interesting cases, as ribosomal gene duplicates were previously thought to not be viable due to dosage constraints (Uechi et al., 2001; Yoshihama et al., 2002; Zhang et al., 2002).

Rather than being ‘dead on arrival’, doomed to pseudogenisation and loss, retrogenes are now seen to make a significant contribution to molecular evolution (Brosius, 1991). This functionality of retrocopies does not only extend to housekeeping and testis-specific genes however, and several well-known phenotypic and disease associated retrogenes exist. In domestic dogs for example, all short-legged breeds carry a retrogene copy of FGF4, and it is known to be responsible for chondrodysplasia in these breeds (Parker et al., 2009). The vertebrate RHOB gene, a tumour suppressor gene (Prendergast, 2001), also originates from retroposition early in vertebrate evolution (Sakai et al., 2007). Two other notable retrogenes have become so indispensable that mutations in these genes cause a severe disease phenotype. *TACTSTD2* is one of these, where mutation causes gelatinous drop-like corneal dystrophy, an autosomal disease that leads to blindness (Tsujikawa et al., 1999). Likewise, UTP14B is another disease-causing retrogene that originated out of the X-chromosome and plays an essential role in mammalian spermatogenesis, and deletion causes a severe recessive spermatogenic defect. Interestingly, the parent gene has undergone independent duplications in both mouse and human (UTP14C) (Bradley et al., 2004).

Since many examples exist of retrogenes becoming bona fide genes in their own right, it is maybe unsurprising that several of these retrogenes have in fact replaced the original gene, in function as well as resulting in a loss of the parental gene from the genome. This interesting phenomenon is known as retrogene replacement (Krasnov et al., 2005), or ‘orphaned retrogenes’ (Ciomborowska et al., 2013), and has been documented largely by the observation of single exon gene copies but a lack of multi-exonic parent copy. This retrogene replacement has been well documented in the case of the *Iroquois*-*Sowah* locus of bilaterians (Maeso et al., 2012). This syntenic block, in which *Iroquois* genes are linked to the ankyrin-repeat-containing *Sowah* in most bilaterians despite 600 million years of evolution, is seen to be dismantled by retrogene replacement in the tetrapods. Interestingly, despite *Sowah* genes no longer being linked to the *Iroquois* locus in tetrapods, *Irx cis*-regulatory modules are still seen to remain within the pseudogenised remnants of *Sowah* genes within the *Iroquois* loci (Maeso et al., 2012). This retrogene replacement of Sowah from the Iroqouis locus shows how retroduplication of a gene can lead to the loss of a parental gene, despite otherwise strong evolutionary pressures to maintain a genomic position. Retrogene replacement is one mechanism that has been heavily utilised within the genomes of tunicates, whose rapidly evolving genomes have undergone huge compaction and gene loss (Lemaire, 2011; Seo et al., 2001)(reviewed in Berna and Alvarez-Valin (2014)). The process of retrogene replacement is thought to contribute significantly to the loss of introns within tunicate genes. For example, the TEPP proteins, which are expressed within the testis, prostate and placenta of humans, are multi-exonic throughout the metazoans. However, tunicate TEPP genes are intronless, and no multi-exonic parental genes exist in this lineage, suggesting retrogene replacement of these genes (Hahn, 2009). Comparisons between *Ciona intestinalis* and *Ciona savignyi* have suggested that retrogene replacement is a major force for the generation of the large amount of intron-less genes in the tunicate lineage, and also accounts for many differences between the gene content of the two species (Kim et al., 2014).

SCP1 (or SYCP1) belongs to a group of proteins that form the synaptonemal complex (SC), a protein complex that forms a zipper-like structure that aligns homologues and allows crossover recombination during meiosis (Page and Hawley, 2004; Zickler and Kleckner, 1999) and consists of three main proteins in humans; SYCP1, SYCP2 and SYCP3. SYCP1 forms the central element, and is a transverse filament protein made up of coiled-coil domains (Meuwissen et al., 1992), whilst SYCP2 and SYCP3 form the lateral or axial elements of the synaptonemal complex (Dobson et al., 1994; Offenberg et al., 1998). Both SYCP1 and SYCP3 are so far the only structural SC proteins authenticated in mammals (Liu et al., 1996; Yuan et al., 1998), and SYCP1 has been shown to form SC-like structures on its own, forming 'polycomplexes' that represent stacks of SC central regions (Liu et al., 1996). Whilst SYCP1 has been observed within the vertebrates (Casey et al., 2015; Costa et al., 2005; Iwai et al., 2006; Meuwissen et al., 1992; Meuwissen et al., 1997; Qiao et al., 2012; Zheng et al., 2009), the conservation of this gene has since been discovered throughout the Metazoa (Fraune et al., 2012), with the presence of SYCP1 proteins within the Cnidaria (*Hydra*) and even Porifera (*Amphimedon*). Interestingly, it appears that the ecdysozoans lack SCP1 genes, and have instead convergently evolved functionally similar proteins to fulfil the role of SCP1. These ecdysozoan synaptonemal complex genes are also coiled-coil containing transverse filament proteins, like SCP1, and fulfil the same functional role as SCP1 in other metazoans, (Bogdanov et al., 2003; Page and Hawley, 2001; Schild-Prufert et al., 2011), but are non-homologous to SCP1. More intriguing still is that different ecdysozoan phyla have independently evolved separate, non-homologous proteins to fulfil the role of SCP1. *Drosophila* has one such protein, encoded by the *c(3)G*/*CG17604* gene (Bogdanov Iu et al., 2002; Bogdanov et al., 2003; Page and Hawley, 2001), whilst *C.elegans* has evolved a whole host of proteins, SYP1, SYP2, SYP3 and SYP4, that interact to perform the same function as the single SCP1 and *c(3)G* genes, (Colaiácovo et al., 2003; MacQueen et al., 2002; Schild-Prufert et al., 2011; Smolikov et al., 2007).

Being crucial to the formation of the SC complex during meiosis, it is unsurprising that SYCP1 proteins have been observed to show expression within the germ cells (Iwai et al., 2006; Zheng et al., 2009) and even a promoter fragment of SYCP1 was shown to be sufficient to drive germline expression in zebrafish, without requiring additional regulatory elements (Gautier et al., 2013). Similarly, in *Hydra*, SYCP1 expression was observed within the basal located cells of *Hydra* testis. The importance of SYCP1 to the meiotic function is highlighted in the case of *SYCP1-/-* mice, in which synapses do not form and meiosis does not progress. The expression of SYCP1 in germ cells, especially the testis, makes it a potential target for the 'Out of the testis' route of retrogene production. This is indeed seen to be the case in mice, where at least one *SYCP1* retrocopy, *Sycp1-ps1*, is present in many related *Mus* subspecies, though it has been partially pseudogenised and is no longer transcribed (Sage et al., 1997). Interestingly, in the lab strain of *Mus musculus*, a further retroposition of SYCP1 has occurred, this time resulting in the transcribed *Sycp1-ps2* gene.

The ParaHox cluster of chordates has been shown to be open to invasion by retrogenes (Osborne and Ferrier, 2010; Osborne et al., 2006), perhaps due to Cdx transcription in the germline opening the cluster to these transposable elements (Kurimoto et al., 2008). These transposable elements include many retroposons, of which the most frequent are the BflSINE elements (Holland, 2006; Osborne and Ferrier, 2010; Osborne et al., 2006). Most notably, and unique to the amphioxus ParaHox cluster, is the presence of an intron-less copy of the *SCP1*, or *synaptonemal complex protein 1*, gene, just upstream of *Gsx* (Ferrier et al., 2005). This ParaHox-neighbouring *SCP1* is the only copy within the *B.floridae* genome, and there is no multi-exonic parental copy. This is in contrast to the case in vertebrates (de Vries et al., 2005), and SYCP1 in humans and zebrafish has 32 introns, whilst mouse has 33. In addition, SYCP1 genes within human and mouse are located upon chromosomes 1 and 6 respectively (SYCP1 exon number and chromosomal location from ensemble gene browser, http://www.ensembl.org/), whilst the four ParaHox loci are located on Chr:13, 4, 5 and X in human, and Chr:5, 5, 18 and X in mouse (Ferrier et al., 2005). AmphiSCP1 thus most likely constitutes an example of retrogene replacement. This makes it an excellent case in which to study how the regulation of retrogenes is affected when they enter a new locus, particularly as the ParaHox genes are regulated in a complex fashion, possibly via long range (maybe pan-cluster) mechanisms (Osborne et al., 2009) (reviewed in Garstang and Ferrier (2013)). The likely dense regulatory landscape of the ParaHox cluster gives an opportunity to examine both the regulation of *SCP1* and of the surrounding genes. Has *SCP1* perhaps co-opted local regulatory elements, or is it still limited to germ cell expression as in other species? This may point to pan-cluster regulatory elements, or co-option of nearby non-ParaHox regulatory elements if *SCP1* shows unexpected expression patterns, or perhaps insulating elements if the expression of *SCP1* is limited to germ cells.

Aims

Examine if *SCP1* is present as a retrogene in other amphioxus species, or whether this is a novelty of *B.floridae* and characterise amphioxus *SCP1* expression to observe whether *AmphiSCP1* shows canonical germ cell expression, or if an unexpected expression pattern is observed beyond this. Bioinformatic approaches are used to further characterise the *AmphiSCP1* retrogene, and determine the gene structure using transcriptomic data. Finally, the evolution of SCP1 protein sequence across the metazoa is examined with alignments and phylogenetic analysis.

4.2. Methods

4.2.1 Bioinformatic Approaches

The position of *AmphiSCP1* on the *B.floridae* ParaHox reassembly was confirmed by TBLASTN search against the *B.floridae* ParaHox reassembly, using *M.musculus SYCP1* peptide sequence as a query sequence, and also through a BLASTN search using the previously identified *AmphiSCP1* nucleotide sequence from the *B.floridae* ParaHox PACs. The resulting *B.floridae SCP1* nucleotide and peptide sequence was then used as a query to perform both BLASTN and TBLASTN searches against *B.lanceolatum* Scaffold_0000038 and *B.belcheri* Scaffold_0000020 to confirm the presence of the *AmphiSCP1* retrogene adjacent to the *B.lanceolatum* and *B.belcheri* ParaHox clusters. *B.floridae SCP1* 5’ and 3’ EST reads were obtained through BLASTN searches against the NCBI EST database using the *B.floridae* ParaHox reassembly *SCP1* nucleotide sequence.

SCP1 protein sequences were acquired by either TBLASTN or BLASTP searches using the *B.floridae SCP1* or *M.musculus SYCP1* peptide sequences as a query against protein, transcriptomic shotgun assembly, whole genome shotgun assembly and EST databases using NCBI, UNIPROT and JGI databases. Sequences were then aligned using CLUSTAL Omega within Jalview, using the default settings. An 83 amino acid (aa) ‘CM1’ conserved domain, identified within (Fraune et al., 2012), was extracted and used to determine evolutionary relationships. ProtTest3.3 was used to infer the best-fit model for building phylogenetic trees. Both Neighbour Joining and Maximum Likelihood trees were determined using MEGA6 and PHYML respectively. CDCC39 sequences from human, sea urchin and fruit fly were obtained and used as an outgroup to help root the phylogenetic trees.

4.2.2. Experimental Approaches

B.floridae SCP1 was cloned from amphioxus ParaHox PAC clone 33B4 (Ferrier et al., 2005), whilst *B.lanceolatum SCP1* was cloned from adult cDNA according to sections 2.2.2-2.2.6. The primers detailed in table 2.14 were used to clone *B.floridae* and *B.lanceolatum SCP1*, and ends of transcripts sequenced according to section 2.2.6 using the T7 and SP6 primers (see table 2.9.). The following sequencing primers were used to sequence through the centre of the *B.lanceolatum* transcript.

Table 4.1. SCP1 Sequencing Primers

Primer name	Sequence
B.la SCP1-centre F	**AGTCTCTTCAAGATCAGCTGCAA**
B.la SCP1-centre R	**CTTTATCTTCGATGGTTTTCTTCA**

An antisense RNA probe was synthesised according to section 2.2.9. and hydrolysed according to section 2.2.10. In situ hybridisation of *B.la-SCP1* and *B.fl-SCP1* was then carried out according to section 2.2.12 with *B.lanceolatum* and *B.floridae* embryos respectively.

4.3. Results

4.3.1. Amphioxus *SCP1* is a transcribed retrogene adjacent to the ParaHox cluster that has led to retrogene replacement of the parental copy.

BLASTN searches using the SCP1 coding sequence from the *B.floridae* ParaHox PACs against the *B.floridae* ParaHox reassembly confirmed that *B.floridae* SCP1 was indeed upstream of *Gsx* and present as a single coding exon (figure 4.1 A). BLASTN searches using this *B.floridae SCP1* sequence to search against both *B.lanceolatum* Scaffold 0000038 and *B.belcheri* Scaffold 0000020 also revealed that SCP1 is present in the same location, and as a single coding exon, in both of these species (figure 4.1 B-C), revealing that the SCP1 retrotransposition must have occurred prior to the divergence of the *Branchiostoma* group. BLAST searches against the genomes of these three amphioxus species reveal no other SCP1 copies.

Whilst the whole coding sequence for *SCP1* is present in *B.floridae* and *B.lanceolatum*, *B.belcheri* Sc0000020 contains only the central region of *SCP1* coding sequence as the 5' adjacent sequence does not match SCP1 and seems to be unrelated non-coding sequence, and the 3' adjacent sequence is represented by a string of N's. This is likely due to the low quality sequence in this region, or problems with the assembly within v15h11.r2 rather than *B.belcheri SCP1* being incomplete. The position of amphioxus SCP1 genes is given relative to the surrounding *CHIC* and *Gsx* genes in figure 4.1 for *B.floridae* (figure 4.1 A), *B.lanceolatum* (figure 4.1 B) and *B.belcheri* (figure 4.1 C).

In order to examine whether the amphioxus SCP1 retrogene is also transcribed, BLASTN searches using the obtained *B.floridae* and *B.lanceolatum* coding sequences, as well as surrounding non-coding sequence, were performed against the *B.floridae* EST database and *B.lanceolatum* transcriptomic shotgun assembly database respectively. This revealed transcripts from *B.floridae* (Accession numbers: BW697675.1, BW716295.1) (examined further in section 4.3.3) and *B.lanceolatum* (Accession number: JT853830.1). In addition, to provide independent experimental confirmation of SCP1 expression, *B.lanceolatum* SCP1 was cloned from whole adult cDNA and sequenced. This cloned cDNA sequence, along with the translated peptide sequence, is given in Appendix 7.4.

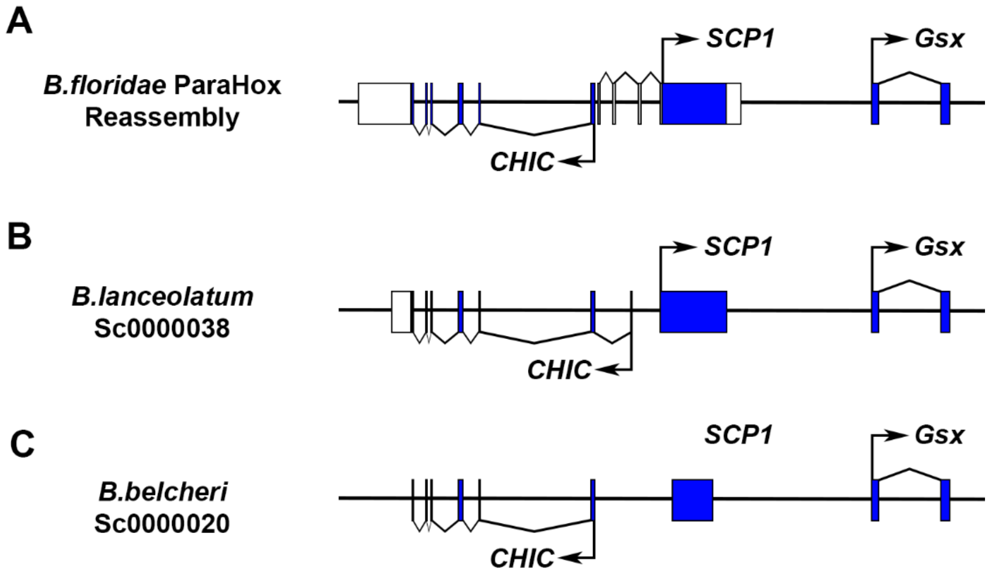


Figure 4.1. Comparison of amphioxus *SCP1* sequences and positions.

(A) A schematic of the *B.floridae SCP1* gene with relative positions of coding sequence and identified 5' and 3' UTRs with respect to the surrounding genes. (B) A schematic of the *B.lanceolatum SCP1* gene with relative positions of coding sequence with respect to the surrounding genes. (C) A schematic of the *B.belcheri SCP1* gene with relative positions of coding sequence with respect to the surrounding genes. *B.belcheri SCP1* is missing both the 3' and 5' ends of the coding sequence. Arrows at right angles (↲) indicate translational start sites and orientation of transcription. Transcription start sites are unknown.

4.3.2 Amphioxus SCP1 displays extensive embryonic expression in somatic tissue.

As experimental confirmation of *AmphiSCP1* expression had been obtained from whole adult cDNA, as well as transcriptomic data (section 4.3.1) the next logical step was to perform *in situ* hybridisation on *AmphiSCP1* in order to visualise the spatio-temporal expression of the transcript. *B.lanceolatum* embryos were used for the majority of the SCP1 *in-situ* hybridisation experiments, as material from this species was more readily available, and the embryos were collected much more recently and of better quality. A time course of amphioxus embryos from mid-gastrula (m.g) to pre-mouth stages were used as this covers stages where typical germ cell markers such as *nanos* and *vasa* are seen to be expressed (Wu et al., 2011b), as well all three ParaHox genes (Osborne et al., 2009). In-situ hybridisation was then carried out upon *B.floridae* embryos in order to confirm that *AmphiSCP1* expression was consistent between the two amphioxus species.

This time course reveals that *B.lanceolatum SCP1* shows unexpected and extensive expression throughout multiple somatic tissues (figure 4.2). Expression is first seen in the mid-late gastrula within the endoderm (black arrowhead) and dorsal mesoderm (white arrowhead) (figure 4.2. A-C), before becoming restricted to the central endoderm and dorsal mesoderm in the early neurula. At this stage, no transcription is seen in either the extreme posterior of extreme anterior of embryos (figure 42 D-E). This endoderm and mesodermic expression pattern continues through the mid neurula stage (figure 4.2 F-H). *SCP1* expression remains absent from the posterior tailbud region. This becomes more apparent at the late-neurula stage where expression is notably absent from the posterior, whilst an anterior patch of expression below the forming cerebral vesicle becomes stronger (black arrow) (figure 4.2 I-K). The premouth stage was the final stage examined, and *SCP1* appears to be expressed throughout the endoderm and mesoderm, though is still excluded from the tailbud, CNS and ectoderm (figure 4.2 L).

B.floridae SCP1 expression was only observable within neurula stages of the embryos examined, as poor staining and bad morphology hampered efforts to obtain a full *B.floridae SCP1* time course. However, reliable expression patterns were obtained for neurula stages. Much as with *B.lanceolatum SCP1*, *B.floridae SCP1* exhibits expression within the endoderm (black arrowhead) and dorsal mesoderm (white arrowhead) in the mid-neurula stage (figure 4.3 A-D), with expression stronger within the anterior, again just below the forming cerebral vesicle (black arrow). This expression pattern carries through to the late neurula, where the presence of expression becomes more apparent in the area below the presumptive cerebral vesicle (figure 4.3 E), and remains clear within the endoderm (figure 4.3 E-F). As observed in *B.lanceolatum*, expression is absent within the tailbud, ectoderm and CNS of *B.floridae* in all stages examined (figure 4.3 A-F).

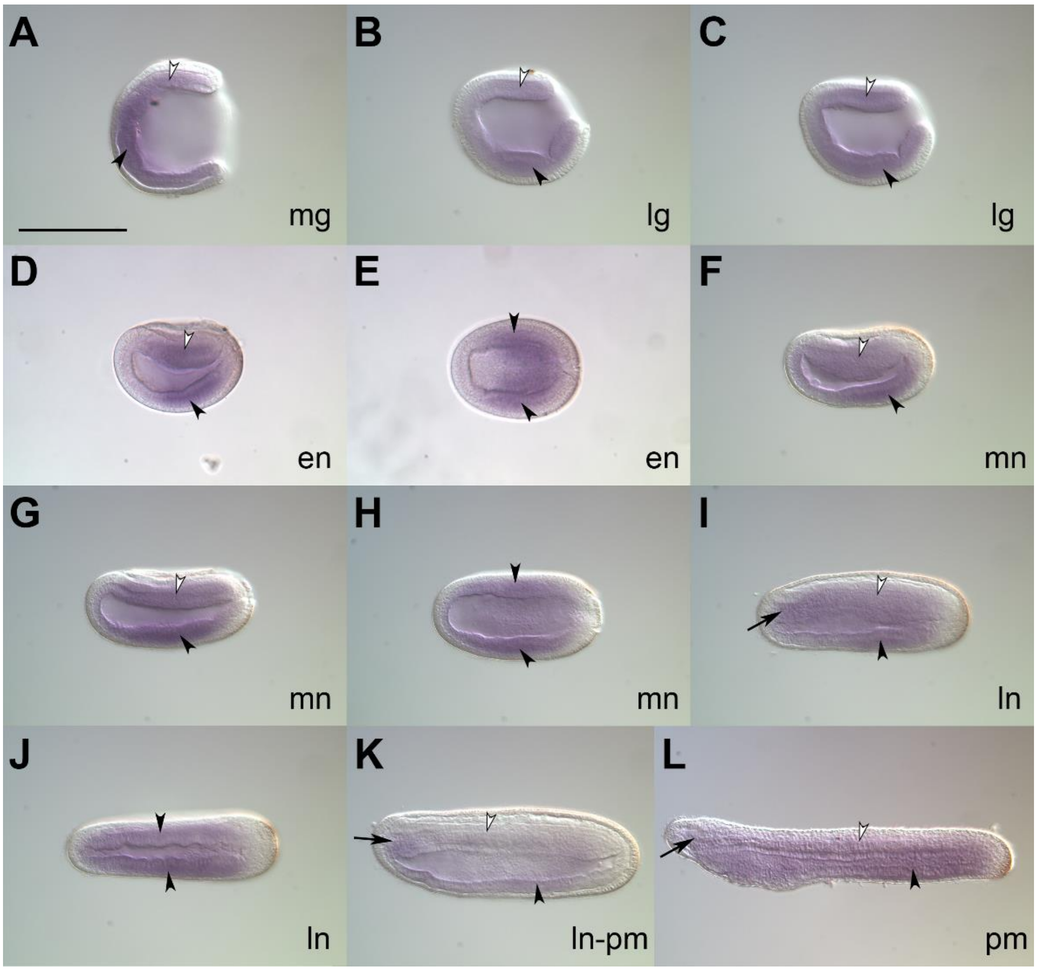


Figure 4.2. Embryonic expression of B.lanceolatum SCP1

B.lanceolatum SCP1 expression begins in the endoderm (black arrowhead) and dorsal mesoderm (white arrowhead) at the mid-gastrula stage through to the late gastrula (A-C) before becoming more refined to the centre of the animal, and excluded from the anterior and posterior poles in the early neurula (D-E). This expression pattern continues into the mid-late neurula (F-J). Expression reaches anteriorly to a region below the forming cerebral vesicle (black arrow) throughout the late neurula-premouth (I-L), whilst expression elsewhere becomes much more diffuse throughout the somites and endoderm (I-L). (A-D, F, G, I, K, L) represent lateral views, whilst (E, H, J) represent dorsal views. mg- mid gastrula, lg- late gastrula, en- early neurula, mn- mid neurula, ln- late neurula, pm-premouth. Scale bar represents 100 µm.

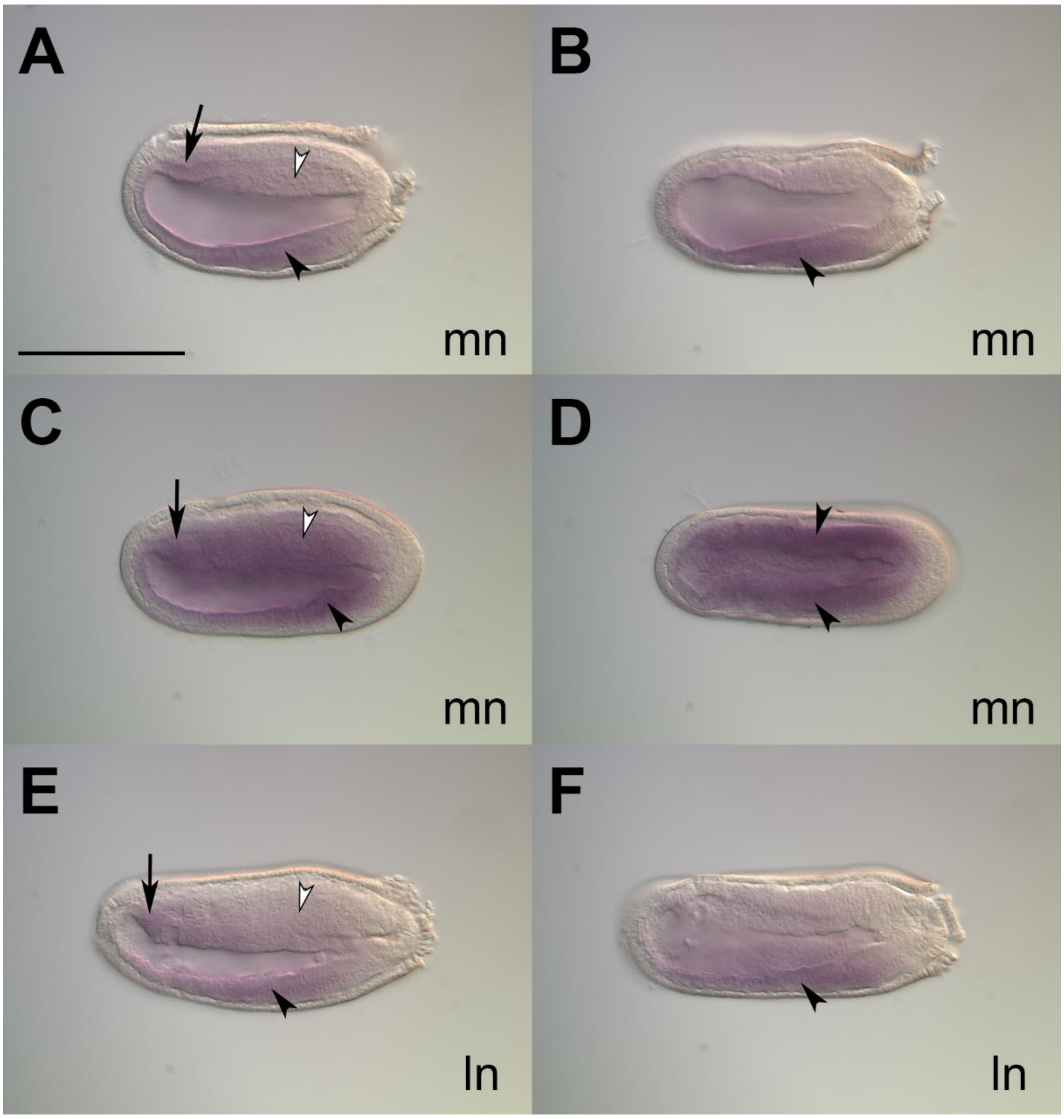


Figure 4.3. Embryonic expression of *B.floridae* SCP1

B.floridae SCP1 expression is observed in the mid-neurula, diffuse throughout the somites (white arrowhead) and endoderm (black arrowhead), with a stronger patch below the presumptive cerebral vesicle in the anterior (black arrow) (A-F). Lateral views are given in (A, C, E), whilst (B, D, and F) represent dorsal views with the focal plane through the endoderm and archentron to highlight endodermal expression. Panels (B) and (F) represent embryos where endodermal expression is much stronger within one side of the embryo, though this is likely an artefact of how the embryos were lying whilst staining, as (B) is in the left whilst (F) is in the right hand side. The embryo shown in (D) exhibits expression in both left and right sides of the endoderm. Earlier and later stages are unavailable due to poor embryo and staining quality. mn- mid neurula, ln-late neurula. Scale bar represents 100 µm.

4.3.3. *B.floridae SCP1* has a 3' UTR and multi-exonic 5' UTR.

B.floridae SCP1 has previously been described as a retrogene, as it contains a single open reading frame with no introns within the amphioxus ParaHox PAC clones 33B4 and 36D2 (Ferrier et al., 2005). As a complementary approach to *in situ* hybridisation, both TBLASTN and BLASTN searches were performed using the *B.floridae SCP1* peptide and coding nucleotide sequences as queries against the *B.floridae* EST database. This revealed a *B.floridae* cDNA clone, bfad022l10, containing 5' and 3' ESTs that align to *B.floridae SCP1* coding sequence, and immediately flanking non-coding sequence (figure 4.4). This EST clone was obtained from whole adult animal, which would be consistent with SCP1 expression within meiotic cells within the testes and ovaries.

The 3' EST, bfad022l10 3' (accession number **BW716295.1**), designates a 685bp 3'UTR immediately adjacent to the coding sequence of *SCP1*. This represents a single exon containing the *SCP1* coding sequence and 3'UTR. As expected, the 5' EST, bfad022l10 5' (accession number **BW697675.1**), aligned to the most 5' coding sequence of *B.floridae SCP1*, with a 334bp alignment covering this region. Additionally, a short 53bp region immediately 5' and adjacent to the coding sequence was also hit by the EST, designating 5' UTR sequence present in the same exon as the coding sequence.

Most interestingly, the 5' EST, bfad022l10 5', also aligned to additional regions upstream of the *SCP1* coding exon, with the mRNA sequence indicating three exons spread throughout the 3259bp between *SCP1* and *CHIC*. The three additional 5' UTR exons were identified with discontiguous megaBLAST, used to align the short regions of these exons that did not show 100% sequence identity to the B.floridae ParaHox reassembly and is simply due to polymorphism over such a small region. In total, only 16 nucleotides across the entire 599bp of bfad022l10 5' did not show a match to the *B.floridae* ParaHox reassembly genomic sequence. Exon lengths and percentage identities for both bfad022l10 3' and 5' hits against the ParaHox reassembly, as well as exon positions, are given in table 4.2. UTR exonic and intronic positions are visualised, along with the relative position of bfad022l10 EST matches in figure 4.4.

Table 4.2. ***B.floridae SCP1*** **EST identities and** ***SCP1*** **exon positions.**

Name	Start position (bp)	End Position (bp)	Length (bp)	Identities	Percentage Identity (%)
Exon 1 (5' UTR)	1603681	1603735	55	55/55	100
Exon 2 (5' UTR)	1604359	1604423	65	59/65	91
Exon 3 (5' UTR)	1605937	1606001	65	65/65	100
Exon 4 (5' UTR)	1606870	1606922	53	50/54	93
Exon 4 (Coding sequence)	1606923	1610063	3141	354/362	98
Exon 4 (3' UTR)	1610064	1610748	685	632/727	87

N.B. Individual exons are indicated by colour shading and heavier borders.

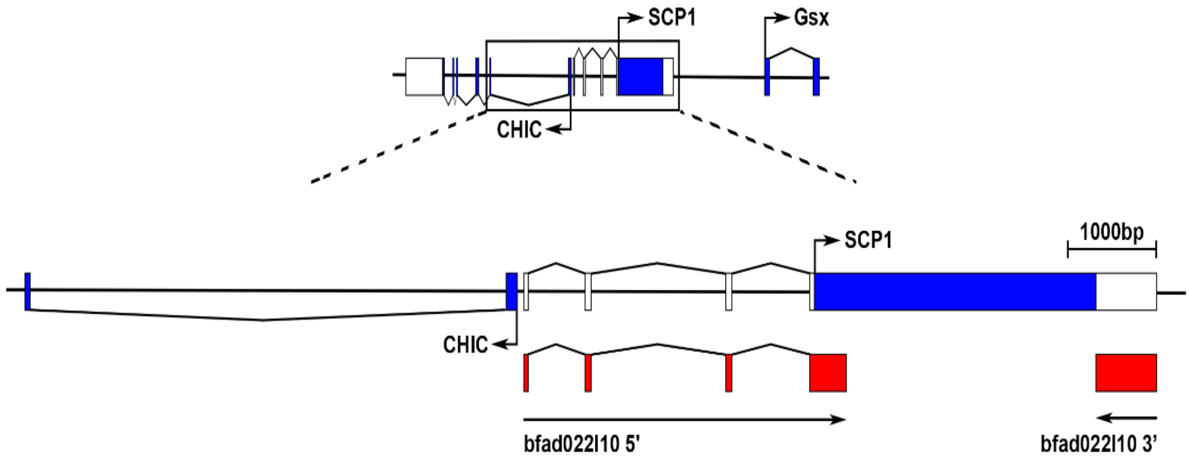


Figure 4.4. ***B.floridae*** **SCP1 has a multi-exonic 5'UTR, and 3'UTR.**

A schematic depicting the relative positions of exons within the *CHIC-SCP1* region of the *B.floridae* ParaHox Reassembly. The EST transcript bfad022|10 identifies both a multi-exonic 5' UTR and 3' UTR that is adjacent to the single exon coding sequence. Blue boxes represent coding sequence, white represents UTR and red represents EST sequence. Arrows at right angles (↲) indicate translational start sites and orientation of transcription. Transcription start sites are unknown

4.3.4. Promoter analysis of the region surrounding the *B.floridae SCP1* 5' UTR.

The ParaHox-neighbouring *AmphiSCP1* gene is the only copy of *SCP1* within the *B.floridae* genome, and consists of a single coding exon. This is in contrast to the case in vertebrates (de Vries et al., 2005), and in non-chordate bilaterians (see section 4.3.5 and table 4.3.) In addition, SYCP1 and ParaHox genes are located upon different chromosomes within vertebrates, with human and mouse SYCP1 genes located upon chromosomes 1 and 6 respectively (chromosomal location from ensemble gene browser, http://www.ensembl.org/), whilst the four ParaHox loci are located on chromosomes 13, 4, 5 and X in human, and chromosomes 5, 5, 18 and X in mouse (Ferrier et al., 2005). This supports the insertion of the *AmphiSCP1* retrogene next to the ParaHox locus of amphioxus, and subsequent loss of a multi-exonic parental copy.

As the 5' UTR of *B.floridae SCP1* must have evolved post-invasion of the ancestral amphioxus SCP1 single-exon retrogene, a promoter region driving the transcription of this 5' UTR sequence must have either been co-opted from an existing nearby promoter sequence, or evolved *de novo*. Promoter prediction was carried using three independent prediction programs; Neural Network Promoter Prediction (NNPP) (Reese et al., 1996; Reese and Eeckman, 1995), TSSW (Solovyev et al., 2010) and WWW Promoter Scan (Prestridge, 1995), which uses ProScan 1.7, promoter prediction programs (See section 4.4.3 for the details as to the differing methods used in each program). A total of 7000bp, starting from within *CHIC* intron 1, covering *CHIC* exon 1, to the end of the *SCP1* coding sequence was analysed for promoter sequences.

A total of five 50bp predicted promoter sequences were identified by NNPP (figure 4.5, positions of predicted promoters indicated by red boxes) (table 4.3), with the prediction with highest support located surrounding the start of *SCP1* 5' UTR exon 1. This region, annotated as NNPP3 in figure 4.5, was the only sequence predicted in all three Promoter prediction programs, NNPP, TSSW and ProScan 1.7, and had the highest support value in both NNPP and ProScan 1.7 (Table 4.3). This was also the only region predicted by TSSW and is identified as 50bp in length using NNPP and 250 bp in ProScan. It also lies on the negative strand and spans the start of *SCP1* 5' UTR exon 1, in the same orientation as the *CHIC* gene, and is located 56bp upstream of *CHIC* (figure 4.5). Predicted promoter sequences can be found on the genepalette ParaHox Reassembly, as well as Appendix 7.5.

Interestingly, both of the ProScan 1.7 predicted promoter regions hit parts of the same region, with the highest supported hit covering the reverse strand in the direction of *CHIC* (figure 4.5 and Table 4.3) whilst the second ProScan hit, covers the positive strand in the direction of *SCP1* 5' UTR, suggesting that a single bi-directional promoter may direct transcription of both *CHIC* and *SCP1*.

Table 4.4. SCP1 5' UTR predicted Promoters

Promoter identifier	DNA Strand	Start Position (bp)	End Position (bp)	Length (bp)	Support Value	Support Threshold
NNPP 1	+	1602839	1602888	50	0.85/1	0.80
NNPP 2	-	1602876	1602925	50	0.94/1	0.80
NNPP 3	-	1603666	1603715	50	0.99/1	0.80
NNPP 4	+	1603987	1604036	50	0.87/1	0.80
NNPP 5	+	1604341	160	50	0.81/1	0.80
TSSW Promoter TSS	-	1603689	1603689	n/a	31.63	0.45
ProScan 1	+	1603548	1603797	250	62.56	53.0
ProScan 2	-	1603635	1603884	250	118.41	53.0

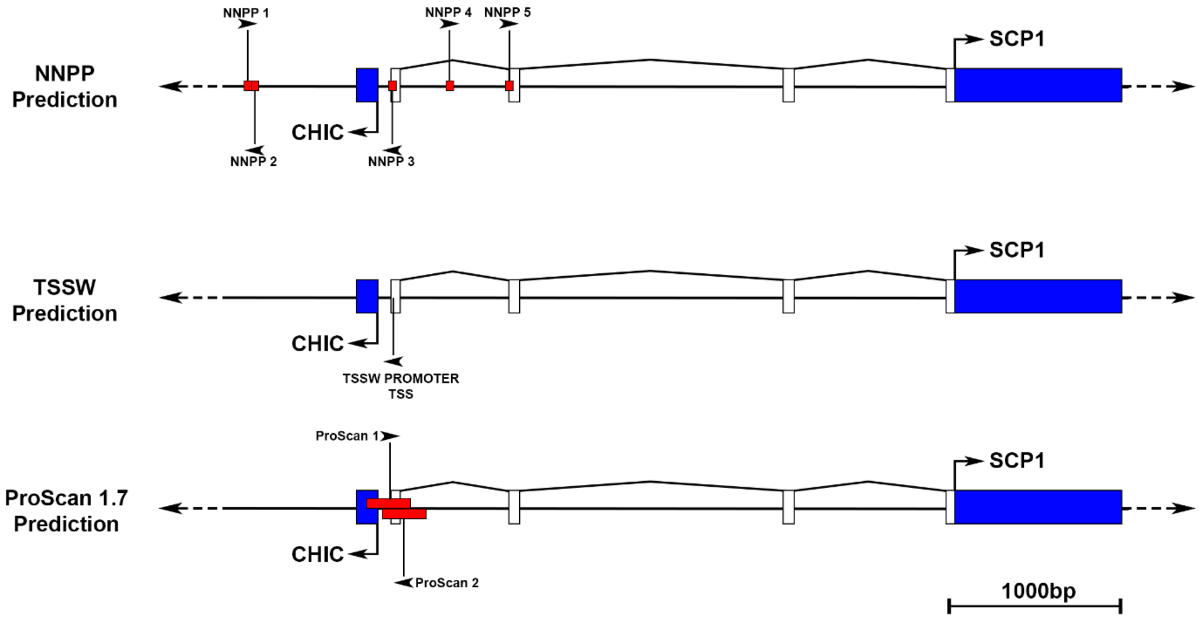


Figure 4.5. A schematic indicating promoters predicted using three different promoter prediction programs.

Promoter sequences predicted by either NNPP, TSSW or ProScan 1.7 are visualised relative to surrounding exons and introns. The size and position of each predicted promoter identified is indicated by a red box and black vertical line. In addition, black arrowheads indicate the direction of the DNA strand the promoter was identified upon. Five promoters were predicted by NNPP (NNPP1, NNPP2, NNPP3, NNPP4, NNPP5), one by TSSW (TSSW PROMOTER TSS), and two by ProScan 1.7 (ProScan1, ProScan 2). Only one promoter region, including NNPP3, TSSW PROMOTER TSS, and PROSCAN 1 and 2 agree across all three prediction models. Blue boxes represent coding exons, whilst white boxes represent UTR exons.

4.3.5. SCP1 has a single coding exon in both amphioxus and *C.intestinalis*, but is a large multi-exonic gene throughout the rest of the Bilateria.

SCP1 mRNA and genomic sequences were collected as detailed in section 4.2. and the length of nucleotide sequence in bp of the mRNA noted, as well as the number of introns present within the gene. Though all SCP1 genes examined this way are of varying size, being between 2200 and 3500bp in length, it remains constant that in the echinoderms, vertebrates, molluscs and polychaetes there are many introns within the SCP1 gene. In fact, SCP1 has between 21 and 33 introns depending on the species, which is in stark contrast to the single coding exons of all three *Branchiostoma* amphioxus species and *Ciona intestinalis*. Table 4.4 shows the species examined, as well as the number of exons and length of the SCP1 transcript for each SCP1 gene. As *B.floridae* SCP1 coding sequence is 3141bp in length, this shows that overall transcript length has been maintained despite the loss of all introns during retrogene formation. Interestingly, Ciona SCP1 has also undergone retrogene replacement, as only the single exon copy of SCP1 can be found within its genome.

Table 4.4. Intron numbers and length of SCP1 genes

Species	No. of *SCP1* introns	Length of transcribed *SCP1* gene (bp)	Total intron + exon length (bp)
Mus musculus	33	3437	117601
Homo sapiens	32	3452	140567
Danio rerio	32	3440	8345
Ciona intestinalis	0	2696	2696
Branchiostoma floridae	0	3141	3141
Strongylocentrotus purpuratus	23	2816	11135
Aplysia californica	21	2196	18402
Capitella teleta	25	3048	5745
Lottia gigantea	21	2214	15448

4.3.6 *SCP1* is conserved across the Metazoa.

In order to assess the extent of SCP1 presence outside of the chordates, SCP1 protein sequences were collected as described in section 4.2. for a wide range of organisms across the metazoa. This list of SCP1 proteins was built both with the aim of being exhaustive and to sample phyla that were underrepresented in previous analyses (Fraune et al., 2012). The chimaera (*Callorhinchus milii*) was added to the vertebrata, as a basal fish lineage, as well as additional echinoderm species, including a second echinoid (*Lytechinus variegatus*) and a member of the less divergent asteroids (*Asterias amurensis*). Additionally, a single hemichordate SCP1 sequence from (*Saccoglossus kowalevskii*) was identified, giving examples of SCP1 from all three main deuterostome phyla. In the Protostomia, an additional lophotrochozoan sequence was obtained from the Mollusca (*Pomacea canaliculata*), though no additional ecdysozoan members were obtained (beyond the highly divergent and short *Petrolisthes cinctipes* sequence). No additional members of the Cnidaria were obtained beyond *Hydra vulgaris* and the short *Nematostella vectensis SCP1* EST. Finally, the poriferan *Amphimedon queenslandica* and the Ctenophore *Pleurobrachia pileus* represent the sole examples of SCP1 so far identified in these phyla. Full species names, groups and accession numbers are given in Appendix 7.6.

Alignment of all identified SCP1 sequences was carried out using CLUSTAL Omega (Sievers et al., 2011) and visualised in Jalview (Waterhouse et al., 2009). The full SCP1 alignment can be found in Appendix 7.7 (figure 7.3). An 83aa motif in the N-terminus has previously been observed to be conserved between hydra and vertebrates (Fraune et al., 2012), and this was seen to be highly conserved across all species examined in this study (figure 4.6). This domain lies within the N-terminal coiled-coil domain and is conserved between rat and hydra (Fraune et al., 2012). However, a few species stand out from this analysis as being divergent. One of these is *Petrolisthes*, which stands out as the only example of SCP1 genes within the Ecdysozoa. Of the two ESTs identified in Fraune et al. (2012), one was not included in this study due to its highly divergent CM1 motif. The remaining *Petrolisthes* sequence included in this study remains divergent, even compared with the cnidarian, poriferan and ctenophoran sequences. It is possible that these are not in fact SCP1 genes, or represent contamination, as other Ecdysozoans have evolved non-homologous genes to fulfil the role of SCP1 . *Capitella SCP1* is another sequence that shows divergence, this time having a deletion of the C-terminal end of the CM1 motif, though further C-terminal sequence picks up again shortly after. This divergence may be an artefact of protein prediction from the genomic shotgun sequence. Finally, *Nematostella* has a very short EST read providing its SCP1 representative and lacks the C-terminal end of the CM1 motif as well as any further C-terminal peptide sequence. Despite these anomalies, across the CM1 motif a 70.5% similarity can be seen between rat and *Pleurobrachia*, with

44.3% identity, showing that this CM1 motif has maintained high conservation across metazoan evolution. An alignment of this CM1 motif is visualised in figure 4.6.

In order to test the phylogenetic relationships of SCP1 proteins, both neighbour joining (NJ) and maximum likelihood (ML) trees were built using the 83aa CM1 conserved motif. The coiled-coil domain containing CCDC39 proteins from *Drosophila*, *Strongylocentrotus* and *human* were used as an outgroup to build the phylogenies, using alignment with the 83aa CM1 motif coiled-coil domain. Phylogenetic models were tested using ProtTest (Abascal et al., 2005). Neighbour joining tree was built using the Poisson model with 1000 bootstraps, and a Maximum likelihood was tree built using the LG+G model with 1000 bootstraps. *Capitella* was removed from the analysis due to the large expanse of missing C-terminal CM1 motif.

Though bootstrap support values are low on many branches, both NJ (figure 4.7 A) and ML (figure 4.7 B) analyses have similar topologies, although the topology does have a good match to the known relationships of the taxa. The vertebrates group together with significant support, as do the different vertebrate groups such as mammals, fish and lizards/birds. Amphioxus *SCP1* consistently groups with the hemichordate *Saccoglossus* rather than the vertebrate chordates, with the asteroid *Asterias* branching further down, making an interesting grouping with regards to evolutionary relationships of the echinoderms, hemichordates and chordates, though this relationship has no support. The tunicates appear to group with the echinoids, though this is a very long branch length and has no support. This grouping could perhaps represent the divergent evolution of both echinoids and tunicates. Strangely, the cnidarian *Nematostella* groups with the lophotrochozoan clade, again with no report. This could be due to the short transcript of *Nematostella* and perhaps loss of phylogenetic signal. Both *Pleurobrachia* and *Amphimedon* branch basal to the other Metazoa as expected, though with long branch lengths and very low bootstrap support. Petrolisthes consistently groups basal to all lineages other than *Pleurobrachia* and *Amphimedon*, including the Cnidaria. This may either represent several different scenarios; either an extremely divergent SCP1 in Petrolisthes, that this short sequence is perhaps not ecdysozoan and in fact represents some contamination, or that this *Petrolisthes* sequence is actually not an SCP1 homologue. It should be pointed out that conclusions about phylogenetic grouping can only really be made for groups with significance support (>70%). In addition, taxa with long branch lengths/divergent sequences being problematic to place reliably.

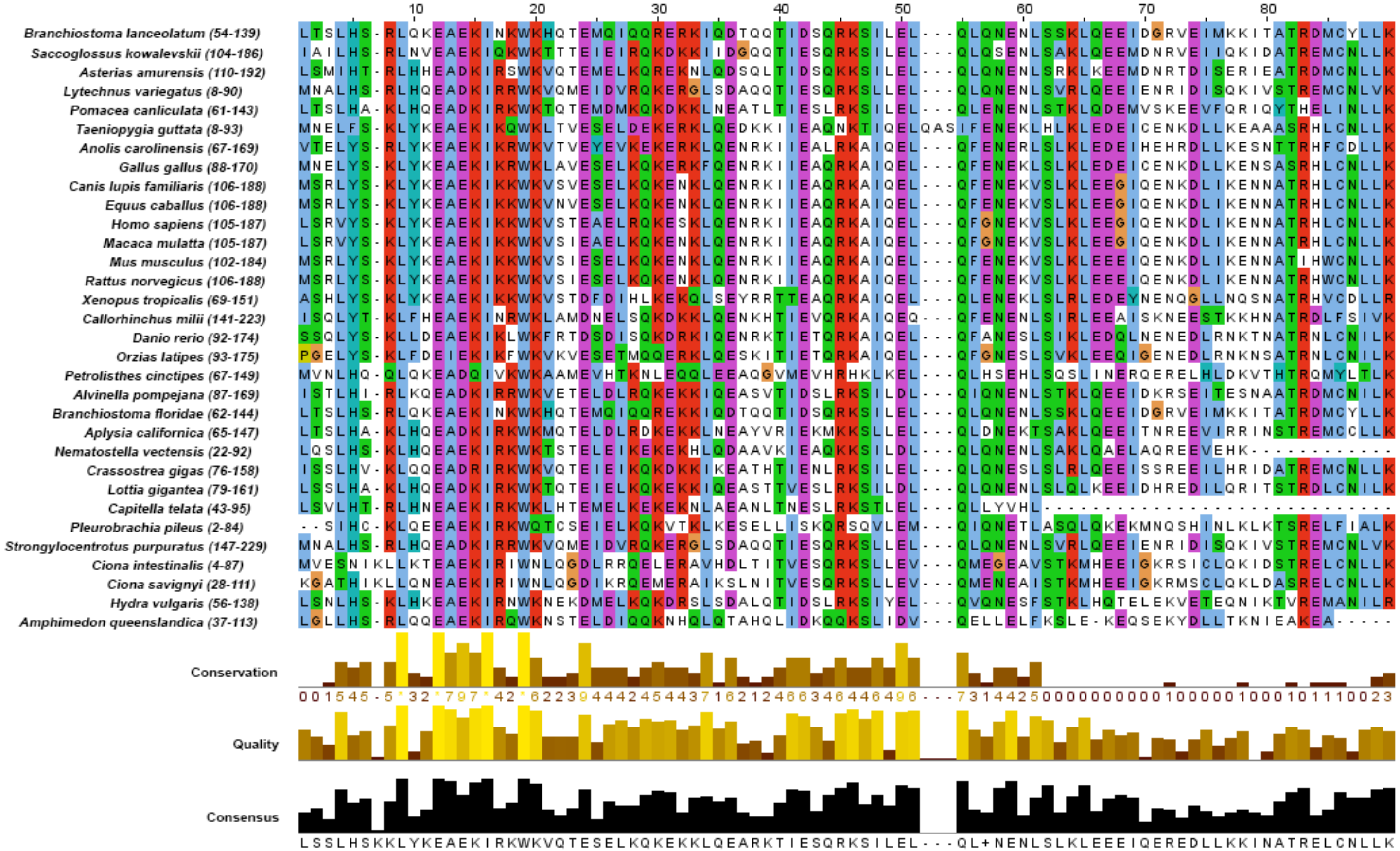

10
20
30
40
50
60
70
80
Branchiostoma lanceolatum (54-139)
Saccoglossus kowalevskii (104-186)
Asterias amurensis (110-192)
Lytechnus variegatus (8-90)
Pomacea canliculata (61-143)
Taeniopygia guttata (8-93)
Anolis carolinensis (67-169)
Gallus gallus (88-170)
Canis lupis familiaris (106-188)
Equus caballus (106-188)
Homo sapiens (105-187)
Macaca mulatta (105-187)
Mus musculus (102-184)
Rattus norvegicus (106-188)
Xenopus tropicalis (69-151)
Callorhinchus milii (141-223)
Danio rerio (92-174)
Orzias latipes (93-175)
Petrolisthes cinctipes (67-149)
Alvinella pompejana (87-169)
Branchiostoma floridae (62-144)
Aplysia californica (65-147)
Nematostella vectensis (22-92)
Crassostrea gigas (76-158)
Lottia gigantea (79-161)
Capitella telata (43-95)
Pleurobrachia pileus (2-84)
Strongylocentrotus purpuratus (147-229)
Ciona intestinalis (4-87)
Ciona savignyi (28-111)
Hydra vulgaris (56-138)
Amphimedon queenslandica (37-113)
Conservation
Quality
Consensus
LSSLHSKKLYKEAEKIRKWKVQTESELKQKEKKLQEARKTIESQRKSILEL---QL+NENLSLKLEEEIQEREDLLKKINATRELCNLLK

Figure 4.6. The SCP1 protein CM1 motif is conserved across the Metazoa.

A CLUSTAL Omega protein multiple alignment of the CM1 domains of SCP1 shows a high level of conservation across an 83 aa motif across the metazoan species examined. Conservation is visualised with false colour using the ClustalX colour table for amino acids. Conservation is given below as a score out of 10 across all aligned sequences in yellow-brown. The same is given for the quality of alignment, represented by the sequence similarity. Finally a consensus sequence made up of the most abundant amino acid for each position is given in black. Names of species used are given to the left of the alignment. Numbers in parenthesis indicate the position of the CM1 motif amino acids within the native peptide sequence.

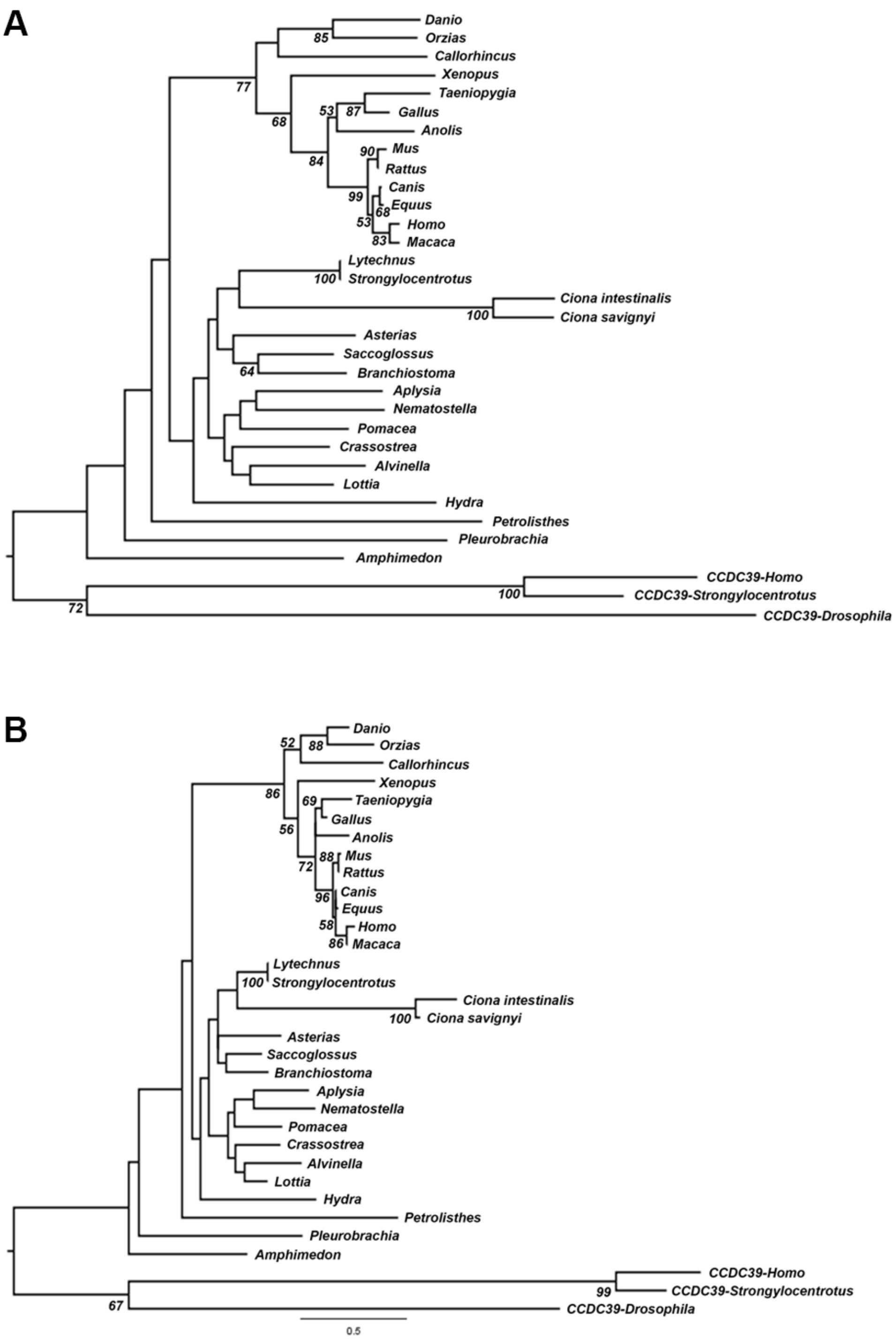


Figure 4.7. Phylogeny of metazoan SCP1 proteins

(A) Neighbour joining tree built using the 83aa CM1 domain of SCP1 proteins, using the POISSON matrix and 1000 bootstraps. (B) Maximum likelihood tree built using the 83aa CM1 domain of SCP1 proteins, using the LG+G model with 1000 bootstraps. CCDC39 Proteins were used as an outgroup to SCP1. Bootstrap values over 50% are given. Longer branch lengths equate to a further evolutionary distance between nodes. Trees were built using MEGA6.

4.4. Discussion

4.4.1. Amphioxus SCP1 is a transcribed retrogene that replaced its parental multi-exonic copy before the divergence of the *Branchiostoma* genus.

Comparisons between the three amphioxus genomes show that amphioxus SCP1 is present as a single coding exon within *B.floridae*, *B.lanceolatum* and *B.belcheri* (figure 4.1). B.belcheri is the sister to B.lanceolatum and B.floridae, and the two groups are estimated to have diverged 112 million years ago (Nohara et al., 2004). Thus we can conclude that an SCP1 retrogene must have been present upstream of the ParaHox cluster, between *CHIC* and *Gsx*, before the divergence of these three species. As such, it would be very interesting to examine the ParaHox cluster of both *Aysmmetron* (Yue et al., 2014) and *Epigonichthys* (Nohara et al., 2005) as the only two other amphioxus groups currently known and observe if SCP1 is also present as a retrogene upstream of Gsx in these more distantly related amphioxus species. The presence of multi-exonic SCP1 genes throughout the rest of the Bilateria, within the vertebrates, echinoderms and Lophotrochozoa (table 4.4) makes it highly likely that both amphioxus and *Ciona* SCP1 genes evolved via retrotransposition and replaced a multi-exonic ancestral parent gene. In addition, the existence of SCP1 retrogene copies within mouse also suggests that SCP1 is prone to retrotransposition, as it has occurred twice within the mouse lineage, though this time retaining the parental multi-exonic copy (Sage et al., 1997). The expression of SCP1 within germ cells may very well make SCP1 a target for the 'out of the testis' route of retrogene production (Kleene et al., 1998; Vinckenbosch et al., 2006) and the gene and eventual replacement of the parent gene by the retrocopy (Ciomborowska et al., 2013).

4.4.2. Expression of *AmphiSCP1* is much broader than expected for a meiosis gene.

It is clear from the in situ hybridisation of amphioxus SCP1 that expression is by no means limited to the germ cells, and typical germ cell markers such as *nanos* and *vasa* show markedly different embryonic expression patterns to SCP1 (Wu et al., 2011b). As SCP1 expression is limited to meiotic cells in both vertebrates ((Casey et al., 2015; de Vries et al., 2005; Iwai et al., 2006), including primordial germ cells (Zheng et al., 2009)), and *Hydra* (Fraune et al., 2012), it was expected that no embryonic expression would be observed, as the testis and ovaries have not yet formed in amphioxus, or that SCP1 would display *nanos*/*vasa*-like germ cell expression (Wu et al., 2011b). Furthermore, if *AmphiSCP1* had transposed along with its own regulatory elements, such as a promoter region, it might even be expected that germ cell expression is the most likely outcome, as previous work has shown the zebrafish SYCP1 promoter region to be sufficient to drive GFP transgenes within germ cells (Gautier et al., 2013).

Amphioxus *SCP1* drives expression in the endoderm and mesoderm in a broad pattern throughout these tissues, and also seems to exhibit spatio-temporal changes in expression. *AmphiSCP1* is notably absent from the ectoderm and the posterior tailbud, but also from the extreme anterior in all stages (figures 5.2 and 5.3). This expression pattern, which is much broader than expected for SCP1, suggests that *AmphiSCP1* has co-opted regulatory elements from its new genomic locus. It does not appear to have come under the influence of ParaHox regulatory elements however, as the broad expression pattern is not reminiscent of ParaHox expression, and there appears to be no neural expression, whilst CNS expression is a hallmark of ParaHox genes (Brooke et al., 1998; Osborne et al., 2009). It may, however have co-opted regulatory elements from the adjacent *AmphiCHIC* gene.

Amphioxus CHIC expression has not yet been examined as far as we are aware, and very little expression data exists even for the vertebrate genes. Having said that, *CHIC1* and *CHIC2* were both originally identified as *Brain x-linked protein* (*Brx*) and *BrX-like translocated in leukemia* (*BTL*) respectively, and data does exist describing roles in the regulation of nuclear hormone receptors (Kino et al., 2006) and exocytosis (Cools et al., 2001). Expression of vertebrate CHIC genes was first identified in the brain, though both CHIC1 and CHIC2 also exhibit expression in; the testis, ovary, uterus, endomesoderm, intestine, ectoderm, many secretary organs of the digestive tract, thyroid, prostrate and pineal gland (http://www.proteinatlas.org/ (Uhlén et al., 2015)). CHIC genes seem to show expression in a range of tissues but it stands out that many, if not all, have secretory functions, which may well be linked to the described role in plasma membranes and vesicles, and exocytosis (Cools et al., 2001). This expression also holds true for the protostome CHIC homologues *TAG-266* (*C.elegans*) (UNIPROT) (Consortium, 1998) and *CG5938* (UNIPROT) (*D.melanogaster*) (Hoskins et al., 2015). Since bilaterian CHIC genes are expressed in the testis and ovaries, co-option of *CHIC* regulatory elements would still allow *AmphiSCP1* to carry out its meiotic function, and also give the potential to evolve new expression domains within somatic tissues. The expression of CHIC in the digestive tract, and endomesoderm would match with the expression seen in figures 5.2 and 5.3. However, CHIC expression would first need to be characterised in amphioxus in order to tell if the expression matches with *AmphiSCP1*.

One other bilaterian SCP1 expression peculiarity is noteworthy. In the sea urchin *S.purpuratus*, SCP1 is found to be expressed in the larvae throughout the adult rudiment (Yajima et al., 2013). This structure goes on to form most of the adult animal and the larvae is largely cast off or reabsorbed (Minsuk and Raff, 2002; Wilt, 2002; Yajima and Kiyomoto, 2006; Yajima et al., 2013). The expression of SCP1, along with other meiotic genes, throughout the adult rudiment is perhaps suggestive of an ability to function in the normal division of cells, or even another function during larval development. It is also possible that transcription of SCP1 is not indicative of any function in somatic cells. Mammalian studies have indicated that meiotic genes can be activated in initially broad domains and only later become restricted to germ cells (Saitou et al., 2002; Saitou et al., 2003), with transcription often beginning prior to the initiation of meiotic events (Kimble and Page, 2007). As such, it is entirely possible that the somatic expression of *AmphiSCP1* transcripts merely represent non-functional transcription. It is also possible that SCP1 transcription is allowed to proceed in somatic tissues as it has no negative effect, or that the improvement to transcription in target tissues granted by co-opted regulatory elements outweighs any transcriptional costs in somatic tissues.

4.4.3. *B.floridae SCP1* has evolved a de novo multi-exonic 5' UTR that may originate from a co-opted bi-directional *CHIC* promoter.

Transcriptomic data supports the presence of a multi-exonic 5' UTR stretching upstream from the *SCP1* coding sequence between *SCP1* and *CHIC* (figure 4.4). Promoter analysis revealed no promoter present immediately upstream of the *SCP1* coding region (figure 4.5). However, one promoter, lying upstream of *CHIC* exon 1, was identified with high support values in all three of the programs used for prediction. Three promoter prediction programs were used here in order to provide different methods of support for putative promoter sequences. NNPP characterises promoter regions using a neural network to predict the interaction of TATA and initiator (*Inr*) binding sites that make up eukaryotic polymerase II (Pol II) promoters, which has then been 'trained' upon both human and *Drosophila* promoter datasets (Reese, 2001). TSSW is the most modern of the promoter prediction programs used, and is designed to distinguish between and identify both TATA positive, and TATA negative promoters using a variety of characteristics including, but not limited to, the presence of: Hexaplets, TATA box content and score, Triplets around the TSS, Sp-1 motif content and CpG content, with different characteristic scores used for TATA+ and TATA- promoters (Solovyev et al., 2006; Solovyev et al., 2010). Finally ProScan1.7 predicts Promoter regions based on scoring homologies with putative eukaryotic Pol II promoter sequences, using a ratio of the density of transcription factor binding sites (using all mammalian transcription factors from transcription factors from the Ghosh TFD database (Ghosh, 1992)) along with TATA binding site weight matrices (Prestridge, 1995). All three use programs thus use slightly different approaches, and each gives different rates of true site, and positive site prediction (Solovyev et al., 2010). Thus, using the three programs in conjunction is likely to mitigate the downfalls of each program, and in this case leads to the prediction of one promoter region in all three cases, making it much more likely to be a bona fide promoter sequence than if only one program had been used. This predicted promoter region spans the start of the first (5') exon of the SCP1 5'UTR transcript and lies on the same strand as *CHIC*, within 56bp of *CHIC* exon 1 (figure 4.5), and may even cover both *CHIC* exon 1 and *SCP1* 5' UTR exon 1 (Proscan 1.7 predictions in figure 4.5). Even more intriguing is that ProScan 1.7 predicts a promoter on both positive and negative strands at this region, raising the possibility that this may be a bidirectional promoter. The presence of this promoter overlapping the first exons of both CHIC and SCP1 5' UTR is certainly consistent with this.

This raises the possibility of a very interesting evolutionary scenario, in which *SCP1* has co-opted a *CHIC* promoter, and also allowed germ cell expression as discussed in section 4.4.2. SCP1 would then have either evolved its own de-novo 5' UTR in order to take advantage of this bidirectional promoter. It is likely that the orientation of the two genes, and position of the predicted promoter sequence, precludes the co-option of 5' UTR elements from *CHIC*. Whilst it may seem a large evolutionary leap for a retrogene to evolve a 5' UTR, or co-opt an existing nearby regulatory element, this has been seen to occur with other bilaterian retrogenes. For example, a genome-wide screen of retrogenes within *Drosophila* revealed that several regulatory motifs were over-represented in the cis-regulatory elements of testis-expressed retrogenes, and that specific regulatory motifs had been selectively recruited by retrogenes from their new genomic loci (Bai et al., 2009). Another key study selectively looked at the evolution of introns within retrogenes of mammals and found that most introns found associated with retrogenes occurred in the 5' flanking sequence to the retrogene insertion site (Fablet et al., 2009). The analysis also showed that retrogenes with introns display higher transcription levels and broader expression patterns than those without. Fablet et al. (2009) propose a scenario where 5'exon-intron structures evolve de novo or through fusion to the 5' UTR of a neighbouring gene as a direct link to the recruitment of a distant promoter by a retrogene. It should also be noted that of those recruited by distant promoters, and which gained 5' exon-intron UTR structures, most were recruited by bidirectional CpG promoters (Fablet et al., 2009). Several other more recent studies have also examined these phenomena of retrogenes recruiting regulatory elements from regions flanking their insert site, as well as retrogenes gaining introns, and it seems that these phenomena may not be as rare as they once seemed (Kang et al., 2012; Matsumura et al., 2014; Sorourian et al., 2014). There is an abundance of general transcription occurring within cells to which no functional role can be attributed, and lots of non-coding, non-functional, RNA is produced (reviewed in Struhl (2007)). It is entirely possible that retrogenes could be co-opting the sequences involved in this so-called 'junk' transcription to facilitate their own transcription as part of retrogene evolution.

The combination of 5' UTR transcript, perfect placement of a predicted promoter (perhaps bidirectional) adjacent to both *CHIC* exon 1 and *SCP1* 5' UTR exon 1, and broad somatic expression of *AmphiSCP1* in embryos are all consistent with recruitment of a *CHIC* promoter by the amphioxus SCP1 retrogene. SCP1 would then have evolved a de novo 5' intron-exon structure to make use of the distant promoter. A preliminary check for CpG islands within the CHIC-SCP1 5' UTR region yielded no results, but the identified promoter region could nonetheless still display bidirectionality. Further work should examine this promoter region in a reporter background to test both its bidirectionality as well as similarity to *AmphiSCP1* expression. In addition, *in situ* hybridisation of *AmphiCHIC* should be carried out in order to compare CHIC and SCP1 expression. It would also be useful to identify if *AmphiCHIC* itself has any 5' UTR sequence and identify whether the transcriptional start site (TSS) for *AmphiCHIC* does indeed overlap with that of *AmphiSCP1* , perhaps using RACE to identify the TSS of both *SCP1* and *CHIC*.

4.4.4. SCP1 proteins show deep conservation of an 83aa CM1 motif, and have representatives in all major phyla except perhaps those of the Ecdysozoa.

As Fraune et al (2012) showed, SCP1 proteins are much more deeply conserved across the metazoa than previously believed. Prior to the work of Fraune et al (2012) it was widely thought that SCP1 homologues were limited to the mammals, and was only reported within a non-mammalian vertebrate in 2006 (Iwai et al., 2006). However, both *Hydra* (Fraune et al., 2012) and sea urchin (Yajima et al., 2013) synaptonemal proteins have recently been characterised, confirming expression and roles in meiosis throughout the Eumetazoa. This amphioxus work has sought to build upon the work of Fraune et al (2012), identifying SCP1 genes and proteins throughout the Metazoa, utilising the wealth of new genome projects that have become available. This has allowed a broader sampling of SCP1 from within the non-chordate deuterostomes, namely with the addition of another echinoid, an asteroid and also a hemichordate sequence from the Ambulacraria. This has provided at least one example of SCP1 from each deuterostome phylum (figure 4.6, Appendix 7.6.). Strangely, no further ecdysozoan sequences beyond *Petrolisthes* were obtained from TBLASTN or BLASTP searches against protein, transcriptomic shotgun assembly, whole genome shotgun assembly and EST databases contained within the NCBI, UNIPROT and JGI databases. *B.floridae SCP1*, *M.musculus SYCP1,* or *Hydra SYCP1* peptide sequences were all used as queries when looking for ecdysozoan sequences. In light of this, along with the short and highly divergent EST hits coding *Petrolisthes*

SCP1, as well as the tendency for *Petrolisthes SCP1* to be unstable when building phylogenies (fig 4.6-4.7), it would be wise to leave these *Petrolisthes* sequences from future SCP1 analyses until other Ecdysozoan sequences are obtained, which can either confirm or refute their SCP1 affinity. It remains suspicious that no other ecdysozoan group gives even a partial hit to SCP1, particularly in a more basal arthropod such as with the Myriapod *Strigamia* (Chipman et al., 2014). This is even more relevant in light of the lineage specific synaptonemal complexes of well-studied ecdysozoans such as *Drosophila* and *C.elegans*, and both have individually evolved functionally similar, but novel synaptonemal complex proteins that fulfil the same functional role as SCP1 in other metazoans, (Bogdanov Iu et al., 2002; Bogdanov et al., 2003; Colaiácovo et al., 2003; MacQueen et al., 2002; Page and Hawley, 2001; Schild-Prufert et al., 2011; Smolikov et al., 2007), but show no homology to those of vertebrates, or even between *Drosophila* and *C.elegans*. The complete lack of SYCP proteins in any other Ecdysozoan, and evolution of lineage specific synaptonemal proteins in both *Drosophila* and *C.elegans*, suggest that the Petrolisthes sequence could be a case of mis-indentification. Since the *Petrolisthes* ESTs were obtained from heart, gills and whole crab, contamination from another source is a distinct possibility. It is also possible that the 'SCP1' hits are not, in fact, SCP1 and that a longer sequence would reveal a lack of homology. Indeed, iterations of the alignment carried out with CLUSTALW and MUSCLE did not align the *Petrolisthes* ESTs to the conserved CM1 domain at all, and instead they aligned further towards the coiled-coil containing C-terminus of other SCP1 proteins.

More complete sequences would improve the phylogeny in several cases, and *Capitella* was removed from the alignment for the purposes of tree building due to the instability attributed to a lack of the C-terminus of the CM1 region. A similar case may be true of *Nematostella* and to some extent *Amphimedon*. Overall poor support for clades was a common theme both in this study and likely in the Bayesian inference trees of Fraune et al. (2012), where very few branches actually have significant support values. This could be improved with better taxon sampling to help resolve phylogenetic relationships, though a general problem could be that fast evolving lineages such as the tunicates cause long-branch attraction issues.

Aside from these issues, we do see grouping of the vertebrates, Lophotrochozoa and non-vertebrate deuterstomes into monophyletic clades, and the conservation of the CM1 domain is clear in the alignment between such divergent animals as ctenophores and humans. However, in reality very little can be concluded from the phylogeny about the relationships of SCP1 proteins, as support values are very low for all groups outside of the vertebrates.

Chapter 5. TCF/Lef regulates the Gsx ParaHox gene in central nervous system development of invertebrate chordates

5.1. Introduction

The understanding of the regulation of amphioxus ParaHox genes is so far limited to two studies in which response to potential regulators was directly screened in embryos. One of these studies examined the response of *AmphiCdx* to Wnt signalling (Onai et al., 2009). Treatment with Li^+, which upregulates Wnt/β-catenin signalling via inhibition of *GSK3β*, causes embryos to gain an ectopic anterior domain as well as a reduction of the CNS domain in the neurula and possible expansion in the hindgut. The other study was targeted more specifically to the ParaHox genes, and their response to retinoic acid (RA) as well as an RA signalling inhibitor (BMS009) (Osborne et al., 2009). Treatment with RA or BMS009 caused a dramatic shift in the anterior/posterior boundaries of all three ParaHox genes, having a particularly strong effect on the A/P position of the boundary between the *Xlox* and *Cdx* expression domains in the mid-hindgut, as well as the early *Gsx* neural domain. Such pharmacological treatments can have undesired effects however, especially on non-model organisms such as amphioxus, and can lead to deformed embryos or secondary effects. In order to overcome this problem, Osborne et al. also examined the response of reporter constructs containing amphioxus ParaHox retinoic acid response elements (RAREs) to induction by retinoic acid in cell culture. This showed direct response of the ParaHox cluster to RA treatment, but avoided secondary embryonic effects. Whilst these types of study are useful for identifying the response of genes to a regulatory pathway, studies that examine the regulatory regions imparting these responses are needed in order for us to understand the mechanisms underlying the regulation of developmental genes and their evolution.

Although there are some initial results illustrating that reporter gene analyses can be done in amphioxus, the technique is currently still much more challenging in this species. The first *in vivo* studies of amphioxus regulatory elements focussed upon the *FoxD* (Yu et al., 2004) and *engrailed* (Beaster-Jones et al., 2007) genes, injecting regulatory DNA into amphioxus embryos in order to identify those which harbour regulatory function. Though this approach was able to identify regulatory elements that could function *in vivo*, injection of regulatory DNA into amphioxus embryos caused embryos to often display deformities, and expression efficiency of reporter genes within the correct tissues was low.

Cross-species transgenesis has, however, provided an alternative route to rapidly analysing putative amphioxus regulatory elements (Beaster-Jones et al., 2007; Natale et al., 2011; Wada et al., 2005). Amphioxus Hox genes have been one such target for such cross-species studies, and regulatory regions for amphioxus Hox 1-3 were first identified through analysing expression of reporters in mouse and chick (Manzanares et al., 2000), as well as in transgenic *C. intestinalis* (Natale et al., 2011; Wada et al., 2005). Further study went on to identify RARE and *Ets* binding sites, which have been identified within regulatory regions for amphioxus *Hox1* and *Hox3* (Wada et al., 2006) and *Hox2* (Wada et al., 2005) respectively. Though expression may not perfectly mirror that of the endogenous gene, these cross-species regulatory studies often at least recapitulate expression within the same tissues. Expression of the amphioxus *engrailed* regulatory regions are one example of this, and though *Cin-engrailed* is not expressed in conserved domains, *AmphiEn* constructs drove expression within the muscle cells of *C.intestinalis*, showing that the transcriptional pathways needed to activate the *AmphiEn* regulatory region must still be conserved (Beaster-Jones et al., 2007). Regardless of tissue and/or cell type homology (or the lack of it) between *Ciona* and amphioxus, the molecular make-up of Ciona has been studied to unprecedented detail, beyond that of vertebrate models, and the cell lineage of Ciona can be traced accurately (as in the ANISEED database (Tassy et al., 2010)). This means that any restricted, reproducible reporter expression can allow insights into what transcription factors are likely to be controlling the reporter expression, just by examining those expressed within the relevant cell lineages.

C.intestinalis provides a system that is highly amenable to analysis of cis-regulatory elements via electroporation of reporter constructs, allowing the generation of large numbers of transgenic embryos (Corbo et al., 1997). Though the system may not prove amenable for some regulatory elements, such as many of those originally identified in (Manzanares et al., 2000) (Wada et al., 2006), the ability to screen large numbers of embryos aids in dissecting the function of those regulatory elements that do show cross-species function in *Ciona*. In this case, though most of the Hox1-3 constructs did not show regulatory function in *Ciona* (Wada et al., 2006), one 113bp regulatory element downstream of *AmphiHox2* was well characterised using *Ciona* transgenics. Here, Ets binding sites were shown to have function and help drive reporter expression in the sensory vesicle, oral syphon and palps, consistent with the expression of *AmphiEts1/2* in the pharyngeal endoderm and preoral pit prior to *AmphiHox2* expression. Compared to the difficulty of injecting and examining regulatory elements within amphioxus (Beaster-Jones et al., 2007; Yu et al., 2004), the *Ciona* system has shown that it is ideal for both the initial wider screening for functional regulatory elements (Di Gregorio and Levine, 2002; Harafuji et al., 2002), as well as more detailed studies to dissect regulatory element function and input (Bertrand et al., 2003; Chen et al., 2014; Kanda et al., 2013).

It is important to recognise that whilst conserved expression may be identified in cross-species regulatory studies, the presence of conserved binding sites does not necessarily indicate that these sites are functional. This is highlighted well by (Chen et al., 2014), where their *Ci-βγ-crystallin* regulatory element is able to drive expression both in the otolith of *C.intestinalis* and the lens of zebrafish. However, mutation of binding sites that causes abolition of expression in *Ciona* has little to no effect on the expression of the reporter in zebrafish. One explanation posed for this is that it is important to look across multiple binding sites when carrying out cross-species transgenesis, as a cumulative effect may be contributing to regulatory element function. Another is that the specificity of transcription factors for different binding sites can change across species, and a site that was previously non-functional in one species may then become functional within another. Hence it is important to consider the cooperative function of transcription factor binding sites as well as less canonical sites during cross species transgenesis.

Currently, Gsx is by far the least well examined of the ParaHox genes and no direct regulation of Gsx genes has yet been observed, though several examples of regulation of Gsx genes exist. One promising factor is Pax6, which has been shown to act in a mutually repressive manner with Gsh2 within the telencephalon (forebrain) of mice (Corbin et al., 2003; Toresson et al., 2000) (discussed further in section 5.4.3). Wnt signalling may also potentially target Gsx expression within the telencephalon (forebrain) of *Platynereis dumerilii*, and down regulated within Azakenpaullone (which upregulates Wnt signalling (Schneider and Bowerman, 2007)) treated embryos (Tomer et al., 2010). It is unclear though whether this is a direct or secondary effect upon *Gsx* expression. One study highlighted a potential regulatory network surrounding *Gsh2* within mouse dorsal interneurons. Here, *Gsh2* expression was induced by the transient overexpression of *Mash1*, yet was repressed by overexpression of *Ngn1* (Kriks et al., 2005). There may also be repression of *Gsh1* by *Gsh2*, as loss of *Gsh2* shifts *Gsh1* expression dorsally. In addition to these mouse studies, the loss of Gsh2 expression within mouse *Shh* mutants is detailed in (Corbin et al., 2000), suggesting the activation of *Gsh2* by *Sonic hedgehog* signalling. Within the neural tube, there appears to be a conserved regulatory pathway involving Gsx within primary neurogenesis, where *Vnd* (Nkx) represses *Ind* (Gsx) expression, and *Ind* represses *Msh* (Msx) within the drosophila neurectoderm (Cowden and Levine, 2003), with a modified version of the pathway present in *Xenopus*. Here, *Dbx1* serves as an intermediate between *Nkx6.1* and *Gsh2*, with *Dbx1* and *Gsh2* mutually repressing each other (Winterbottom et al., 2010). Finally, only one study has so far examined Gsx regulation within amphioxus, and *Gsx* was found to be regulated by RA, with altered RA signalling causing A/P shifts in the early neural tube domain of *Gsx* expression (Osborne et al., 2009). An increase in RA resulted in both additional *Gsx* positive cells within this early neural tube domain as well as an anterior shift in the expression domain. Conversely this early neural tube domain is abolished in BMS009 (RA signalling inhibitor) treated animals, corresponding to a decrease in *Gsx* positive cells and posterior shift along the A/P axis. The later cerebral vesicle domain of amphioxus Gsx is unaffected by altered RA signalling.

Though no direct regulation of Gsx has yet been observed, the expression of Gsx genes within the anterior CNS of Bilateria (as examined in chapter 1), particularly within the chordates, holds promise for the identification of conserved regulatory pathways involved in the regulation of Gsx. Looking to the regulatory mechanisms controlling Gsx regulation, conserved regulatory elements for vertebrate Gsx genes have been identified (Dimitrieva and Bucher, 2013; Engstrom et al., 2008; Pennacchio et al., 2006; Woolfe et al., 2005), with one *Gsh1* and one *Gsh2* regulatory element from human both driving expression in the CNS of mouse (Pennacchio et al., 2006; Visel et al., 2008). This preliminary data from vertebrates, suggesting conserved regulation of Gsx expression within the CNS, makes amphioxus *Gsx* an excellent candidate for identifying what factors may be regulating conserved or ancestral chordate Gsx expression within the CNS.

One factor examined more closely in this chapter is the transcription factor TCF/Lef, which binds DNA through a High-mobility-group domain (HMG-box) at the DNA minor groove (Love et al., 1995). The widespread conserved nature of TCF/Lef proteins was first characterised with the identification of the Lef-1 homologue *Pangolin* within *Drosophila* (Brunner et al., 1997) and TCF/Lef proteins are present across the metazoa, within species as evolutionarily distant as vertebrates (Behrens et al., 1996; Faro et al., 2009; Huber et al., 1996; Korinek et al., 1998; Molenaar et al., 1998; Roël et al., 2003; Schmidt et al., 2004; Young et al., 2002), echinoderms (Huang et al., 2000), Ecdysozoa (Brunner et al., 1997; Herman, 2001), Cnidaria (Hobmayer et al., 2000)(reviewed in Lee et al. (2006)) and even the Porifera (Adamska et al., 2010; Adell et al., 2007). The conserved binding of these TCF/Lef proteins to nuclear β-catenin (Behrens et al., 1996; Brunner et al., 1997; Huber et al., 1996) is also widespread and is highly studied within the literature. The TCF/Lef/nuclear β-catenin dimer is then able to bind conserved TCF/Lef binding sites (CTTTG A/T A/T) and activate transcription of target genes (Faro et al., 2009; Galceran et al., 1999; Huber et al., 1996; Korinek et al., 1998). This interaction in particular has been well studied as the nuclear effector complex of canonical Wnt signalling (reviewed in Clevers and Nusse (2012)), and TCF/Lef has many actions within both development and disease (reviewed in Arce et al. (2006)). TCF/Lef has been shown to act as both an activator and repressor of this signalling pathway, dependent upon its co-factors. In the absence of Wnt signalling and nuclear β-catenin, TCF/Lef instead binds with the co-factor Groucho, and forms a complex that represses the expression of Wnt gene targets (Brantjes et al., 2001). Groucho is then displaced by nuclear β-catenin in the presence of Wnt signalling (Daniels and Weis, 2005).

TCF/Lef factors are expressed relatively broadly within many tissues at different stages. However, during development, expression is present a higher levels within certain tissues and regions of the embryo. There appears to be a conserved role for TCF/Lef expression within the posterior of the embryo, the endoderm, and the neural tube, particularly within the chordates. Vertebrate TCF/Lef proteins consist of several family members, and show expression within at high levels throughout the neural tube, posterior growth zone, limb buds, somites (Schmidt et al., 2004) and developing gut (Theodosiou and Tabin, 2003) of chick, with similar patterns observed within mice (Ah Cho and Dressler, 1998; Oosterwegel et al., 1993), *Xenopus* (Molenaar et al., 1998; Roël et al., 2003), Zebrafish (Dorsky et al., 1999; Pelegri and Maischein, 1998). Within the invertebrate chordates, *Ci-TCF/Lef* is expressed within the sensory vesicle, nerve cord, and lateral mesoderm (section 5.3.10 and (Imai et al., 2004), whilst amphioxus *TCF/Lef* expression is very similar to that of the vertebrates, with weak expression throughout but stronger expression domains within the neural tube, cerebral vesicle, somites, endoderm and far posterior of the embryo (Lin et al., 2006). Looking outside of the chordates, echinoderm TCF/Lef expression *S.purpuratus* is found to play an important role within the development of the endoderm, mesenchyme and aboral ectoderm (Huang et al., 2000). Within the Ecdysozoa, *Drosophila Pangolin* is expressed throughout the germband embryo as a maternal transcript, though involvement within the posterior elongation of short germ arthropods such as the beetle *Tribolium castaneum* (Bolognesi et al., 2008) and the milkweed bug *Oncopeltus fasciatus* (Angelini and Kaufman, 2005) suggest a role for TCF/Lef in the posterior growth zone, whilst maternal transcripts of *Tc-Pangolin* are found in the anterior of the Tribolium embryo (Bucher et al., 2005). It is clear that TCF/Lef proteins have a wide range of effects, though the conserved expression of chordate TCF/Lef throughout the CNS holds key implications for the development of this tissue and the potential for the regulation of genes expressed within the CNS, such as the chordate ParaHox genes.

Aims: To use *C. intestinalis* as a system in which to test the function of amphioxus ParaHox regulatory elements using cross-species transgenesis, assessing the ability of ParaHox regulatory elements to function across chordate sub-phyla with the aim of identifying functionally conserved regulatory mechanisms. This system will be used to dissect the function of the upstream region of the ParaHox gene *Gsx* of *B. floridae* (Bf-Gsx) using deletion analysis, along with mutagenesis of specific transcription factor binding sites to characterise regulatory function. Comparative genomic techniques will also be used to identify conservation of amphioxus *Gsx* regulatory elements across three species of amphioxus; *B. floridae, B. lanceolatum and B. belcheri.*

5.2. Methods

5.2.1. Bioinformatics

Sequence data for bioinformatic analysis was obtained from both the *B.floridae* PAC clone covering Gsx (PAC-33B4) and initially Scaffold_116 from the *B. floridae* genome (accessed at JGI, http://genome.jgi.doe.gov/Brafl1/Brafl1.home.html). Subsequently, a reassembly (carried out by Nick Putnam) of the JGI trace reads, centred upon the *B.floridae* ParaHox cluster was used. This reassembly is referred to as the *B.floridae* ParaHox reassembly. For both *B.lanceolatum* and *B.belcheri*, sequence data was obtained from the draft genomes of both species (not available for public use at the date of analysis), to which access was kindly granted by Hector Escriva and Anlong Xu respectively. Note that the *B.belcheri* genome has been subsequently released for public access (Huang et al., 2012; Huang et al., 2014). For Phylogenetic footprinting, VISTA (Frazer et al., 2004; Mayor et al., 2000) (using the alignment program Shuffle-LAGAN) was used with the default settings of 70% identity across a window size of 100bp. MULAN (Ovcharenko et al., 2005) was also used, utilising the MultiTF program with both the default 85% similarity threshold and a lower 70% similarity threshold to try to compensate for vertebrate weight matrices. Further transcription factor binding site prediction software used included PROMO (Farre et al., 2003; Messeguer et al., 2002). PROMO currently uses version 8.3 of TRANSFAC (Matys et al., 2006), whilst TESS (Schug, 2008) (support for this program has since been withdrawn due to the author leaving the group and it is now unavailable) was using the TRANSFAC 7.0 (2005) public database (Matys et al., 2006) as well as the JASPAR database (Mathelier et al., 2014).

5.2.2. Experimental Approaches

A 2.1Kb upstream region of *B. floridae Gsx* was obtained by PCR from PAC clone -33B4 (Ferrier et al., 2005) using the Bf-Gsx-up-poly-3Fa and Bf-Gsx-up-poly-1R primers. The 1.7kb Bf-Gsx-Up-Proximal regulatory region (figure 5.1A) was then obtained by further PCR from this upstream region (Osborne 2009 unpublished data), with subsequent smaller regions obtained by PCR from this 1.7kb Bf-Gsx-Up region. All primers used for cloning *Bf-Gsx* regulatory regions are shown in Table 5.1. All cloning of regulatory elements was carried out using the High fidelity Pwo-polymerase with products subsequently A-tailed according to Section 2.2.2. Primers for Bf-Gsx-UpProximal and smaller regions contained a 5’ PstI site on the forward primer and 3’ BamHI site on the reverse primer. These restriction sites facilitated directional cloning into the multiple cloning site of the pCES expression vector with minimal flanking sequence, after shuttling through pGEM-T Easy (Promega). Mutation of TCF/Lef sites within constructs was carried out as described in to section 2.2.14 within the pGEM-T Easy vector before cloning into pCES. Electroporation of pCES constructs was carried out as described in section 2.2.13. All electroporations were carried out in triplicate to ensure repeatable results. In addition, the Bf-Gsx-Up-Proximal regions of 7 different *B. floridae* individuals were isolated from gDNA, cloned and sequenced using the methods described in section 2.2.2-6. These sequences have been deposited in Genbank (accession numbers Bf2_Gsx_Upstream: **KP739759,** Bf4_Gsx_Upstream: **KP739757**, Bf5_Gsx_Upstream: **KP739758**, Bf7_5_Gsx_Upstream: **KP739762**, Bf7_6_Gsx_Upstream: **KP739761**, Bf8_Gsx_Upstream: **KP739763**, Bf9_Gsx_Upstream: **KP739764**, Bf10_Gsx_Upstream: **KP739760**).

Table 5.1. Primers and modifications used to clone regulatory regions.

Identifier	Sequence	Annealing Temp (°C)	Primer Modification
Bf-Gsx-up-poly-3Fa	GTGTCGGATGTTTGCCTTTT	45	n/a
Bf-Gsx-up-poly-1R	AAGTGGCTGTGTCCTGTGGT		n/a
Bf-Gsx-upF (Proximal)	GGATCCTGGGGGAAGAAGAACAA	55	BamHI site
Bf-Gsx-upR (Proximal)	GGATCCCTTGAGTCGACTTCGGTGAC		BamHI site
Bf-Gsx up3F	GCTGCAGAACGCAGCATACAA	52	PstI Site
Bf-Gsx up3R	GCGGATCCACTTTGCCACCA		BamHI Site
Bf Gsx up2F (Gsx-Up2a)	CTGCAGTTGCATGGTGGCAAA	52	PstI site
Bf-Gsx up 2aR (Gsx-Up2a)	TGGGATCCAGGAGAAGGTAAACA		BamHI site
Bf-Gsx up 2bF (Gsx-Up2b)	CTGCAGCCATTCATGCCCGTT	54	PstI site
Bf Gsx up2R (Gsx-Up2b)	GGATCCAGTAGGAGTGAGGAC		BamHI site
Bf Gsx up1F (Gsx-Up1a)	CGCCTGCAGTCCTCACTCCTACT	50	PstI site
BfGsxup 1aR (Gsx-Up1a)	CTGCTCGGATCCTTTACTGCT		BamHI site
BfGsxup 1bF (Gsx-Up1b)	TGCTGCAGTAAAGGTCCGAGCAG	50	PstI site
BfGsxup 1bR (Gsx-Up1b)	TAGGATCCTTGAGTCGACTTCGGTGAC		BamHI site
Bf-Gsx up1cF (Gsx-Up1c)	CTGCAGAAAGGGCCTCTATTGCTTTC	56	PstI site
Bf-Gsx up1cR (Gsx-Up1c)	GGATCCAGCCCTTGCCAATGAAAAA		BamHI site

5.3. Results

5.3.1. An amphioxus *Gsx* regulatory element drives expression of a LacZ reporter throughout the neural tube of *C. intestinalis.*

To screen for potential amphioxus ParaHox gene regulatory elements the work detailed here has taken advantage of the ability to rapidly transform *C. intestinalis* embryos via electroporation. A 1.7kb upstream region of *B. floridae* Gsx, Bf-Gsx-Up-Proximal, spanning from -1667bp to +69bp from the translational start site, was cloned into the MCS of the pCES LacZ reporter (figure 5.1) and found to reliably drive expression of LacZ throughout the central nervous system of *C. intestinalis* embryos. The Bf-Gsx-Up-Proximal driven expression throughout the central nervous system was first detected in the neural plate of early stages (figure 5.1B-E) and then throughout the neural tube, except for the most anterior region of the sensory vesicle (figure 5.1 F-M). This expression was found to be highly reproducible, notwithstanding the fact that not all embryos expressed LacZ within all cells of the CNS due to the mosaic and transient nature of *C. intestinalis* electroporation-mediated transgenesis.

5.3.2. Deletion analysis of the Bf-Gsx-Up-Proximal regulatory element.

With Bf-Gsx-Up-Proximal producing strong, specific neural tube expression, deletion analysis was used as an approach to try and find the minimal region required for neural tube expression, cutting down the original region into several smaller stretches. This began by creating three smaller, overlapping constructs covering the length of the proximal region; Bf-Gsx-Up1, Bf-Gsx-Up2 and Bf-Gsx-Up3 (Osborne 2009, unpublished data) (Figure 5.2). Of these three constructs, only Bf-Gsx-Up1 produced the CNS expression seen in the full Bf-Gsx-Up-Proximal region, with Bf-Gsx-Up2 and Bf-Gsx-Up3 producing no LacZ expression in the CNS. Using the large numbers of embryos afforded by *Ciona* electroporation, it was possible to identify three distinct patterns of LacZ expression within the CNS of *C. intestinalis.* The first of these was expression throughout both the nerve cord and sensory vesicle (Figure 5.2 D,G), mirroring the expression first identified in the Bf-Gsx-Up-Proximal region. The second and third are partial expression patterns, covering the nerve cord only (Figure 5.2 B, E) and sensory vesicle only (Figure 5.2 C,F). For the Bf-Gsx-Up1 regulatory element, there is a 14.1% incidence of the ‘full’ nerve cord + sensory vesicle expression pattern, whilst nerve cord only and sensory vesicle only show an incidence of 6.5% and 6.9% respectively.

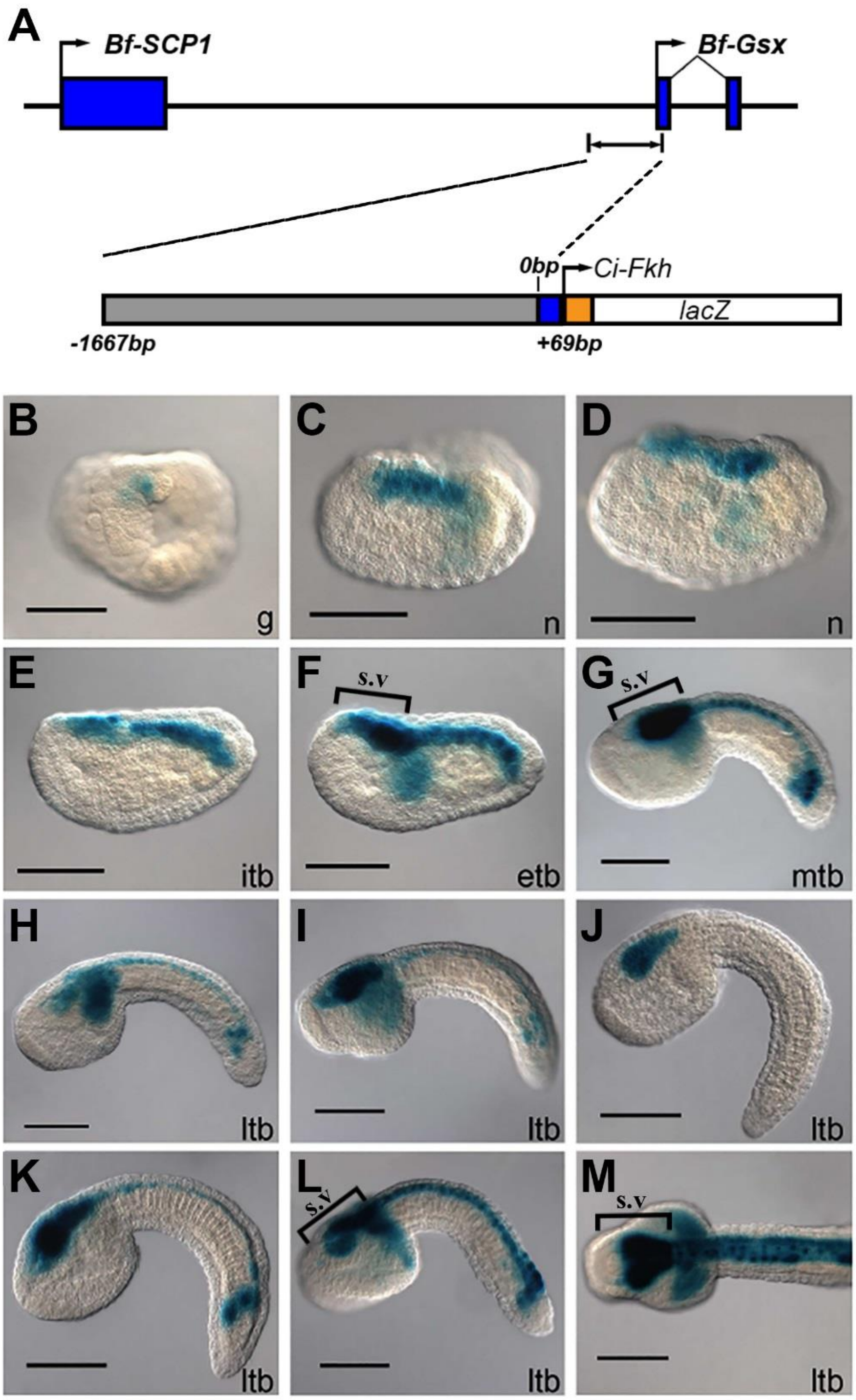


Figure 5.1. Expression of the Bf-Gsx-UpProximal construct in Ciona intestinalis.

(A) Genomic map of the region comprising the Bf-Gsx-UpProximal regulatory element, with pCES LacZ reporter schematic. (B-D) LacZ expression is observed from the earliest collected stages in neural plate cells. (D-L) Tailbud stages: expression can be observed in the mid-posterior of the sensory vesicle (S.V), the visceral ganglion and in every cell of the tail nerve cord. Only the very anterior tip of the SV does not express LacZ. (J) ltb embryo with expression only in the SV. (M) Anterior region of a *Ciona* embryo displaying strong expression in all four rows (dorsal, ventral, left and right) of the tail nerve cord. All embryos are lateral views (except B and M, which are dorsal views) with anterior to left. Lower case lettering refers to the stage of development; g, gastrula; n, neurula; itb, initial tailbud; etb, early tailbud; mtb, mid tailbud; ltb, late tailbud. Scale bars represent 100 µm. (Figure adapted from Osborne 2009, Unpublished data)

5.3.3. Bf-Gsx-Up1c is the minimal enhancer required for nerve cord expression.

As the Bf-Gsx-Up1 region was still producing robust CNS expression further deletion analysis was carried out upon this regulatory element, further dividing this in half again, into Bf-Gsx-Up1a and Bf-Gsx-Up1b. However, when split in this way, neither Up1a nor Up1b showed any CNS expression in *C. intestinalis* embryos (figure 5.2A). This suggested that some sequence in the centre of Bf-Gsx-Up1 was crucial to its regulatory function. Therefore, in order to examine if this break between Up1a and Up1b had disrupted a crucial element in the centre of Bf-Gsx-Up1, a further region, Bf-Gsx-Up1c, was created spanning the centre region of Up1 and bridging the Up1a-Up1b break. With this new construct, Bf-Gsx-Up1c, LacZ expression was detected in the nerve cord and visceral ganglion but not the sensory vesicle (figure 5.2A), albeit at a lower efficiency of 4.8% in the nerve cord only domain, showing it was able to function independently of the surrounding sequence. As such, it can be concluded that Bf-Gsx-Up1c, a region of 215bp (-236 to -21bp from the translational start site (TSS)), is the minimal regulatory region required for nerve cord expression in *C. intestinalis* embryos. Neither sensory vesicle only, nor the combined nerve cord + sensory vesicle expression patterns were present in any of the Bf-Gsx-Up1c embryos, and further sequence, most likely 5' of Up1c, may be required to drive LacZ expression within the sensory vesicle.

5.3.4. The Bf-Gsx-Up1-2 region displays high levels of conservation both within *B.floridae* and across the wider *Branchiostoma* genus.

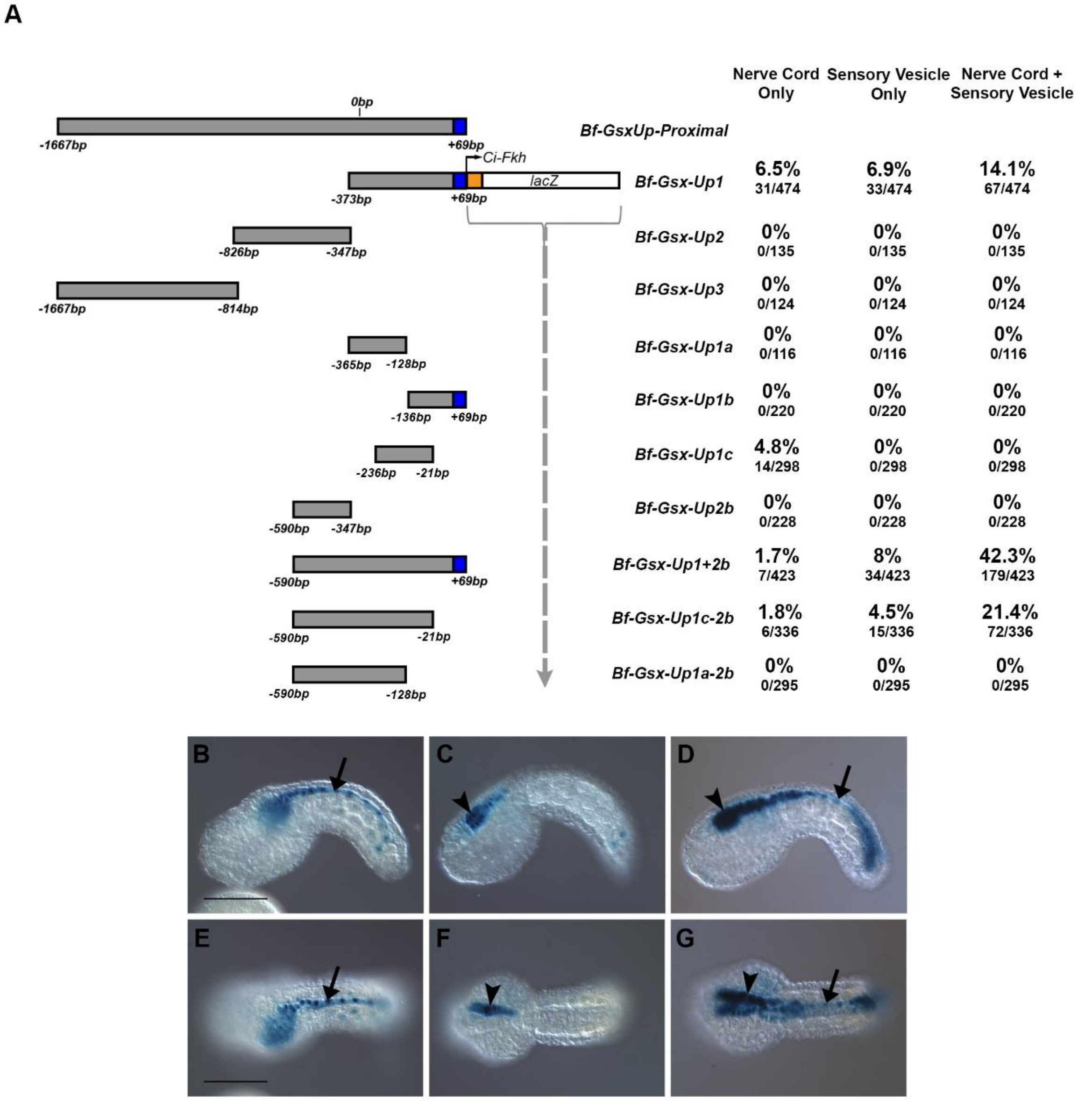
In order to identify regions outside of Bf-Gsx-Up1 that may also contribute to CNS expression, 11 Bf-Gsx-UpProximal sequences, from 7 *B.floridae* individuals, as well as those of the *B.floridae* PAC-33B4 and *B.floridae* scaffold_116, were aligned using Bioedit (Hall, 1999) and ClustalW (Larkin et al., 2007) to examine sequence conservation. These sequences were then aligned and visualised using rVISTA (Frazer et al., 2004; Mayor et al., 2000), using PAC-33B4 as a base sequence (Figure 5.3). Upon comparing the VISTA plots of each individual, it is evident that conservation is very high within the 3'half of the Bf-Gsx-Up Proximal region, but less so in the 5'. Specifically, the Bf-Gsx-Up1 region and the 3'of the Bf-Gsx-Up2 region, show high sequence conservation across all individuals. Within the Bf-Gsx-Up2 region, polymorphism becomes much greater and sequence conservation drops dramatically at the 5' of this region, with polymorphism remaining high throughout the Bf-Gsx-Up3 region. This dramatic difference in sequence conservation and high polymorphism between the 3' and 5' halves of the Bf-Gsx-Up Proximal region suggests that regulatory function is likely to be localised to the Bf-Gsx-Up1 and 3' of the Bf-Gsx-Up2 region.

With the high conservation of the 3’ region of Bf-Gsx-Up Proximal across multiple *B.floridae* individuals, it was possible that deeper conservation of this regulatory region was present across amphioxus species, which would lend further support to it functioning in the regulation of amphioxus *Gsx*. With the recent release of both the *B.lanceolatum* and *B.belcheri* draft genomes, which the Ferrier lab has kindly been granted access to, the Gsx-Up Proximal region was examined across these three different, but closely related amphioxus species (Figure 5.4). In place of the B.fl scaffold_116, the Gsx-Up Proximal region was taken from the *B.floridae* ParaHox reassembly. Using PAC-33B4 as a base sequence to align the others to, the observations highlighted in the polymorphism study become even more apparent when looking between amphioxus species.

Whilst the *B.floridae* ParaHox reassembly shows a VISTA profile very similar to that of the *B.floridae* indivduals, both *B.lanceolatum* and *B.belcheri* have little sequence conservation with *B.floridae* throughout the 5’ of the Gsx-Up Proximal sequence. However, the 3’ region again shows high conservation across the three species, showing high levels of sequence similarity within both the Gsx-Up1 region and 3’ of Gsx-Up2. Looking at both *B.lanceolatum* and *B.belcheri*, the same drop in sequence conservation within the 5’ of the Gsx-Up2 region is seen, though this contrast is far starker in the cross-species comparison due to the lack of sequence conservation in the rest of the 5’ Gsx-Up-3 region.

5.3.5. Addition of the Bf-Gsx-2b region increases expression efficiency, but still requires the Bf-Gsx-Up1c region in order to drive CNS expression.

Though Bf-Gsx-Up2 is unable to drive CNS expression alone (Figure 5.2), VISTA analysis reveals that the 3’ of this region is also highly conserved, along with the functional Bf-Gsx-Up1 region, across both *B.floridae* individuals and *Branchiostoma* species. To confirm that this 3’ region of Bf-Gsx-Up2, or Bf-Gsx-Up2b (-590 to -347bp), is unable to drive reporter expression when isolated from the non-conserved region, the Bf-Gsx-Up2b construct was electroporated to examine LacZ expression. As expected, Bf-Gsx-Up2b transgenic embryos exhibited no LacZ expression in any of the CNS domains (Figure 5.2). However, knowing that the Bf-Gsx-Up1 region is also highly conserved, and does indeed direct strong CNS LacZ expression, the Bf-Gsx-Up2b region could instead provide additional function, or improved efficiency, to the Bf-Gsx-Up1 region, rather than acting independently. In order to examine the function of this region, a longer construct was produced, Bf-Gsx-Up1+2b (-590 to +69bp). When electroporated, it was observed that this region produced lower numbers of individuals with nerve cord only expression, at 1.7% (down from 6.5% with Bf-Gsx-Up1), similar numbers of individuals with only sensory vesicle expression (8% compared to 6.9%), whilst having drastically increased numbers of animals showing the nerve cord plus sensory vesicle expression, up from 14.1% in Bf-Gsx-Up1 to 42.3% of embryos in Bf-Gsx-Up1+2b (figure 5.2). This increase in the prevalence of full CNS expression suggested that the additional conserved sequence found in the Bf-Gsx-Up2b region is in fact important to the function of the regulatory region, in contrast to the results suggested by the lack of CNS expression in the Bf-Gsx-Up2 construct alone (figure 5.2)**.** To identify if this expanded region was still dependant on the minimal enhancer, Bf-Gsx-Up1c, or if the additional sequence added redundancy, two more constructs were made, as it was possible that region important for CNS expression had been split when making Bf-Gsx-Up1 and -Up2, as had happened with Bf-Gsx-Up1a and -Up1b. The first construct, Bf-Gsx-Up1c-2b (-590 to -21bp), was identical to Bf-Gsx-Up1+2b in all respects except that it stopped at the 3' boundary of the Up1c region. This construct still produced the full nerve cord plus sensory vesicle expression seen in both Bf-Gsx-UpProximal and Bf-Gsx-Up1, at higher numbers than Bf-Gsx-Up1 (21.4% up from 14.1% in Bf-Gsx-Up1), but less than that of Bf-Gsx-Up1+2b (42.3%) (figure 2A). However, the second construct, Bf-Gsx-Up1a-2b (-590 to -128bp) had the same 3' boundary as Gsx-Up1a, mirroring the split in Bf-Gsx-Up1a/Bf-Gsx-Up1b and breaking of the Up1c region seen earlier in (figure 2A). Bf-Gsx-Up1a-2b, abolished both nerve cord and sensory vesicle expression, and in fact showed no CNS expression at all, as expected if the intact Gsx-Up1c region is integral to regulatory function.



	Nerve Cord Only	Sensory Vesicle Only	Nerve Cord + Sensory Vesicle
Bf-GsxUp-Proximal			
Bf-Gsx-Up1	6.5% 31/474	6.9% 33/474	14.1% 67/474
Bf-Gsx-Up2	0% 0/135	0% 0/135	0% 0/135
Bf-Gsx-Up3	0% 0/124	0% 0/124	0% 0/124
Bf-Gsx-Up1a	0% 0/116	0% 0/116	0% 0/116
Bf-Gsx-Up1b	0% 0/220	0% 0/220	0% 0/220
Bf-Gsx-Up1c	4.8% 14/298	0% 0/298	0% 0/298
Bf-Gsx-Up2b	0% 0/228	0% 0/228	0% 0/228
Bf-Gsx-Up1+2b	1.7% 7/423	8% 34/423	42.3% 179/423
Bf-Gsx-Up1c-2b	1.8% 6/336	4.5% 15/336	21.4% 72/336
Bf-Gsx-Up1a-2b	0% 0/295	0% 0/295	0% 0/295

Figure 5.2. Deletion analysis of the Bf-Gsx-UpProximal construct.

(A) Deletion map showing the deletion analysis of the Bf-Gsx-UpProximal construct with the numbers of embryos exhibiting LacZ expression in the nerve cord only, sensory vesicle only or both nerve cord and sensory vesicle for each construct recorded both as a percentage of the total number of embryos that developed, and as raw numbers of embryos expressing LacZ for each domain alongside the total number of embryos that developed. Grey regions indicate the relative positions of each construct compared to the Bf-Gsx-UpProximal construct, with 5’ and 3’ limits denoted in number of base pairs from the *B.floridae Gsx* translational start site. Blue regions denote coding sequence, whereas orange regions indicate the pCES Forkhead promoter. The Grey dashed-arrow indicates that the pCES Forkhead-LacZ construct directly abuts the displayed regulatory region in each reporter construct. (B) Lateral view of a mid tailbud *Ciona* embryo displaying nerve cord only LacZ expression. (C) Lateral view of a mid tailbud *Ciona* embryo displaying sensory vesicle only LacZ expression. (D) Lateral view of a mid tailbud *Ciona* embryo displaying the combined nerve cord + sensory vesicle LacZ expression pattern. (E) Dorsal view of a mid tailbud *Ciona* embryo displaying nerve cord only LacZ expression. (F) Dorsal view of a mid tailbud *Ciona* embryo displaying sensory vesicle only LacZ expression. (G) Dorsal view of a mid tailbud *Ciona* embryo displaying nerve cord + sensory vesicle LacZ expression pattern. Black arrows denote LacZ expression within the nerve cord. Black arrowheads denote LacZ expression in the sensory vesicle. Scale bar represents 100 µm.

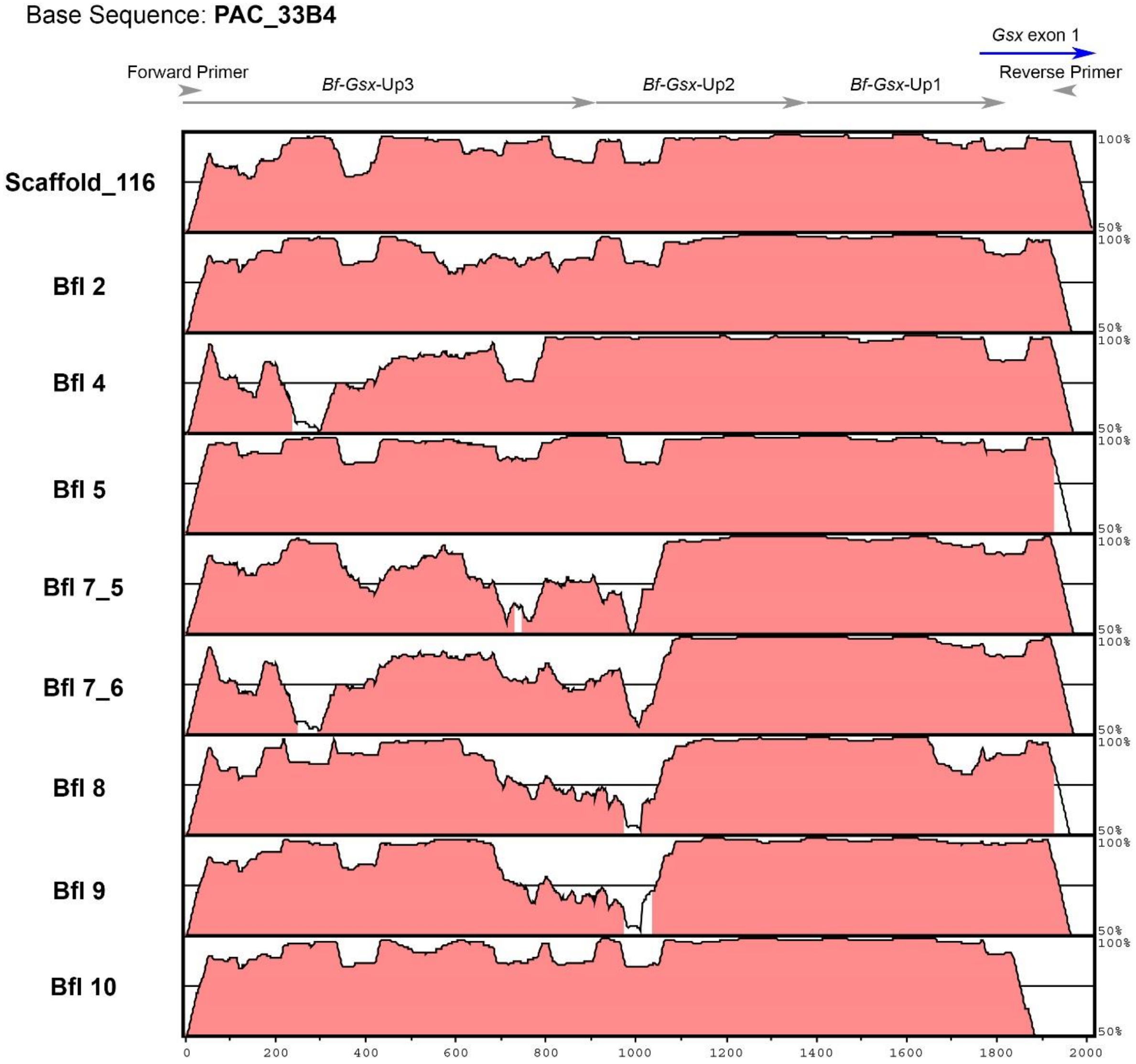


Figure 5.3. VISTA analysis of the polymorphic Bf-Gsx-Upstream Proximal region.

Sequences from multiple individuals were compared to the amphioxus ParaHox PAC sequence. The regions corresponding to the deletion constructs *Gsx*-Up3 and the 5' half of *Gsx*-Up2 are the most variable, while the region covering *Gsx*-Up1 and the 3' half of *Gsx*-Up2 is the most conserved. Note that Bf7_5 and Bf7_6 are different haplotypes from the same individual. Accession numbers for these sequences are found in section 5.2.1. (Figure adapted from (Osborne, 2009 Unpublished data)

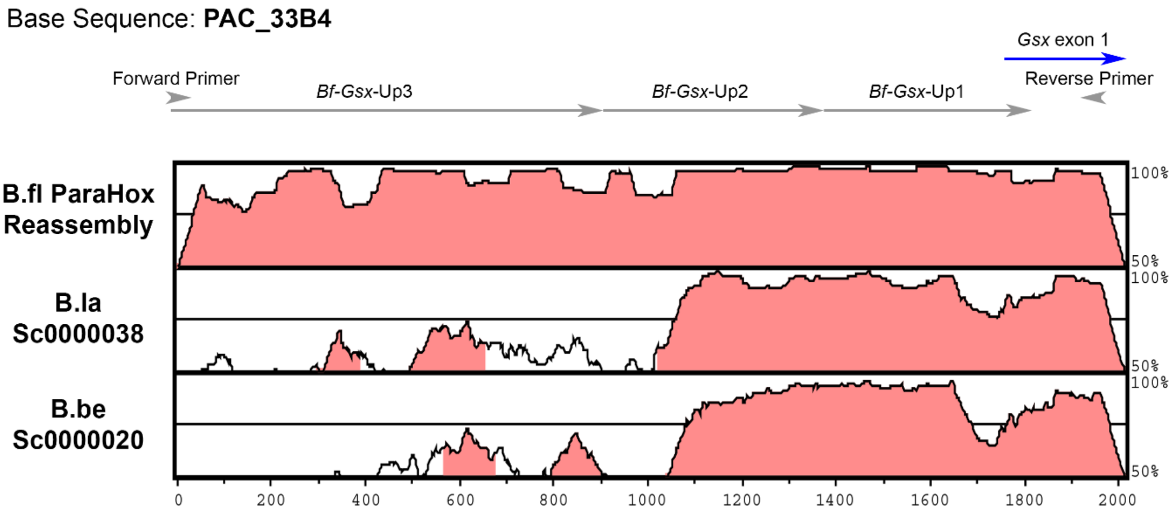


Figure 5.4. VISTA analysis of the Gsx-Upstream Proximal region of three amphioxus species. Sequences from a *B.floridae* reassembly, *B.lanceolatum* and *B.belcheri* were compared to the *B.floridae* ParaHox PAC sequence. The regions corresponding to the deletion constructs *Gsx*-Up3 and the 5' half of *Gsx*-Up2 are the most variable, while the region covering *Gsx*-Up1 and the 3' half of *Gsx*-Up2 is the most conserved.

5.3.6. CNS expression is dependent on the function of TCF/Lef binding sites

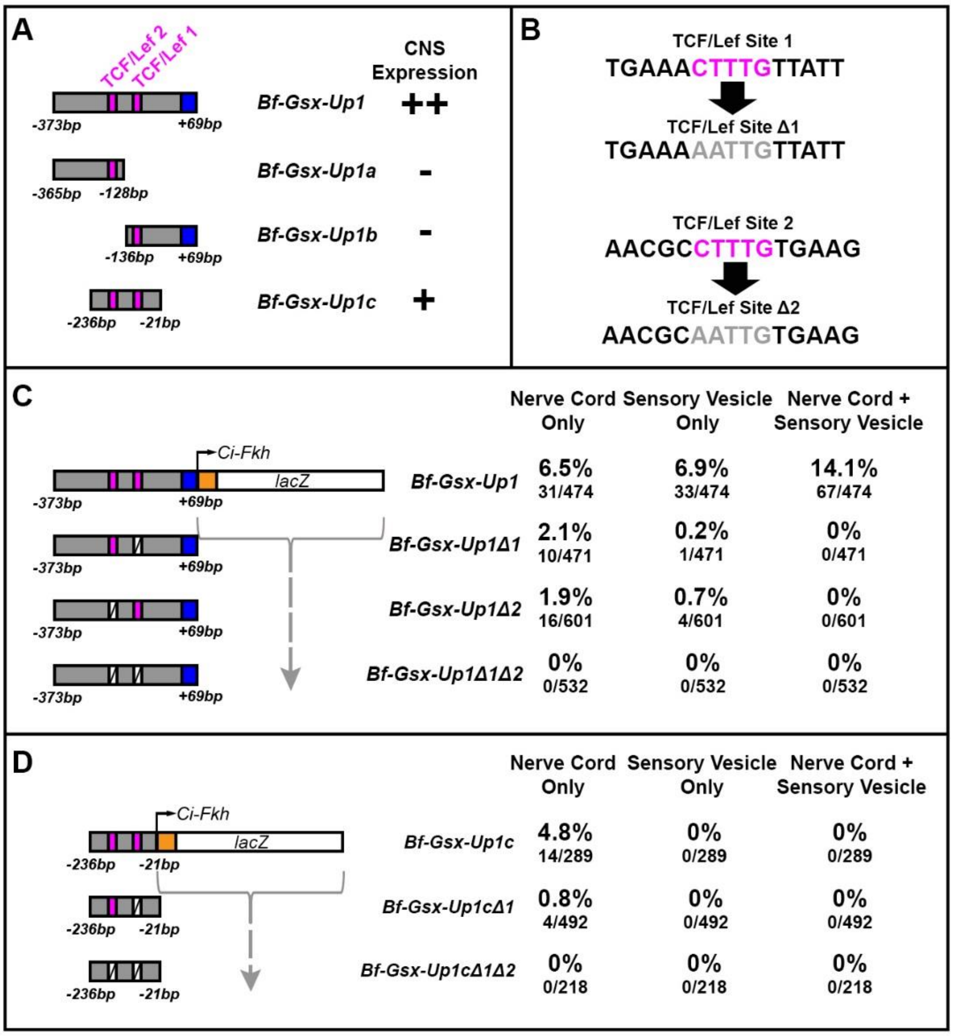
As neither Bf-Gsx-Up1a nor Up1b show CNS expression, yet both Up1 and Up1c do, it was decided to look more closely at what might be so crucial about this central region to the activity of this regulatory region. With expression being abolished when Up1 was split in half, it was hypothesized that at least two binding sites, on either side of the Up1a/1b split, could be functioning in conjunction with one another and that when this region was broken they were not able to drive expression alone. In order to identify transcription factors that may be coordinating the function of this regulatory region, polymorphisms within the Bf-Gsx-Up regulatory region were analysed in order to identify conserved transcription factor binding sites (TFBs). The 11 independent Bf-Gsx-Up-proximal sequences used in the VISTA analysis (Figure 5.3) were submitted to MULAN and analysed using the multiTF program (using vertebrate TFBs)(Ovcharenko et al., 2005), identifying a series of 87 conserved potential binding sites across the Bf-Gsx-Up1 construct. These were then cross-referenced against the ANISEED database (Tassy et al., 2010) to leave a list of 7 transcription factors expressed throughout the entire neural tube of *C. intestinalis*. Of these 7, there were 3 Ets binding factors (Ci-Ets, Cin-ERF and Ets79D), SoxC, Hunchback-like, RAR, and TCF/Lef. Of these factors, all seven are strongly expressed in other tissues that do not express the Bf-Gsx-Up construct apart from TCF/Lef. In order to confirm the presence of TCF/Lef sites both TESS (Schug, 2008) and PROMO (Farre et al., 2003; Messeguer et al., 2002) were used to specifically search for TCF/Lef sites, with TCF/Lef sites being identified at a maximum dissimilarity rate of 15% using vertebrate weight matrices.

In further support for the role of TCF/Lef as a possible activator of the regulatory region, TCF/Lef binding sites are also located on either side of the Up1a/1b boundary, present as a pair in both Bf-Gsx-Up1 and the minimal enhancer Bf-Gsx-Up1c (figure 5.5 A). TCF/Lef binding sites have been well characterised and the consensus 5'-CTTTG[A/T][A/T]-3' is widely accepted as being TCF/Lef specific. Mutagenesis of the two sites identified in both Bf-Gsx-Up1 and the minimal enhancer Bf-Gsx-Up1c, was then performed to see if these sites are involved in driving CNS expression. One of these sites matches the consensus with CTTTGTT, whilst the second site has the slightly divergent CTTTGTG. This second site was not discounted as the flanking sequence, CTTTGTGAA, shows similarity to TCF/Lef sites and it also occupies the functionally relevant location on one side of the Gsx-Up1a/1b split. Indeed, it may be that a single G has been inserted within this particular site, but has not disrupted the CTTTG core and so it could remain functional.

The mutagenesis focused on the core CTTTG element of each TCF/Lef binding site, which has no redundancy in the consensus. In order to alter this core element in each site the following TCF/Lef site mutations were produced; SiteΔ1: TGAAA**AATTG**TTATT, Site Δ2: AACGC**AATTG**TGAAG (figure 5.5 B). These mutations were carried out both separately and as a double TCF/Lef site mutation, with the numbers of animals expressing either nerve cord, sensory vesicle, or both nerve cord and sensory vesicle expression noted for each resulting construct (figure 5.5 C). With both Bf-Gsx-Up1Δ1 and Bf-Gsx-Up1Δ2, mutation of either site alone abolishes the sensory vesicle with nerve cord expression, and reduces individual nerve cord expression, from 6.5% to 2.1% in Δ1 and 1.9% in Δ2, and sensory vesicle expression, from 6.9% to 0.2% in Δ1 and 0.7% in Δ2 (figure 5.5 C). This implies that TCF/Lef sites are contributing to the expression in the CNS, and could be functioning cumulatively. With the double TCF/Lef site mutation, Bf-Gsx-UpΔ1Δ2, complete abolition of CNS expression is observed, with no animals showing either nerve cord or sensory vesicle expression, confirming that these TCF/Lef sites are crucial to the function of the regulatory region (figure 5.5 C). In order to test if the minimal enhancer, Bf-Gsx-Up1c, is indeed functioning as suspected, by allowing these two TCF/Lef sites to interact and drive nerve cord expression, a function lost in Bf-Gsx-Up1a and Bf-Gsx-Up1b where these TCF/Lef sites are separated, same Δ1 and Δ2 mutations were introduced into the minimal enhancer region. Again, numbers of embryos expressing LacZ in either nerve cord, sensory vesicle, or both nerve cord and sensory vesicle were noted (figure 5.5 D). Both Bf-Gsx-Up1cΔ1 and Bf-Gsx-Up1cΔ1Δ2 were cloned into pCES successfully, but unfortunately Bf-Gsx-Up1cΔ2 was refractory to cloning into pCES. Nevertheless, as with Bf-Gsx-Up1, the single site mutation of Bf-Gsx-Up1Δ1 reduced efficiency of nerve cord expression (figure 5.5 D). Also, the double mutation in Bf-Gsx-Up1cΔ1Δ2 completely abolishes nerve cord expression for this minimal region as it does in the larger Bf-Gsx-Up1 (figure 5.5 C, D).

5.3.7. Gsx-Up1c TCF/Lef sites are highly conserved in both position and sequence across amphioxus species.

With functional TCF/Lef binding sites identified within the Bf-Gsx-Up1/Up1c sequence, it was then conjectured whether these binding sites were also conserved within the two other amphioxus species, *B.lanceolatum* and *B.belcheri*. With the previous VISTA alignment showing that the Up1+2b region was highly conserved across *Branchiostoma* species, the alignment was carried out again, this time through MULAN using multiTF with vertebrate TCF/Lef binding site weight matrices, and a similarity threshold of 85%. This enabled identification TCF/Lef binding sites across the three species. This approach identified 4 sites within the Gsx-UpProximal sequences, two of which were identified by vertebrate Lef-1 matrices, (**GCCTTTGTGA** and **AACTTTGTTA**) and two by vertebrate TCF4 matrices (**ATAAAAGC** and **TTCAAAGG**). Of the sites identified, only three contained the CTTTG (or the reverse CAAAG) motif that was previously established as one of the parameters for TCF/Lef binding sites. Two of these were the 'Lef-1' sites, and one a 'TCF4' site, though this 'TCF4' site is present within the *Gsx* coding sequence and so was discounted from my analysis. This left two TCF/Lef binding sites meeting all of the criteria; those identified by the Lef-1 weight matrices. These sites not only had conserved sequence, but also conserved position within the Gsx-Up1c regulatory region (Figure 5.6). This, along with the highly conserved sequence across the Gsx-Up1+2b region (Section 5.3.4, Figures 5.3 & 5.4), may indicate that the Bf-Gsx-Up1+2b region carries regulatory function and responds to TCF/Lef binding across all three amphioxus species.

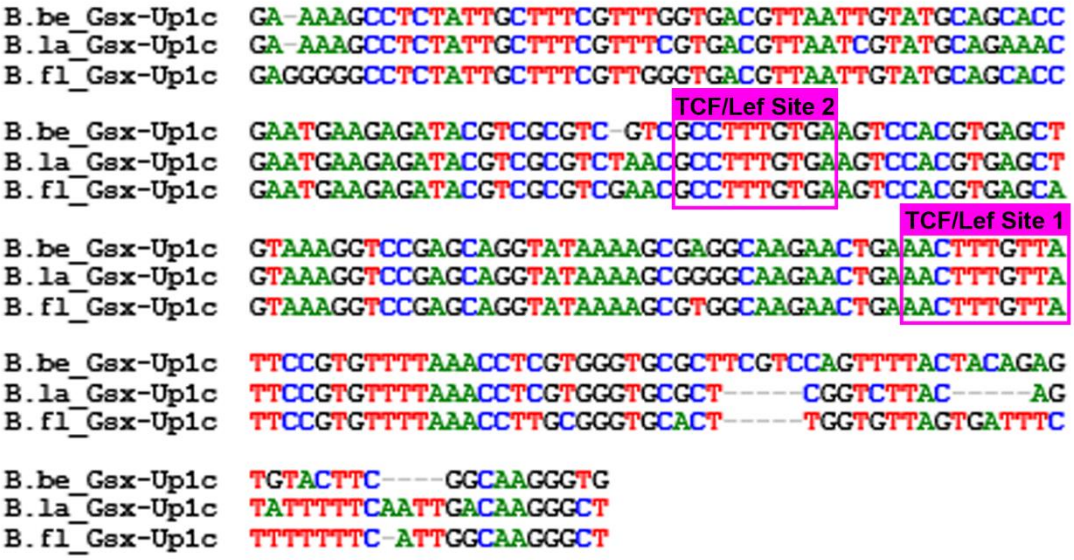


	Nerve Cord Only	Sensory Vesicle Only	Nerve Cord + Sensory Vesicle
Bf-Gsx-Up1	6.5% 31/474	6.9% 33/474	14.1% 67/474
Bf-Gsx-Up1Δ1	2.1% 10/471	0.2% 1/471	0% 0/471
Bf-Gsx-Up1Δ2	1.9% 16/601	0.7% 4/601	0% 0/601
Bf-Gsx-Up1Δ1Δ2	0% 0/532	0% 0/532	0% 0/532

	Nerve Cord Only	Sensory Vesicle Only	Nerve Cord + Sensory Vesicle
Bf-Gsx-Up1c	4.8% 14/289	0% 0/289	0% 0/289
Bf-Gsx-Up1cΔ1	0.8% 4/492	0% 0/492	0% 0/492
Bf-Gsx-Up1cΔ1Δ2	0% 0/218	0% 0/218	0% 0/218

Figure 5.5. Mutation of TCF/Lef sites within the Bf-Gsx-Up1 and Bf-Gsx-Up1c constructs.

(A) Relative positions of TCF/Lef sites within the Bf-Gsx-Up1, Bf-Gsx-Up1a, Bf-Gsx-Up1b and Bf-Gsx-Up1c regulatory regions. TCF/Lef sites lie either side of the Bf-Gsx-Up1a/Bf-Gsx-Up1b split. + symbols denote the presence of LacZ expression in the corresponding construct, with + denoting low LacZ expression and ++ high LacZ expression, whereas – denotes the absence of LacZ expression. (B) Schematic showing the DNA sequence of TCF/Lef site1 and TCF/Lef site2 before and after mutations were carried out. Pink sequence denotes the TCF/Lef site 'core' sequence before mutation, whereas light grey sequence denotes the TCF/Lef site 'core' sequence after mutation. (C) Comparison of CNS expression incidence in the Bf-Gsx-Up1 construct with TCF/Lef binding motif mutants. The numbers of embryos displaying either nerve cord only, sensory vesicle only, or the nerve cord with sensory vesicle LacZ expression patterns have been recorded both as a percentage of the total number of embryos that developed and as raw numbers of embryos expressing LacZ for each domain alongside the total number of embryos that developed. Pink boxes denote the positions of intact TCF/Lef sites, whereas white crossed boxes indicate the positions of mutated TCF/Lef sites. (D) Comparison of CNS expression incidence in the Bf-Gsx-Up1c 'minimal enhancer' construct with TCF/Lef binding motif mutants. The numbers of embryos displaying either nerve cord only, sensory vesicle only, or the nerve cord with sensory vesicle LacZ expression patterns have been recorded both as a percentage of the total number of embryos that developed and as raw numbers of embryos expressing LacZ for each domain alongside the total number of embryos that developed. Pink boxes denote the positions of intact TCF/Lef sites, whereas white crossed boxes indicate the positions of mutated TCF/Lef sites.



```
B.be_Gsx-Up1c  GA-AAAGCCTCTATTGCTTTCGTTTGGTGACGTTAATTGTATGCAGCACC
B.la_Gsx-Up1c  GA-AAAGCCTCTATTGCTTTCGTTTCGTGACGTTAATCGTATGCAGAAAC
B.fl_Gsx-Up1c  GAGGGGGCCTCTATTGCTTTCGTTGGGTGACGTTAATTGTATGCAGCACC

B.be_Gsx-Up1c  GAATGAAGAGATACGTCGCGTC-GTCGCCTTTGTGAAGTCCACGTGAGCT
B.la_Gsx-Up1c  GAATGAAGAGATACGTCGCGTCTAACGCCTTTGTGAAGTCCACGTGAGCT
B.fl_Gsx-Up1c  GAATGAAGAGATACGTCGCGTCGAACGCCTTTGTGAAGTCCACGTGAGCA

B.be_Gsx-Up1c  GTAAAGGTCCGAGCAGGTATAAAAGCGAGGCAAGAACTGAAACTTTGTTA
B.la_Gsx-Up1c  GTAAAGGTCCGAGCAGGTATAAAAGCGGGGCAAGAACTGAAACTTTGTTA
B.fl_Gsx-Up1c  GTAAAGGTCCGAGCAGGTATAAAAGCGTGGCAAGAACTGAAACTTTGTTA

B.be_Gsx-Up1c  TTCCGTGTTTTAAACCTCGTGGGTGCGCTTCGTCCAGTTTTACTACAGAG
B.la_Gsx-Up1c  TTCCGTGTTTTAAACCTCGTGGGTGCGCT-----CGGTCTTAC-----AG
B.fl_Gsx-Up1c  TTCCGTGTTTTAAACCTTGCGGGTGCACT-----TGGTGTTAGTGATTTC

B.be_Gsx-Up1c  TGTACTTC----GGCAAGGGTG
B.la_Gsx-Up1c  TATTTTTCAATTGACAAGGGCT
B.fl_Gsx-Up1c  TTTTTTTC-ATTGGCAAGGGCT
```

Figure 5.6. Alignment showing conservation of TCF/Lef sites within the Gsx-Up1c regulatory region of 3 amphioxus species.

Both *B.lanceolatum* and *B.belcheri* sequences are aligned against the *Branchiostoma floridae* Gsx-Up1c sequence. Bases are coloured as follows to allow viewing of conservation: G=black, A=Green, T=Red, C=Blue. Conserved TCF/Lef sites are indicated by pink boxes.

5.3.8. TCF/Lef sites show unequal contribution to CNS expression domains

The role of TCF/Lef in the function of the Bf-Gsx-Up1 and Up1c regulatory regions prompted the search for further sites that may be contributing towards the increase in expression efficiency seen in those constructs also containing the 2b region. As the Bf-Gsx-Up1/Up1c regions showed at least some collaborative effect between TCF/Lef sites, a third site located within the Gsx-Up2b region indicated a good target for further mutagenesis (figure 5.7 A). By analysing the effect of mutation on this third site, again as both a single mutation and in all possible permutations with the existing 'core' TCF/Lef Δ1 and Δ2 mutations, it was hoped to examine if TCF/Lef site function was acting cumulatively and could account for the increase in expression seen in Bf-Gsx-Up1+2b and Up1c-2b, or if this third site could perhaps buffer against mutations in the 'core' Up1c region, providing a level of redundancy. Thus, the SiteΔ3 mutation, GTAGG**AATTG**ATGAA was produced (figure 5.7 B). The first of this set of Gsx-Up1+2b constructs, Bf-Gsx-Up1+2bΔ1, contains a mutation of the first 'Core' TCF/Lef site as carried out in the other constructs. This produced a dramatic decrease in CNS expression overall, though it is most apparent in the combined sensory vesicle with nerve cord expression, which decreases from 42.3% in the wild type Bf-Gsx-Up1+2b to 10% in the Bf-Gsx-Up1+2bΔ1 mutant (Figure 5.7 B) . Interestingly, the Bf-Gsx-Up1+2bΔ2 mutated construct showed a less significant decrease in expression in the combined pattern (23.5%), though nerve cord (0.4%) and sensory vesicle (6.7%) individually show comparable results to that of the Bf-Gsx-Up1+2bΔ1 construct (0.2% and 4.3% respectively) (figure 5.7 C). These results seem to show a disparity in the contribution to regulatory function between site1 and site2, perhaps explained by the non-canonical binding sequence of site2, and the ability of site3 to compensate for this lower affinity site. However, if both site1 and site2 are mutated, as in the Bf-Gsx-Up1+2bΔ1Δ2 construct, the nerve cord with sensory vesicle expression decreases dramatically from 42.3% in the WT to 6.5% in the Δ1Δ2 mutation (figure 5.7 C). This is also lower than either single Core site mutation alone, supporting the idea that these TCF/Lef sites are functioning cumulatively. However, what does not happen is a complete abolition as in the Bf-Gsx-Up1Δ1Δ2 and Bf-Gsx-Up1cΔ1Δ2 constructs (figure 5.5 C,D), implying that the third site (TCF/Lef 3) in the Bf-Gsx-Up1+2b region (figure 5.7 A) is able to partially compensate for the lack of TCF/Lef binding in the Core Up1c region (figure 5.7 C).

The 3' deletions of the Gsx-Up1+2b region, in Bf-Gsx-Up1a-2b, do show that the Up1c region is required for CNS expression, even in the presence of a functional site3 binding site in the 2b region, suggesting that other transcription factors with binding sites in the Up1c core may be required to activate expression. Another particularly interesting observation from the Bf-Gsx-Up1+2bΔ1Δ2 construct is that sensory vesicle expression alone is actually increased above that of the WT, from 8% to 12% in the Δ1Δ2 mutation (figure 5.7 C). This increase in sensory vesicle expression is indicative of a loss of nerve cord expression in embryos that would otherwise show the full nerve cord with sensory vesicle expression pattern and that TCF/Lef site3 has a biased expression towards that of sensory vesicle rather than nerve cord. The converse mutation, with site3 mutated and the Core site1 and site 2 intact, shows the opposite outcome to this. Bf-Gsx-Up1+2bΔ3 shows a decrease in both the combined nerve cord with sensory vesicle and sensory vesicle alone categories, but this time has nerve cord only expression increased above that of the WT Bf-Gsx-Up1+2b region. This leads to a model where the Core Up1c TCF/Lef sites contribute more heavily, but not exclusively, to expression in the nerve cord (figure 5.7 D(i)) and TCF/Lef site 3, in the Gsx-Up2b region, contributes more heavily, but again not exclusively, to sensory vesicle expression (figure 5.7 D(ii)). The final construct, with all three TCF/Lef site mutations, Bf-Gsx-Up1+2bΔ1Δ2Δ3 bolsters previous evidence from Bf-Gsx-Up1Δ1Δ2 and Bf-Gsx-Up1c Δ1Δ2, showing that whilst other transcription factors may be involved in refining the output of this regulatory element, if all TCF/Lef binding sites are mutated then CNS expression is completely abolished and it is this transcription factor that provides the principal activation input (figure 5.5 C, D and 5.7 C).

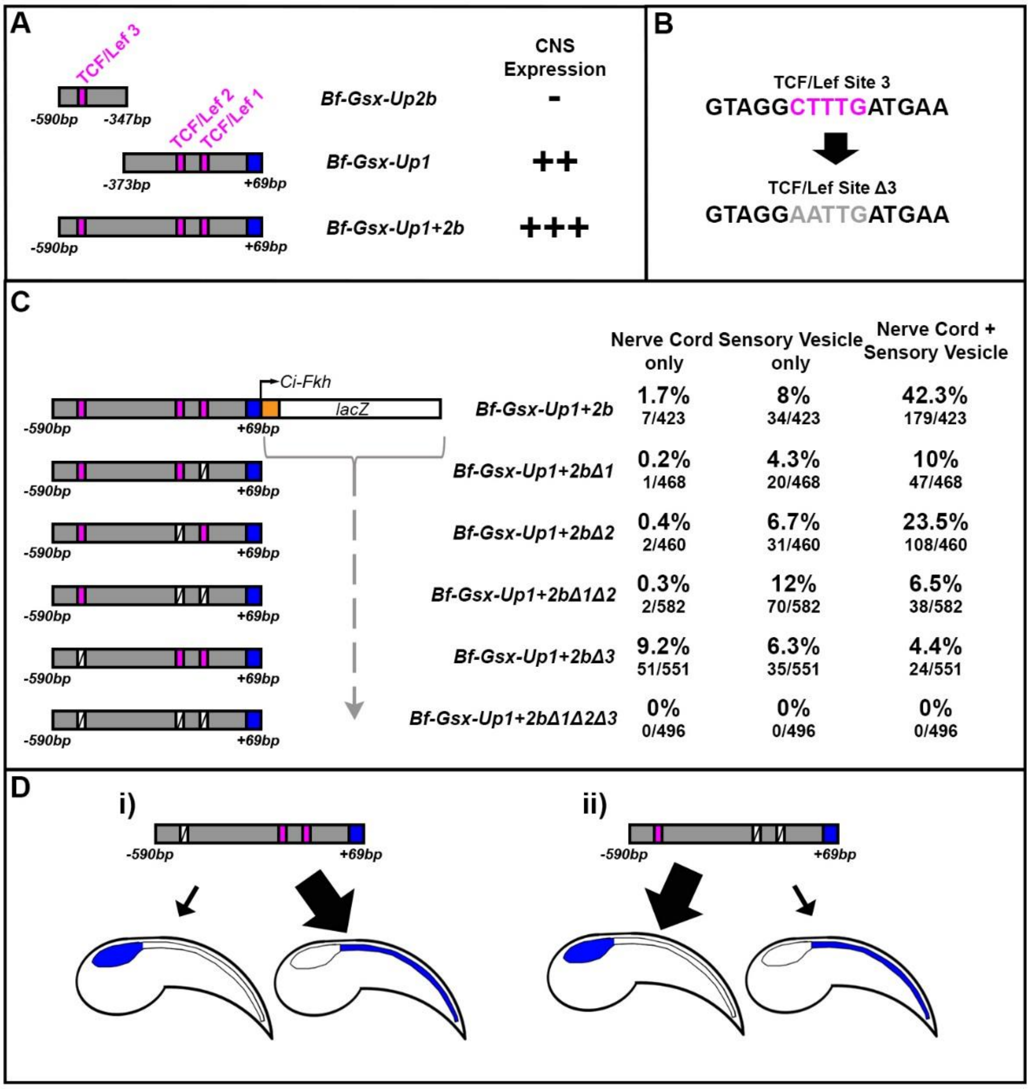

A
TCF/Lef 3
-590bp
-347bp
TCF/Lef 2
TCF/Lef 1
-373bp
+69bp
-590bp
+69bp
CNS Expression
Bf-Gsx-Up2b -
Bf-Gsx-Up1 ++
Bf-Gsx-Up1+2b +++
B
TCF/Lef Site 3
GTAGGCTTTGATGAA
TCF/Lef Site Δ3
GTAGGAATTGATGAA
C
Ci-Fkh
lacZ
-590bp
+69bp
Nerve Cord only
Sensory Vesicle only
Nerve Cord + Sensory Vesicle
Bf-Gsx-Up1+2b 1.7% 7/423 8% 34/423 42.3% 179/423
Bf-Gsx-Up1+2bΔ1 0.2% 1/468 4.3% 20/468 10% 47/468
Bf-Gsx-Up1+2bΔ2 0.4% 2/460 6.7% 31/460 23.5% 108/460
Bf-Gsx-Up1+2bΔ1Δ2 0.3% 2/582 12% 70/582 6.5% 38/582
Bf-Gsx-Up1+2bΔ3 9.2% 51/551 6.3% 35/551 4.4% 24/551
Bf-Gsx-Up1+2bΔ1Δ2Δ3 0% 0/496 0% 0/496 0% 0/496
D
i)
-590bp
+69bp
ii)
-590bp
+69bp

Figure 5.7. Mutation of TCF/Lef sites within the Bf-Gsx-Up1+2b construct.

(A) Relative positions of TCF/Lef sites within the Bf-Gsx-Up2b, Bf-Gsx-Up1 and Bf-Gsx-Up1+2b regulatory regions. An additional TCF/Lef site within the 2b region is added in the Bf-Gsx-Up1+2b construct. + symbols denote the presence of LacZ expression in the corresponding construct, with ++ denoting high LacZ expression and +++ very high LacZ expression, whereas – denotes the absence of LacZ expression. (B) Schematic showing the DNA sequence of TCF/Lef site 3 before and after mutation was carried out. Pink sequence denotes the TCF/Lef site 'core' sequence before mutation, whereas light gray sequence denotes the TCF/Lef site 'core' sequence after mutation. (C) Comparison of CNS expression incidence in the Bf-Gsx-Up1+2b construct with TCF/Lef binding motif mutants. The numbers of embryos displaying either nerve cord only, sensory vesicle only, or the nerve cord with sensory vesicle LacZ expression patterns have been recorded both as a percentage of the total number of embryos that developed and as raw numbers of embryos expressing LacZ for each domain alongside the total number of embryos that developed. Pink boxes denote the positions of intact TCF/Lef sites, whereas white crossed boxes indicate the positions of mutated TCF/Lef sites. (D) Schematic showing the partial division of function into 'nerve cord' and 'sensory vesicle' domains across the Bf-Gsx-Up1+2b regulatory element. (D(i)) Shows the bias of the 3' region and TCF/Lef sites 1 and 2 to drive nerve cord expression over sensory vesicle expression whilst (D(ii)) shows the bias of the 5' region and TCF/Lef site 3 to drive sensory vesicle expression over nerve cord expression.

5.3.9. Mutation of TCF/Lef sites unmasks a latent repressive function

A final observation, made when comparing expression in constructs with all present TCF/Lef sites mutated, was that the background head mesenchyme and tail muscle expression inherent to the pCES construct was different between Bf-Gsx-Up1cΔ1Δ2, Bf-Gsx-Up1Δ1Δ2 and Bf-Gsx-Up1+2bΔ1Δ2Δ3. It was observed that as the TCF/Lef mutant constructs became longer, from Bf-Gsx-Up1cΔ1Δ2 as the smallest to Bf-Gsx-Up1+2bΔ1Δ2Δ3 as the longest, the pCES background expression also decreased. This led to, alongside the lack of CNS expression in all of these constructs, high pCES background in Bf-Gsx-Up1cΔ1Δ2 (52.2% of embryos), a small amount of pCES background in Bf-Gsx-Up1Δ1Δ2 (9.6% of embryos), and a complete abolition of any expression, in the longer Bf-Gsx-Up1+2bΔ1Δ2Δ3 (0% of embryos) (figure 5.8). Numbers of embryos displaying background pCES expression within the wild-type constructs are as follows; Bf-Gsx-Up1c, 88.2% of embryos (255/289), Bf-Gsx-Up1, 45.5% of embryos (156/343), and Bf-Gsx-Up1+2b, 10.9% of embryos (46/423). This then shows a decrease in the levels of pCES background expression in direct response to TCF/Lef site mutation, with Bf-Gsx-Up1c > Bf-Gsx-Up1cΔ1Δ2 (decreasing from 88.2% to 52.2%), Bf-Gsx-Up1 > Bf-Gsx-Up1Δ1Δ2 (decreasing from 45.5% to 9.6%), and Bf-Gsx-Up1+2b > Bf-Gsx-Up1+2bΔ1Δ2Δ3 (decreasing from 10.9% to 0%). It should be noted that pCES background expression levels in embryos containing the empty pCES vector (i.e. the reporter with no regulatory element insertion) lies at 94.6% of embryos (142/150), whilst a long but non-functional region such as Bf-Gsx-Up2 (479bp) has pCES background within 88.8% of embryos (120/135). This suggests that the decrease in pCES background seen in response to increased construct length is also specific to the Bf-Gsx-Up1+2b region. It was therefore hypothesised that by removing TCF/Lef activation, a latent repressive function may be unmasked that is spread throughout the regulatory region. Thus, in the absence of TCF/Lef binding, as the region inserted into the pCES multiple cloning site increases in size, it becomes more able to repress the background activity of the forkhead promoter. This repression may be mediated by the binding of, currently unknown, repressive transcription factors. A combination of activation and repression effects, likely mediated via transcription factor binding, would then allow the tight control of expression within different tissues and developmental stages via the presence of different sets of transcription factors. For developmental genes, where ectopic or 'leaky' expression could cause dramatic phenotypes in embryogenesis, this kind of mechanism would help greatly in narrowing expression of a gene to specific regions and times.

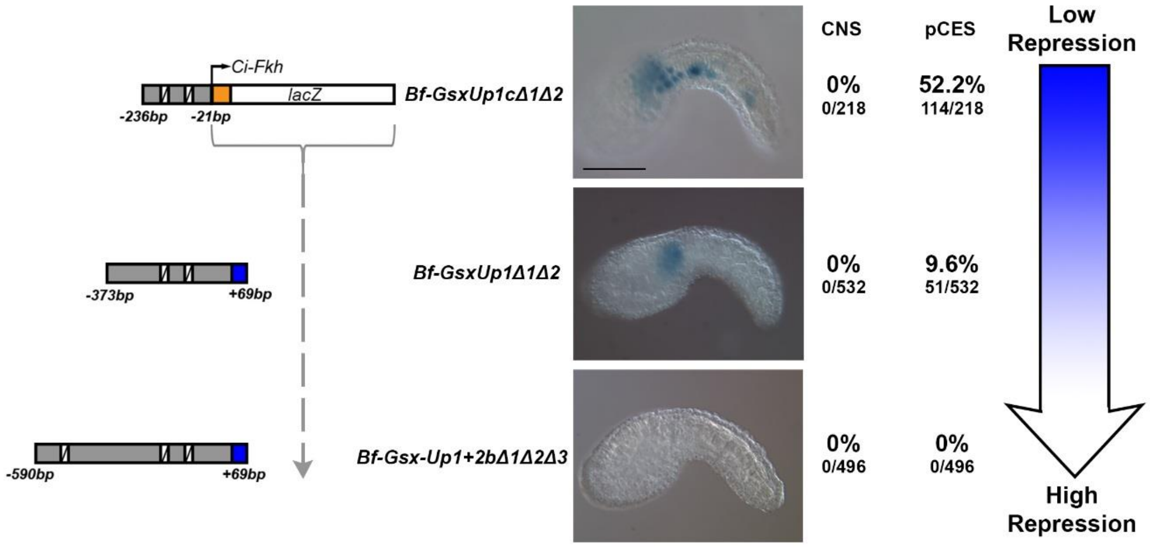


Figure 5.8. Mutation of all TCF/Lef sites abolishes CNS expression and reveals a latent repressive function that increases with regulatory element length.
Images of mid tailbud *Ciona* embryos represent the maximal example of pCES 'background' LacZ expression for Bf-Gsx-Up1cΔ1Δ2, Bf-Gsx-Up1Δ1Δ2 and Bf-Gsx-Up1+2bΔ1Δ2Δ3 respectively. The numbers of embryos displaying either of the CNS LacZ expression domains examined in figures 2, 4 and 5 along with those displaying only pCES background expression. Numbers of embryos showing either pattern have been recorded both as a percentage of the total number of embryos that developed and as raw numbers of embryos expressing LacZ for each domain alongside the total number of embryos that developed. White crossed boxes indicate the positions of mutated TCF/Lef sites. The blue graduated arrow represents the decrease in pCES background expression associated with an increase in construct length from Bf-Gsx-Up1cΔ1Δ2 to Bf-Gsx-Up1Δ1Δ2, to Bf-Gsx-Up1+2bΔ1Δ2Δ3.

5.3.10. *Ci-TCF/Lef in-situ* hybridisation

To confirm that *Ciona TCF/Lef* is present within the tissues that the Bf-Gsx-Up reporters are expressed, *in-situ* hybridisation of *Ci-TCF/Lef* was carried out, analysing a time course around the stages in which the Bf-Gsx-UpProximal reporter is activated (figure 5.9). Expression begins in mesenchymal and neural plate cells (figure 5.9 A,E), before becoming more widespread in the neural plate and head/lateral trunk mesenchyme (figure 5.9 B,C,D,F,G,H), with some staining also in the endoderm (G,H). In tailbud stages, staining becomes more refined in the head/lateral trunk mesenchyme and weakly throughout the nerve cord. Weak staining can also be seen in the endodermal strand in initial and early tailbud stages (I,J,M,N). From early tailbud onwards, a strong domain of expression is seen within the centre of the sensory vesicle (figure 5.9 J, K, L, N, O, P), which remains even when staining becomes weaker in the rest of the CNS in the mid-late tailbud (figure 5.9 O-T). This time course confirms that *C. intestinalis TCF/Lef* is expressed in the developing neural plate during the mid-gastrula through neurula stages (figure 5.9A-G) when the Bf-Gsx-Up reporter is first activated, and also in the sensory vesicle and more weakly throughout the neural tube in the early and mid tailbud stages (figure 5.9 Q-T).

Ci-TCF/Lef is also expressed strongly in the lateral head mesenchyme (figure 5.9 D,G,H-O), which happens to correlate with a particular component of the pCES background expression that is seen in some of the reporter constructs (figure 7.4 in appendix 7). This 'background' expression is inherent to the forkhead promoter of pCES and has been well characterised as LacZ expression in the head/neck mesenchyme and tail muscle cells (figure 7.4 C-H in appendix 7), with some animals also showing LacZ in the centre of the sensory vesicle (figure 7.4 C,E,G,H in appendix 7)**.** This background becomes much weaker or is completely abolished when a regulatory element is driving the promoter. The head/neck mesenchyme and sensory vesicle pCES background expression appears to follow a similar pattern to that of *C.intestinalis TCF/Lef*. The sensory vesicle expression that is seen as pCES background is easy to distinguish from the sensory vesicle expression seen with the Bf-Gsx-Up constructs, as the Bf-Gsx-Up regulatory elements drive expression much more expansively throughout the sensory vesicle.

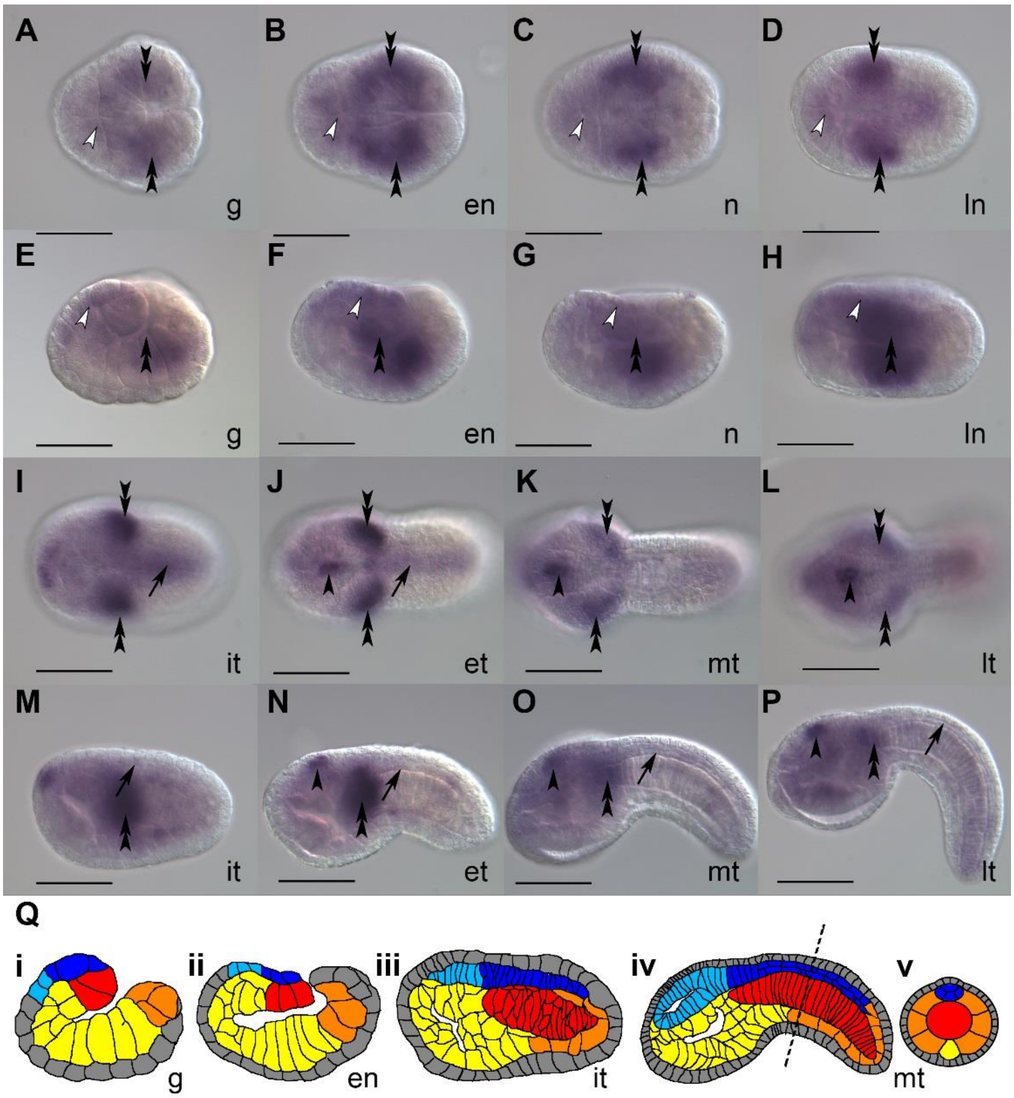


Figure 5.9. Expression of Ciona intestinalis TCF/Lef. In situ hybridization of Ci-TCF/Lef mRNA. A-D and I-L represent dorsal views, whilst E-H and M-P are lateral views. Expression begins in mesenchymal and neural plate cells (A, E), before becoming more widespread in the neural plate (white arrow heads) and head/lateral trunk mesenchyme (black double arrowheads) (B, C, D, F, G, H), with possibly some staining also in the endoderm (G, H). In tailbud stages staining becomes more refined in the head/lateral trunk mesenchyme and weakly throughout the nerve cord (black arrow). Weak staining can also be seen in the endodermal strand in initial and early tailbud stages (I, J, M, N). From early tailbud onwards, a strong domain of expression exists within the centre of the sensory vesicle (black single arrowhead) (J, K, L, N, O, P), which remains even when staining becomes weaker in the rest of the CNS in the mid-late tailbud (K, L, O, P). (Qi-iv) Schematic showing cell and tissue fates through developing embryos. Schematics are made from traces of the embryos in (E, F, M, O), though numbers of cells displayed may not be absolutely accurate due to cell membranes not being visible in different focal planes. (v) Represents a transverse section through the plane shown by the dotted line in (iv). Presumptive notochord cells are shown in red, endoderm yellow, muscle orange, epidermis grey, nerve cord dark blue and sensory vesicle light blue. Lower case lettering refers to the stage of development; g, gastrula; en, early neurula; n, neurula; ln, late neurula; it, initial tailbud; et, early tailbud; mt, mid tailbud; lt, late tailbud. A-D and I-L represent dorsal views, whilst E-H and M-P are lateral views. Scale bars represent 100 µm.

5.3.12. *Ci-Gsx In-situ* Hybridisation

Whilst amphioxus *Gsx* expression has been well characterised (Osborne et al., 2009), *C.intestinalis Gsx* has been examined in relatively few embryonic stages, having been first described in (Hudson and Lemaire, 2001). Thus, a full embryonic time course was produced so that the expression of *Ci-Gsx* could be compared to that of both amphioxus and the Bf-Gsx-Up reporters in *C.intestinalis*. *Ci-Gsx* is first observed in mid-gastrula stage, as faint expression within the a9.33 cell pair (figure 5.10 A). It then becomes stronger throughout the late gastrula stage and very early neurula stages, becoming expressed in both daughter cells, the a10.65 and a10.66 cell pairs (figure 5.10 B-C). These go on to the neurula stage, with the a10.74 cell pair also showing some expression, though it is weak in the embryos examined here (figure 5.10 D). Data from ANISEED (Tassy et al., 2010), also corroborates this a10.74 expression.

From the initial tailbud to early tailbud stages, expression becomes split into a strong domain covering the presumptive posterior sensory vesicle, and also a weaker domain in part of the presumptive anterior sensory vesicle (figure 5.10 E-F). As embryos develop through to mid and late tailbud stages, these two domains become stronger and more defined, with expression in the posterior sensory vesicle and in the part of the ventral anterior sensory vesicle (figure 5.10 G-L). However, a peculiarity arises at the late tailbud stage. In almost all late tailbud stage embryos, *Ci-Gsx* is expressed in only one half of the sensory vesicle across the sagittal plane (figure 5.10 J, L). This half differs between embryos, with some individuals displaying expression only in the right half of the brain whilst others have expression only in the left half. Numbers of late tailbud embryos and the location of Gsx expression (right half, left half or both) are given in Figure 5.11. It should be noted that the embryo showing expression in both halves of the sensory vesicle lacks expression within the anterior of the right half of the sensory vesicle, whereas it displays both anterior and posterior sensory vesicle expression within the left half.

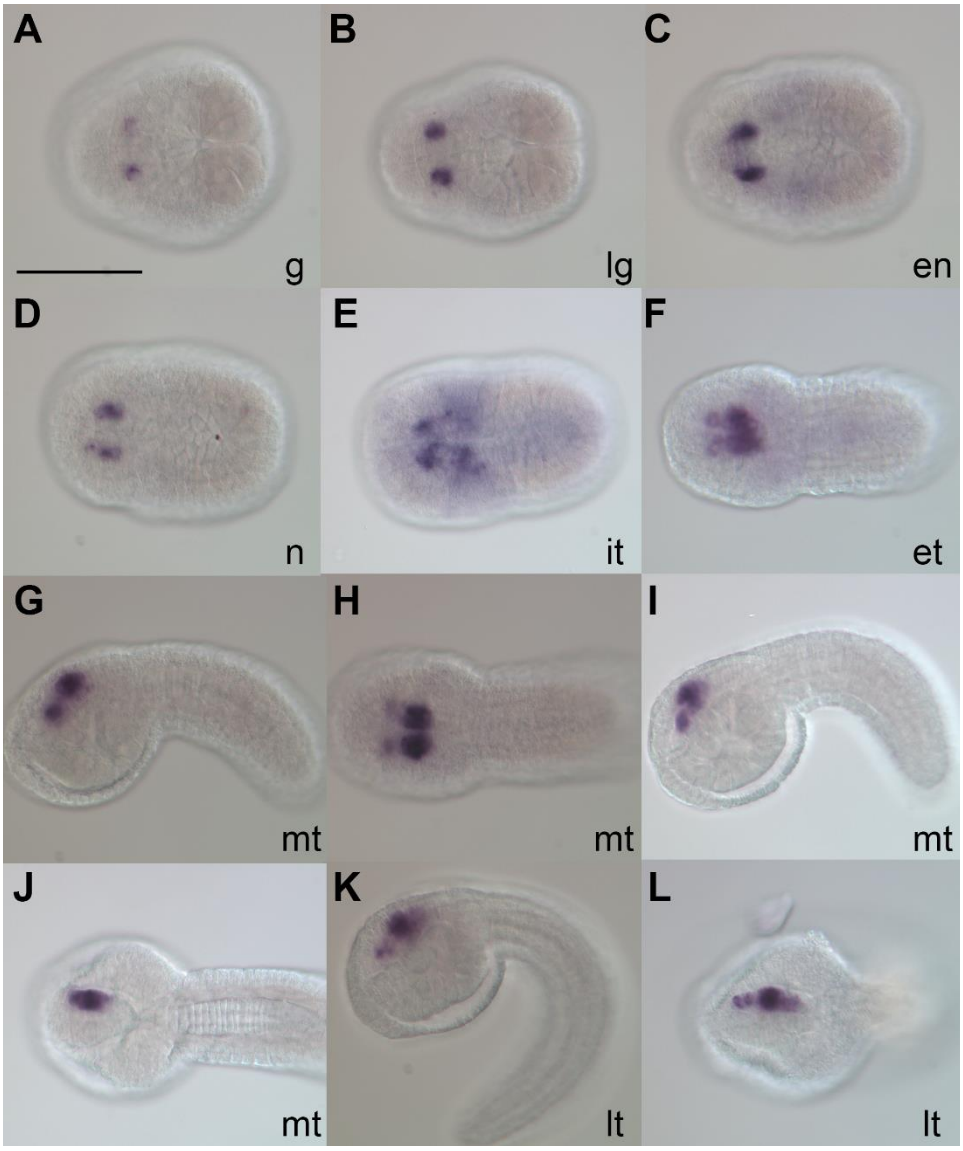
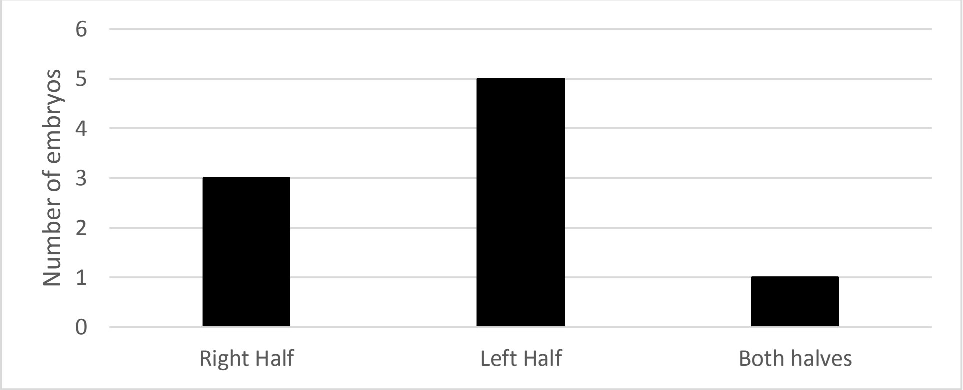


Figure 5.10. Expression of *Ciona intestinalis Gsx*.

(A) Expression begins in the a9.33 cell pair in the mid-late gastrula stage, before later becoming expressed in the two daughter a10.65 and a10.66 cell pairs in the late gastrula to neurula stages (B-C). At the neurula stage, weak expression is also seen in the a10.74 cell pair. Expression expands into the presumptive sensory vesicle in the initial tailbud stage (it), becoming more defined into a strong posterior sensory vesicle domain, and a second smaller domain within the ventral anterior sensory vesicle (F-H) in the early-mid tailbud stage. These two sensory vesicle domains become more defined in the mid tailbud, and expression appears to become restricted to one sagittal half of the sensory vesicle, though this differs in different embryos (I-J). This pattern continues into the late tailbud, becoming more distinct within the posterior sensory vesicle and ventral anterior sensory vesicle and continuing to show restriction to one sagittal half of the sensory vesicle K-L. g, gastrula; en, early neurula; n, neurula; it, initial tailbud; et, early tailbud; mt, mid tailbud; lt, late tailbud. A-F,H,J and L represent dorsal views. G, I and K represent lateral views. Scale bar represents 100μm.

Figure 5.11. Numbers of late tailbud embryos showing *Ci-Gsx* expression within the sensory vesicle.



5.4. Discussion

5.4.1. *C. intestinalis* electroporation provides an amenable system for examining chordate ParaHox regulation.

In this chapter, a regulatory region upstream of the amphioxus ParaHox gene Gsx that drives reporter gene expression in the CNS of *C.intestinalis* has been identified and characterised. *C.intestinalis* is currently a much more tractable system than amphioxus in which to perform reporter transgenics and rapidly characterise gene regulatory regions (Di Gregorio and Levine, 2002; Harafuji et al., 2002). The ability of an amphioxus Gsx regulatory element to drive strong, efficient and reproducible expression in *C.intestinalis* is thus promising as a system for examining ParaHox regulation via this cross-species transgenic approach. The conservation of ParaHox gene expression in the CNS (and gut) allow us to dissect ParaHox regulation in *Ciona*, as although the larvae lack a gut, the CNS of *Ciona* shows clear similarities in gene expression to the wider Chordata, perhaps even the Bilateria (Holland et al., 2013; Wada et al., 1998). In addition, the tunicates also retain many of the signalling pathways involved in Hox and ParaHox regulation, such as RA (Hinman and Degnan, 1998; Kanda et al., 2013; Katsuyama et al., 1995), FGF (Bertrand et al., 2003; Imai et al., 2009; Satou et al., 2002), BMP (Christiaen et al., 2010; Darras and Nishida, 2001), Wnts (Hino et al., 2003; Sasakura et al., 1998) and hedgehog (Hino et al., 2003; Islam et al., 2010) (Imai et al., 2004). By using *C. intestinalis*, it is possible to quickly assess hundreds of embryos at a time, allowing the identification of even weak regulatory elements, e.g. Bf-Gsx-Up1c. With this approach, even subtle differences in expression between different mutations are observed, allowing a much more detailed characterisation of regulatory function.

5.4.2 Bf-Gsx-Up1+2b is a highly conserved regulatory element within amphioxus.

The Bf-Gsx-Up1+2b element has been shown to be the strongest regulatory element with regards to functionality (Figure 5.2). In addition, this region also displays high levels of sequence conservation not only within *B.floridae* individuals (Figure 5.3), but also across the *Branchiostoma* genus, with the Gsx-Up1+2b regulatory region also showing conservation between *B.floridae*, *B.lanceolatum* and *B.belcheri* (Figure 5.4). This observation suggests that the Gsx-Up1+2b region is likely to have similar functional capabilities across the 3 amphioxus species. This region is also functionally dependant on TCF/Lef binding sites (Figures 5.5, 5.7), and TCF/Lef binding sites within the 'minimal enhancer' Up1c are conserved across the three amphioxus species examined (Figure 5.6). Considering these results, and the CNS expression pattern observed in *Ciona* transgenics, it is highly probable that the Bf-Gsx-Up regulatory element is also functional in the *in vivo* regulation of amphioxus *Gsx*, and could be functioning through a TCF/Lef driven mechanism in all three amphioxus species. This would need to be tested via amphioxus transgenics, and further mutagenesis using Gsx-Up1+2b regions taken from *B.lanceolatum* and *B.belcheri*, to confirm this hypothesis.

In addition, the comparative genomic methods used here to compare regulatory sequence across the three amphioxus species could now be applied quickly and efficiently throughout the amphioxus ParaHox cluster. The availability of three amphioxus species would allow the identification of conserved regulatory elements not only within the ParaHox cluster, but for other developmental genes with highly conserved expression. This would open up the use of other phylogenetic techniques that could identify even deeper conservation of regulatory elements, perhaps also within the vertebrates or wider deuterostoma.

The ongoing *Branchiostoma lanceolatum* genome project is making use of techniques such as Chip-seq to identify various methylation states indicative of regulatory domains (Johnson et al., 2007), ATAC-seq to spot promoters (Buenrostro et al., 2001), and 3C and 4C sequencing methods (reviewed in Dekker et al. (2013) to examine chromatin contacts and regulatory regions. The transgenic approach used in this chapter provides an essential supplement to all of the above, providing *in vivo* functional confirmation of bone fide regulatory elements and allowing the detailed analysis of regulatory inputs in regards to regulatory element function.

5.4.3. The amphioxus Gsx-Up regulatory region recapitulates aspects of conserved chordate Gsx expression in the CNS.

The ability of the Bf-Gsx-Up regulatory regions to drive LacZ expression throughout the CNS of *C.intestinalis* is intriguing, as these reporters show LacZ expression in homologous tissues to those expressing *Gsx* in amphioxus, i.e the neural tube and the cerebral vesicle (Osborne et al., 2009). The partial division of function observed in the Bf-Gsx-Up1+2b region into nerve cord and sensory vesicle domains may be linked to the two domains of native amphioxus expression. In this case, the visceral ganglion and nerve cord domain produced by the Bf-Gsx-Up region could correspond to the early domain of amphioxus *Gsx*, expressed at the level of somite 5 (Osborne et al., 2009), whilst the observed sensory vesicle domain might then correspond to the later cerebral vesicle domain of amphioxus *Gsx*. The lack of expression in the most anterior cerebral vesicle region in both cases supports this and may indicate a defined boundary that is present in both the *C.intesinalis* and amphioxus anterior CNS.

One factor that could be playing a role in the exclusion of Gsx expression from this anterior most-domain is the transcription factor Pax6. The *Pax6* and *Gsh2* genes are mutually repressive, and exhibit complementarity domains within the telencephalon of mice (Toresson et al., 2000; Yun et al., 2001), though both *Pax6* (Bel-Vialar et al., 2007; Ericson et al., 1997; Goulding et al., 1993; Walther and Gruss, 1991) and Gsh genes (Hsiehli et al., 1995; Valerius et al., 1995) are expressed within the spinal cord within mouse. Looking to the expression of *Pax6* within *Ciona*, expression is split into three domains; one within the tail nerve cord, a small domain within the visceral ganglion and then another domain within the sensory vesicle (Irvine et al., 2008). Interestingly, it can be observed at the mid-tailbud stage that the sensory vesicle expression is clearly split into two domains, one covering the posterior sensory vesicle and part of the anterior sensory vesicle, and then a second, distinct domain covering the far anterior sensory vesicle. This far anterior domain is in the same region that lacks Bf-Gsx-Up reporter expression, and could be homologous to the *Pax6* expression within the telencephalon (Toresson et al., 2000; Yun et al., 2001), the far-most anterior CNS region, of mouse than exhibits Pax6/Gsx mutual repression. Amphiouxs *Pax6,* is also expressed throughout the cerebral vesicle, but not within the posterior neural tube (Glardon et al., 1998), whilst *AmphiGsx* expression is present within a small number of cells in the central region of the cerebral vesical (Osborne et al., 2009). A small region devoid of *AmphiPax6* expression does exist in the centre of the cerebral vesicle, between the presumptive frontal eye and posterior cerebral vesicle, though it is unknown as to how late *AmphiGsx* expression maps to this *AmphiPax6* expression. Double *in situ* hybridisation of *AmphiGsx* and *AmphiPax6* would have to be carried out in order to examine the regionalisation of these two genes with respect to one another within the cerebral vesicle. The

Pax6/Gsx interaction is particularly relevant in light of the presence of several potential Pax6 binding sites within the Bf-Gsx-Up1+2b regulatory element with similarities to vertebrate Pax6 sites of 70% (5 sites) and 75% (1 site), though these have not yet been analysed functionally. Given the expression of the Bf-Gsx-Up reporters, as well as Pax6 in *Ciona* and amphioxus, and the presence of Pax6 binding sites within the Bf-Gsx-Up1+2b regulatory element it is likely that there is some regulatory input of Pax6 into *AmphiGsx* expression, be it involved in repression or activation of *AmphiGsx*.

Two further transcription factors may be playing a role in the regionalisation of Bf-Gsx-Up reporter expression. The first of these again relates to the lack of expression within the anterior-most CNS. Retinal homeobox protein (Rx) exhibits conserved expression across chordates and is expressed within the anterior-most region of the CNS only (Holland et al., 2013). Expression of *Ciona Rx* is localised to the anterior most part of the sensory vesicle, as well as within the central region of the sensory vesicle during the mid-tailbud stage, between the two 'lobes' of the sensory vesicle (D'Aniello et al., 2006). Both of these regions are devoid of Bf-Gsx-Up reporter expression in *Ciona*, and so this may represent a repressive factor. Indeed a similar pattern can be seen for *Amphi-Rx*, with expression again present in the anterior-most part of the cerebral vesicle (Vopalensky et al., 2012). Two Rx binding sites have been identified within the Bf-Gsx-Up1+2b regulatory element, through the JASPAR vertebrate weight matrix, with the high similarity threshold value of 80%. Given the expression of *AmphiRx* within the anterior cerebral vesicle, or presumptive 'frontal eye' region, which is also the region corresponding to the 'frontal eye' expression of *AmphiPax6*, and the corresponding expression of these genes within *Ciona*, it could be hypothesised that a combination of Rx and Pax6 might work to repress Gsx in the anterior most sensory vesicle of chordates, though further work would need to test this hypothesis. The second factor that may be involved in Gsx regionalisation is Pax2/5/8. In amphioxus, Pax2/5/8 expression begins in a small cluster of cells within the neural tube at the level of somite 5, before expanding to cover the neural tube, but not the anterior sensory vesicle. This expression domain is thought to expand to the midbrain-hindbrain boundary at the anterior (Kozmik et al., 1999). A later domain within the larva is then present in the anterior-most region of the cerebral vesicle, with expression lying dorsally within this region. Within *Ciona*, expression of Pax2/5/8 is present within the visceral ganglion (Mazet et al., 2003; Wada et al., 1998), which is also thought to correspond to the midbrain-hindbrain boundary. Pax2/5/8 has also been shown to be crucial for the formation of the vertebrate midbrain (Schwarz et al., 1999), and this may represent a conserved pattern of chordate Gsx regulation at the midbrain-hindbrain boundary, with the somite 5 region of the amphioxus *AmphiPax2/5/8* expression also abutting this boundary (Holland et al., 2013; Wada et al., 1998). This expression may be particularly relevant to the early domain of *AmphiGsx* expression, which is also present at the level of somite 5 within a few cells of the amphioxus neural tube, and represents a 'hindbrain' domain of expression (Osborne et al., 2009). This may overlap with, or lie adjacent to AmphiPax2/5/8 expression. Two potential Pax 2/5/8 binding sites, one with 80% threshold value and another with a 75% threshold value, are present within the Bf-Gsx-Up1+2b region and again may serve as a direct regulator of *AmphiGsx* given the regional CNS expression of *AmphiPax2/5/8* and similar location as the amphioxus early *Gsx* domain.

The Pax genes in general pose as excellent candidates for involvement in the conserved regulation and regionalisation of Gsx expression. Along with Pax6 and Pax2/5/8, Pax3/7 may also play a role in Gsx regulation, as it is expressed broadly throughout the amphioxus neural tube in the early-mid neurula, and is then restricted to the dorsal cerebral vesicle in the late neurula onwards (Holland et al., 1999). The late domain of AmphiGsx is also expressed within the dorsal cerebral vesicle at similar stages (Osborne et al., 2009). Pax3/7 is also present in the neck region and anterior-most sensory vesicle of *Ciona* (Mazet et al., 2003). Intriguingly, mouse *Pax3* has been shown to directly bind with *Lef-1*, a member of the TCF/Lef family of transcription factors, and Pax3 is able to enhance Lef-1 mediated, and Lef-1/β-catenin mediated transcription via *Lef-1*-bound TCF/Lef binding sites (Christova et al., 2010). A similar study also found that *Pax5* and *Lef-1* interact physically to activate the *RAG-2* promoter in immature B cells (Jin et al., 2002). This is immediately relevant in light of the importance of TCF/Lef binding sites in the expression of the Bf-Gsx-Up reporter (see results sections 5.3.6-5.3.9), and thus a scenario may exist where Pax genes are acting to regulate *AmphiGsx* via interaction with TCF/Lef. Thus, the involvement of Pax and Rx genes in the conserved partitioning of the CNS (D'Aniello et al., 2006; Holland et al., 2013; Mazet et al., 2003; Schwarz et al., 1999) places both of these gene families at an ideal position to inform the regulation of Gsx within different domains of the CNS, and Gsx expression could potentially lie downstream of this initial CNS partitioning.

Whilst there are obvious similarities between the Bf-Gsx-Up reporter expression in *C. intestinalis* and native amphioxus *Gsx* expression, it does remain more expansive within the *C.intestinalis* CNS than would be expected from amphioxus *Gsx* and endogenous *C.intestinalis Gsx* expression, specifically throughout the tail nerve cord. One explanation for this much broader reporter expression could be that the Bf-Gsx-Up regulatory element functions in conjunction with additional repressive elements that would otherwise spatiotemporally restrict expression. Such repressive elements would lie outside of the Bf-Gsx-Up region and so not lend function to the Bf-Gsx-Up reporters. Alternatively, it is possible that the repressive transcription factor system functioning to restrict expression in amphioxus does not exist in *C.intestinalis*, or is too divergent between the two species to provide function to the Bf-Gsx-Up reporter. Divergence of transcription factors and their binding sites is one obvious limitation of performing cross-species transgenesis.

Within the wider chordate phylum, it is possible that there is some conserved regulation of *Gsx*, as there is conservation of expression within the anterior CNS, particularly the 'hindbrain' and mid-forebrain as distinct domains of *Gsx* expression. In amphioxus, the early domain of *Gsx* expression is at the level of somite 5, a region thought to have homology to the vertebrate hindbrain (Holland and Holland, 1996; Holland and Garcia-Fernàndez, 1996). Indeed, in vertebrate Gsx genes, the hindbrain domain is also the first to be expressed, as seen in medaka (Deschet et al., 1998), *Xenopus* (Illes et al., 2009), and mouse (Hsiehli et al., 1995; Valerius et al., 1995). This suggests that there is a conserved regulatory program within the chordates that leads to the expression of this initial hindbrain domain. In addition, the expression of Gsx genes (Illes et al., 2009) within the mid-forebrain of vertebrates (Deschet et al., 1998; Hsiehli et al., 1995), the cerebral vesicle of amphioxus (Osborne et al., 2009), and the sensory vesicle of *C. intestinalis* (Figure 5.10) (Hudson and Lemaire, 2001) again hints at conserved regulation of Gsx within the chordates. This, in conjunction with the ability of the Bf-Gsx-Up reporter to drive expression throughout the CNS of *C.intestinalis*, suggests that a conserved regulatory pathway may be driving chordate Gsx expression.

5.4.4. A role for TCF/Lef in chordate Gsx regulation

The Bf-Gsx-Up regulatory region shows a clear and strong response to the mutation of TCF/Lef binding sites, requiring them to be intact in order to drive CNS expression in *C. intestinalis*. In concordance with this, native *C.intestinalis TCF/Lef* expression is consistent with a role for TCF/Lef in the direct regulation of Bf-Gsx-Up, as expression is present in the neural plate, then later in both the sensory vesicle and throughout the tail nerve cord and visceral ganglion, albeit weakly (figure 5.9), which are all domains that the Bf-Gsx-Up regulatory regions drive expression in. In addition to these CNS domains, *C. intestinalis TCF/Lef* is also expressed in the head/neck mesenchyme, where the background pCES expression remains high even in the presence of a strong enhancer such as Bf-Gsx-Up1a-2b, suggesting that regulation of the reporters by TCF/Lef could be allowing this head/neck mesenchyme domain to persist when a weak regulatory element is present. Though some expression data for *Ci-TCF/Lef* is available on the ANISEED database (Tassy et al., 2010) (Imai et al., 2004), here, a more detailed characterisation of *Ci-TCF/Lef* expression, particularly in the stages in which the Bf-Gsx-Up reporter constructs are expressed, is provided.

In the case of amphioxus, there is strong expression of *TCF/Lef* within the cerebral vesicle, correlating with stages where *Gsx* is expressed in the same tissue (Lin et al., 2006). This raises the distinct possibility that the Bf-Gsx-Up region responds to the same signal in its native environment. Though no strong *TCF/Lef* domain has yet been observed in the position of the early amphioxus *Gsx* domain (Lin et al., 2006; Osborne et al., 2009), it is possible that weak *TCF/Lef* expression present in the nerve cord of amphioxus may be sufficient to allow Gsx expression, with other transcription factors acting to restrict the expression domain.

In the vertebrates, the expression of TCF/Lef family members are similar to that of *C. intestinalis TCF/Lef*, with strong expression of TCF/Lef family members present throughout the neural tube (Schmidt et al., 2004). The presence of TCF/Lef expression in the CNS of all three chordate sub-phyla shows a conservation of expression within these tissues, which along with the association of *TCF/Lef* with amphioxus *Gsx* regulation in the Bf-Gsx-Up reporters, implies an ancestral role for TCF/Lef in the direct regulation of chordate Gsx. In order to test this hypothesis further, future work would be required to establish whether there is similar direct regulation of *C.intestinalis* and vertebrate Gsx genes by TCF/Lef. Currently, whilst it is known that the mutation of TCF/Lef binding sites within the Bf-Gsx-Up regulatory elements causes an abolition of CNS expression, it has not yet been confirmed that TCF/Lef is actually binding the sequence, though it is likely. As such, the direct binding of TCF/Lef to the Bf-Gsx-Up regulatory elements could be examined by electrophoretic mobility shift assay (EMSA). This approach would then confirm if a DNA/protein complex was forming between TCF/Lef and the Bf-Gsx-Up regulatory DNA. As a complimentary approach, TCF/Lef morpholinos, or a TCF/Lef antibody, could be used to knockdown *Ci-TCF/Lef*, producing a background in which it could be examined whether expression of Bf-Gsx-Up reporters is also knocked down. In addition, a heat-shock promoter coupled to the *Ci-TCF/Lef* or *AmphiTCF/Lef* protein could be used, with a Hsp70-TCF/Lef construct being electroporated alongside the Bf-Gsx-Up1+2b construct. In this case, heat induced TCF/Lef overexpression would also cause ectopic expression of the Bf-Gsx-Up1+2b construct in the case of direct TCF/Lef binding. The mutant Bf-Gsx-Up1+2bΔ1Δ2Δ3 construct could then be used to determine if ectopic expression is abolished in the presence of mutated TCF/Lef sites.

5.4.5. Unravelling complex cis-regulatory function and multiple levels of regulation.

Though the Up1c minimal enhancer region is sufficient to drive CNS expression, if only in the nerve cord, the addition of further sequence, up to Bf-Gsx-Up1+2b, vastly improves both general expression efficiency as well as the incidence of the full CNS ‘nerve cord + sensory vesicle’ expression pattern. In addition, intact TCF/Lef binding sites are crucial to the function of this regulatory element. Within the Bf-Gsx-Up1+2b regulatory element, a differential response to the mutation of TCF/Lef sites across the regulatory region. Sites within the 3’ minimal Up1c region contribute with a bias toward nerve cord expression whereas the third site, within the 5’ Up2b region, contributes with a bias toward sensory vesicle expression (figure 5.7 D). The presence of a single intact TCF/Lef site still allows for expression of the whole CNS expression pattern even with this bias present (figure 5.7 C), however, suggesting that TCF/Lef is required to activate expression, but not necessarily specify more restricted expression domains. TCF/Lef can thus be seen as permissive for Gsx activation, with further spatial restriction coming from other factors that are presumably bound to the Bf-Gsx-Up regulatory region.

The ability of the Bf-Gsx-Up regulatory region to not only drive expression in the presence of intact TCF/Lef sites but also actively repress expression, if these sites do not remain intact, presents another layer of control within this regulatory region. In addition to the partial division of regulatory function into different domains, once TCF/Lef binding is abolished a latent repressive function is unmasked and currently unknown repressive factors are able to silence gene expression, preventing any background ectopic pCES transcription in the absence of TCF/Lef (figure 5.8). This repressive state may even be the ‘default’ for Bf-Gsx-Up, which would then switch from repressor to enhancer in the presence of TCF/Lef binding (figure 5.12).

The observation of two different CNS domains, nerve cord and sensory vesicle, responding in a partially independent manner suggests that further transcription factors are involved in the specification of these two domains, and further work could aim to identify these factors. Proteomics of Isolated Chromatin segments (PiCh) (or variations such as ChAP-MS, GENECAPP, iChIP, Hy-CCAPP) is a technique that has been used in several contexts to characterise the proteins that bind to a particular DNA sequence, from several biological samples including human cell lines, yeast and chicken cells (Déjardin and Kingston, 2009; Fujita and Fujii, 2014; Kennedy-Darling et al., 2014; Smith et al., 2011; Wu et al., 2011a). This could be adapted to use on the Bf-Gsx-Up reporters, using the large numbers of *Ciona* embryos to isolate and pull down the reporter along with any bound proteins, which could then be identified via Mass spectrometry.

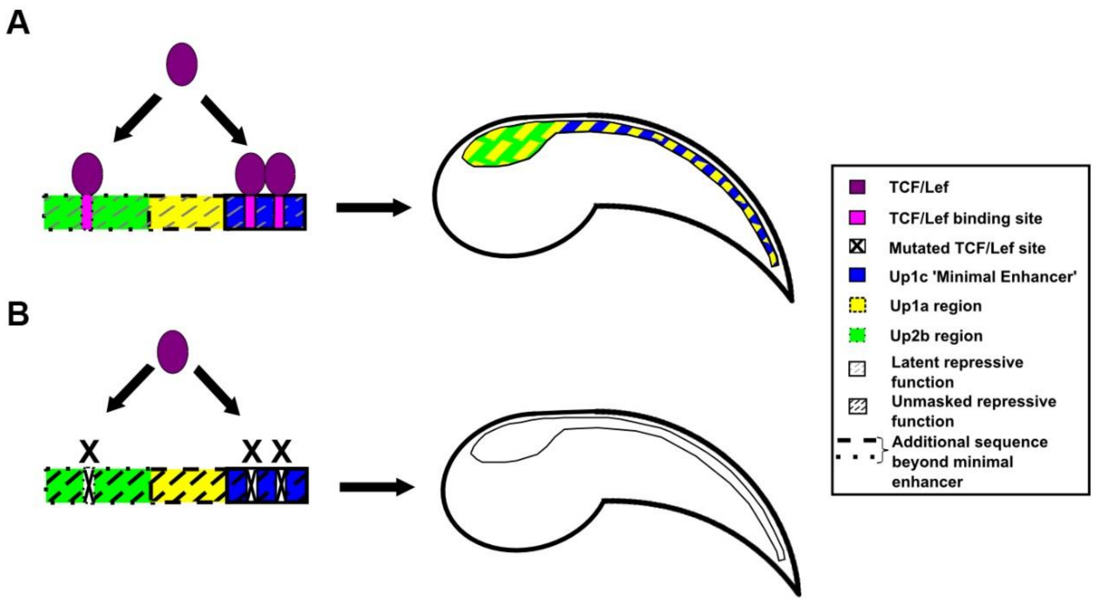


Figure 5.12. Model for the mode of action of the Bf-Gsx-Up1+2b regulatory element.

(A) In the presence of TCF/Lef binding, CNS expression is activated. Additional sequence beyond the Up1c 'minimal enhancer' both increases the efficiency of CNS expression and reveals a partial division of function into the Up1c 'minimal enhancer' nerve cord domain and the Up2b region sensory vesicle domain. The intermediate Up1a region contributes partially to both expression domains. (B) In the absence of TCF/Lef binding, a latent repressive function is unmasked, preventing CNS expression.

5.4.7. *Ci-Gsx* down regulation indicates left-right differences in the 'brain'.

Though *Ci-Gsx in-situ* hybrisation has been carried out before, It requires searching through several different studies (Hudson and Lemaire, 2001; Imai et al., 2004) in order to obtain an overview of *Ci-Gsx* across different stages of embryonic development. Even then, not all stages are covered within that data. Thus a comprehensive time course of *Ci-Gsx* embryonic expression was carried out, in particular covering the stages that could be compared to the Bf-Gsx-Up reporter constructs. Though no new data was gained from the early stages, the late tailbud stage expression proved to be intriguing, with expression limited to one sagittal half of the sensory vesicle, though not the identical half in all embryos. Though there was no time to examine the late tailbud *Ci-Gsx* expression further, it remains an interesting observation that should be examined in future, perhaps representing anti-symmetric expression. Alternatively, *Gsx* expression could be becoming down-regulated in a stochastic fashion at this late tailbud stage, and without any functional consequences.

In this case, the observations made would just represent expression partway through this down-regulation process. Regardless of this it is clear that the left and right halves of the sensory vesicle are not absolutely equivalent in molecular terms at this stage of development. One clue as whether this may be a functional molecular and regulatory event brings us back to the examination of the Ci-Rx transcription factor. It was earlier noted that Ci-Rx may represent a factor involved in the regulation of Gsx, specifically in the exclusion of Gsx expression, and Bf-Gsx-Up reporter expression, from the anterior-most part of the anterior sensory vesicle. The expression of Ci-Rx is present in this anterior most region, and throughout early development is present in both 'halves' of the anterior tip of the sensory vesicle in a symmetric fashion (D'Aniello et al., 2006). Intriguingly, this transcription factor also becomes restricted to one half of the sensory vesicle at the late tailbud stage, though is restricted to the right side of the sensory vesicle only. It is not noted how many embryos were examined here, though a subsequent study does go on to examine this restriction process in more detail. It was determined that the downregulation of *Ci-Rx* within the left side of the sensory vesicle was due to the expression of, and regulation by *Ci-Nodal*, a signalling factor heavily involved with left-right asymmetry across the chordates (Boorman and Shimeld, 2002). Though *Ci-Gsx* does not always show expression within the right as with *Ci-Rx*, it is possible that the antisymmetrical expression observed does represent some interaction with the Nodal/Lefty left-right asymmetry pathways, perhaps with inputs from both Nodal and Lefty (reviewed in Shen (2007)) resulting in either the left, or right sided expression becoming more dominant in different embryos. As stated, further examination of Ci-Gsx late expression is required before any true hypothesis can be made.

5.4.8. Building a picture of ancestral ParaHox regulation.

The results here show that TCF/Lef is likely to be directly regulating amphioxus *Gsx*, and though further study is needed to confirm TCF/Lef binding, this is the first evidence of such an interaction with Gsx. If Gsx is directly regulated by TCF/Lef, this now provides examples of TCF/Lef regulation across all three ParaHox genes within the chordates (See general discussion). These data, in addition to that described in this thesis, suggests that all three ParaHox genes may have been ancestrally regulated by TCF/Lef. Indeed TCF/Lef expression patterns in vertebrates, as well as amphioxus, show strong expression in both the CNS and in the gut (Gregorieff et al., 2004; Lee et al., 1999; Lin et al., 2006; Schmidt et al., 2004). The presence of amphioxus *TCF/Lef* in this posterior gut region (Lin et al., 2006) makes experiments to determine if amphioxus *Xlox* and *Cdx* are also regulated by this transcription factor an interesting prospect.

Whilst evidence for direct TCF/Lef involvement in the regulation of ParaHox genes is currently limited to the chordates, data suggests that Wnt signalling may also play a wider role in the regulation of ParaHox genes within the Bilateria. This holds particular relevance as TCF/Lef family members have been shown to act in a complex with β-catenin as the downstream transcriptional activator of canonical Wnt signalling during embryogenesis (Brunner et al., 1997; Korinek et al., 1998). This pathway has also been shown to be crucial to the proper development of the gut and neural tube (Faro et al., 2009; Galceran et al., 1999; Ikeya et al., 1997), making it an excellent candidate pathway for ParaHox gene regulation (See general discussion (Section 6.2.) for discussion of TCF/Lef and Wnt signalling regulation of Xlox and Cdx genes).

Little work has been done on the regulation of Gsx genes in general, but a study on gene expression within the brain of *Drosophila melanogaster* suggests that Wnt signalling may indeed also be playing a role in the regulation of Gsx, with *Wingless* (*Wg*) active within *Intermediate neuroblasts defective* (*Ind*) (Gsx) positive brain neuroblasts. There is also evidence that Gsx expression within the telencephalon of *Platynereis dumerilii* is down regulated within Azakenpaullone (which upregulates Wnt signalling (Schneider and Bowerman, 2007)) treated embryos (Tomer et al., 2010), though other factors such as Rx, which is excluded from Gsx positive cells in vertebrates, are not so in *Platynereis*. It is possible that Wnt signalling could be functioning upstream of the Bf-Gsx-Up regulatory region, acting via the TCF/Lef mediated canonical Wnt pathway, and indeed Wnt ligands are expressed in the neural tube of both *C.intestinalis* (Imai et al., 2004) and amphioxus (Schubert et al., 2000a; Schubert et al., 2001). In particular Wnt 4 and Wnt7b are expressed in the sensory vesicle at the same time as *Gsx* in amphioxus (Schubert et al., 2000a). Wnt7 is also expressed throughout the nerve cord in a manner similar to the Bf-Gsx-Up reporter constructs in *Ciona* (Imai et al., 2004). From the ANISEED database, it also seems that *Wnt2* (*Orphan Wntb*) is expressed throughout the CNS of *Ciona* (Imai et al., 2004). If these are cases of direct regulation of Gsx by Wnt signalling, it is possible that a Wnt-Gsx regulatory pathway was present at the base of the Bilateria. Wnt signalling, via TCF/Lef, may even function as a pan-cluster regulatory mechanism within the ParaHox cluster, acting upon all three ParaHox genes, and further regulatory studies across distant taxa and multiple ParaHox genes would help to elucidate this Wnt-TCF/Lef ParaHox regulatory hypothesis. Considering the regulation of other ParaHox genes by Wnts, an interesting avenue to explore would be to examine if Wnt is coordinating TCF binding in the Bf-Gsx-Up regulatory region. A heat shock promoter coupled to *Ci-Wnt7* may be one method of identifying if Wnts are regulating the Bf-Gsx-Up regulatory elements. Knockdowns of Wnt genes would be difficult to carry out as they tend to have widespread effects, and it would be difficult to tell whether the effect was specific upon the reporter genes or a consequence of widespread tissue specification issues. The same would be true of

Lithium treatments to upregulate Wnt signalling. In order to test whether TCF/Lef, or even Wnt, was regulating all three amphioxus ParaHox genes, a construct containing all three amphioxus ParaHox genes, such as the amphioxus ParaHox PACs -33B4 and 36D2 (Ferrier et al., 2005), could be introduced into a vertebrate, or eventually amphioxus, cell line. This could then be treated with TCF/Lef and Wnt proteins or inhibitors without the worry of secondary axial patterning effects, and the expression of the amphioxus ParaHox genes examined in response to these treatments. One cell line that may suit this purpose is a neural stem-cell line, due to the expression of all three amphioxus ParaHox genes in neural tissues (Osborne et al., 2009), or alternatively pancreatic beta-cell lines, which have been previously shown to express all three ParaHox genes (Rosanas-Urgell et al., 2008).

Looking even further back into the evolution of animals, the involvement of Wnt signalling in anterior-posterior patterning also appears to play a role within the Cnidaria (Hobmayer et al., 2000; Kusserow et al., 2005) and even sponges (Adamska et al., 2007; Adamska et al., 2010; Adell et al., 2003; Adell et al., 2007; Lapebie et al., 2009). Given that the ParaHox genes are now thought to have originated in the last common ancestor of animals (Fortunato et al., 2014; Garstang and Ferrier, 2013; Ramos et al., 2012) and the Wnt system is similarly ancient, then there may well have been a direct ParaHox-TCF/Lef link from the earliest stages of animal evolution and development.

Chapter 6. General Discussion

6.1. The Amphioxus ParaHox Cluster: A key model for developmental gene cluster regulation and evolution.

The phylogenetic placement of amphioxus as the basal chordate lineage, and the presence of developmental stages that bridge echinoderm-like (radial cleavage and enterocoely) and vertebrate-like characteristics mark amphioxus at a key evolutionary position to study the evolution of both chordates and the wider Deuterostomia. This transitional position between invertebrate deuterostome and vertebrate development and morphology makes understanding the regulation of those genes controlling amphioxus development particularly interesting, providing both a deeper evolutionary view into chordate and deuterostome developmental regulatory mechanisms, as well as providing a much simpler genome (pre-2R duplication) in which to study regulatory mutations that cause vertebrate developmental abnormalities and disease phenotypes (Putnam et al., 2008).

Until recently, only the chordates had been shown to have intact, collinear ParaHox clusters (Brooke et al., 1998; Ferrier et al., 2005; Osborne et al., 2009) though this has now been shown to extend to the base of the deuterostomes, with members of the Ambulacraria also possessing intact, collinear ParaHox clusters (Annunziata et al., 2013; Ikuta et al., 2013). The amphioxus ParaHox cluster holds particular value as a model gene cluster as the large number of regulatory studies on the vertebrate Hox cluster can be applied to the ParaHox cluster, but it also provides a much simpler cluster in which to identify regulatory mechanisms, which can then be applied back to the Hox and ParaHox clusters of vertebrates as well as the wider Deuterostomia. In addition to the collinearity observed in these ParaHox clusters, the presence of an Xlox-Cdx midgut-hindgut boundary in the sea urchin (Annunziata and Arnone, 2014; Cole et al., 2009), acorn worm (Ikuta et al., 2013) and amphioxus (Osborne et al., 2009) is highly suggestive of conserved regulatory mechanisms across the Deuterostomia. Work within both vertebrates (Bayha et al., 2009; Chen et al., 2004; Kinkel et al., 2008; Kumar et al., 2003; Stafford and Prince, 2002) and amphioxus (Osborne et al., 2009) even suggests that RA signalling may play a role in the regulation of the midgut-hindgut boundary. Again, this highlights the important role amphioxus can play in bridging the gap between the detailed but vertebrate specific studies and answering broader evolutionary questions regarding ParaHox gene regulation.

With the release of the *B.floridae* (Putnam et al., 2008) and *B.belcheri* genomes (Huang et al., 2012; Huang et al., 2014), and preliminary access to the B.lanceolatum genome (unpublished), the genomic resources available for amphioxus now allow comparative genomics approaches previously only possible between vertebrates. This approach has already been applied between the amphioxus and vertebrate Hox clusters, identifying several conserved regulatory elements (Amemiya et al., 2008; Pascual-Anaya et al., 2008). Still, the overall sequence similarity between vertebrate and amphioxus non-coding sequence is very poor, and no resolution can be seen between amphioxus and vertebrate ParaHox clusters (section 3.3.5), which may have been compounded by substantial gene loss across three of the vertebrate ParaHox loci (Ferrier et al., 2005). However, with access to three amphioxus genomes, comparisons can now be carried out across amphioxus species, highlighting regions of potential regulatory interest that were not observed with vertebrate-amphioxus comparisons alone.

This has been applied to the amphioxus ParaHox cluster in chapter 3, and has revealed not only conserved genomic architecture, but also a remarkably well conserved regulatory landscape across the amphioxus ParaHox clusters. The use of VISTA analysis has resulted in a comprehensive map of prospective regulatory elements, in the form of conserved non-coding regions, across the amphioxus ParaHox cluster (section 3.3.4). The genomic comparison can then be fine-tuned for specific regions of high conservation (section 5.3.3) to provide a precise candidate region, to provide a focus for functional analyses. These methods have been successfully applied to the identification and subsequent functional analysis of the Bf-Gsx-Up1+2b regulatory element (chapter 5), highlighting the potential for this approach in the identification and subsequent analysis of regulatory elements more generally.

The ever-growing abundance of amphioxus transcriptomic data can also be incorporated into such comparative regulatory analysis. This has been particularly useful in the identification of ParaHox UTRs, in particular a large Xlox 3' UTR (section 3.3.3) and a 5' SCP1 UTR that has likely evolved to make use of nearby CHIC regulatory elements (chapter 4). The identification of UTR sequences provides a further avenue for regulatory studies, as UTRs often have significant regulatory functions. One potential avenue for further investigation of such regions, particularly within long 3' UTRs, would be to look for the presence of micro-RNA (miRNA) target sites (Barrett et al., 2012). These sites have been shown to be present in both 5' and 3' UTRs (Lytle et al., 2007), and many ubiquitously expressed housekeeping genes are thought to have evolved shortened 3'UTRs specifically to avoid miRNA targeting (Stark et al., 2005). miRNAs are known to have diverse functions in animal development, gene regulation, and disease, and act to restrict gene expression by targeting mRNA transcripts for destruction, or block translation (Alvarez-Garcia and Miska, 2005; Ambros, 2004). This may hold particular relevance for the ParaHox cluster, as the Hox cluster is well known for both its conserved miRNAs, miR-10 and miR-196, (Hui et al., 2009a; McGlinn et al., 2009;

Woltering and Durston, 2008; Yekta et al., 2004), and their involvement in Hox gene regulation and posterior prevalence (Yekta et al., 2008).

Taken together, the growing abundance of amphioxus genomic and transcriptomic data is providing ever greater insights into gene regulation in amphioxus and the ancestral chordate, and amphioxus is clearly a highly valuable model organism for the study of gene regulation in evo-devo.

6.2. ParaHox gene regulation by TCF/Lef

The studies carried out in chapter 5 of this thesis provide the first evidence for the regulation of the ParaHox gene *Gsx* by TCF/Lef, and also of ParaHox gene regulation by TCF/Lef within the cephalochordate amphioxus. The mutation of discrete TCF/Lef binding sites within the *Bf-Gsx-Up cis*-regulatory element and subsequent loss of LacZ reporter gene expression in response to these mutations suggests the direct binding of *TCF/Lef* activates *AmphiGsx* expression. The expression of the *B.floridae Gsx-Up* regulatory element within the CNS of *C.intestinalis* also suggests conserved signalling mechanisms present in the chordate ancestor that allows for the neural expression of Gsx, and amphioxus (Osborne et al., 2009), Ciona (Hudson and Lemaire, 2001) (also see section 5.3.11) and the vertebrates (Cheesman and Eisen, 2004; Deschet et al., 1998; Hsiehli et al., 1995; Illes et al., 2009; Li et al., 1996; Toresson and Campbell, 2001; Valerius et al., 1995) all display Gsx expression within multiple regions of the CNS. The appearance of an early 'hindbrain' domain, followed by later mid-forebrain expression, within both amphioxus and the vertebrates, is also suggestive of a conserved chordate regulatory mechanism controlling this expression pattern. The presence of TCF/Lef throughout the neural tube of *Ciona* (Section 5.3.10 and (Imai et al., 2004)) and vertebrates (Korinek et al., 1998; Molenaar et al., 1998; Roël et al., 2003; Schmidt et al., 2004), and throughout the neural plate and later the cerebral vesicle amphioxus (Lin et al., 2006), is also suggestive that similar signalling mechanisms controlling Gsx expression may be present across the chordate phylum. The *Bf-Gsx-Up1+2b* reporter is currently being tested within transgenic zebrafish embryos to examine whether the vertebrates have also maintained the signalling pathways necessary to drive CNS expression of this construct. If so, this would provide an even stronger basis for the analysis of vertebrate Gsx regulatory elements and TCF/Lef regulation of vertebrate Gsx.

In addition to the work described in chapter 5, conserved TCF/Lef binding sites have also been identified across the ParaHox clusters of three species of amphioxus in chapter 3. This both supports previous studies and provides novel data regarding the abundance of TCF/Lef binding sites within the amphioxus ParaHox cluster. Within a study looking at the effects of altered Wnt and RA signalling upon amphioxus axial patterning, Onai et al. (2009) briefly mention that they had identified three TCF/Lef binding sites upstream of *AmphiCdx*, though this data is not shown and no functional analysis was carried out. In section 3.3.8, the presence of TCF/Lef binding sites upstream of *AmphiCdx* is confirmed and expanded upon, with a total of five conserved TCF/Lef binding sites identified. Though no direct regulation of amphioxus *Cdx* by TCF/Lef has yet been observed, several studies have observed direct regulation of vertebrate Cdx genes by TCF/Lef mediated through TCF/Lef binding sites (Beland et al., 2004; Gaunt et al., 2003; Gaunt and Paul, 2014; Pilon et al., 2006; Pilon et al., 2007). The presence of several conserved amphioxus TCF/Lef binding sites upstream of *Cdx*, as well as the abundance of vertebrate literature describing direct TCF/Lef activation of Cdx, as well as by TCF/Lef mediated Wnt signalling, make amphioxus *Cdx* an excellent candidate for further experimental studies.

Conserved TCF/Lef sites can be found upstream of all three ParaHox genes, as well as in clustered 'islands'. TCF/Lef is a regulator of both Xlox (Lee et al., 1999) and Cdx (Gaunt et al., 2003; Gregorieff et al., 2004; Pilon et al., 2006) within the vertebrates, and the identification of conserved TCF/Lef binding sites in section 3.3.8 suggests that this may also extend to amphioxus. The conserved position of the identified TCF/Lef sites across all three *Branchiostoma* species examined also greatly increases the possibility that there is functional relevance to these binding sites, particularly as many lie within very highly conserved 'peak' regions of amphioxus ParaHox non-coding DNA, as well as in upstream regions that typically hold active regulatory elements. The functionality of most of these sites has not yet been experimentally tested, though TCF/Lef binding site 5 (figure 3.10) was experimentally tested within chapter 5, and corresponds to the TCF/Lef site 1 within this more detailed functional study. Mutagenesis of this conserved site did have a significant effect upon *LacZ* reporter activity, and so the method used within section 3.3.8 to identify conserved sites is able to reveal functionally relevant binding sites. This suggests that all three ParaHox genes, not only *Gsx*, may be under direct regulation by TCF/Lef within amphioxus, and the studies detailed here show that this type of preliminary comparative genomic data can be successfully utilised to inform more in depth functional analysis.

The implied direct TCF/Lef regulation of amphioxus *Gsx* and other ParaHox genes raises another question: are ParaHox genes being regulated by Wnt signalling? This is particularly relevant as TCF/Lef is known to be a downstream nuclear effector of the canonical Wnt signalling pathway, where nuclear β-catenin interacts with TCF/Lef (Behrens et al., 1996; Huber et al., 1996), forming a protein dimer that then binds DNA via TCF/Lef binding sites (CTTTG A/T A/T) and activates transcription of target genes (Faro et al., 2009; Galceran et al., 1999; Huber et al., 1996; Korinek et al., 1998). TCF/Lef binding site multimers are even used in reporter constructs such as TOPFLASH as an established assay for the presence of Wnt signalling (Barolo, 2006), though these reporters may activate under other non-Wnt, TCF/Lef driven signalling events. This pathway has been shown to be crucial to the proper development of the gut and neural tube (Faro et al., 2009; Galceran et al., 1999; Ikeya et al., 1997), and has been well studied as a regulator of vertebrate Cdx genes in these tissues (Deschamps and van de Ven, 2012; Ikeya and Takada, 2001; Lickert et al., 2000; Lickert and Kemler, 2002; Pilon et al., 2006; Shimizu et al., 2005). This regulation of vertebrate Cdx genes is also direct regulation, and work on mouse *Cdx1* promoters has shown that TCF/Lef- β-catenin complex binding is required to activate *Cdx1* expression and embryos suffer abrogated *Cdx1* expression in the small intestine in response to *TCF4* null mutants (Lickert et al., 2000). In addition, Cdx genes have been shown to act as the mediator for the activation of Hox genes by Wnt signalling (Gaunt et al., 2003; Gaunt and Paul, 2014; Pilon et al., 2006; Pilon et al., 2007). It has not yet been determined within other phyla displaying this Wnt>Cdx posterior patterning whether direct regulation is present. It is likely that direct regulation of *Cdx* by Wnt is also occurring within amphioxus, as Lithium treatment, Li$_+$ treatment of embryos (which upregulates Wnt/β-catenin signalling via inhibition of GSK3β), causes an ectopic anterior *AmphiCdx* domain as well as reduction of the CNS domain and expansion in the hindgut domain of *AmphiCdx* (Onai et al., 2009).

Though lithium treatment is highly toxic, the presence of several upstream TCF/Lef binding sites is highly suggestive of Wnt>*TCF/Lef* direct regulation of *Cdx* within amphioxus as well. The presence of Wnt signalling throughout the neural tube in amphioxus (*Wnt3, Wnt4, Wnt5, Wnt6, Wnt7b*)(Schubert et al., 2000a; Schubert et al., 2001) is consistent with a role for the canonical Wnt signalling pathway in the regulation of *Cdx*. Indeed the conserved expression of Wnt signalling in the posterior of amphioxus (*Wnt3, 4, 5, 6*) indicates that the posterior Wnt>TCF/Lef>Cdx pathway seen in vertebrates could also be present in amphioxus. Thus, it is possible that Wnt signalling may also be involved in the direct regulation of amphioxus *Cdx* via *TCF/Lef*.

Whilst evidence for direct TCF/Lef involvement in the regulation of ParaHox genes is currently limited to the chordates, data suggests that Wnt signalling may play a wider role in the regulation of Cdx within the Bilateria. Indeed Cdx has also been shown to respond to Wnt signalling in the posterior growth zones of both the beetle *Tribolium castaneum* (Oberhofer et al., 2014) and the spider *Parasteatoda tepidariorum* (McGregor et al., 2008; McGregor et al., 2009), suggesting that the regulation of Cdx by Wnt signalling may have been present at the base of the Bilateria, and is in concordance with the prevalence of Wnt signalling as a posterior axial patterning signal across many phyla (reviewed in Martin and Kimelman (2009)).

The role of Wnt signalling and TCF/Lef upon Xlox is more complex than with Cdx, as current vertebrate studies are largely restricted to effects on pancreas growth and adult islet cells. Studies upon the roles of Wnt signalling in this region are also conflicting, with both repression by Wnt/β-catenin signalling (McLin et al., 2007; Zhang et al., 2013) and activation by Wnt/β-catenin signalling (Papadopoulou and Edlund, 2005) described as being required for pancreatic growth. The TCF/Lef family proteins however, do appear to be required for these for the development of the pancreas and expression of *IPF1/Pdx1* (Xlox). In addition, *Xenopus Xlhbox8* (Xlox) has been shown to be induced by mouse *TCF4* and human *Lef-1* in animal cap explants, where it normally plays a role in the specification of the duodenum and pancreatic tissues (Lee et al., 1999). The presence and expression of several members of the Canonical Wnt signalling pathway, including *Lef1*, *TCF4* and *frizzled1* and *8* (Wnt co-receptors), within the developing chick duodenum and pyloric sphincter (Theodosiou and Tabin, 2003) suggests that Wnt signalling may indeed be playing a role in the development of these tissues at the midgut/hindgut boundary. As is evident, the role of Wnt signalling within the pancreas and duodenum of vertebrates is complex, and studies of Xlox regulation are largely biased towards its roles in the vertebrate pancreas. It could be that the duodenum and pyloric sphincter (i.e the stomach/intestine boundary) expression of Xlox is more representative of the ancestral expression, as the pancreas is a vertebrate innovation and has likely co-opted many signalling pathways and transcription factors for its development and function. The origin of Xlox regulation by Wnt-TCF/Lef is much harder to determine, as Xlox appears to have been lost from the ecdysozoans sampled so far (with the possible exception of *Strigamia maritima* (Chipman et al., 2014)), but studies of the Xlox/Cdx boundary in the gut of deuterostomes may yield insights. One study has identified a putative midgut-hindgut GRN within the sea urchin, and the interaction causing the restriction of *Xlox* expression by *Cdx* appears to be mediated via repression of *Xlox* through *Wnt10*, though there are unknown mediators thought to lie both upstream and downstream of *Wnt10* (Annunziata and Arnone, 2014). Studies within amphioxus may help elucidate some of this confusion and help resolve the ancestral gut-patterning mechanisms underpinning the midgut-hindgut boundary, as well as Xlox regulation by TCF/Lef, due to the much simpler amphioxus gut and lack of TCF/Lef and Wnt signalling paralogues to confuse matters. In particular, functional analysis of the TCF/Lef binding site 12-13 and 14 conserved regions, both surrounding amphioxus *Xlox*, would make excellent starting points for initial examination into the regulation of *Xlox* by *TCF/Lef*. It is also possible that Wnt may interact with RA to determine this boundary, as altered RA signalling has been shown to alter the position of this Xlox/Cdx boundary in amphioxus (Osborne et al., 2009), and Cdx1 in vertebrates is known to respond directly to combinatorial inputs from both the RA and Wnt signalling pathways

(Pilon et al., 2007). This could help inform regulatory studies, as a similar scenario with interacting RARE and TCF/Lef response elements could be present in amphioxus *Xlox* or *Cdx* regulatory regions.

Finally, it cannot be ruled out that Wnt signalling is not playing any role in the activation of amphioxus *Gsx*. TCF/Lef is known to interact with proteins other than β-catenin to regulate gene transcription. The most well studied of these is Groucho/TLE, which acts as a corepressor, forming a TCF/Lef:Groucho dimer that represses genes in the absence of Wnt signalling (Brantjes et al., 2001; Daniels and Weis, 2005). TCF/Lef:Groucho corepression is unlikely to be the case in the *Bf-Gsx-Up* reporter, as this effect is still mediated through TCF/Lef binding sites, and a repression activity is uncovered upon mutation of these sites within the *Bf-Gsx-Up* reporter. However, there may be other non-Wnt dependent partners involved, and TCF/Lef has been shown to interact with other transcription factors such as *Cdx1*, (Beland et al., 2004) or *ALY*, which forms a complex with *Lef-1* and *AML-1*, acting as a context-dependant activation complex (Bruhn et al., 1997). As such, it is clear that further work will be required to investigate the role of both TCF/Lef and Wnt signalling in the regulation of the ParaHox genes of amphioxus, particularly in the identification of TCF/Lef binding partners and analysis of TCF/Lef binding site-containing regulatory regions surrounding *Xlox* and *Cdx*.

6.3. Maintenance of the chordate ParaHox Cluster

6.3.1. Regulation is likely Key

Previous work has suggested that the ParaHox cluster is under less constraint to remain as a gene cluster than its sister the Hox cluster. Whilst four Hox clusters have been maintained within the vertebrates, and though there are four ParaHox loci, only a single ParaHox cluster is maintained (Ferrier et al., 2005), and even this has disintegrated within the teleost lineage during another round of WGD (Mulley et al., 2006; Siegel et al., 2007). It has been suggested that the degeneration of the ParaHox cluster within teleosts may be due to the acquisition of redundancy of both genes and regulatory elements, increasing the probability of cluster degeneration due to gene loss. Previously interdigitated regulatory elements would now be present on multiple loci and still remain able to properly regulate the remaining ParaHox genes despite the loss of other ParaHox genes at that locus (Mulley et al., 2006). It has been suggested by Mulley et al. (2006) that the ParaHox cluster is not under the same sequential regulatory activation as the Hox cluster, such as by the ELCR, POST and GCR elements (Kmita et al., 2000; Lehoczky et al., 2004; Spitz et al., 2003; Spitz et al., 2005; Tarchini and Duboule, 2006), where cluster integrity is necessary for correct gene control.

Though no such elements have yet been described for the ParaHox cluster, the ParaHox cluster is now known to be much more highly retained as an intact cluster than once thought. These intact ParaHox clusters also display the regulatory phenomena of collinearity, in which genes are activated in the same order as they occur along the chromosome, and can be found across the deuterostome phyla, with examples in the chordates (Brooke et al., 1998; Osborne et al., 2009), echinoderms (Annunziata et al., 2013) and hemichordates (Ikuta et al., 2013). Not only does the ParaHox cluster display collinearity, but it appears that this regulatory phenomenon plays a key role in the maintenance of an intact ParaHox cluster, with temporal collinearity possibly key to this (Ferrier and Minguillon, 2003; Garstang and Ferrier, 2013), as this form of collinearity is only observed within intact ParaHox clusters, and likewise intact ParaHox clusters are only observed where temporal collinearity is maintained. There is the possibility that this constraint may have been overcome within the teleost lineage, whilst still maintaining regulatory mechanisms, due to the redundancy allowed by the further 3R genome duplication of teleosts (Meyer and Van de Peer, 2005; Mulley et al., 2006). Proper analysis of ParaHox temporal collinearity has not been examined in a single study within teleosts and it is not necessarily clear how to compare time across degenerate and potentially redundant cluster loci.

This in turn means that there are likely to be regulatory mechanisms directing ParaHox expression, both spatially and temporally, that are vital to the maintenance of the ParaHox cluster. RA signalling may be one regulatory mechanism, and some of the earliest data on the regulation of Hox genes revealed a role for RA in sequential temporal activation (Simeone et al., 1990) and the direct regulation of Hox genes by RA is well established. Intriguingly, RA regulates all of the ParaHox genes in amphioxus and several RAREs have been identified within the amphioxus ParaHox cluster that may mediate this response (Osborne et al., 2009). A link between RA signalling and intact Hox clusters has been proposed (Canestro and Postlethwait, 2007), which could just as well extend to the ParaHox genes. Either way it is clear that pan-cluster regulatory mechanisms are likely involved in the regulation of the ParaHox cluster, and that these mechanisms are probably key to the maintenance of the ParaHox genes as an intact cluster. Regions that mediate response to classical axial patterning signals, such as the RARE elements identified within Osborne *et al*. (Osborne et al., 2009), but also other axial morphogen gradients such as Wnt signalling, are good candidates for further functional regulatory studies. The conserved TCF/Lef binding sites identified across the amphioxus ParaHox cluster are suggestive of pan-cluster regulation, probably by Wnt signalling. The further examination of these sites would give more insight into TCF/Lef and Wnt input into ParaHox regulation. In this context, the presence of multiple TCF/Lef sites in the region upstream of *Cdx* (figure 3.11, TCF18-22) could be interesting, as Cdx is both the first ParaHox gene activated as well as the most posterior and therefore subject to the highest levels of Wnt signalling. The Hox genes are activated sequentially by RA signalling, which is expressed at highest levels in the anterior and induces sequential activation with the anterior Hox genes being expressed first and the posterior last, from Hoxb1-b9 (Simeone et al., 1990). It is possible then that the reversed temporal collinearity of the ParaHox cluster, with *Cdx* (posterior) first and *Gsx* (anterior) last, may be activated by a signal that originates in the posterior, such as Wnt signalling. If so, the TCF/Lef site rich region upstream of *Cdx* would be a good candidate for investigation of this.

The presence of GRBs and shared regulatory elements is likely to play a role in the maintenance of an intact ParaHox cluster, providing a constraint for the ParaHox genes to remain clustered together. The presence of GRBs, where regulatory elements for one gene are interspersed amongst nearby bystander genes, may account for the close association of the ParaHox cluster with genes such as CHIC, PRHOXNB and FLT1 over vast evolutionary distances (Ferrier et al., 2005; Ikuta et al., 2013). Indeed a GRB surrounding the ParaHox cluster has been observed (Kikuta et al., 2007), and a *Cdx2* regulatory element has been identified within the third intron of PRHOXNB (Benahmed et al., 2008). The mapping of amphioxus ParaHox non-coding conservation beyond the ParaHox cluster proper, including the neighbouring *CHIC* and *PRHOXNB* genes, will be key to the identification of any ParaHox regulatory elements that may form an ancestral ParaHox GRB.

Shared regulatory elements within the ParaHox cluster may account for an additional constraint on cluster maintenance. A conserved non-coding region between vertebrate Gsh1 and Pdx1 is one such region that may hold shared regulatory potential, though functional analysis has not yet been carried out upon this region (Mulley et al., 2006). The adjacent neural expression of the amphioxus Gsx early domain and Xlox neural domain, at the level of the presumptive pigment spot, would also support a regulatory interaction between the two genes at least, if not a shared regulatory element (Osborne et al., 2009). The amphioxus ParaHox non-coding landscape detailed within chapter 3 has revealed several regions that may be suitable candidates for such a shared Gsx-Xlox regulatory element. In particular, two regions stand out within the amphioxus ParaHox cluster in this respect. The first is a discrete peak with high conservation across amphioxus species (figure 3.4, present at ~39Kb), that lies roughly equidistant between Gsx and Xlox. This region contains a conserved CTCF binding site (figure 3.8, CTCF site 10), and may hold potential as a functional regulatory element involved in the regulation of multiple genes via modification of chromatin domains, as observed in the Hox cluster (Narendra et al., 2015). The second is a region just downstream of *Gsx* that contains a cluster of conserved *TCF/Lef* binding sites (figure 3.10 TCF 6-9). We have already established that such sites are crucial to the expression of the *Bf-Gsx-Up* reporters within the CNS (chapter 5), and that amphioxus *Xlox* expression is present in an adjacent neural domain to *Gsx*. Since the posterior boundary of amphioxus *Gsx* overlaps with the anterior boundary of *Xlox* expression, it is possible that the same signals are involved in the expression of these neural domains. Thus a regulatory region located between the two genes, with a high density of *TCF/Lef* binding sites is potentially an excellent candidate region for a shared regulatory element co-ordinating these neural *Gsx/Xlox* domains.

6.3.2. The ParaHox cluster is, in fact, insulated from outside genomic influence to at least some extent.

Transposable elements are another factor that may greatly effect ParaHox cluster integrity. The presence of TEs can lead to the breakup of gene clusters, as suggested for ecdysozoan Hox (Fried et al., 2004) clusters, and also cause rearrangement events as in the mammalian MHC cluster (Childers et al., 2006). Previous work has identified that although intact Hox clusters appear to exclude TEs, the chordate ParaHox cluster does not and even be a hotspot for TE insertion (Osborne and Ferrier, 2010). Even if the ParaHox cluster in total is a hotspot for TE insertion, a more detailed examination within chapter 3 reveals specific regions within the cluster that do seem to exclude TEs. The high density of TEs flanking the ParaHox cluster proper does suggest that there is some exclusion of TEs, and those that do invade the ParaHox cluster itself are instead presumably targeted to regions less important to ParaHox regulation. This partial exclusion could then prevent TEs from disrupting regulatory mechanisms vital to ParaHox cluster integrity.

Another example of the ParaHox cluster being resistant to TE invasion, at least in some regions and perhaps keeping its regulatory elements tightly guarded, is observed in the case of the retrogene *SCP1*. With its insertion upstream of *Gsx*, between *CHIC* and *Gsx* (Ferrier et al., 2005), it could be expected that *SCP1* may fall under the regulation of nearby ParaHox regulatory elements. This does not appear to be the case however, and *SCP1* instead appears to have possibly hijacked *CHIC* regulatory elements (chapter 4). There is so far no hint of any ParaHox-like expression or regulation, and mechanisms may exist that prevent nearby genes from both being affected by, and affecting ParaHox gene expression.

One mechanism that may be intrinsically tied to ParaHox cluster insulation is the action of CTCF. This transcription factor is known to be key to the function of insulator elements, which prevent enhancer function across a 'border' (Bell et al., 1999)(reviewed in Wallace and Felsenfeld (2007)). The function of insulators has been well characterised within the Hox cluster and is key to the precise regulation of Hox genes and prevention of errant transcription (Kmita et al., 2002; Moon et al., 2005) (reviewed in Herold et al. (2012)). Within the amphioxus ParaHox cluster we see many CTCF binding sites present within highly conserved regions that may represent potential insulator elements (figure 3. 9). In particular, three sites (sites 2-4) exist between the start of *SCP1* and *Gsx*. Of these, site 4 (figure 3.9) lies equidistant from the start codon of *Gsx* and the stop codon of *SCP1* and is located within a discrete highly conserved region. This marks it as a good target for a potential insulator and could be involved in blocking enhancer function across the ParaHox cluster boundary between *SCP1* and *Gsx*. Many other sites exist within the ParaHox cluster that may also be involved in the fine-tuning of ParaHox expression, and several sites upstream of Cdx (figure 3.9 sites 20-22) and within the first two introns of *PRHOXNB* (figure 3.9 sites 23-25) could represent potential ParaHox boundary insulators at the *Cdx* end of the cluster. In addition to its function within insulator elements, CTCF may also be carrying out another function vital to the regulation, and perhaps integrity of the ParaHox cluster. CTCF is also known to be crucial to the formation of transcription activation domains, which mark domains of chromatin that loop and interact together as a transcriptional unit. The orientation of CTCF sites has been shown to be key to this function and is evolutionarily conserved (Gomez-Marin et al., 2015; Vietri Rudan et al., 2015). Several such oppositely orientated CTCF sites exist within, and surrounding, the amphioxus ParaHox cluster that may mark TAD boundaries (figures 3.7-3.9). Such TADs would be key to ensuring the tight regulation of the ParaHox genes and preventing genes outside of the ParaHox cluster from both interacting with ParaHox regulatory elements, and also ParaHox regulatory elements from influencing genes outside of the cluster. This formation of transcriptional domains may also be key to maintaining an intact ParaHox cluster, encouraging the evolution of ParaHox genes as a regulatory unit. This is certainly observed within the Hox cluster, and various enhancers and even nearby Hox long non-coding-RNAs outside of the Hox cluster have been observed to be transcriptionally linked together via TADs (Delpretti et al., 2013). This formation of TADs is also linked to the association of CTCF with regions of the repressive chromatin modification H3K27me3 (Gomez-Marin et al., 2015; Narendra et al., 2015) , where CTCF is enriched at H3K27me3-rich boundaries (Cuddapah et al., 2009). This is particularly relevant as the presence of euchromatin and heterochromatin domains in transcriptionally active and transcriptionally repressive regions may be involved in determining where TEs are able to invade. It is thought that TEs may be able to invade the ParaHox cluster due to the activity of Cdx in the germline (Osborne and Ferrier, 2010), which would require open chromatin domains. Indeed, differing chromatin states within germ cells is thought to be one of the key factors in the differing Hox and ParaHox TE content (Osborne and Ferrier, 2010). The activation of Cdx in the germline could explain the high density of TEs in the region upstream of amphioxus *Cdx* and surrounding *PRHOXNB*. The presence of context specific TADs across the region covering other areas of the ParaHox cluster could potentially help exclude TEs from more important regions of the ParaHox cluster by creating domains of heterochromatin that TEs are unable to invade.

6.4. Using *Ciona intestinalis* as a 'living test tube' for the analysis of ParaHox regulatory elements

Though experimental manipulation of amphioxus itself is difficult, *C.intestinalis* can be used as a living 'test tube' for the analysis of amphioxus ParaHox regulatory elements. Whilst the ever-growing genomic and transcriptomic resources available for amphioxus have allowed improved identification of potential regulatory elements, functional studies of regulatory elements within amphioxus itself are difficult (Beaster-Jones et al., 2007; Holland et al., 2008; Yu et al., 2004)(reviewed in Beaster-Jones (2012). Analysis of reporter constructs within *Ciona* serves as an alternative and offers several advantages. *Ciona* transgenics are created via electroporation of plasmid constructs, which generates large numbers of transgenic embryos very rapidly. It is therefore much more rapid than microinjection approaches that are used in other species, and the mutagenised plasmid constructs used in *Ciona* transgenics are much quicker and easier to create than the genetic mutants of other species. The large numbers of transformed embryos that can be rapidly generated permits robust construction of expression patterns from even weakly expressing mosaic reporters and can be applied to rapidly analyse a large number of regulatory elements (Harafuji et al., 2002). This efficiency advantage is extremely important when performing cross-species reporter transgenics, as the reporters usually operate with reduced efficiency compared to intra-species reporters.

Cross-species transgenics implicitly enables a focus on transcription factors that are conserved between species. Thus, they are more likely to also be conserved to vertebrates and even more widely. *Ciona* is being used as a 'living test tube' within this thesis, with no real requirements for being able to draw conclusions about homologous relationships between specific tissues (or even cells). What is far more important is that *Ciona* transcription factor expression patterns have been mapped out to an unprecedented level of detail, and can be mapped to specific cells and lineages within the development of the *Ciona* embryo, which cannot be matched in any vertebrate system (e.g. the ANISEED database (Tassy et al., 2010)). The conservation of many of these developmental transcription factors across chordate evolution means that cross-species reporter constructs usually work reliably (Natale et al., 2011; Wada et al., 2005) and the *Ciona* expression data provides a magnificent system in which to rapidly interpret the developmental readout from reporters into terms of likely transcription factors acting upon them.

This has been applied to the analysis of the *Bf-Gsx-Up* regulatory element and the large numbers involved have allowed the differentiation of subtle regulatory changes between different mutants. In situ hybridisation of *Ci-TCF/Lef* was also carried out to build upon preliminary expression data pulled from the ANISEED database, in order to properly characterise *Ci-TCF/Lef* expression across the developmental stages relevant to *Bf-Gsx-Up* reporter expression. The phenotypic effect of TCF/Lef binding site mutations upon *Bf-Gsx-Up* reporter expression also highlights the conservation of developmental transcription factor function between the chordates.

Future Work

Several questions have arisen from this work that are readily approachable and will further examine the regulation of the amphioxus ParaHox cluster, and use the studies carried out here as a basis. Building upon the work examining the *Bf-Gsx-Up* reporter expression, one line of experimentation would be to examine if the TCF/Lef binding sites that control *AmphiGsx* are indeed binding *TCF/Lef*, through the use of Electromobility shift assays (EMSAs), and if *in vivo* manipulation of TCF/Lef alters both *Bf-Gsx-Up* reporter expression and ParaHox expression in both *Ciona* and amphioxus. The identification of conserved *TCF/Lef* binding sites across the ParaHox cluster would make analysing amphioxus ParaHox gene response to *TCF/Lef* and Wnt signalling manipulation an interesting line of study and address whether *TCF/Lef* is a direct, pan-cluster regulator of the chordate ParaHox cluster. Further ParaHox reporter constructs containing these potential regulatory elements could then be designed based upon the work in chapter 3 to test the direct action of *TCF/Lef* on regulatory elements across the ParaHox cluster.

It is currently unknown what other proteins may be binding TCF/Lef to activate Bf-Gsx-Up reporter expression, and identifying these would be key in determining what signalling pathways are directing this *TCF/Lef* binding and subsequent ParaHox expression. For example, if β-catenin is identified bound in conjunction with *TCF/Lef* then it would greatly strengthen the argument for the involvement of Wnt signalling. This could also be used to identify candidate transcription factors mediating the interesting repressive response of *Bf-Gsx-Up* regulatory elements in the absence of *TCF/Lef* binding.

The availability of an accurate map of conserved non-coding regions covering and extending beyond the amphioxus ParaHox cluster will greatly aid in the identification of further ParaHox regulatory elements, and could even be used to identify long range enhancers within neighbouring genes, that may be involved in the formation of a ParaHox GRB and conservation of ParaHox gene neighbours.

Finally, the function of CTCF within insulator elements could also be examined using the *Ciona* transgenics system. Potential insulator elements, demarcated by conserved non-coding regions containing CTCF binding sites, could be used in conjunction with known reliable enhancer elements, such as Bf-Gsx-Up1, to test insulator function in a reporter. This would use combinations of one or two different enhancers in combination with varying insulator placement to test if the potential insulator is able to block the function of one or more enhancer driven expression patterns.

Conclusions

Throughout the work detailed in this thesis, the examination of the regulatory landscape surrounding the amphioxus ParaHox cluster has been a common thread. Though experimental manipulation of amphioxus itself is difficult, *C.intestinalis* can be used as a living 'test tube' for the rapid analysis of regulatory elements across the amphioxus ParaHox cluster, and can be combined with amphioxus genomics and transcriptomics to investigate the evolution and regulation of the ParaHox cluster through multiple avenues.

Several approaches have been utilised during this work in order to examine a wide range of regulatory mechanisms. Comparative genomics between amphioxus species has allowed the mapping of conserved non-coding elements across the ParaHox cluster for the first time, and will greatly aid in the identification of functional regulatory elements. In order to build upon this, several classical Hox regulatory inputs have been examined in conjunction with this, and both conserved CTCF and TCF/Lef binding regions have been identified.

This comparative genomics approach has then been successfully utilised to inform the functional analysis of individual ParaHox gene regulatory elements to provide insight into the signalling inputs and mechanisms controlling the expression of ParaHox genes. Deletion analysis of the *Bf-Gsx-Up* regulatory element has identified a minimal regulatory element required to drive CNS expression within transgenic *Ciona* (*Bf-Gsx-Up1c*), as well as the longer more efficient, and highly conserved *Bf-Gsx-Up1+2b* region. Mutagenesis has also revealed that *TCF/Lef* is likely a direct regulator of *AmphiGsx*, and the identification of conserved *TCF/Lef* binding sites throughout the ParaHox cluster suggests that *TCF/Lef* may exhibit a regulatory effect across the ParaHox cluster.

A combination of studies looking at ParaHox cluster integrity have revealed that the ParaHox cluster may be more resilient to outside influence than was thought. Transposable elements, rather than being present throughout the ParaHox cluster, appear to be under a certain level of exclusion from regions presumed to exhibit regulatory potential. In conjunction with this, the retrogene *SCP1*, rather than having invaded the ParaHox cluster, also appears to have been excluded from the ParaHox transcriptional landscape. CTCF binding sites across the ParaHox cluster may be involved in maintaining the ParaHox cluster as a cohesive discrete transcriptional unit through the formation of TADs, and also through insulator elements that block enhancer function across specific boundaries.

Bibliography

A.F.A. Smit, R. Hubley, Green, P., unpublished data. RepeatMasker at http://repeatmasker.org

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104-2105.

Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P., Inoko, H., 2002. Evidence of en bloc duplication in vertebrate genomes. Nature Genetics 31, 100-105.

Abzhanov, A., Kaufman, T.C., 2000. Embryonic expression patterns of the Hox genes of the crayfish Procambarus clarkii (Crustacea, Decapoda). Evolution & Development 2, 271-283.

Adamska, M., Degnan, S.M., Green, K.M., Adamski, M., Craigie, A., Larroux, C., Degnan, B.M., 2007. Wnt and TGF-beta Expression in the Sponge Amphimedon queenslandica and the Origin of Metazoan Embryonic Patterning. Plos One 2.

Adamska, M., Larroux, C., Adamski, M., Green, K., Lovas, E., Koop, D., Richards, G.S., Zwafink, C., Degnan, B.M., 2010. Structure and expression of conserved Wnt pathway components in the demosponge Amphimedon queenslandica. Evolution & Development 12, 494-518.

Adell, T., Nefkens, I., Muller, W.E.G., 2003. Polarity factor 'Frizzled' in the demosponge Suberites domuncula: identification, expression and localization of the receptor in the epithelium/pinacoderm. Febs Letters 554, 363-368.

Adell, T., Thakur, A.N., Mueller, W.E.G., 2007. Isolation and characterization of Wnt pathway-related genes from Porifera. Cell Biology International 31, 939-949.

Agostini, F., Zanzoni, A., Klus, P., Marchese, D., Cirillo, D., Tartaglia, G.G., 2013. catRAPID omics: a web server for large-scale prediction of protein–RNA interactions. Bioinformatics 29, 2928-2930.

Ah Cho, E., Dressler, G.R., 1998. TCF-4 binds β-catenin and is expressed in distinct regions of the embryonic brain and limbs. Mechanisms of Development 77, 9-18.

Akiyama-Oda, Y., Oda, H., 2003. Early patterning of the spider embryo: a cluster of mesenchymal cells at the cumulus produces Dpp signals received by germ disc epithelial cells. Development 130, 1735-1747.

Albalat, R., Canestro, C., 2009. Identification of Aldh1a, Cyp26 and RAR orthologs in protostomes pushes back the retinoic acid genetic machinery in evolutionary time to the bilaterian ancestor. Chemico-Biological Interactions 178, 188-196.

Alipour, E., Marko, J.F., 2012. Self-organization of domain structures by DNA-loop-extruding enzymes. Nucleic Acids Research 40, 11202-11212.

Alvarez-Garcia, I., Miska, E.A., 2005. MicroRNA functions in animal development and human disease. Development 132, 4653-4662.

Ambros, V., 2004. The functions of animal microRNAs. Nature 431, 350-355.

Amemiya, C.T., Prohaska, S.J., Hill-Force, A., Cook, A., Wasserscheid, J., Ferrier, D.E.K., Pascual-Anaya, J., Garcia-Fernandez, J., Dewar, K., Stadler, P.F., 2008. The amphioxus Hox cluster: Characterization, comparative genomics, and evolution. Journal of Experimental Zoology Part B-Molecular and Developmental Evolution 310B, 465-477.

Amores, A., Force, A., Yan, Y.L., Joly, L., Amemiya, C., Fritz, A., Ho, R.K., Langeland, J., Prince, V., Wang, Y.L., Westerfield, M., Ekker, M., Postlethwait, J.H., 1998. Zebrafish hox clusters and vertebrate genome evolution. Science 282, 1711-1714.

Angelini, D.R., Kaufman, T.C., 2005. Functional analyses in the milkweed bug Oncopeltus fasciatus (Hemiptera) support a role for Wnt signaling in body segmentation but not appendage development. Developmental Biology 283, 409-423.

Annunziata, R., Arnone, M.I., 2014. A dynamic regulatory network explains ParaHox gene control of gut patterning in the sea urchin. Development 141, 2462-2472.

Annunziata, R., Martinez, P., Arnone, M.I., 2013. Intact cluster and chordate-like expression of ParaHox genes in a sea star. Bmc Biology 11, 14.

Arce, L., Yokoyama, N.N., Waterman, M.L., 2006. Diversity of LEF//TCF action in development and disease. Oncogene 25, 7492-7504.

Arnone, M.I., Rizzo, F., Annunciata, R., Cameron, R.A., Peterson, K.J., Martinez, P., 2006. Genetic organization and embryonic expression of the ParaHox genes in the sea urchin S-purpuratus: Insights into the relationship between clustering and colinearity. Developmental Biology 300, 63-73.

Arnosti, D.N., Barolo, S., Levine, M., Small, S., 1996. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. Development 122, 205-214.

Ashworth, A., Skene, B., Swift, S., Lovell-Badge, R., 1990. Zfa is an expressed retroposon derived from an alternative transcript of the Zfx gene. Embo j 9, 1529-1534.

Ayoub, N., Richler, C., Wahrman, J., 1997. Xist RNA is associated with the transcriptionally inactive XY body in mammalian male meiosis. Chromosoma 106, 1-10.

Baguna, J., Riutort, M., 2004. The dawn of bilaterian animals: the case of acoelomorph flatworms. Bioessays 26, 1046-1057.

Bai, Y., Casola, C., Feschotte, C., Betran, E., 2007. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. Genome Biology 8, R11.

Bai, Y.S., Casola, C., Betran, E., 2009. Quality of regulatory elements in Drosophila retrogenes. Genomics 93, 83-89.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S., 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research 37, W202-W208.

Bao, W.D., Kojima, K.K., Kohany, O., 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6, 6.

Barolo, S., 2006. Transgenic Wnt/TCF pathway reporters: all you need is Lef? Oncogene 25, 7505-7511.

Barrett, L., Fletcher, S., Wilton, S., 2012. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. Cellular and Molecular Life Sciences 69, 3613-3634.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., Zhao, K., 2007. High-resolution profiling of histone methylations in the human genome. Cell 129, 823-837.

Baughman, K.W., McDougall, C., Cummins, S.F., Hall, M., Degnan, B.M., Satoh, N., Shoguchi, E., 2014. Genomic Organization of Hox and ParaHox Clusters in the Echinoderm, Acanthaster planci. Genesis 52, 952-958.

Bayha, E., Jørgensen, M.C., Serup, P., Grapin-Botton, A., 2009. Retinoic Acid Signaling Organizes Endodermal Organ Specification along the Entire Antero-Posterior Axis. PLoS ONE 4, e5845.

Beaster-Jones, L., 2012. Cis-regulation and conserved non-coding elements in amphioxus. Briefings in Functional Genomics 11, 118-130.

Beaster-Jones, L., Schubert, M., Holland, L.Z., 2007. Cis-regulation of the amphioxus engrailed gene: Insights into evolution of a muscle-specific enhancer. Mechanisms of Development 124, 532-542.

Beck, F., Erler, T., Russell, A., James, R., 1995. Expression of Cdx-2 in the mouse embryo and placenta: possible role in patterning of the extra-embryonic membranes. Dev Dyn 204, 219-227.

Behrens, J., vonKries, J.P., Kuhl, M., Bruhn, L., Wedlich, D., Grosschedl, R., Birchmeier, W., 1996. Functional interaction of beta-catenin with the transcription factor LEF-1. Nature 382, 638-642.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., Haussler, D., 2004. Ultraconserved elements in the human genome. Science 304, 1321-1325.

Bel-Vialar, S., Medevielle, F., Pituello, F., 2007. The on/off of Pax6 controls the tempo of neuronal differentiation in the developing spinal cord. Developmental Biology 305, 659-673.

Beland, M., Lohnes, D., 2005. Chicken ovalbumin upstream promoter-transcription factor members repress retinoic acid-induced Cdx1 expression. Journal of Biological Chemistry 280, 13858-13862.

Beland, M., Pilon, N., Houle, M., Oh, K., Sylvestre, J.R., Prinos, P., Lohnes, D., 2004. Cdx1 autoregulation is governed by a novel Cdx1-LEF1 transcription complex. Molecular and Cellular Biology 24, 5028-5038.

Bell, A.C., Felsenfeld, G., 2000. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature 405, 482-485.

Bell, A.C., West, A.G., Felsenfeld, G., 1999. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell 98, 387-396.

Belozerov, V.E., Majumder, P., Shen, P., Cai, H.N., 2003. A novel boundary element may facilitate independent gene regulation in the Antennapedia complex of Drosophila. Embo Journal 22, 3113-3121.

Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., Dekker, J., 2012. Hi-C: A comprehensive technique to capture the conformation of genomes. Methods 58, 268-276.

Ben-Shushan, E., Marshak, S., Shoshkes, M., Cerasi, E., Melloul, D., 2001. A pancreatic beta-cell-specific enhancer in the human PIX-1 gene is regulated by hepatocyte nuclear factor 3 beta (HNF-3 beta), HNF-1 alpha, and SPs transcription factors. Journal of Biological Chemistry 276, 17533-17540.

Benahmed, F., Gross, I., Gaunt, S.J., Beck, F., Jehan, F., Domon-Dell, C., Martin, E., Kedinger, M., Freund, J.-N., Duluc, I., 2008. Multiple regulatory regions control the complex expression pattern of the mouse Cdx2 homeobox gene. Gastroenterology 135, 1238-1247.

Bender, W., Akam, M., Karch, F., Beachy, P.A., Peifer, M., Spierer, P., Lewis, E.B., Hogness, D.S., 1983. Molecular Genetics of the Bithorax Complex in Drosophila melanogaster. Science 221, 23-29.

Benito-Gutierrez, E., Weber, H., Bryant, D.V., Arendt, D., 2013. Methods for Generating Year-Round Access to Amphioxus in the Laboratory. Plos One 8, 7.

Berna, L., Alvarez-Valin, F., 2014. Evolutionary Genomics of Fast Evolving Tunicates. Genome Biology and Evolution 6, 1724-1738.

Bertrand, V., Hudson, C., Caillol, D., Popovici, C., Lemaire, P., 2003. Neural tissue in ascidian embryos is induced by FGF9/16/20, acting via a combination of maternal GATA and Ets transcription factors. Cell 115, 615-627.

Betran, E., Thornton, K., Long, M., 2002. Retroposed new genes out of the X in Drosophila. Genome Research 12, 1854-1859.

Bhatia, S., Monahan, J., Ravi, V., Gautier, P., Murdoch, E., Brenner, S., van Heyningen, V., Venkatesh, B., Kleinjan, D.A., 2014. A survey of ancient conserved non-coding elements in the PAX6 locus reveals a landscape of interdigitated cis-regulatory archipelagos. Developmental Biology 387, 214-228.

Biemont, C., Vieira, C., 2005. What transposable elements tell us about genome organization and evolution: the case of Drosophila. Cytogenetic and Genome Research 110, 25-34.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., Kuehn, M.S., Taylor, C.M., Neph, S., Koch, C.M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J.A., Andrews, R.M., Flicek, P., Boyle, P.J., Cao, H., Carter, N.P., Clelland, G.K., Davis, S., Day, N., Dhami, P., Dillon, S.C., Dorschner, M.O., Fiegler, H., Giresi, P.G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K.D., Johnson, B.E., Johnson, E.M., Frum, T.T., Rosenzweig, E.R., Karnani, N., Lee, K., Lefebvre, G.C., Navas, P.A., Neri, F., Parker, S.C.J., Sabo, P.J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F.S., Dekker, J., Lieb, J.D., Tullius, T.D., Crawford, G.E., Sunyaev, S., Noble, W.S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I.L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H.A., Sekinger, E.A., Lagarde, J., Abril, J.F., Shahab, A., Flamm, C., Fried, C., Hackermueller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J.S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M.C., Thomas, D.J., Weirauch, M.T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K.G., Sung, W.-K., Ooi, H.S., Chiu, K.P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M.L., Valencia, A., Choo, S.W., Choo, C.Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T.G., Brown, J.B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C.N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J.S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R.M., Rogers, J., Stadler, P.F., Lowe, T.M., Wei, C.-L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S.E., Fu, Y., Green, E.D., Karaoez, U., Siepel, A., Taylor, J., Liefer, L.A., Wetterstrand, K.A., Good, P.J., Feingold, E.A., Guyer, M.S., Cooper, G.M., Asimenos, G., Dewey, C.N., Hou, M., Nikolaev, S., Montoya-Burgos, J.I., Loeytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N.R., Holmes, I., Mullikin, J.C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W.J., Stone, E.A., Batzoglou, S., Goldman, N., Hardison, R.C., Haussler, D., Miller, W., Sidow, A., Trinklein, N.D., Zhang, Z.D., Barrera, L., Stuart, R., King, D.C., Ameur, A., Enroth, S., Bieda, M.C., Kim, J., Bhinge, A.A., Jiang, N., Liu, J., Yao, F., Vega, V.B., Lee, C.W.H., Ng, P., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M.J., Inman, D., Singer, M.A., Richmond, T.A., Munn, K.J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J.C., Couttet, P., Bruce, A.W., Dovey, O.M., Ellis, P.D., Langford, C.F., Nix, D.A., Euskirchen, G., Hartman, S., Urban, A.E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T.H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C.K., Rosenfeld, M.G., Force Aldred, S., Cooper, S.J., Halees, A., Lin, J.M., Shulha, H.P., Zhang, X., Xu, M., Haidar, J.N.S., Yu, Y., Iyer, V.R., Green, R.D., Wadelius, C., Farnham, P.J., Ren, B., Harte, R.A., Hinrichs, A.S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A.S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R.M., Karolchik, D., Armengol, L., Bird, C.P., de Bakker, P.I.W., Kern, A.D., Lopez-Bigas, N., Martin, J.D., Stranger, B.E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrimsdottir, I.B., Huppert, J., Zody, M.C., Abecasis, G.R., Estivill, X., Bouffard, G.G., Guan, X., Hansen, N.F., Idol, J.R., Maduro, V.V.B., Maskeri, B., McDowell, J.C., Park, M., Thomas, P.J., Young, A.C., Blakesley, R.W., Muzny, D.M., Sodergren, E., Wheeler, D.A., Worley, K.C., Jiang, H., Weinstock, G.M., Gibbs, R.A., Graves, T., Fulton, R., Mardis, E.R., Wilson, R.K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D.B., Chang, J.L., Lindblad-Toh, K., Lander, E.S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., de Jong, P.J., Consortium, E.P., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799-816.

Bogdanov Iu, F., Grishaeva, T.M., Dadashev, S., 2002. [CG17604 gene from Drosophila melanogaster--possible functional homolog of the yeast ZIP1 and SCP1 (SYCP1) mammalian genes, coding for synaptonemal complex proteins. Genetika 38, 108-112.

Bogdanov, Y.F., Dadashev, S.Y., Grishaeva, T.M., 2003. In silico search for functionally similar proteins involved in meiosis and recombination in evolutionarily distant organisms. In Silico Biology 3, 173-185.

Bolognesi, R., Farzana, L., Fischer, T.D., Brown, S.J., 2008. Multiple Wnt Genes Are Required for Segmentation in the Short-Germ Embryo of Tribolium castaneum. Current Biology 18, 1624-1629.

Bonhomme, C., Calon, A., Martin, E., Robine, S., Neuville, A., Kedinger, M., Domon-Dell, C., Duluc, I., Freund, J.N., 2008. Cdx1, a dispensable homeobox gene for gut development with limited effect in intestinal cancer. Oncogene 27, 4497-4502.

Boorman, C.J., Shimeld, S.M., 2002. The evolution of left–right asymmetry in chordates. BioEssays 24, 1004-1011.

Bourlat, S.J., Juliusdottir, T., Lowe, C.J., Freeman, R., Aronowicz, J., Kirschner, M., Lander, E.S., Thorndyke, M., Nakano, H., Kohn, A.B., Heyland, A., Moroz, L.L., Copley, R.R., Telford, M.J., 2006. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature 444, 85-88.

Bradley, J., Baltus, A., Skaletsky, H., Royce-Tolland, M., Dewar, K., Page, D.C., 2004. An X-to-autosome retrogene is required for spermatogenesis in mice. Nature Genetics 36, 872-876.

Brantjes, H., Roose, J., van de Wetering, M., Clevers, H., 2001. All Tcf HMG box transcription factors interact with Groucho-related co-repressors. Nucleic Acids Research 29, 1410-1419.

Brink, C., 2003. Promoter elements in endocrine pancreas development and hormone regulation. Cellular and Molecular Life Sciences 60, 1033-1048.

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

Brosius, J., 1991. Retroposons--seeds of evolution. Science 251, 753.

Brudno, M., Malde, S., Poliakov, A., Do, C.B., Couronne, O., Dubchak, I., Batzoglou, S., 2003. Glocal alignment: finding rearrangements during alignment. Bioinformatics 19, i54-i62.

Bruhn, L., Munnerlyn, A., Grosschedl, R., 1997. ALY, a context-dependent coactivator of LEF-1 and AML-1, is required for TCRalpha enhancer function. Genes & Development 11, 640-653.

Brunner, E., Peter, O., Schweizer, L., Basler, K., 1997. pangolin encodes a Lef-1 homologue that acts downstream of Armadillo to transduce the Wingless signal in Drosophila. Nature 385, 829-833.

Brusca, R.C., Brusca, G.J., 2003. Invertebrates, 2nd ed. Sinauer Associates, Sunderland, MA,USA.

Bucher, G., Farzana, L., Brown, S.J., Klingler, M., 2005. Anterior localization of maternal mRNAs in a short germ insect lacking bicoid. Evolution & Development 7, 142-149.

Buenrostro, J.D., Wu, B., Chang, H.Y., Greenleaf, W.J., 2001. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide, Current Protocols in Molecular Biology. John Wiley & Sons, Inc.

Byrne, M., Nakajima, Y., Chee, F.C., Burke, R.D., 2007. Apical organs in echinoderm larvae: insights into larval evolution in the Ambulacraria. Evolution & Development 9, 432-445.

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

Canestro, C., Postlethwait, J.H., 2007. Development of a chordate anterior-posterior axis without classical retinoic acid signaling. Developmental Biology 305, 522-538.

Carrasco, M., Delgado, I., Soria, B., Martin, F., Rojas, A., 2012. GATA4 and GATA6 control mouse pancreas organogenesis. Journal of Clinical Investigation 122, 3504-3515.

Cartwright, P., Bowsher, J., Buss, L.W., 1999. Expression of a Hox gene, Cnox-2, and the division of labor in a colonial hydroid. Proceedings of the National Academy of Sciences of the United States of America 96, 2183-2186.

Casey, A.E., Daish, T.J., Grutzner, F., 2015. Identification and characterisation of synaptonemal complex genes in monotremes. Gene 567, 146-153.

Chalmers, A.D., Slack, J.M.W., Beck, C.W., 2000. Regional gene expression in the epithelia of the Xenopus tadpole gut. Mechanisms of Development 96, 125-128.

Charite, J., de Graaff, W., Vogels, R., Meijlink, F., Deschamps, J., 1995. Regulation of the Hoxb-8 gene: synergism between multimerized cis-acting elements increases responsiveness to positional information. Dev Biol 171, 294-305.

Chawengsaksophak, K., James, R., Hammond, V.E., Kontgen, F., Beck, F., 1997. Homeosis and intestinal tumours in Cdx2 mutant mice. Nature 386, 84-87.

Cheesman, S.E., Eisen, J.S., 2004. gsh1 demarcates hypothalamus and intermediate spinal cord in zebrafish. Gene Expression Patterns 5, 107-112.

Chen, W.-C., Pauls, S., Bacha, J., Elgar, G., Loose, M., Shimeld, S.M., 2014. Dissection of a Ciona regulatory element reveals complexity of cross-species enhancer activity. Developmental Biology 390, 261-272.

Chen, Y., Pan, F.C., Brandes, N., Afelik, S., Sölter, M., Pieler, T., 2004. Retinoic acid signaling is essential for pancreas development and promotes endocrine at the expense of exocrine cell differentiation in Xenopus. Developmental Biology 271, 144-160.

Chiba, S., Sasaki, A., Nakayama, A., Takamura, K., Satoh, N., 2004. Development of Ciona intestinalis juveniles (through 2nd ascidian stage). Zoolog Sci 21, 285-298.

Childers, C.P., Newkirk, H.L., Honeycutt, D.A., Ramlachan, N., Muzney, D.M., Sodergren, E., Gibbs, R.A., Weinstock, G.M., Womack, J.E., Skow, L.C., 2006. Comparative analysis of the bovine MHC class IIb sequence identifies inversion breakpoints and three unexpected genes. Animal Genetics 37, 121-129.

Chiori, R., Jager, M., Denker, E., Wincker, P., Da Silva, C., Le Guyader, H., Manuel, M., Queinnec, E., 2009. Are Hox Genes Ancestrally Involved in Axial Patterning? Evidence from the Hydrozoan Clytia hemisphaerica (Cnidaria). Plos One 4, 14.

Chipman, A.D., Arthur, W., Akam, M., 2004. A double segment periodicity underlies segment generation in centipede development. Current Biology 14, 1250-1255.

Chipman, A.D., Ferrier, D.E.K., Brena, C., Qu, J.X., Hughes, D.S.T., Schroder, R., Torres-Oliva, M., Znassi, N., Jiang, H.Y., Almeida, F.C., Alonso, C.R., Apostolou, Z., Aqrawi, P., Arthur, W., Barna, J.C.J., Blankenburg, K.P., Brites, D., Capella-Gutierrez, S., Coyle, M., Dearden, P.K., Du Pasquier, L., Duncan, E.J., Ebert, D., Eibner, C., Erikson, G., Evans, P.D., Extavour, C.G., Francisco, L., Gabaldon, T., Gillis, W.J., Goodwin-Horn, E.A., Green, J.E., Griffiths-Jones, S., Grimmelikhuijzen, C.J.P., Gubbala, S., Guigo, R., Han, Y., Hauser, F., Havlak, P., Hayden, L., Helbing, S., Holder, M., Hui, J.H.L., Hunn, J.P., Hunnekuhl, V.S., Jackson, L., Javaid, M., Jhangiani, S.N., Jiggins, F.M., Jones, T.E., Kaiser, T.S., Kalra, D., Kenny, N.J., Korchina, V., Kovar, C.L., Kraus, F.B., Lapraz, F., Lee, S.L., Lv, J., Mandapat, C., Manning, G., Mariotti, M., Mata, R., Mathew, T., Neumann, T., Newsham, I., Ngo, D.N., Ninova, M., Okwuonu, G., Ongeri, F., Palmer, W.J., Patil, S., Patraquim, P., Pham, C., Pu, L.L., Putman, N.H., Rabouille, C., Ramos, O.M., Rhodes, A.C., Robertson, H.E., Robertson, H.M., Ronshaugen, M., Rozas, J., Saada, N., Sanchez-Gracia, A., Scherer, S.E., Schurko, A.M., Siggens, K.W., Simmons, D., Stief, A., Stolle, E., Telford, M.J., Tessmar-Raible, K., Thornton, R., van der Zee, M., von Haeseler, A., Williams, J.M., Willis, J.H., Wu, Y.Q., Zou, X.Y., Lawson, D., Muzny, D.M., Worley, K.C., Gibbs, R.A., Akam, M., Richards, S., 2014. The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede Strigamia maritima. Plos Biology 12, 24.

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

Christiaen, L., Jaszczyszyn, Y., Kerfant, M., Kano, S., Thermes, V., Joly, J.S., 2007. Evolutionary modification of mouth position in deuterostomes. Semin Cell Dev Biol 18, 502-511.

Christiaen, L., Stolfi, A., Levine, M., 2010. BMP signaling coordinates gene expression and cell migration during precardiac mesoderm development. Developmental Biology 340, 179-187.

Christova, T., Mojtahedi, G., Hamel, P.A., 2010. Lymphoid enhancer factor-1 mediates loading of Pax3 to a promoter harbouring lymphoid enhancer factor-1 binding sites resulting in enhancement of transcription. The International Journal of Biochemistry & Cell Biology 42, 630-640.

Ciomborowska, J., Rosikiewicz, W., Szklarczyk, D., Makalowski, W., Makalowska, I., 2013. "Orphan" Retrogenes in the Human Genome. Molecular Biology and Evolution 30, 384-396.

Clevers, H., Nusse, R., 2012. Wnt/β-Catenin Signaling and Disease. Cell 149, 1192-1205.

Colaiácovo, M.P., MacQueen, A.J., Martinez-Perez, E., McDonald, K., Adamo, A., La Volpe, A., Villeneuve, A.M., 2003. Synaptonemal Complex Assembly in C. elegans Is Dispensable for Loading Strand-Exchange Proteins but Critical for Proper Completion of Recombination. Developmental Cell 5, 463-474.

Cole, A.G., Arnone, M.L., 2009. Fluorescent in situ hybridization reveals multiple expression domains for SpBrn1/2/4 and identifies a unique ectodermal cell type that co-expresses the ParaHox gene SpLox. Gene Expression Patterns 9, 324-328.

Cole, A.G., Rizzo, F., Martinez, P., Fernandez-Serra, M., Arnone, M.I., 2009. Two ParaHox genes, SpLox and SpCdx, interact to partition the posterior endoderm in the formation of a functional gut. Development 136, 541-549.

Colleypriest, B.J., Farrant, J.M., Slack, J.M.W., Tosh, D., 2010. The role of Cdx2 in Barrett's metaplasia. Biochemical Society Transactions 38, 364-369.

Conklin, E.G., 1905. The organization and cell-lineage of the ascidian egg.

Conklin, E.G., 1932. The embryology of Amphioxus. Journal of Morphology 54, 69-140.

Conlon, R.A., Rossant, J., 1992. Exogenous retinoic acid rapidly induces anterior ectopic expression of murine Hox-2 genes in vivo. Development 116, 357-368.

Consortium, C.e.S., 1998. Genome sequence of the nematode C-elegans: A platform for investigating biology. Science 282, 2012-2018.

Cook, C.E., Jimenez, E., Akam, M., Salo, E., 2004. The Hox gene complement of acoel flatworms, a basal bilaterian clade. Evolution & Development 6, 154-163.

Cools, J., Mentens, N., Marynen, P., 2001. A new family of small, palmitoylated, membrane-associated proteins, characterized by the presence of a cysteine-rich hydrophobic motif. Febs Letters 492, 204-209.

Copf, T., Rabet, N., Celniker, S.E., Averof, M., 2003. Posterior patterning genes and the identification of a unique body region in the brine shrimp Artemia franciscana. Development 130, 5915-5927.

Corbin, J.G., Gaiano, N., Machold, R.P., Langston, A., Fishell, G., 2000. The Gsh2 homeodomain gene controls multiple aspects of telencephalic development. Development 127, 5007-5020.

Corbin, J.G., Rutlin, M., Gaiano, N., Fishell, G., 2003. Combinatorial function of the homeodomain proteins Nkx2.1 and Gsh2 in ventral telencephalic patterning. Development 130, 4895-4906.

Corbo, J.C., Levine, M., Zeller, R.W., 1997. Characterization of a notochord-specific enhancer from the Brachyury promoter region of the ascidian, Ciona intestinalis. Development 124, 589-602.

Corrado, M., Aniello, F., Fucci, L., Branno, M., 2001. Ci-IPF1, the pancreatic homeodomain transcription factor, is expressed in neural cells of Ciona intestinalis larva. Mechanisms of Development 102, 271-274.

Costa, Y., Speed, R., Ollinger, R., Alsheimer, M., Semple, C.A., Gautier, P., Maratou, K., Novak, I., Hoog, C., Benavente, R., Cooke, H.J., 2005. Two novel proteins recruited by synaptonemal complex protein 1 (SYCP1) are at the centre of meiosis. Journal of Cell Science 118, 2755-2762.

Cowden, J., Levine, M., 2003. Ventral dominance governs sequential patterns of gene expression across the dorsal-ventral axis of the neuroectoderm in the Drosophila embryo. Developmental Biology 262, 335-349.

Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.-Y., Cui, K., Zhao, K., 2009. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Research 19, 24-32.

D'Aniello, S., D'Aniello, E., Locascio, A., Memoli, A., Corrado, M., Russo, M.T., Aniello, F., Fucci, L., Brown, E.R., Branno, M., 2006. The ascidian homolog of the vertebrate homeobox gene Rx is essential for ocellus development and function. Differentiation 74, 222-234.

Daniels, D.L., Weis, W.I., 2005. [beta]-catenin directly displaces Groucho/TLE repressors from Tcf/Lef in Wnt-mediated transcription activation. Nat Struct Mol Biol 12, 364-371.

Darras, S., Nishida, H., 2001. The BMP/CHORDIN antagonism controls sensory pigment cell specification and differentiation in the ascidian embryo. Developmental Biology 236, 271-288.

Dasilva, C., Hadji, H., Ozouf-Costaz, C., Nicaud, S., Jaillon, O., Weissenbach, J., Crollius, H.R., 2002. Remarkable compartmentalization of transposable elements and pseudogenes in the heterochromatin of the Tetraodon nigroviridis genome. Proceedings of the National Academy of Sciences 99, 13636-13641.

Davidson, A.J., Zon, L.I., 2006. The caudal-related homeobox genes cdx1a and cdx4 act redundantly to regulate hox gene expression and the formation of putative hematopoietic stem cells during zebrafish embryogenesis. Developmental Biology 292, 506-518.

Davidson, E.H., 2001. Genomic regulatory systems: development and evolution. Genomic regulatory systems: development and evolution., i-xii, 1-261.

Davidson, E.H., 2010. Emerging properties of animal gene regulatory networks. Nature 468, 911-920.

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

de Rosa, R., Prud'homme, B., Balavoine, G., 2005. caudal and even-skipped in the annelid Platynereis dumerilii and the ancestry of posterior growth. Evolution & Development 7, 574-587.

de Vries, F.A.T., de Boer, E., van den Bosch, M., Baarends, W.M., Ooms, M., Yuan, L., Liu, J.G., van Zeeland, A.A., Heyting, C., Pastink, A., 2005. Mouse Sycp1 functions in synaptonemal complex assembly, meiotic recombination., and XY body formation. Genes & Development 19, 1376-1389.

Dearden, P.K., Akam, M., 2001. Early embryo patterning in the grasshopper, Schistocerca gregaria: wingless, decapentaplegic and caudal expression. Development 128, 3435-3444.

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

Dekker, J., Marti-Renom, M.A., Mirny, L.A., 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet 14, 390-403.

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

Delpretti, S., Montavon, T., Leleu, M., Joye, E., Tzika, A., Milinkovitch, M., Duboule, D., 2013. Multiple Enhancers Regulate Hoxd Genes and the Hotdog LncRNA during Cecum Budding. Cell Reports 5, 137-150.

Delsuc, F., Brinkmann, H., Chourrout, D., Philippe, H., 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439, 965-968.

Denes, A.S., Jekely, G., Steinmetz, P.R.H., Raible, F., Snyman, H., Prud'homme, B., Ferrier, D.E.K., Balavoine, G., Arendt, D., 2007. Molecular architecture of annelid nerve cord supports common origin of nervous system centralization in bilateria. Cell 129, 277-288.

Dermitzakis, E.T., Kirkness, E., Schwarz, S., Birney, E., Reymond, A., Antonarakis, S.E., 2004. Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. Genome Research 14, 852-859.

Deschamps, J., van de Ven, C., 2012. Concerted involvement of Cdx/Hox genes and Wnt signalling in morphogenesis of the caudal neural tube and cloacal derivatives from the posterior growth zone. ArrayExpress Archive.

Deschet, K., Bourrat, F., Chourrout, D., Joly, J.S., 1998. Expression domains of the medaka (Oryzias latipes) Ol-Gsh 1 gene are reminiscent of those of clustered and orphan homeobox genes. Development Genes and Evolution 208, 235-244.

Di Gregorio, A., Levine, M., 2002. Analyzing gene regulation in ascidian embryos: new tools for new perspectives. Differentiation 70, 132-139.

Dimitrieva, S., Bucher, P., 2013. UCNEbase-a database of ultraconserved non-coding elements and genomic regulatory blocks. Nucleic Acids Research 41, D101-D109.

Dobson, M.J., Pearlman, R.E., Karaiskakis, A., Spyropoulos, B., Moens, P.B., 1994. Synaptonemal complex proteins: occurrence, epitope mapping and chromosome disjunction. J Cell Sci 107 (Pt 10), 2749-2760.

Dorsky, R.I., Snyder, A., Cretekos, C.J., Grunwald, D.J., Geisler, R., Haffter, P., Moon, R.T., Raible, D.W., 1999. Maternal and embryonic expression of zebrafish lef1. Mechanisms of Development 86, 147-150.

Doxiadis, G.G.M., de Groot, N., Bontrop, R.E., 2008. Impact of endogenous intronic retroviruses on major histocompatibility complex class II diversity and stability. Journal of Virology 82, 6667-6677.

Duboule, D., 1994. Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. Dev Suppl, 135-142.

Duboule, D., 2007. The rise and fall of Hox gene clusters. Development 134, 2549-2560.

Dupe, V., Davenne, M., Brocard, J., Dolle, P., Mark, M., Dierich, A., Chambon, P., Rijli, F.M., 1997. In vivo functional analysis of the Hoxa-1 3' retinoic acid response element (3'RARE). Development 124, 399-410.

Déjardin, J., Kingston, R.E., 2009. Purification of Proteins Associated with Specific Genomic Loci. Cell 136, 175-186.

Edgar, L.G., Carr, S., Wang, H., Wood, W.B., 2001. Zygotic expression of the caudal homolog pal-1 is required for posterior patterning in Caenorhabditis elegans embryogenesis. Developmental Biology 229, 71-88.

Edvardsen, R.B., Seo, H.C., Jensen, M.F., Mialon, A., Mikhaleva, J., Bjordal, M., Cartry, J., Reinhardt, R., Weissenbach, J., Wincker, P., Chourrout, D., 2005. Remodelling of the homeobox gene complement in the tunicate Oikopleura dioica. Current Biology 15, R12-R13.

Ehrman, L.A., Yutzey, K.E., 2001. Anterior expression of the caudal homologue cCdx-B activates a posterior genetic program in avian embryos. Developmental Dynamics 221, 412-421.

Emerson, J.J., Kaessmann, H., Betran, E., Long, M.Y., 2004. Extensive gene traffic on the mammalian X chromosome. Science 303, 537-540.

Engel, N., Bartolomei, M.S., 2003. Mechanisms of insulator function in gene regulation and genomic imprinting. International Review of Cytology - a Survey of Cell Biology, Vol 232 232, 89-+.

Engstrom, P.G., Fredman, D., Lenhard, B., 2008. Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. Genome Biology 9.

Ericson, J., Rashbass, P., Schedl, A., Brenner-Morton, S., Kawakami, A., van Heyningen, V., Jessell, T.M., Briscoe, J., 1997. Pax6 Controls Progenitor Cell Identity and Neuronal Fate in Response to Graded Shh Signaling. Cell 90, 169-180.

Fablet, M., Bueno, M., Potrzebowski, L., Kaessmann, H., 2009. Evolutionary Origin and Functions of Retrogene Introns. Molecular Biology and Evolution 26, 2147-2156.

Fang, R., Olds, L.C., Sibley, E., 2006. Spatio-temporal patterns of intestine-specific transcription factor expression during postnatal mouse gut development. Gene Expr Patterns 6, 426-432.

Faro, A., Boj, S.F., Ambrosio, R., van den Broek, O., Korving, J., Clevers, H., 2009. T-Cell Factor 4 (tcf7l2) Is the Main Effector of Wnt Signaling During Zebrafish Intestine Organogenesis. Zebrafish 6, 59-68.

Farre, D., Roset, R., Huerta, M., Adsuara, J.E., Rosello, L., Alba, M.M., Messeguer, X., 2003. Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN. Nucleic Acids Research 31, 3651-3653.

Fatica, A., Bozzoni, I., 2014. Long non-coding RNAs: new players in cell differentiation and development. Nat Rev Genet 15, 7-21.

Fedoriw, A.M., Stein, P., Svoboda, P., Schultz, R.M., Bartolomei, M.S., 2004. Transgenic RNAi reveals essential function for CTCF in H19 gene imprinting. Science 303, 238-240.

Ferraiuolo, M.A., Rousseau, M., Miyamoto, C., Shenker, S., Wang, X.Q.D., Nadler, M., Blanchette, M., Dostie, J., 2010. The three-dimensional architecture of Hox cluster silencing. Nucleic Acids Research 38, 7472-7484.

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

Ferrier, D.E.K., 2010. Evolution of Hox Complexes, in: Deutsch, J.S. (Ed.), Hox Genes: Studies from the 20th to the 21st Century. Springer-Verlag Berlin, Berlin, pp. 91-100.

Ferrier, D.E.K., (in revision, 2015). The origin of the Hox/ParaHox genes, the Ghost Locus Hypothesis and the complexity of the first animal Briefings in Functional Genomics

Ferrier, D.E.K., Holland, P.W.H., 2001a. Ancient origin of the Hox gene cluster. Nature Reviews Genetics 2, 33-38.

Ferrier, D.E.K., Holland, P.W.H., 2001b. Sipunculan ParaHox genes. Evolution & Development 3, 263-270.

Ferrier, D.E.K., Holland, P.W.H., 2002. Ciona intestinalis ParaHox genes: evolution of Hox/ParaHox cluster integrity, developmental mode, and temporal colinearity. Molecular Phylogenetics and Evolution 24, 412-417.

Ferrier, D.E.K., Minguillon, C., 2003. Evolution of the Hox/ParaHox gene clusters. International Journal of Developmental Biology 47, 605-611.

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

Filippova, G.N., 2008. Genetics and epigenetics of the multifunctional protein CTCF. Current Topics in Developmental Biology, Vol 80 80, 337-360.

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

Finnerty, J.R., Pang, K., Burton, P., Paulson, D., Martindale, M.Q., 2004. Origins of Bilateral Symmetry: Hox and Dpp Expression in a Sea Anemone. Science 304, 1335-1337.

Finnerty, J.R., Paulson, D., Burton, P., Pang, K., Martindale, M.Q., 2003. Early evolution of a homeobox gene: the parahox gene Gsx in the Cnidaria and the Bilateria. Evolution & Development 5, 331-345.

Flores, M.V.C., Hall, C.J., Davidson, A.J., Singh, P.P., Mahagaonkar, A.A., Zon, L.I., Crosier, K.E., Crosier, P.S., 2008. Intestinal Differentiation in Zebrafish Requires Cdx1b, a Functional Equivalent of Mammalian Cdx2. Gastroenterology 135, 1665-1675.

Fortunato, S.A.V., Adamski, M., Ramos, O.M., Leininger, S., Liu, J., Ferrier, D.E.K., Adamska, M., 2014. Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. Nature 514, 620-+.

Fraune, J., Alsheimer, M., Volff, J.N., Busch, K., Fraune, S., Bosch, T.C.G., Benavente, R., 2012. Hydra meiosis reveals unexpected conservation of structural synaptonemal complex proteins across metazoans. Proceedings of the National Academy of Sciences of the United States of America 109, 16588-16593.

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M., Dubchak, I., 2004. VISTA: computational tools for comparative genomics. Nucleic Acids Research 32, W273-W279.

Fried, C., Prohaska, S.J., Stadler, P.F., 2004. Exclusion of repetitive DNA elements from gnathostome Hox clusters. Journal of Experimental Zoology Part B-Molecular and Developmental Evolution 302B, 165-173.

Frobius, A.C., Seaver, E.C., 2006. ParaHox gene expression in the polychaete annelid Capitella sp I. Development Genes and Evolution 216, 81-88.

Fröbius, A.C., Seaver, E.C., 2006. ParaHox gene expression in the polychaete annelid Capitella sp I. Development Genes and Evolution 216, 81-88.

Fuentes, M., Benito, E., Bertrand, S., Paris, M., Mignardot, A., Godoy, L., Jimenez-Delgad, S., Oliveri, D., Candiani, S., Hirsinger, E., D'Aniello, S., Pascual-Anaya, J., Maeso, I., Pestarino, M., Vernier, P., Nicolas, J.F., Schubert, M., Laudet, V., Geneviere, A.M., Albalat, R., Fernandez, J.G., Holland, N.D., Escriva, H., 2007. Insights into spawning behavior and development of the European amphioxus (Branchiostoma lanceolatum). Journal of Experimental Zoology Part B-Molecular and Developmental Evolution 308B, 484-493.

Fujita, T., Fujii, H., 2014. Efficient isolation of specific genomic regions retaining molecular interactions by the iChIP system using recombinant exogenous DNA-binding proteins. BMC Molecular Biology 15, 26.

Fukuda, A., Kawaguchi, Y., Furuyama, K., Kodama, S., Horiguchi, M., Kuhara, T., Kawaguchi, M., Terao, M., Doi, R., Wright, C.V.E., Hoshino, M., Chiba, T., Uemoto, S., 2008. Reduction of Ptf1a gene dosage causes pancreatic hypoplasia and diabetes in mice. Diabetes 57, 2421-2431.

Galceran, J., Farinas, I., Depew, M.J., Clevers, H., Grosschedl, R., 1999. Wnt3a(-/-)-like phenotype and limb deficiency in Lef1(-/-)Tcf1(-/-) mice. Genes & Development 13, 709-717.

Gamer, L.W., Wright, C.V., 1993. Murine Cdx-4 bears striking similarities to the Drosophila caudal gene in its homeodomain sequence and early expression pattern. Mech Dev 43, 71-81.

Gannon, M., Ables, E.T., Crawford, L., Lowe, D., Offield, M.F., Magnuson, M.A., Wright, C.V.E., 2008. pdx-1 function is specifically required in embryonic cells to generate appropriate numbers of endocrine cell types and maintain glucose homeostasis. Developmental Biology 314, 406-417.

Gao, N., LeLay, J., Vatamaniuk, M.Z., Rieck, S., Friedman, J.R., Kaestner, K.H., 2008. Dynamic regulation of Pdx1 enhancers by Foxa1 and Foxa2 is essential for pancreas development. Genes & Development 22, 3435-3448.

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

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

Garcia-Fernàndez, J., 2005. Hox, ParaHox, ProtoHox: facts and guesses. Heredity 94, 145-152.

Garstang, M., Ferrier, D.E.K., 2013. Time is of the essence for ParaHox homeobox gene clustering. Bmc Biology 11.

Gaszner, M., Felsenfeld, G., 2006. Insulators: exploiting transcriptional and epigenetic mechanisms. Nature Reviews Genetics 7, 703-713.

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

Gaunt, S.J., 1988. Mouse homeobox gene transcripts occupy different but overlapping domains in embryonic germ layers and organs: a comparison of Hox-3.1 and Hox-1.5. Development 103, 135-144.

Gaunt, S.J., Drage, D., Cockley, A., 2003. Vertebrate caudal gene expression gradients investigated by use of chick cdx-A/lacZ and mouse cdx-1/lacZ reporters in transgenic mouse embryos: evidence for an intron enhancer. Mechanisms of Development 120, 573-586.

Gaunt, S.J., Drage, D., Trubshaw, R.C., 2008. Increased Cdx protein dose effects upon axial patterning in transgenic lines of mice. Development 135, 2511-2520.

Gaunt, S.J., Paul, Y.L., 2014. Synergistic action in P19 pluripotential cells of retinoic acid and Wnt3a on Cdx1 enhancer elements. International Journal of Developmental Biology 58, 307-314.

Gautier, A., Goupil, A.S., Le Gac, F., Lareyre, J.J., 2013. A Promoter Fragment of the sycp1 Gene Is Sufficient to Drive Transgene Expression in Male and Female Meiotic Germ Cells in Zebrafish. Biology of Reproduction 89.

Gerasimova, T.I., Byrd, K., Corces, V.G., 2000. A chromatin insulator determines the nuclear localization of DNA. Molecular Cell 6, 1025-1035.

Gerrish, K., Cissell, A.A., Stein, R., 2001. The role of hepatic nuclear factor 1 alpha and PDX-1 in transcriptional regulation of the pdx-1 gene. Journal of Biological Chemistry 276, 47775-47784.

Geyer, P.K., Corces, V.G., 1992. DNA position-specific repression of transcription by a Drosophila zinc finger protein. Genes and Development.

Geyer, P.K., Spana, C., Corces, V.G., 1986. On the molecular mechanism of *Gypsy*-induced mutations at the Yellow locus of *Drosophila melanogaster*. Embo Journal 5, 2657-2662.

Ghosh, D., 1992. TFD: The transcription factors database. Nucleic Acids Research 20, 2091-2093.

Glardon, S., Holland, L.Z., Gehring, W.J., Holland, N.D., 1998. Isolation and developmental expression of the amphioxus Pax-6 gene (AmphiPax-6): insights into eye and photoreceptor evolution. Development 125, 2701-2710.

Glazov, E.A., Pheasant, M., McGraw, E.A., Bejerano, G., Mattick, J.S., 2005. Ultraconserved elements in insect genomes: A highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. Genome Research 15, 800-808.

Goetz, P., Chandley, A.C., Speed, R.M., 1984. Morphological and temporal sequence of meiotic prophase development at puberty in the male mouse. Journal of Cell Science 65, 249-263.

Gomez-Marin, C., Tena, J.J., Acemel, R.D., Lopez-Mayorga, M., Naranjo, S., de la Calle-Mustienes, E., Maeso, I., Beccari, L., Aneas, I., Vielmas, E., Bovolenta, P., Nobrega, M.A., Carvajal, J., Gomez-Skarmeta, J.L., 2015. Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. Proceedings of the National Academy of Sciences of the United States of America 112, 7542-7547.

Gompel, N., Prud'homme, B., Wittkopp, P.J., Kassner, V.A., Carroll, S.B., 2005. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature 433, 481-487.

Gould, A., Morrison, A., Sproat, G., White, R.A.H., Krumlauf, R., 1997. Positive cross-regulation and enhancer sharing: Two mechanisms for specifying overlapping Hox expression patterns. Genes & Development 11, 900-913.

Goulding, M.D., Lumsden, A., Gruss, P., 1993. Signals from the notochord and floor plate regulate the region-specific expression of two Pax genes in the developing spinal cord. Development 117, 1001-1016.

Gradstein, F.M., Ogg, J.G., 2004. Geologic Time Scale 2004 - why, how, and where next! Lethaia 37, 175-181.

Grainger, S., Savory, J.G.A., Lohnes, D., 2010. Cdx2 regulates patterning of the intestinal epithelium. Developmental Biology 339, 155-165.

Grant, C.E., Bailey, T.L., Noble, W.S., 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017-1018.

Gregorieff, A., Grosschedl, R., Clevers, H., 2004. Hindgut defects and transformation of the gastrointestinal tract in Tcf4(-/-)/Tcf1(-/-) embryos. Embo Journal 23, 1825-1833.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, David U., Jung, I., Wu, H., Zhai, Y., Tang, Y., Lu, Y., Wu, Y., Jia, Z., Li, W., Zhang, Michael Q., Ren, B., Krainer, Adrian R., Maniatis, T., Wu, Q., 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell 162, 900-910.

Hadrys, T., Punnamoottil, B., Pieper, M., Kikuta, H., Pezeron, G., Becker, T.S., Prince, V., Baker, R., Rinkwitz, S., 2006. Conserved co-regulation and promoter sharing of hoxb3a and hoxb4a in zebrafish. Developmental Biology 297, 26-43.

Hahn, Y., 2009. Molecular Evolution of TEPP Protein Genes in Metazoans. Biochemical Genetics 47, 651-664.

Hall, T., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. , Nucleic Acids Symposium Series pp. 95-98.

Hamada, M., Kiryu, H., Sato, K., Mituyama, T., Asai, K., 2009. Prediction of RNA secondary structure using generalized centroid estimators. Bioinformatics 25, 465-473.

Handoko, L., Xu, H., Li, G.L., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W.H., Ye, C.P., Ping, J.L.H., Mulawadi, F., Wong, E., Sheng, J.P., Zhang, Y.B., Poh, T., Chan, C.S., Kunarso, G., Shahab, A., Bourque, G., Cacheux-Rataboul, V., Sung, W.K., Ruan, Y.J., Wei, C.L., 2011. CTCF-mediated functional chromatin interactome in pluripotent cells. Nature Genetics 43, 630-U198.

Harafuji, N., Keys, D.N., Levine, M., 2002. Genome-wide identification of tissue-specific enhancers in the Ciona tadpole. Proceedings of the National Academy of Sciences of the United States of America 99, 6802-6805.

Haremaki, T., Tanaka Y., Hongo, I., Yuge, M., Okamoto, H., 2003. Integration of multiple signal transducing pathways on Fgf response elements of the Xenopus caudal homologue Xcad3. Development 130, 4907-4917.

Hark, A.T., Schoenherr, C.J., Katz, D.J., Ingram, R.S., Levorse, J.M., Tilghman, S.M., 2000. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. Nature 405, 486-489.

Harrison, P.M., Milburn, D., Zhang, Z., Bertone, P., Gerstein, M., 2003. Identification of pseudogenes in the Drosophila melanogaster genome. Nucleic Acids Research 31, 1033-1037.

Harrison, P.M., Zheng, D.Y., Zhang, Z.L., Carriero, N., Gerstein, M., 2005. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic Acids Research 33, 2374-2383.

Hatsheck, B., 1893. The Amphioxus and its development. Swan Sonnenschein & Co, London, UK.

Hayward, D.C., Catmull, J., Reece-Hoyes, J.S., Berghammer, H., Dodd, H., Hann, S.J., Miller, D.J., Ball, E.E., 2001. Gene structure and larval expression of cnox-2Am from the coral Acropora millepora. Dev Genes Evol 211, 10-19.

Heath, H., de Almeida, C.R., Sleutels, F., Dingjan, G., van de Nobelen, S., Jonkers, I., Ling, K.-W., Gribnau, J., Renkawitz, R., Grosveld, F., Hendriks, R.W., Galjart, N., 2008. CTCF regulates cell cycle progression of alpha beta T cells in the thymus. Embo Journal 27, 2839-2850.

Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E., Wiehe, T., 2012. The chromatin insulator CTCF and the emergence of metazoan diversity. Proceedings of the National Academy of Sciences of the United States of America 109, 17507-17512.

Hejnol, A., Martindale, M.Q., 2008. Acoel development indicates the independent evolution of the bilaterian mouth and anus. Nature 456, 382-U345.

Herman, M., 2001. C. elegans POP-1/TCF functions in a canonical Wnt pathway that controls cell migration and in a noncanonical Wnt pathway that controls cell polarity. Development 128, 581-590.

Herold, M., Bartkuhn, M., Renkawitz, R., 2012. CTCF: insights into insulator function during development. Development 139, 1045-1057.

Hinman, V.F., Becker, E., Degnan, B.M., 2000. Neuroectodermal and endodermal expression of the ascidian Cdx gene is separated by metamorphosis. Development Genes and Evolution 210, 212-216.

Hinman, V.F., Degnan, B.M., 1998. Retinoic acid disrupts anterior ectodermal and endodermal development in ascidian larvae and postlarvae. Development Genes and Evolution 208, 336-345.

Hinman, V.F., Nguyen, A.T., Cameron, R.A., Davidson, E.H., 2003. Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. Proceedings of the National Academy of Sciences of the United States of America 100, 13356-13361.

Hino, K., Satou, Y., Yagi, K., Satoh, N., 2003. A genomewide survey of developmentally relevant genes in Ciona intestinalis - VI. Genes for Wnt, TGF beta, Hedgehog and JAK/STAT signaling pathways. Development Genes and Evolution 213, 264-272.

Hirano, T., Nishida, H., 1997. Developmental fates of larval tissues after metamorphosis in ascidian Halocynthia roretzi. I. Origin of mesodermal tissues of the juvenile. Dev Biol 192, 199-210.

Hobmayer, B., Rentzsch, F., Kuhn, K., Happel, C.M., von Laue, C.C., Snyder, P., Rothbacher, U., Holstein, T.W., 2000. WNT signalling molecules act in axis formation in the diploblastic metazoan Hydra. Nature 407, 186-189.

Hoffmann, F.G., Opazo, J.C., Storz, J.F., 2012. Whole-Genome Duplications Spurred the Functional Diversification of the Globin Gene Superfamily in Vertebrates. Molecular Biology and Evolution 29, 303-312.

Holland, L.Z., 2006. A SINE in the genome of the cephalochordate amphioxus is an Alu element. International Journal of Biological Sciences 2, 61-65.

Holland, L.Z., Albalat, R., Azumi, K., Benito-Gutiérrez, È., Blow, M.J., Bronner-Fraser, M., Brunet, F., Butts, T., Candiani, S., Dishaw, L.J., Ferrier, D.E.K., Garcia-Fernàndez, J., Gibson-Brown, J.J., Gissi, C., Godzik, A., Hallböök, F., Hirose, D., Hosomichi, K., Ikuta, T., Inoko, H., Kasahara, M., Kasamatsu, J., Kawashima, T., Kimura, A., Kobayashi, M., Kozmik, Z., Kubokawa, K., Laudet, V., Litman, G.W., McHardy, A.C., Meulemans, D., Nonaka, M., Olinski, R.P., Pancer, Z., Pennacchio, L.A., Pestarino, M., Rast, J.P., Rigoutsos, I., Robinson-Rechavi, M., Roch, G., Saiga, H., Sasakura, Y., Satake, M., Satou, Y., Schubert, M., Sherwood, N., Shiina, T., Takatori, N., Tello, J., Vopalensky, P., Wada, S., Xu, A., Ye, Y., Yoshida, K., Yoshizaki, F., Yu, J.-K., Zhang, Q., Zmasek, C.M., de Jong, P.J., Osoegawa, K., Putnam, N.H., Rokhsar, D.S., Satoh, N., Holland, P.W.H., 2008. The amphioxus genome illuminates vertebrate origins and cephalochordate biology. Genome Research 18, 1100-1111.

Holland, L.Z., Carvalho, J.E., Escriva, H., Laudet, V., Schubert, M., Shimeld, S.M., Yu, J.K., 2013. Evolution of bilaterian central nervous systems: a single origin? Evodevo 4, 20.

Holland, L.Z., Holland, N.D., 1996. Expression of AmphiHox-1 and AmphiPax-1 in amphioxus embryos treated with retinoic acid: Insights into evolution and patterning of the chordate nerve cord and pharynx. Development 122, 1829-1838.

Holland, L.Z., Holland, N.D., 1998. Developmental gene expression in amphioxus: New insights into the evolutionary origin of vertebrate brain regions, neural crest, and rostrocaudal segmentation. American Zoologist 38, 647-658.

Holland, L.Z., Holland, N.D., 2007. A revised fate map for amphioxus and the evolution of axial patterning in chordates. Integrative and Comparative Biology 47, 360-372.

Holland, L.Z., Holland, P.W.H., Holland, N.D., 1996. Revealing homologies between body parts of distantly related animals by in situ hybridization to developmental genes: Amphioxus versus vertebrates, in: Ferraris, J.D., Palumbi, S.R. (Eds.), Molecular zoology: Advances, strategies, and protocols. Wiley-Liss, Inc., 605 Third Avenue, New York, New York 10158-0012, USA; Wiley-Liss, Ltd., Chichester, England, pp. 267-282.

Holland, L.Z., Laudet, V., Schubert, M., 2004. The chordate amphioxus: an emerging model organism for developmental biology. Cellular and Molecular Life Sciences 61, 2290-2308.

Holland, L.Z., Schubert, M., Kozmik, Z., Holland, N.D., 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, 153-165.

Holland, N.D., Holland, L.Z., Heimberg, A., 2015. Hybrids Between the Florida Amphioxus (Branchiostoma floridae) and the Bahamas Lancelet (Asymmetron lucayanum): Developmental Morphology and Chromosome Counts. Biological Bulletin 228, 13-24.

Holland, P.W., Garcia-Fernandez, J., Williams, N.A., Sidow, A., 1994. Gene duplications and the origins of vertebrate development. Dev Suppl, 125-133.

Holland, P.W., Holland, L.Z., Williams, N.A., Holland, N.D., 1992. An amphioxus homeobox gene: sequence conservation, spatial expression during development and insights into vertebrate evolution. Development 116, 653-661.

Holland, P.W.H., Garcia-Fernàndez, J., 1996. Hox genes and chordate evolution. Developmental Biology 173, 382-395.

Hoskins, R.A., Carlson, J.W., Wan, K.H., Park, S., Mendez, I., Galle, S.E., Booth, B.W., Pfeiffer, B.D., George, R.A., Svirskas, R., Krzywinski, M., Schein, J., Accardo, M.C., Damia, E., Messina, G., Mendez-Lago, M., de Pablos, B., Demakova, O.V., Andreyeva, E.N., Boldyreva, L.V., Marra, M., Carvalho, A.B., Dimitri, P., Villasante, A., Zhimulev, I.F., Rubin, G.M., Karpen, G.H., Celniker, S.E., 2015. The Release 6 reference sequence of the Drosophila melanogaster genome. Genome Research 25, 445-458.

Hotta, K., Mitsuhara, K., Takahashi, H., Inaba, K., Oka, K., Gojobori, T., Ikeo, K., 2007. A web-based interactive developmental table for the ascidian Ciona intestinalis, including 3D real-image embryo reconstructions: I. From fertilized egg to hatching larva. Developmental Dynamics 236, 1790-1805.

Houle, M., Prinos, P., Iulianella, A., Bouchard, N., Lohnes, D., 2000. Retinoic acid regulation of Cdx1: an indirect mechanism for retinoids and vertebral specification. Molecular and Cellular Biology 20, 6579-6586.

Houle, M., Sylvestre, J.R., Lohnes, D., 2003. Retinoic acid regulates a subset of Cdx1 function in vivo. Development 130, 6555-6567.

Hozumi, A., Yoshida, R., Horie, T., Sakuma, T., Yamamoto, T., Sasakura, Y., 2013. Enhancer activity sensitive to the orientation of the gene it regulates in the chordategenome. Developmental Biology 375, 79-91.

Hsiehli, H.M., Witte, D.P., Szucsik, J.C., Weinstein, M., Li, H., Potter, S.S., 1995. Gsh-2, A murine Homeobox gene expressed in the developing brain. Mechanisms of Development 50, 177-186.

Huang, L., Li, X.T., El-Hodiri, H.M., Dayal, S., Wikramanayake, A.H., Klein, W.H., 2000. Involvement of Tcf/Lef in establishing cell types along the animal-vegetal axis of sea urchins. Development Genes and Evolution 210, 73-81.

Huang, S.F., Chen, Z.L., Huang, G.R., Yu, T., Yang, P., Li, J., Fu, Y.G., Yuan, S.C., Chen, S.W., Xu, A.L., 2012. HaploMerger: Reconstructing allelic relationships for polymorphic diploid genome assemblies. Genome Research 22, 1581-1588.

Huang, S.F., Chen, Z.L., Yan, X.Y., Yu, T., Huang, G.R., Yan, Q.Y., Pontarotti, P.A., Zhao, H.C., Li, J., Yang, P., Wang, R.H., Li, R., Tao, X., Deng, T., Wang, Y.Q., Li, G., Zhang, Q.J., Zhou, S.S., You, L.M., Yuan, S.C., Fu, Y.G., Wu, F.F., Dong, M.L., Chen, S.W., Xu, A.L., 2014. Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes. Nature Communications 5, 12.

Huber, O., Korn, R., McLaughlin, J., Ohsugi, M., Herrmann, B.G., Kemler, R., 1996. Nuclear localization of beta-catenin by interaction with transcription factor LEF-1. Mechanisms of Development 59, 3-10.

Hudson, C., Lemaire, P., 2001. Induction of anterior neural fates in the ascidian Ciona intestinalis. Mechanisms of Development 100, 189-203.

Hudson, C., Lotito, S., Yasuo, H., 2007. Sequential and combinatorial inputs from Nodal, Delta2/Notch and FGF/MEK/ERK signalling pathways establish a grid-like organisation of distinct cell identities in the ascidian neural plate. Development 134, 3527-3537.

Hui, J.H.L., Griffiths-Jones, S., Ronshaugen, M., 2009a. The evolution of miR-10 family function in animal development. Mechanisms of Development 126, S56-S56.

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

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

Hui, J.H.L., Raible, F., Korchagina, N., Dray, N., Samain, S., Magdelenat, G., Jubin, C., Segurens, B., Balavoine, G., Arendt, D., Ferrier, D.E.K., 2009b. Features of the ancestral bilaterian inferred from Platynereis dumerilii ParaHox genes. Bmc Biology 7, 13.

Ikeya, M., Lee, S.M.K., Johnson, J.E., McMahon, A.P., Takada, S., 1997. Wnt signalling required for expansion of neural crest and CNS progenitors. Nature 389, 966-970.

Ikeya, M., Takada, S., 2001. Wnt-3a is required for somite specification along the anteroposterior axis of the mouse embryo and for regulation of cdx-1 expression. Mechanisms of Development 103, 27-33.

Ikuta, T., Chen, Y.-C., Annunziata, R., Ting, H.-C., Tung, C.-h., Koyanagi, R., Tagawa, K., Humphreys, T., Fujiyama, A., Saiga, H., Satoh, N., Yu, J.-K., Arnone, M.I., Su, Y.-H., 2013. Identification of an intact ParaHox cluster with temporal colinearity but altered spatial colinearity in the hemichordate Ptychodera flava. Bmc Evolutionary Biology 13.

Ikuta, T., Satoh, N., Saiga, H., 2010. Limited functions of Hox genes in the larval development of the ascidian Ciona intestinalis. Development 137, 1505-1513.

Ikuta, T., Yoshida, N., Satoh, N., Saiga, H., 2004. Ciona intestinalis Hox gene cluster: Its dispersed structure and residual colinear expression in development. Proceedings of the National Academy of Sciences of the United States of America 101, 15118-15123.

Illes, J.C., Winterbottom, E., Isaacs, H.V., 2009. Cloning and Expression Analysis of the Anterior ParaHox Genes, Gsh1 and Gsh2 From Xenopus tropicalis. Developmental Dynamics 238, 194-203.

Imai, K.S., Hino, K., Yagi, K., Satoh, N., Satou, Y., 2004. Gene expression profiles of transcription factors and signaling molecules in the ascidian embryo: towards a comprehensive understanding of gene networks. Development 131, 4047-4058.

Imai, K.S., Stolfi, A., Levine, M., Satou, Y., 2009. Gene regulatory networks underlying the compartmentalization of the Ciona central nervous system. Development 136, 285-293.

Inoue, H., Nojima, H., Okayama, H., 1990. High-efficiency transformation of Escherichia coli with Plasmids. Gene 96, 23-28.

Irvine, S.Q., Fonseca, V.C., Zompa, M.A., Antony, R., 2008. Cis-regulatory organization of the Pax6 gene in the ascidian Ciona intestinalis. Developmental Biology 317, 649-659.

Isaacs, H.V., Andreazzoli, M., Slack, J.M.W., 1999. Anteroposterior patterning by mutual repression of orthodenticle and caudal-type transcription factors. Evolution & Development 1, 143-152.

Isaacs, H.V., Pownall, M.E., Slack, J.M.W., 1998. Regulation of Hox gene expression and posterior development by the Xenopus caudal homologue Xcad3. Embo Journal 17, 3413-3427.

Islam, A., Moly, P.K., Miyamoto, Y., Kusakabe, T.G., 2010. Distinctive Expression Patterns of Hedgehog Pathway Genes in the Ciona intestinalis Larva: Implications for a Role of Hedgehog Signaling in Postembryonic Development and Chordate Evolution. Zoological Science 27, 84-90.

Iwai, T., Yoshii, A., Yokota, T., Sakai, C., Hori, H., Kanamori, A., Yamashita, M., 2006. Structural components of the synaptonemal complex, SYCP1 and SYCP3, in the medaka fish Oryzias latipes. Experimental Cell Research 312, 2528-2537.

Iwata, T.N., Cowley, T.J., Sloma, M., Ji, Y.W., Qi, L., Lee, S.S., 2013. The Transcriptional Co-Regulator HCF-1 Is Required for INS-1 beta-cell Glucose-Stimulated Insulin Secretion. Plos One 8, 6.

Jackman, W.R., Langeland, J.A., Kimmel, C.B., 2000. islet reveals segmentation in the amphioxus hindbrain homolog. Developmental Biology 220, 16-26.

Jaillon, O., Aury, J.M., Brunet, F., Petit, J.L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, R., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biemont, C., Skalli, Z., Cattolico, L., Poulain, J., de Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J.P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K.J., McEwan, P., Bosak, S., Kellis, M., Volff, J.N., Guigo, R., Zody, M.C., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quetier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E.S., Weissenbach, J., Crollius, H.R., 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431, 946-957.

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

Janvier, P., 2003. Vertebrate characters and the Cambrian vertebrates. Comptes Rendus Palevol 2, 523-531.

Jeffery, W.R., Swalla, B.J., 1997. Tunicates. Embryology: constructing the organism., 331-364.

Jimenez-Guri, E., Paps, J., Garcia-Fernandez, J., Salo, E., 2006. Hox and ParaHox genes in Nemertodermatida, a basal bilaterian clade. International Journal of Developmental Biology 50, 675-679.

Jin, T.R., Li, H.Q., 2001. POU homeodomain protein OCT1 is implicated in the expression of the caudal-related homeobox gene Cdx-2. Journal of Biological Chemistry 276, 14752-14758.

Jin, Z.-X., Kishi, H., Wei, X.-C., Matsuda, T., Saito, S., Muraguchi, A., 2002. Lymphoid Enhancer-Binding Factor-1 Binds and Activates the Recombination-Activating Gene-2 Promoter Together with c-Myb and Pax-5 in Immature B Cells. The Journal of Immunology 169, 3783-3792.

Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B., 2007. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science 316, 1497-1502.

Jonsson, J., Carlsson, L., Edlund, T., Edlund, H., 1994. Insulin-Promoter-Factor-1 is required for Pancreas development in mice. Nature 371, 606-609.

Jordan, I.K., Rogozin, I.B., Glazko, G.V., Koonin, E.V., 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends in Genetics 19, 68-72.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research 110, 462-467.

Jurka, J., Klonowski, P., Dagman, V., Pelton, P., 1996. Censor - A program for identification and elimination of repetitive elements from DNA sequences. Computers & Chemistry 20, 119-121.

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

Kanda, M., Ikeda, T., Fujiwara, S., 2013. Identification of a retinoic acid-responsive neural enhancer in the Ciona intestinalis Hox1 gene. Development Growth & Differentiation 55, 260-269.

Kang, L.F., Zhu, Z.L., Zhao, Q., Chen, L.Y., Zhang, Z., 2012. Newly evolved introns in human retrogenes provide novel insights into their evolutionary roles. Bmc Evolutionary Biology 12, 10.

Karch, F., Weiffenbach, B., Peifer, M., Bender, W., Duncan, I., Celniker, S., Crosby, M., Lewis, E.B., 1985. The abdominal region of the bithorax complex. Cell 43, 81-96.

Katsuyama, Y., Sato, Y., Wada, S., Saiga, H., 1999. Ascidian tail formation requires caudal function. Developmental Biology 213, 257-268.

Katsuyama, Y., Wada, S., Yasugi, S., Saiga, H., 1995. Expression of the Labial group Hox gene HrHox-1 and its alteration induced by retinoic acid in development of the ascidian Halocynthia roretzi. Development 121, 3197-3205.

Keenan, I.D., Sharrard, R.M., Isaacs, H.V., 2006. FGF signal transduction and the regulation of Cdx gene expression. Developmental Biology 299, 478-488.

Keller, D.M., McWeeney, S., Arsenlis, A., Drouin, J., Wright, C.V.E., Wang, H., Wollheim, C.B., White, P., Kaestner, K.H., Goodman, R.H., 2007. Characterization of pancreatic transcription factor pdx-1 binding sites using promoter Microarray and serial analysis of chromatin occupancy. Journal of Biological Chemistry 282, 32084-32092.

Kellum, R., Schedl, P., 1992. A group of SCS elements functions as domain boundaries in an enhancer-blocking assay. Molecular and Cellular Biology 12, 2424-2431.

Kennedy-Darling, J., Holden, M.T., Shortreed, M.R., Smith, L.M., 2014. Multiplexed Programmable Release of Captured DNA. ChemBioChem 15, 2353-2356.

Kessel, M., 1992. Respecification of vertebral identities by retinoic acid. Development 115, 487-501.

Kidwell, M.G., 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica 115, 49-63.

Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A.Z., Engstrom, P.G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K., Ghislain, J., Pezeron, G., Mourrain, P., Ellingsen, S., Oates, A.C., Thisse, C., Thisse, B., Foucher, I., Adolf, B., Geling, A., Lenhard, B., Becker, T.S., 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Research 17, 545-555.

Kim, D.S., Wang, Y., Oh, H.J., Choi, D., Lee, K., Hahn, Y., 2014. Retroduplication and loss of parental genes is a mechanism for the generation of intronless genes in Ciona intestinalis and Ciona savignyi. Development Genes and Evolution 224, 255-260.

Kim, J.H., Waterman, M.S., Li, L.M., 2007a. Diploid genome reconstruction of Ciona intestinalis and comparative analysis with Ciona savignyi. Genome Research 17, 1101-1110.

Kim, S., Domon-Dell, C., Wang, Q.D., Chung, D.H., Di Cristofano, A., Pandolfi, P.P., Freund, J.N., Evers, B.M., 2002. PTEN and TNF-alpha regulation of the intestinal-specific Cdx-2 homeobox gene through a PI3K, PKB/Akt, and NF-kappa B-dependent pathway. Gastroenterology 123, 1163-1178.

Kim, S.K., Hebrok, M., Melton, D.A., 1997. Notochord to endoderm signaling is required for pancreas development. Development 124, 4243-4252.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., Ren, B., 2007b. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell 128, 1231-1245.

Kimble, J., Page, D.C., 2007. The mysteries of sexual identity: The germ cell's perspective. Science 316, 400-401.

Kinkel, M.D., Eames, S.C., Alonzo, M.R., Prince, V.E., 2008. Cdx4 is required in the endoderm to localize the pancreas and limitβ -cell number. Development 135, 919-929.

Kino, T., Souvatzoglou, E., Charmandari, E., Ichijo, T., Driggers, P., Mayers, C., Alatsatianos, A., Manoli, I., Westphal, H., Chrousos, G.P., Segars, J.H., James, H.S., 2006. Rho family guanine nucleotide exchange factor Brx couples extracellular signals to the glucocorticoid signaling system. Journal of Biological Chemistry 281, 9118-9126.

Kleene, K.C., Mulligan, E., Steiger, D., Donohue, K., Mastrangelo, M.A., 1998. The mouse gene encoding the testis-specific isoform of poly(A) binding protein (Pabp2) is an expressed retroposon: Intimations that gene expression in spermatogenic cells facilitates the creation of new genes. Journal of Molecular Evolution 47, 275-281.

Kmita, M., Kondo, T., Duboule, D., 2000. Targeted inversion of a polar silencer within the HoxD complex re-allocates domains of enhancer sharing. Nature Genetics 26, 451-454.

Kmita, M., Tarchini, B., Duboule, D., Herault, Y., 2002. Evolutionary conserved sequences are required for the insulation of the vertebrate Hoxd complex in neural cells. Development 129, 5521-5528.

Kon, T., Nohara, M., Yamanoue, Y., Fujiwara, Y., Nishida, M., Nishikawa, T., 2007. Phylogenetic position of a whale-fall lancelet (Cephalochordata) inferred from whole mitochondrial genome sequences. Bmc Evolutionary Biology 7, 12.

Koop, D., Holland, N.D., Semon, M., Alvarez, S., Rodriguez de Lera, A., Laudet, V., Holland, L.Z., Schubert, M., 2010. Retinoic acid signaling targets Hox genes during the amphioxus gastrula stage: Insights into early anterior-posterior patterning of the chordate body plan. Developmental Biology 338, 98-106.

Koressaar, T., Remm, M., 2007. Enhancements and modifications of primer design program Primer3. Bioinformatics 23, 1289-1291.

Korinek, V., Barker, N., Willert, K., Molenaar, M., Roose, J., Wagenaar, G., Markman, M., Lamers, W., Destree, O., Clevers, H., 1998. Two members of the Tcf family implicated in Wnt/beta-catenin signaling during embryogenesis in the mouse. Molecular and Cellular Biology 18, 1248-1256.

Kosaka-Suzuki, N., Suzuki, T., Pugacheva, E.M., Vostrov, A.A., Morse, H.C., Loukinov, D., Lobanenkov, V., 2011. Transcription Factor BORIS (Brother of the Regulator of Imprinted Sites) Directly Induces Expression of a Cancer-Testis Antigen, TSP50, through Regulated Binding of BORIS to the Promoter. Journal of Biological Chemistry 286, 27378-27388.

Kostrouch, Z., Kostrouchova, M., Love, W., Jannini, E., Piatigorsky, J., Rall, J.E., 1998. Retinoic acid X receptor in the diploblast, Tripedalia cystophora. Proceedings of the National Academy of Sciences of the United States of America 95, 13442-13447.

Kozmik, Z., Holland, N.D., Kalousova, A., Paces, J., Schubert, M., Holland, L.Z., 1999. Characterization of an amphioxus paired box gene, AmphiPax2/5/8: developmental expression patterns in optic support cells, nephridium, thyroid-like structures and pharyngeal gill slits, but not in the midbrain-hindbrain boundary region. Development 126, 1295-1304.

Krasnov, A.N., Kurshakova, M.M., Ramensky, V.E., Mardanov, P.V., Nabirochkina, E.N., Georgieva, S.G., 2005. A retrocopy of a gene can functionally displace the source gene in evolution. Nucleic Acids Research 33, 6654-6661.

Kriks, S., Lanuza, G.M., Mizuguchi, R., Nakafuku, M., Goulding, M., 2005. Gsh2 is required for the repression of Ngn1 and specification of dorsal interneuron fate in the spinal cord. Development 132, 2991-3002.

Kulakova, M.A., Cook, C.E., Andreeva, T.F., 2008. ParaHox gene expression in larval and postlarval development of the polychaete Nereis virens (Annelida, Lophotrochozoa). Bmc Developmental Biology 8.

Kumano, G., Nishida, H., 2007. Ascidian embryonic development: an emerging model system for the study of cell fate specification in chordates. Dev Dyn 236, 1732-1747.

Kumar, M., Jordan, N., Melton, D., Grapin-Botton, A., 2003. Signals from lateral plate mesoderm instruct endoderm toward a pancreatic fate. Developmental Biology 259, 109-122.

Kurimoto, K., Yabuta, Y., Ohinata, Y., Shigeta, M., Yamanaka, K., Saitou, M., 2008. Complex genome-wide transcription dynamics orchestrated by Blimp1 for the specification of the germ cell lineage in mice. Genes & Development 22, 1617-1635.

Kusakabe, T., Yoshida, R., Kawakami, I., Kusakabe, R., Mochizuki, Y., Yamada, L., Shin-i, T., Kohara, Y., Satoh, N., Tsuda, M., Satou, Y., 2002. Gene Expression Profiles in Tadpole Larvae of Ciona intestinalis. Developmental Biology 242, 188-203.

Kusserow, A., Pang, K., Sturm, C., Hrouda, M., Lentfer, J., Schmidt, H.A., Technau, U., von Haeseler, A., Hobmayer, B., Martindale, M.Q., Holstein, T.W., 2005. Unexpected complexity of the Wnt gene family in a sea anemone. Nature 433, 156-160.

Lacalli, T., 2003. Developmental biology: A larval revelation. Nature 421, 120-121.

Lander, E.S., Int Human Genome Sequencing, C., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H.M., Yu, J., Wang, J., Huang, G.Y., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S.Z., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H.Q., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W.H., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J.R., Slater, G., Smit, A.F.A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., Int Human Genome Sequencing, C., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Lanfear, R., Bromham, L., 2008. Statistical Tests between Competing Hypotheses of Hox Cluster Evolution. Systematic Biology 57, 708-718.

Langston, A.W., Thompson, J.R., Gudas, L.J., 1997. Retinoic acid-responsive enhancers located 3' of the Hox A and Hox B homeobox gene clusters - Functional analysis. Journal of Biological Chemistry 272, 2167-2175.

Lapebie, P., Gazave, E., Ereskovsky, A., Derelle, R., Bezac, C., Renard, E., Houliston, E., Borchiellini, C., 2009. WNT/beta-Catenin Signalling and Epithelial Patterning in the Homoscleromorph Sponge Oscarella. Plos One 4.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and clustal X version 2.0. Bioinformatics 23, 2947-2948.

Le Gouar, M., Lartillot, N., Adoutte, A., Vervoort, M., 2003. The expression of a caudal homologue in a mollusc, Patella vulgata. Gene Expression Patterns 3, 35-37.

Lee, P.N., Pang, K., Matus, D.Q., Martindale, M.Q., 2006. A WNT of things to come: Evolution of Wnt signaling and polarity in cnidarians. Seminars in Cell & Developmental Biology 17, 157-167.

Lee, T.I., Young, R.A., 2000. Transcription of eukaryotic protein-coding genes. Annual Review of Genetics 34, 77-137.

Lee, Y.J., Swencki, B., Shoichet, S., Shivdasani, R.A., 1999. A possible role for the high mobility group box transcription factor Tcf-4 in vertebrate gut epithelial cell differentiation. Journal of Biological Chemistry 274, 1566-1572.

Lehoczky, J.A., Williams, M.E., Innis, J.W., 2004. Conserved expression domains for genes upstream and within the HoxA and HoxD clusters suggests a long-range enhancer existed before cluster duplication. Evolution & Development 6, 423-430.

Leininger, S., Adamski, M., Bergum, B., Guder, C., Liu, J., Laplante, M., Brate, J., Hoffmann, F., Fortunato, S., Jordal, S., Rapp, H.T., Adamska, M., 2014. Developmental gene expression provides clues to relationships between sponge and eumetazoan body plans. Nature Communications 5.

Lemaire, P., 2011. Evolutionary crossroads in developmental biology: the tunicates. Development 138, 2143-2152.

Lengerke, C., Schmitt, S., Bowman, T.V., Jang, I.H., Maouche-Chretien, L., McKinney-Freeman, S., Davidson, A.J., Hammerschmidt, M., Rentzsch, F., Green, J.B.A., Zon, L.I., Daley, G.Q., 2008. BMP and Wnt specify hematopoietic fate by activation of the Cdx-Hox pathway. Cell Stem Cell 2, 72-82.

Lengronne, A., Katou, Y., Mori, S., Yokabayashi, S., Kelly, G.P., Ito, T., Watanabe, Y., Shirahige, K., Uhlmann, F., 2004. Cohesin relocation from sites of chromosomal loading to places of convergent transcription. Nature 430, 573-578.

Levine, M., Cattoglio, C., Tjian, R., 2014. Looping Back to Leap Forward: Transcription Enters a New Era. Cell 157, 13-25.

Lewis, E.B., 1978. A gene complex controlling segmentation in Drosophila. Nature 276, 565-570.

Li, H., Zeitler, P.S., Valerius, M.T., Small, K., Potter, S.S., 1996. Gsh-1, an orphan Hox gene, is required for normal pituitary development. Embo Journal 15, 714-724.

Li, X.L., Cao, X., 2003. BMP signaling and Hox transcription factors in limb development. Frontiers in Bioscience 8, S805-S812.

Lickert, H., Domon, C., Huls, G., Wehrle, C., Duluc, I., Clevers, H., Meyer, B.I., Freund, J.N., Kemler, R., 2000. Wnt/beta-catenin signaling regulates the expression of the homeobox gene Cdx1 in embryonic intestine. Development 127, 3805-3813.

Lickert, H., Kemler, R., 2002. Functional analysis of cis-regulatory elements controlling initiation and maintenance of early Cdx1 gene expression in the mouse. Developmental Dynamics 225, 216-220.

Lin, H.-C., Holland, L.Z., Holland, N.D., 2006. Expression of the AmphiTcf gene in amphioxus: Insights into the evolution of the TCF/LEF gene family during vertebrate evolution. Developmental Dynamics 235, 3396-3403.

Liu, J.G., Yuan, L., Brundell, E., Bjorkroth, B., Daneholt, B., Hoog, C., 1996. Localization of the N-terminus of SCP1 to the central element of the synaptonemal complex and evidence for direct interactions between the N-termini of SCP1 molecules organized head-to-head. Experimental Cell Research 226, 11-19.

Lobanenkov, V.V., Nicolas, R.H., Adler, V.V., Paterson, H., Klenova, E.M., Polotskaja, A.V., Goodwin, G.H., 1990. A Novel Squence-specific DNA-binding Protein which Interacts with 3 Regularly Spaced Direct Repeats of the CCCTC-motif in the 5'-flanking Sequence of the Chicken C-MYC Gene. Oncogene 5, 1743-1753.

Long, M., Langley, C.H., 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science 260, 91-95.

Loose, M., Patient, R., 2004. A genetic regulatory network for Xenopus mesendoderm formation. Developmental Biology 271, 467-478.

Love, J.J., Li, X., Case, D.A., Giese, K., Grosschedl, R., Wright, P.E., 1995. Structural basis for DNA bending by the architectural transcription factor LEF-1. Nature 376, 791-795.

Lytle, J.R., Yario, T.A., Steitz, J.A., 2007. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′ UTR as in the 3′ UTR. Proceedings of the National Academy of Sciences 104, 9667-9672.

Ma, J., Chen, M., Wang, J., Xia, H.H.X., Zhu, S., Liang, Y., Gu, Q., Qiao, L., Dai, Y., Zou, B., Li, Z., Zhang, Y., Lan, H., Wong, B.C.Y., 2008. Pancreatic duodenal homeobox-1 (PDX1) functions as a tumor suppressor in gastric cancer. Carcinogenesis 29, 1327-1333.

Macdonald, P.M., Struhl, G., 1986. A molecular gradient in early Drosophila embryos and its role in specifying the body pattern. Nature 324, 537-545.

MacQueen, A.J., Colaiácovo, M.P., McDonald, K., Villeneuve, A.M., 2002. Synapsis-dependent and -independent mechanisms stabilize homolog pairing during meiotic prophase in C. elegans. Genes & Development 16, 2428-2442.

Maeso, I., Irimia, M., Tena, J.J., Gonzalez-Perez, E., Tran, D., Ravi, V., Venkatesh, B., Campuzano, S., Gomez-Skarmeta, J.L., Garcia-Fernandez, J., 2012. An ancient genomic regulatory block conserved across bilaterians and its dismantling in tetrapods by retrogene replacement. Genome Research 22, 642-655.

Makunin, I.V., Shloma, V.V., Stephen, S.J., Pheasant, M., Belyakin, S.N., 2013. Comparison of Ultra-Conserved Elements in Drosophilids and Vertebrates. Plos One 8, 12.

Mallatt, J., Chen, J.Y., 2003. Fossil sister group of craniates: Predicted and found. Journal of Morphology 258, 1-31.

Mallatt, J., Holland, N., 2013. Pikaia gracilens Walcott: Stem Chordate, or Already Specialized in the Cambrian? Journal of Experimental Zoology Part B-Molecular and Developmental Evolution 320B, 247-271.

Manzanares, M., Wada, H., Itasaki, N., Trainor, P.A., Krumlauf, R., Holland, P.W.H., 2000. Conservation and elaboration of Hox gene regulation during evolution of the vertebrate head. Nature 408, 854-857.

Marlor, R.L., Parkhurst, S.M., Corces, V.G., 1986. The *Drosophila melanogaster Gypsy* Transposable Element encodes putative gene-products homologous to Retroviral proteins. Molecular and Cellular Biology 6, 1129-1134.

Marom, K., Shapira, E., Fainsod, A., 1997. The chicken caudal genes establish an anterior-posterior gradient by partially overlapping temporal and spatial patterns of expression. Mechanisms of Development 64, 41-52.

Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A., Kaessmann, H., 2005. Emergence of young human genes after a burst of retroposition in primates. Plos Biology 3, 1970-1979.

Marshak, S., Benshushan, E., Shoshkes, M., Havin, L., Cerasi, E., Melloul, D., 2000. Functional conservation of regulatory elements in the pdx-1 gene: PDX-1 and hepatocyte nuclear factor 3 beta transcription factors mediate beta-cell-specific expression. Molecular and Cellular Biology 20, 7583-7590.

Marshall, H., Morrison, A., Studer, M., Popperl, H., Krumlauf, R., 1996. Retinoids and Hox genes. Faseb Journal 10, 969-978.

Marshall, H., Studer, M., Popperl, H., Aparicio, S., Kuroiwa, A., Brenner, S., Krumlauf, R., 1994. A conserved retinoic acid response element required for early expression of the homeobox gene Hoxb-1. Nature 370, 567-571.

Martin, B.L., Kimelman, D., 2009. Wnt Signaling and the Evolution of Embryonic Posterior Development. Current Biology 19, R215-R219.

Martin, M., Gallego-Llamas, J., Ribes, V., Kedinger, M., Niederreither, K., Chambon, P., Dolle, P., Gradwohl, G., 2005. Dorsal pancreas agenesis in retinoic acid-deficient Raldh2 mutant mice. Developmental Biology 284, 399-411.

Martinelli, C., Spring, J., 2004. Expression pattern of the homeobox gene Not in the basal metazoan Trichoplax adhaerens. Gene Expression Patterns 4, 443-447.

Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C.-y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., Wasserman, W.W., 2014. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Research 42, D142-D147.

Matsumura, K., Nishiyama, E., Ohshima, K., 2014. Identification of regulatory sequences involved in transcriptional regulation of a young retrogene ? Genes & Genetic Systems 89, 340-340.

Matsunami, M., Sumiyama, K., Saitou, N., 2010. Evolution of Conserved Non-Coding Sequences Within the Vertebrate Hox Clusters Through the Two-Round Whole Genome Duplications Revealed by Phylogenetic Footprinting Analysis. Journal of Molecular Evolution 71, 427-436.

Matsuo, K., Yoshida, H., Shimizu, T., 2005. Differential expression of caudal and dorsal genes in the teloblast lineages of the oligochaete annelid Tubifex tubifex. Development Genes and Evolution 215, 238-247.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E., Wingender, E., 2006. TRANSFAC (R) and its module TRANSCompel (R): transcriptional gene regulation in eukaryotes. Nucleic Acids Research 34, D108-D110.

Mavilio, F., Simeone, A., Boncinelli, E., Andrews, P.W., 1988. Activation of four homeobox gene clusters in human embryonal carcinoma cells induced to differentiate by retinoic acid. Differentiation 37, 73-79.

Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., Dubchak, I., 2000. VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16, 1046-1047.

Mazet, F., Hutt, J.A., Millard, J., Shimeld, S.M., 2003. Pax gene expression in the developing central nervous system of Ciona intestinalis. Gene Expression Patterns 3, 743-745.

Mazzoni, E.O., Mahony, S., Peljto, M., Patel, T., Thornton, S.R., McCuine, S., Reeder, C., Boyer, L.A., Young, R.A., Gifford, D.K., Wichterle, H., 2013. Saltatory remodeling of Hox chromatin in response to rostrocaudal patterning signals. Nature Neuroscience 16, 1191-U1150.

McCarrey, J.R., Thomas, K., 1987. Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. Nature 326, 501-505.

McCarrey, J.R., Watson, C., Atencio, J., Ostermeier, G.C., Marahrens, Y., Jaenisch, R., Krawetz, S.A., 2002. X-chromosome inactivation during spermatogenesis is regulated by an Xist/Tsix-independent mechanism in the mouse. genesis 34, 257-266.

McGlinn, E., Yekta, S., Mansfield, J.H., Soutschek, J., Bartel, D.P., Tabin, C.J., 2009. In ovo application of antagomiRs indicates a role for miR-196 in patterning the chick axial skeleton through Hox gene regulation. Proceedings of the National Academy of Sciences of the United States of America 106, 18610-18615.

McGregor, A.P., Pechmann, M., Schwager, E.E., Feitosa, N.M., Kruck, S., Aranda, M., Damen, W.G.M., 2008. Wnt8 Is Required for Growth-Zone Establishment and Development of Opisthosomal Segments in a Spider. Current Biology 18, 1619-1623.

McGregor, A.P., Pechmann, M., Shwager, E.E., Damen, W.G.M., 2009. An ancestral regulatory network for posterior development in arthropods. Communicative & Integrative Biology 2, 174-176.

McLin, V.A., Rankin, S.A., Zorn, A.M., 2007. Repression of Wnt/beta-catenin signaling in the anterior endoderm is essential for liver and pancreas development. Development 134, 2207-2217.

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

Messeguer, X., Escudero, R., Farre, D., Nunez, O., Martinez, J., Alba, M., 2002. PROMO: detection of known transcription regulatory elements using species-tailored searches. Bioinformatics 18, 333-334.

Meuwissen, R.L., Offenberg, H.H., Dietrich, A.J., Riesewijk, A., van Iersel, M., Heyting, C., 1992. A coiled-coil related protein specific for synapsed regions of meiotic prophase chromosomes. Embo j 11, 5091-5100.

Meuwissen, R.L.J., Meerts, I., Hoovers, J.M.N., Leschot, N.J., Heyting, C., 1997. Human synaptonemal complex protein 1 (SCP1): Isolation and characterization of the cDNA and chromosomal localization of the gene. Genomics 39, 377-384.

Meyer, A., Van de Peer, Y., 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). Bioessays 27, 937-945.

Meyer, B.I., Gruss, P., 1993. Mouse Cdx-1 expression during gastrulation. Development 117, 191-203.

Mighell, A.J., Smith, N.R., Robinson, P.A., Markham, A.F., 2000. Vertebrate pseudogenes. Febs Letters 468, 109-114.

Milewski, W.M., Duguay, S.J., Chan, S.J., Steiner, D.F., 1998. Conservation of PDX-1 structure, function, and expression in zebrafish. Endocrinology 139, 1440-1449.

Miller, W.J., McDonald, J.F., Nouaud, D., Anxolabéhère, D., 1999. Molecular domestication - More than a sporadic episode in evolution. Genetica 107, 197-207.

Minsuk, S.B., Raff, R.A., 2002. Pattern formation in a pentameral animal: Induction of early adult rudiment development in sea urchins. Developmental Biology 247, 335-350.

Miyatsuka, T., Matsuoka, T.-a., Shiraiwa, T., Yamamoto, T., Kojima, I., Kaneto, H., 2007. Ptf1a and RBP-J cooperate in activating Pdx1 gene expression through binding to Area III. Biochemical and Biophysical Research Communications 362, 905-909.

Molenaar, M., Roose, J., Peterson, J., Venanzi, S., Clevers, H., Destree, O., 1998. Differential expression of the HMG box transcription factors XTcf-3 and XLef-1 during early Xenopus development. Mechanisms of Development 75, 151-154.

Molotkov, A., Molotkova, N., Duester, G., 2005. Retinoic acid generated by Raldh2 in mesoderm is required for mouse dorsal Endodermal pancreas development. Developmental Dynamics 232, 950-957.

Moon, H., Filippova, G., Loukinov, D., Pugacheva, E., Chen, Q., Smith, S.T., Munhall, A., Grewe, B., Bartkuhn, M., Arnold, R., Burke, L.J., Renkawitz-Pohl, R., Ohlsson, R., Zhou, J.M., Renkawitz, R., Lobanenkov, V., 2005. CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator. Embo Reports 6, 165-170.

Moreno, E., Permanyer, J., Martinez, P., 2011. The Origin of Patterning Systems in Bilateria-Insights from the Hox and ParaHox Genes in Acoelomorpha. Genomics Proteomics & Bioinformatics 9, 65-76.

Moroz, L.L., Kocot, K.M., Citarella, M.R., Dosung, S., Norekian, T.P., Povolotskaya, I.S., Grigorenko, A.P., Dailey, C., Berezikov, E., Buckley, K.M., Ptitsyn, A., Reshetov, D., Mukherjee, K., Moroz, T.P., Bobkova, Y., Yu, F., Kapitonov, V.V., Jurka, J., Bobkov, Y.V., Swore, J.J., Girardo, D.O., Fodor, A., Gusev, F., Sanford, R., Bruders, R., Kittler, E., Mills, C.E., Rast, J.P., Derelle, R., Solovyev, V.V., Kondrashov, F.A., Swalla, B.J., Sweedler, J.V., Rogaev, E.I., Halanych, K.M., Kohn, A.B., 2014. The ctenophore genome and the evolutionary origins of neural systems. Nature 510, 109-114.

Morris, S.C., Caron, J.B., 2012. Pikaia gracilens Walcott, a stem-group chordate from the Middle Cambrian of British Columbia. Biological Reviews 87, 480-512.

Mulley, J.F., Chiu, C.H., Holland, P.W.H., 2006. Breakup of a homeobox cluster after genome duplication in teleosts. Proceedings of the National Academy of Sciences of the United States of America 103, 10369-10372.

Munro, E., Robin, F., Lemaire, P., 2006. Cellular morphogenesis in ascidians: how to shape a simple tadpole. Current Opinion in Genetics & Development 16, 399-405.

Nakano, H., Hibino, T., Oji, T., Hara, Y., Amemiya, S., 2003. Larval stages of a living sea lily (stalked crinoid echinoderm). Nature 0, 158-160.

Nakazawa, K., Yamazawa, T., Moriyama, Y., Ogura, Y., Kawai, N., Sasakura, Y., Saiga, H., 2013. Formation of the digestive tract in Ciona intestinalis includes two distinct morphogenic processes between its anterior and posterior parts. Developmental Dynamics 242, 1172-1183.

Narendra, V., Rocha, P.P., An, D.S., Raviram, R., Skok, J.A., Mazzoni, E.O., Reinberg, D., 2015. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. Science 347, 1017-1021.

Natale, A., Sims, C., Chiusano, M.L., Amoroso, A., D'Aniello, E., Fucci, L., Krumlauf, R., Branno, M., Locascio, A., 2011. Evolution of anterior Hox regulatory elements among chordates. Bmc Evolutionary Biology 11, 19.

Nohara, M., Nishida, M., Manthacitra, V., Nishikawa, T., 2004. Ancient Phylogenetic Separation between Pacific and Atlantic Cephalochordates as Revealed by Mitochondrial Genome Analysis. Zoological Science 21, 203-210.

Nohara, M., Nishida, M., Miya, M., Nishikawa, T., 2005. Evolution of the mitochondrial genome in Cephalochordata as inferred from complete nucleotide sequences from two Epigonichthys species. Journal of Molecular Evolution 60, 526-537.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N.L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., Heard, E., 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485, 381-385.

Nordstrom, U., Maier, E., Jessell, T.M., Edlund, T., 2006. An early role for Wnt signaling in specifying neural patterns of Cdx and Hox gene expression and motor neuron subtype identity. Plos Biology 4, 1438-1452.

Nowickyj, S.M., Chithalen, J.V., Cameron, D., Tyshenko, M.G., Petkovich, M., Wyatt, G.R., Jones, G., Walker, V.K., 2008. Locust retinoid X receptors: 9-Cis-retinoic acid in embryos from a primitive insect. Proceedings of the National Academy of Sciences of the United States of America 105, 9540-9545.

Oberhofer, G., Grossmann, D., Siemanowski, J.L., Beissbarth, T., Bucher, G., 2014. Wnt/beta-catenin signaling integrates patterning and metabolism of the insect growth zone. Development 141, 4740-4750.

Offenberg, H.H., Schalk, J.A.C., Meuwissen, R.L.J., van Aalderen, M., Kester, H.A., Dietrich, A.J.J., Heyting, C., 1998. SCP2: a major protein component of the axial elements of synaptonemal complexes of the rat. Nucleic Acids Research 26, 2572-2579.

Offield, M.F., Jetton, T.L., Labosky, P.A., Ray, M., Stein, R.W., Magnuson, M.A., Hogan, B.L.M., Wright, C.V.E., 1996. PDX-1 is required for pancreatic outgrowth and differentiation of the rostral duodenum. Development 122, 983-995.

Ohlsson, H., Karlsson, K., Edlund, T., 1993. IPF1, a homeodomain-containing transactivator of the insulin gene. The EMBO Journal 12, 4251-4259.

Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y., Okada, N., 2003. Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. Genome Biology 4, R74.

Olesnicky, E.C., Brent, A.E., Tonnes, L., Walker, M., Pultz, M.A., Leaf, D., Desplan, C., 2006. A caudal mRNA gradient controls posterior development in the wasp Nasonia. Development 133, 3973-3982.

Onai, T., Lin, H.C., Schubert, M., Koop, D., Osborne, P.W., Alvarez, S., Alvarez, R., Holland, N.D., Holland, L.Z., 2009. Retinoic acid and Wnt/beta-catenin have complementary roles in anterior/posterior patterning embryos of the basal chordate amphioxus. Developmental Biology 332, 223-233.

Ong, C.-T., Corces, V.G., 2014. CTCF: an architectural protein bridging genome topology and function. Nat Rev Genet 15, 234-246.

Oosterwegel, M., van de Wetering, M., Timmerman, J., Kruisbeek, A., Destree, O., Meijlink, F., Clevers, H., 1993. Differential expression of the HMG box factors TCF-1 and LEF-1 during murine embryogenesis. Development 118, 439-448.

Osborne, P.W., 2009. Evolution of Chordate ParaHox Gene Regulation, Department of Zoology. University of Oxford, Linacre College, Oxford, p. 298.

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

Osborne, P.W., Ferrier, D.E.K., 2010. Chordate Hox and ParaHox Gene Clusters Differ Dramatically in Their Repetitive Element Content. Molecular Biology and Evolution 27, 217-220.

Osborne, P.W., Luke, G.N., Holland, P.W.H., Ferrier, D.E.K., 2006. Identification and Characterisation of five novel Miniature Inverted-repeat Transposable Elements (MITEs) in amphioxus (Branchiostoma floridae). International Journal of Biological Sciences 2, 54-65.

Osigus, H.J., Eitel, M., Schierwater, B., 2013. Chasing the urmetazoon: striking a blow for quality data? Mol Phylogenet Evol 66, 551-557.

Ovcharenko, I., Loots, G.G., Giardine, B.M., Hou, M.M., Ma, J., Hardison, R.C., Stubbs, L., Miller, W., 2005. Mulan: Multiple-sequence local alignment and visualization for studying function and evolution. Genome Research 15, 184-194.

Page, S.L., Hawley, R.S., 2001. c(3)G encodes a Drosophila synaptonemal complex protein. Genes & Development 15, 3130-3143.

Page, S.L., Hawley, R.S., 2004. The genetics and molecular biology of the synaptonemal complex. Annual Review of Cell and Developmental Biology 20, 525-558.

Pai, C.Y., Lei, E.P., Ghosh, D., Corces, V.G., 2004. The centrolsomal protein CP190 is a component of the gypsy chromatin insulator. Molecular Cell 16, 737-748.

Panopoulou, G.D., Clark, M.D., Holland, L.Z., Lehrach, H., Holland, N.D., 1998. AmphiBMP2/4, an amphioxus bone morphogenetic protein closely related to Drosophila decapentaplegic and vertebrate BMP2 and BMP4: Insights into evolution of dorsoventral axis specification. Developmental Dynamics 213, 130-139.

Papadopoulou, S., Edlund, H., 2005. Attenuated Wnt Signaling Perturbs Pancreatic Growth but Not Pancreatic Function. Diabetes 54, 2844-2851.

Parker, H.G., VonHoldt, B.M., Quignon, P., Margulies, E.H., Shao, S., Mosher, D.S., Spady, T.C., Elkahloun, A., Cargill, M., Jones, P.G., Maslen, C.L., Acland, G.M., Sutter, N.B., Kuroki, K., Bustamante, C.D., Wayne, R.K., Ostrander, E.A., 2009. An Expressed Fgf4 Retrogene Is Associated with Breed-Defining Chondrodysplasia in Domestic Dogs. Science 325, 995-998.

Pascual-Anaya, J., D'Aniello, S., Garcia-Fernandez, J., 2008. Unexpectedly large number of conserved noncoding regions within the ancestral chordate Hox cluster. Development Genes and Evolution 218, 591-597.

Pelegri, F., Maischein, H.-M., 1998. Function of zebrafish β-catenin and TCF-3 in dorsoventral patterning. Mechanisms of Development 77, 63-74.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., Plajzer-Frick, I., Akiyama, J., De Val, S., Afzal, V., Black, B.L., Couronne, O., Eisen, M.B., Visel, A., Rubin, E.M., 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444, 499-502.

Perea-Atienza, E., Gavilan, B., Chiodin, M., Abril, J.F., Hoff, K.J., Poustka, A.J., Martinez, P., 2015. The nervous system of Xenacoelomorpha: a genomic perspective. Journal of Experimental Biology 218, 618-628.

Perez-Villamil, B., Schwartz, P.T., Vallejo, M., 1999. The pancreatic homeodomain transcription factor IDX1/IPF1 is expressed in neural cells during brain development. Endocrinology 140, 3857-3860.

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

Petrov, D.A., 2002. DNA loss and evolution of genome size in Drosophila. Genetica 115, 81-91.

Petrov, D.A., Hartl, D.L., 1998. High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. Molecular Biology and Evolution 15, 293-302.

Phan-Hug, F., Guimiot, F., Lelievre, V., Delezoide, A.-L., Czernichow, P., Breant, B., Blondeau, B., 2008. Potential role of glucocorticoid signaling in the formation of pancreatic islets in the human fetus. Pediatric Research 64, 346-351.

Philippe, H., Brinkmann, H., Copley, R.R., Moroz, L.L., Nakano, H., Poustka, A.J., Wallberg, A., Peterson, K.J., Telford, M.J., 2011. Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature 470, 255-+.

Phillips, J.E., Corces, V.G., 2009. CTCF: Master Weaver of the Genome. Cell 137, 1194-1211.

Pilon, N., Oh, K., Sylvestre, J.R., Bouchard, N., Savory, J., Lohnes, D., 2006. Cdx4 is a direct target of the canonical Wnt pathway. Developmental Biology 289, 55-63.

Pilon, N., Oh, K., Sylvestre, J.R., Savory, J.G.A., Lohnes, D., 2007. Wnt signaling is a key mediator of Cdx1 expression in vivo. Development 134, 2315-2323.

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

Popperl, H., Featherstone, M.S., 1993. Identification of a retinoic acid response element upstream of the murine Hox-4.2 gene. Mol Cell Biol 13, 257-265.

Prendergast, G.C., 2001. Actin' up: RhoB in cancer and apoptosis. Nature Reviews Cancer 1, 162-168.

Prestridge, D.S., 1995. Predicting Pol II promoter sequences using transcription factor binding sites. J Mol Biol 249, 923-932.

Pugacheva, E.M., Rivero-Hinojosa, S., Espinoza, C.A., Méndez-Catalá, C.F., Kang, S., Suzuki, T., Kosaka-Suzuki, N., Robinson, S., Nagarajan, V., Ye, Z., Boukaba, A., Rasko, J.E., Strunnikov, A.V., Loukinov, D., Ren, B., Lobanenkov, V.V., 2015. Comparative analyses of CTCF and BORIS occupancies uncover two distinct classes of CTCF binding genomic regions. Genome Biology 16, 161.

Pugacheva, E.M., Suzuki, T., Pack, S.D., Kosaka-Suzuki, N., Yoon, J., Vostrov, A.A., Barsov, E., Strunnikov, A.V., Morse, H.C., Loukinov, D., Lobanenkov, V., 2010. The Structural Complexity of the Human BORIS Gene in Gametogenesis and Cancer. Plos One 5.

Punnamoottil, B., Herrmann, C., Pascual-Anaya, J., D'Aniello, S., Garcia-Fernandez, J., Akalin, A., Becker, T.S., Rinkwitz, S., 2010. Cis-regulatory characterization of sequence conservation surrounding the Hox4 genes. Developmental Biology 340, 269-282.

Putnam, N.H., Butts, T., Ferrier, D.E.K., Furlong, R.F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J.K., Benito-Gutierrez, E., Dubchak, I., Garcia-Fernandez, J., Gibson-Brown, J.J., Grigoriev, I.V., Horton, A.C., de Jong, P.J., Jurka, J., Kapitonov, V.V., Kohara, Y., Kuroki, Y., Lindquist, E., Lucas, S., Osoegawa, K., Pennacchio, L.A., Salamov, A.A., Satou, Y., Sauka-Spengler, T., Schmutz, J., Shin-I, T., Toyoda, A., Bronner-Fraser, M., Fujiyama, A., Holland, L.Z., Holland, P.W.H., Satoh, N., Rokhsar, D.S., 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453, 1064-U1063.

Qian, J., Kaytor, E.N., Towle, H.C., Olson, L.K., 1999. Upstream stimulatory factor regulates Pdx-1 gene expression in differentiated pancreatic beta-cells. Biochemical Journal 341, 315-322.

Qiao, H., Chen, J.K., Reynolds, A., Hoog, C., Paddy, M., Hunter, N., 2012. Interplay between Synaptonemal Complex, Homologous Recombination, and Centromeres during Mammalian Meiosis. Plos Genetics 8.

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

Rabet, N., Gibert, J.M., Queinnec, E., Deutsch, J.S., Mouchel-Vielh, E., 2001. The caudal gene of the barnacle Sacculina carcini is not expressed in its vestigial abdomen. Development Genes and Evolution 211, 172-178.

Raff, R.A., Byrne, M., 2006. The active evolutionary lives of echinoderm larvae. Heredity 97, 244-252.

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

Rao, Suhas S.P., Huntley, Miriam H., Durand, Neva C., Stamenova, Elena K., Bochkov, Ivan D., Robinson, James T., Sanborn, Adrian L., Machol, I., Omer, Arina D., Lander, Eric S., Aiden, Erez L., 2014. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 159, 1665-1680.

Ray, R., Capecchi, M., 2008. An examination of the Chiropteran HoxD locus from an evolutionary perspective. Evolution & Development 10, 657-670.

Rebeiz, M., Posakony, J.W., 2004. GenePalette: a universal software tool for genome sequence visualization and analysis. Developmental Biology 271, 431-438.

Reece-Hoyes, J.S., Keenan, I.D., Pownall, M.E., Isaacs, H.V., 2005. A consensus Oct1 binding site is required for the activity of the Xenopus Cdx4 promoter. Developmental Biology 282, 509-523.

Reece-Hoyes, J.S., Keenan, L.D., Isaacs, H.V., 2002. Cloning and expression of the Cdx family from the frog Xenopus tropicalis. Developmental Dynamics 223, 134-140.

Reese, M., Harris, N., Eeckman, F., 1996. Large scale sequencing specific neural networks for promoter and splice site recognition, in: Lawrence Hunter, Klein, T.E. (Eds.), Biocomputing: Proceedings of the 1996 pacific symposium. World Scientific Publishing Co, Singapore, 1996, January 2-7,.

Reese, M.G., 2001. Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. Computers & Chemistry 26, 51-56.

Reese, M.G., Eeckman, F.H., 1995. Novel Neural Network Algorithms for Improved Eukaryotic Promoter Site Recognition, The Seventh International Genome Sequencing and Analysis Conference, Hilton Head Island, South Carolina.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: The European molecular biology open software suite. Trends in Genetics 16, 276-277.

Richler, C., Ast, G., Goitein, R., Wahrman, J., Sperling, R., Sperling, J., 1994. Splicing components are excluded from the transcriptionally inactive XY body in male meiotic nuclei. Molecular Biology of the Cell 5, 1341-1352.

Roberts, D.J., Johnson, R.L., Burke, A.C., Nelson, C.E., Morgan, B.A., Tabin, C., 1995. Sonic hedgehog is an endodermal signal inducing Bmp-4 and Hox genes during induction and regionalization of the chick hindgut. Development 121, 3163-3174.

Rosanas-Urgell, A., Garcia-Fernandez, J., Marfany, G., 2008. ParaHox genes in pancreatic cell cultures: effects on the insulin promoter regulation. International Journal of Biological Sciences 4, 48-57.

Rothbacher, U., Bertrand, V., Lamy, C., Lemaire, P., 2007. A combinatorial code of maternal GATA, Ets and beta-catenin-TCF transcription factors specifies and patterns the early ascidian ectoderm. Development 134, 4023-4032.

Roël, G., van den Broek, O., Spieker, N., Peterson-Maduro, J., Destrée, O., 2003. Tcf-1 expression during Xenopus development. Gene Expression Patterns 3, 123-126.

Ryan, J.F., Burton, P.M., Mazza, M.E., Kwong, G.K., Mullikin, J.C., Finnerty, J.R., 2006. The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes: evidence from the starlet sea anemone, Nematostella vectensis. Genome Biology 7, 20.

Ryan, J.F., Mazza, M.E., Pang, K., Matus, D.Q., Baxevanis, A.D., Martindale, M.Q., Finnerty, J.R., 2007. Pre-Bilaterian Origins of the Hox Cluster and the Hox Code: Evidence from the Sea Anemone, Nematostella vectensis. Plos One 2, 23.

Saegusa, M., Hashimura, M., Kuwata, T., Hamano, M., Wani, Y., Okayasu, I., 2007. A functional role of Cdx2 in beta-catenin signaling during transdifferentiation in endometrial carcinomas. Carcinogenesis 28, 1885-1892.

Sage, J., Yuan, L., Martin, L., Mattei, M.G., Guenet, J.L., Liu, J.G., Hoog, C., Rassoulzadegan, M., Cuzin, F., 1997. The Sycp1 loci of the mouse genome: Successive retropositions of a meiotic gene during the recent evolution of the genus. Genomics 44, 118-126.

Saitoh, N., Bell, A.C., Recillas-Targa, F., West, A.G., Simpson, M., Pikaart, M., Felsenfeld, G., 2000. Structural and functional conservation at the boundaries of the chicken β-globin domain. EMBO Journal 19, 2315-2322.

Saitou, M., Barton, S.C., Surani, M.A., 2002. A molecular programme for the specification of germ cell fate in mice. Nature 418, 293-300.

Saitou, M., Payer, B., Lange, U.C., Erhardt, S., Barton, S.C., Surani, M.A., 2003. Specification of germ cell fate in mice. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 358, 1363-1370.

Sakai, H., Koyanagi, K.O., Imanishi, T., Itoh, T., Gojobori, T., 2007. Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes. Gene 389, 196-203.

Samadi, L., Steiner, G., 2010. Conservation of ParaHox genes' function in patterning of the digestive tract of the marine gastropod Gibbula varia. Bmc Developmental Biology 10, 15.

Samaras, S.E., Cissell, M.A., Gerrish, K., Wright, C.V.E., Gannon, M., Stein, R., 2002. Conserved sequences in a tissue-specific regulatory region of the pdx-1 gene mediate transcription in pancreatic beta cells: Role for hepatocyte nuclear factor 3 beta and Pax6. Molecular and Cellular Biology 22, 4702-4713.

Samaras, S.E., Zhao, L., Means, A., Henderson, E., Matsuoka, T., Stein, R., 2003. The islet beta cell-enriched RIPE3b1/Maf transcription factor regulates pdx-1 expression. Journal of Biological Chemistry 278, 12263-12270.

Sambrook, J., Fritsch, E.F., Maniatis, T., 1989. Molecular cloning. A laboratory manual (ed. N. Ford). Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York, USA. Illus. Paper.

Sandelin, A., Bailey, P., Bruce, S., Engstrom, P.G., Klos, J.M., Wasserman, W.W., Ericson, J., Lenhard, B., 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. Bmc Genomics 5, 9.

Sasakura, Y., Ogasawara, M., Makabe, K.W., 1998. HrWnt-5: A maternally expressed ascidian Wnt gene with posterior localization in early embryos. International Journal of Developmental Biology 42, 573-579.

Sato, A., Satoh, N., Bishop, J.D., 2012. Field identification of ‘types’ A and B of the ascidian Ciona intestinalis in a region of sympatry. Marine Biology 159, 1611-1619.

Satoh, N., 1994. Developmental biology of ascidians. Developmental and Cell Biology Series 29, i-xv, 1-234.

Satou, Y., Imai, K.S., Satoh, N., 2002. Fgf genes in the basal chordate Ciona intestinalis. Development Genes and Evolution 212, 432-438.

Sauka-Spengler, T., Meulemans, D., Jones, M., Bronner-Fraser, M., 2007. Ancient evolutionary origin of the neural crest gene regulatory network. Developmental Cell 13, 405-420.

Schierwater, B., Desalle, R., 2001. Current problems with the zootype and the early evolution of Hox genes. Journal of Experimental Zoology 291, 169-174.

Schierwater, B., Eitel, M., Jakob, W., Osigus, H.J., Hadrys, H., Dellaporta, S.L., Kolokotronis, S.O., Desalle, R., 2009. Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. PLoS Biol 7, e20.

Schierwater, B., Kamm, K., Srivastava, M., Rokhsar, D., Rosengarten, R.D., Dellaporta, S.L., 2008. The Early ANTP Gene Repertoire: Insights from the Placozoan Genome. Plos One 3, 5.

Schild-Prufert, K., Saito, T.T., Smolikov, S., Gu, Y.J., Hincapie, M., Hill, D.E., Vidal, M., McDonald, K., Colaiacovo, M.P., 2011. Organization of the Synaptonemal Complex During Meiosis in Caenorhabditis elegans. Genetics 189, 411-U437.

Schmidt, M., Patterson, M., Farrell, E., Munsterberg, A., 2004. Dynamic expression of Lef/Tcf family members and beta-catenin during chick gastrulation, neurulation, and early limb development. Developmental Dynamics 229, 703-707.

Schneider, S.Q., Bowerman, B., 2007. β-Catenin Asymmetries after All Animal/Vegetal- Oriented Cell Divisions in Platynereis dumerilii Embryos Mediate Binary Cell-Fate Specification. Developmental Cell 13, 73-86.

Schubert, M., Escriva, H., Xavier-Neto, J., Laudet, V., 2006. Amphioxus and tunicates as evolutionary model systems. Trends in Ecology & Evolution 21, 269-277.

Schubert, M., Holland, L.Z., Holland, N.D., 2000a. Characterization of two amphioxus Wnt genes (AmphiWnt4 and AmphiWnt7b) with early expression in the developing central nervous system. Developmental Dynamics 217, 205-215.

Schubert, M., Holland, L.Z., Jacobs, D.K., Holland, N.D., 2000b. Phylogenetic analysis and expression of amphioxus Wnt genes: A possible ancient function for Wnt1 during gastrulation. Developmental Biology 222, 238-238.

Schubert, M., Holland, L.Z., Stokes, M.D., Holland, N.D., 2001. Three amphioxus Wnt genes (AmphiWnt3, AmphiWnt5, and AmphiWnt6) associated with the tail bud: The evolution of somitogenesis in chordates. Developmental Biology 240, 262-273.

Schubert, M., Yu, J.K., Holland, N.D., Escriva, H., Laudet, V., Holland, L.Z., 2005. Retinoic acid signaling acts via Hox1 to establish the posterior limit of the pharynx in the chordate amphioxus. Development 132, 61-73.

Schug, J., 2008. Using TESS to predict transcription factor binding sites in DNA sequence. Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.] Chapter 2, Unit 2.6.

Schulz, C., Schroder, R., Hausdorf, B., Wolff, C., Tautz, D., 1998. A caudal homologue in the short germ band beetle Tribolium shows similarities to both, the Drosophila and the vertebrate caudal expression patterns. Development Genes and Evolution 208, 283-289.

Schwarz, M., Alvarez-Bolado, G., Dressler, G., Pavel, U., Busslinger, M., Gruss, P., 1999. Pax2/5 and Pax6 subdivide the early neural tube into three domains. Mechanisms of Development 82, 29-39.

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

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

Sharma, S., Leonard, J., Lee, S., Chapman, H.D., Leiter, E.H., Montminy, M.R., 1996. Pancreatic islet expression of the homeobox factor STF-1 relies on an E-box motif that binds USF. Journal of Biological Chemistry 271, 2294-2299.

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

Shen, M.M., 2007. Nodal signaling: developmental roles and regulation. Development 134, 1023-1034.

Shenk, M.A., Bode, H.R., Steele, R.E., 1993a. Expression of Cnox-2, a HOM/HOX homeobox gene in hydra, is correlated with axial pattern formation. Development 117, 657-667.

Shenk, M.A., Gee, L., Steele, R.E., Bode, H.R., 1993b. Expression of Cnox-2, a HOM/HOX gene, is suppressed during head formation in hydra. Dev Biol 160, 108-118.

Sheth, R., Grégoire, D., Dumouchel, A., Scotti, M., Pham, J.M.T., Nemec, S., Bastida, M.F., Ros, M.A., Kmita, M., 2013. Decoupling the function of Hox and Shh in developing limb reveals multiple inputs of Hox genes on limb growth. Development 140, 2130-2138.

Shimizu, T., Bae, Y.K., Muraoka, O., Hibi, M., 2005. Interaction of Wnt and caudal-related genes in zebrafish posterior body formation. Developmental Biology 279, 125-141.

Shinmyo, Y., Mito, T., Matsushita, T., Sarashina, I., Miyawaki, K., Ohuchi, H., Noji, S., 2005. caudal is required for gnathal and thoracic patterning and for posterior elongation in the intermediate-germband cricket Gryllus bimaculatus. Mechanisms of Development 122, 231-239.

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

Shlyueva, D., Stampfel, G., Stark, A., 2014. Transcriptional enhancers: from properties to genome-wide predictions. Nature Reviews Genetics 15, 272-286.

Siegel, N., Hoegg, S., Salzburger, W., Braasch, I., Meyer, A., 2007. Comparative genomics of ParaHox clusters of teleost fishes: gene cluster breakup and the retention of gene sets following whole genome duplications. Bmc Genomics 8, 15.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M.M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., Haussler, D., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 15, 1034-1050.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soeding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7.

Simeone, A., Acampora, D., Arcioni, L., Andrews, P.W., Boncinelli, E., Mavilio, F., 1990. Sequential activation of Hox2 Homeobox genes by retinoic acid in human embryonal carcinoma-cells. Nature 346, 763-766.

Simeone, A., Acampora, D., Nigro, V., Faiella, A., D'Esposito, M., Stornaiuolo, A., Mavilio, F., Boncinelli, E., 1991. Differential regulation by retinoic acid of the homeobox genes of the four HOX loci in human embryonal carcinoma cells. Mech Dev 33, 215-227.

Simon, J., Peifer, M., Bender, W., O'Connor, M., 1990. Regulatory elements of the bithorax complex that control expression along the anterior-posterior axis. Embo j 9, 3945-3956.

Skromne, I., Thorsen, D., Hale, M., Prince, V.E., Ho, R.K., 2007. Repression of the hindbrain developmental program by Cdx factors is required for the specification of the vertebrate spinal cord. Development 134, 2147-2158.

Sleutels, F., Soochit, W., Bartkuhn, M., Heath, H., Dienstbach, S., Bergmaier, P., Franke, V., Rosa-Garrido, M., van de Nobelen, S., Caesar, L., van der Reijden, M., Bryne, J.C., van Ijcken, W., Grootegoed, J.A., Delgado, M.D., Lenhard, B., Renkawitz, R., Grosveld, F., Galjart, N., 2012. The male germ cell gene regulator CTCFL is functionally different from CTCF and binds CTCF-like consensus sites in a nucleosome composition-dependent manner. Epigenetics & Chromatin 5.

Slotkin, R.K., Martienssen, R., 2007. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 8, 272-285.

Small, S., Blair, A., Levine, M., 1996. Regulation of two pair-rule stripes by a single enhancer in the Drosophila embryo. Developmental Biology 175, 314-324.

Smith, C., Heyne, S., Richter, A.S., Will, S., Backofen, R., 2010. Freiburg RNA Tools: a web server integrating INTARNA, EXPARNA and LocARNA. Nucleic Acids Research 38, W373-W377.

Smith, L.M., Shortreed, M.R., Olivier, M., 2011. To understand the whole, you must know the parts: unraveling the roles of protein-DNA interactions in genome regulation. Analyst 136, 3060-3065.

Smolikov, S., Eizinger, A., Hurlburt, A., Rogers, E., Villeneuve, A.M., Colaiácovo, M.P., 2007. Synapsis-Defective Mutants Reveal a Correlation Between Chromosome Conformation and the Mode of Double-Strand Break Repair During Caenorhabditis elegans Meiosis. Genetics 176, 2027-2033.

Soares, M.B., Schon, E., Henderson, A., Karathanasis, S.K., Cate, R., Zeitlin, S., Chirgwin, J., Efstratiadis, A., 1985. RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. Mol Cell Biol 5, 2090-2103.

Solari, A.J., 1974. The behavior of the XY pair in mammals. Int Rev Cytol 38, 273-317.

Solovyev, V., Kosarev, P., Seledsov, I., Vorobyev, D., 2006. Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biology 7.

Solovyev, V.V., Shahmuradov, I.A., Salamov, A.A., 2010. Identification of Promoter Regions and Regulatory Sites, in: Ladunga, I. (Ed.), Computational Biology of Transcription Factor Binding. Humana Press Inc, Totowa, pp. 57-83.

Song, J.H., Xu, Y.Z., Hu, X.X., Choi, B., Tong, Q.C., 2010. Brain Expression of Cre Recombinase Driven by Pancreas-Specific Promoters. Genesis 48, 628-634.

Sorourian, M., Kunte, M.M., Domingues, S., Gallach, M., Ozdil, F., Rio, J., Betran, E., 2014. Relocation Facilitates the Acquisition of Short Cis-Regulatory Regions that Drive the Expression of Retrogenes during Spermatogenesis in Drosophila. Molecular Biology and Evolution 31, 2170-2180.

Soshnikova, N., Montavon, T., Leleu, M., Galjart, N., Duboule, D., 2010. Functional Analysis of CTCF During Mammalian Limb Development. Developmental Cell 19, 819-830.

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., Barthès, P., Kokkinaki, M., Nef, S., Gnirke, A., Dym, M., de Massy, B., Mikkelsen, Tarjei S., Kaessmann, H., 2013. Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. Cell Reports 3, 2179-2190.

Spitz, F., Gonzalez, F., Duboule, D., 2003. A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. Cell 113, 405-417.

Spitz, F., Herkenne, C., Morris, M.A., Duboule, D., 2005. Inversion-induced disruption of the Hoxd cluster leads to the partition of regulatory landscapes. Nature Genetics 37, 889-893.

Splinter, E., Heath, H., Kooren, J., Palstra, R.-J., Klous, P., Grosveld, F., Galjart, N., de Laat, W., 2006. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes & Development 20, 2349-2354.

Srivastava, M., Begovic, E., Chapman, J., Putnam, N.H., Hellsten, U., Kawashima, T., Kuo, A., Mitros, T., Salamov, A., Carpenter, M.L., Signorovitch, A.Y., Moreno, M.A., Kamm, K., Grimwood, J., Schmutz, J., Shapiro, H., Grigoriev, I.V., Buss, L.W., Schierwater, B., Dellaporta, S.L., Rokhsar, D.S., 2008. The Trichoplax genome and the nature of placozoans. Nature 454, 955-U919.

Stafford, D., Prince, V.E., 2002. Retinoic acid signaling is required for a critical early step in zebrafish pancreatic development. Current Biology 12, 1215-1220.

Stanke, M., Morgenstern, B., 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Research 33, W465-W467.

Stark, A., Brennecke, J., Bushati, N., Russell, R.B., Cohen, S.M., 2005. Animal MicroRNAs Confer Robustness to Gene Expression and Have a Significant Impact on 3'UTR Evolution. Cell 123, 1133-1146.

Stoffers, D.A., Ferrer, J., Clarke, W.L., Habener, J.F., 1997. Early-onset type-II diabetes mellitus (MODY4) linked to IPF1. Nature Genetics 17, 138-139.

Stokes, M.D., Holland, N.D., 1995. Embryos and Larvae of a Lancelet, Branchiostoma floridae, from Hatching through Metamorphosis: Growth in the Laboratory and External Morphology. Acta Zoologica 76, 105-120.

Stokes, M.D., Holland, N.D., 1996. Reproduction of the Florida lancelet (Branchiostoma floridae): Spawning patterns and fluctuations in gonad indexes and nutritional reserves. Invertebrate Biology 115, 349-359.

Stolfi, A., Christiaen, L., 2012. Genetic and Genomic Toolbox of the Chordate Ciona intestinalis. Genetics 192, 55-66.

Stornaiuolo, A., Acampora, D., Pannese, M., D'Esposito, M., Morelli, F., Migliaccio, E., Rambaldi, M., Faiella, A., Nigro, V., Simeone, A., et al., 1990. Human HOX genes are differentially activated by retinoic acid in embryonal carcinoma cells according to their position within the four loci. Cell Differ Dev 31, 119-127.

Strick, T.R., Kawaguchi, T., Hirano, T., 2004. Real-time detection of single-molecule DNA compaction by condensin I. Current Biology 14, 874-880.

Struhl, K., 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol 14, 103-105.

Subramanian, V., Meyer, B.I., Gruss, P., 1995. Disruption of the murine homeobox gene Cdx1 affects axial skeletal identities by altering the mesodermal expression domains of Hox genes. Cell 83, 641-653.

Sun, Y., Zhang, L., Gu, H.F., Han, W., Ren, M., Wang, F., Gong, B., Wang, L., Guo, H., Xin, W., Zhao, J., Gao, L., 2008. Peroxisome proliferator-activated receptor-alpha regulates the expression of pancreatic/duodenal homeobox-1 in rat insulinoma (INS-1) cells and ameliorates glucose-induced insulin secretion impaired by palmitate. Endocrinology 149, 662-671.

Suzuki, T., Kosaka-Suzuki, N., Pack, S., Shin, D.M., Yoon, J., Abdullaev, Z., Pugacheva, E., Morse, H.C., Loukinov, D., Lobanenkov, V., 2010. Expression of a Testis-Specific Form of Gal3st1 (CST), a Gene Essential for Spermatogenesis, Is Regulated by the CTCF Paralogous Gene BORIS. Molecular and Cellular Biology 30, 2473-2484.

Takamura, K., Fujimura, M., Yamaguchi, Y., 2002. Primordial germ cells originate from the endodermal strand cells in the ascidian Ciona intestinalis. Dev Genes Evol 212, 11-18.

Tarchini, B., Duboule, D., 2006. Control of Hoxd genes' collinearity during early limb development. Developmental Cell 10, 93-103.

Tassy, O., Dauga, D., Daian, F., Sobral, D., Robin, F., Khoueiry, P., Salgado, D., Fox, V., Caillol, D., Schiappa, R., Laporte, B., Rios, A., Luxardi, G., Kusakabe, T., Joly, J.-S., Darras, S., Christiaen, L., Contensin, M., Auger, H., Lamy, C., Hudson, C., Rothbaecher, U., Gilchrist, M.J., Makabe, K.W., Hotta, K., Fujiwara, S., Satoh, N., Satou, Y., Lemaire, P., 2010. The ANISEED database: Digital representation, formalization, and elucidation of a chordate developmental program. Genome Research 20, 1459-1468.

Theodosiou, N.A., Tabin, C.J., 2003. Wnt signaling during development of the gastrointestinal tract. Developmental Biology 259, 258-271.

Thomas, D.J., Rosenbloom, K.R., Clawson, H., Hinrichs, A.S., Trumbower, H., Raney, B.J., Karolchik, D., Barber, G.P., Harte, R.A., Hillman-Jackson, J., Kuhn, R.M., Rhead, B.L., Smith, K.E., Thakkapallayil, A., Zweig, A.S., Haussler, D., Kent, W.J., Consortium, E.P., 2007. The ENCODE project at UC Santa Cruz. Nucleic Acids Research 35, D663-D667.

Tomer, R., Denes, A.S., Tessmar-Raible, K., Arendt, D., 2010. Profiling by Image Registration Reveals Common Origin of Annelid Mushroom Bodies and Vertebrate Pallium. Cell 142, 800-809.

Toresson, H., Campbell, K., 2001. A role for Gsh1 in the developing striatum and olfactory bulb of Gsh2 mutant mice. Development 128, 4769-4780.

Toresson, H., Potter, S.S., Campbell, K., 2000. Genetic control of dorsal-ventral identity in the telencephalon: opposing roles for Pax6 and Gsh2. Development 127, 4361-4371.

Tschopp, P., Tarchini, B., Spitz, F., Zakany, J., Duboule, D., 2009. Uncoupling Time and Space in the Collinear Regulation of <italic>Hox</italic> Genes. PLoS Genet 5, e1000398.

Tsujikawa, M., Kurahashi, H., Tanaka, T., Nishida, K., Shimomura, Y., Tano, Y., Nakamura, Y., 1999. Identification of the gene responsible for gelatinous drop-like corneal dystrophy. Nature Genetics 21, 420-423.

Uechi, T., Tanaka, T., Kenmochi, N., 2001. A complete map of the human ribosomal protein genes: Assignment of 80 genes to the cytogenetic map and implications for human disorders. Genomics 72, 223-230.

Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C.A.-K., Odeberg, J., Djureinovic, D., Takanen, J.O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J.M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., Pontén, F., 2015. Tissue-based map of the human proteome. Science 347.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., Rozen, S.G., 2012. Primer3-new capabilities and interfaces. Nucleic Acids Research 40.

Urbach, R., Volland, D., Seibert, J., Technau, G.M., 2006. Segment-specific requirements for dorsoventral patterning genes during early brain development in Drosophila. Development 133, 4315-4330.

Valerius, M.T., Li, H., Stock, J.L., Weinstein, M., Kaur, S., Singh, G., Potter, S.S., 1995. Gsh-1 - A novel murine Homeobox gene expressed in the central-nervous-system. Developmental Dynamics 203, 337-351.

van de Ven, C., Bialecka, M., Neijts, R., Young, T., Rowland, J.E., Stringer, E.J., van Rooijen, C., Meijlink, F., Novoa, A., Freund, J.N., Mallo, M., Beck, F., Deschamps, J., 2011. Concerted involvement of Cdx/Hox genes and Wnt signaling in morphogenesis of the caudal neural tube and cloacal derivatives from the posterior growth zone. Development 138, 3451-3462.

van den Akker, E., Forlani, S., Chawengsaksophak, K., de Graaff, W., Beck, F., Meyer, B.I., Deschamps, J., 2002. Cdx1 and Cdx2 have overlapping functions in anteroposterior patterning and posterior axis elongation. Development 129, 2181-2193.

van Nes, J., Graaff, W.C., Lebrin, F., Gerhard, M., Beck, F., Deschamps, J., 2006. The Cdx4 mutation affects axial development and reveals an essential role of Cdx genes in the ontogenesis of the placental labyrinth in mice. Development 133, 419-428.

Van Velkinburgh, J.C., Samaras, S.E., Gerrish, K., Artner, I., Stein, R., 2005. Interactions between areas I and II direct pdx-1 expression specifically to islet cell types of the mature and developing pancreas. Journal of Biological Chemistry 280, 38438-38444.

Vavouri, T., Lehne, B., 2009. Conserved noncoding elements and the evolution of animal body plans. Bioessays 31, 727-735.

Vavouri, T., Walter, K., Gilks, W.R., Lehner, B., Elgar, G., 2007. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. Genome Biology 8, 14.

Vienne, A., Pontarotti, P., 2006. Metaphylogeny of 82 gene families sheds a new light on chordate evolution. International Journal of Biological Sciences 2, 32-37.

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, Duncan T., Tanay, A., Hadjur, S., 2015. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. Cell Reports 10, 1297-1309.

Vinckenbosch, N., Dupanloup, I., Kaessmann, H., 2006. Evolutionary fate of retroposed gene copies in the human genome. Proceedings of the National Academy of Sciences of the United States of America 103, 3220-3225.

Visel, A., Prabhakar, S., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E.M., Pennacchio, L.A., 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nature Genetics 40, 158-160.

Vivat-Hannah, V., Bourguet, W., Gottardis, M., Gronemeyer, H., 2003. Separation of retinoid X receptor homo- and heterodimerization functions. Molecular and Cellular Biology 23, 7678-7688.

Von Frowein, J., Campbell, K., Gotz, M., 2002. Expression of Ngn1, Ngn2, Cash1, Gsh2 and Sfrp1 in the developing chick telencephalon. Mechanisms of Development 110, 249-252.

Vopalensky, P., Pergner, J., Liegertova, M., Benito-Gutierrez, E., Arendt, D., Kozmik, Z., 2012. Molecular analysis of the amphioxus frontal eye unravels the evolutionary origin of the retina and pigment cells of the vertebrate eye. Proceedings of the National Academy of Sciences of the United States of America 109, 15383-15388.

Wada, H., Escriva, H., Zhang, S.C., Laudet, V., 2006. Conserved RARE localization in amphioxus Hox clusters and implications for Hox code evolution in the vertebrate neural crest. Developmental Dynamics 235, 1522-1531.

Wada, H., Garcia-Fernandez, J., Holland, P.W.H., 1999. Colinear and segmental expression of amphioxus Hox genes. Developmental Biology 213, 131-141.

Wada, H., Kobayashi, M., Zhang, S.C., 2005. Ets identified as a trans-regulatory factor of amphioxus Hox2 by transgenic analysis using ascidian embryos. Developmental Biology 285, 524-532.

Wada, H., Saiga, H., Satoh, N., Holland, P.W.H., 1998. Tripartite organization of the ancestral chordate brain and the antiquity of placodes: insights from ascidian Pax-2/5/8, Hox and Otx genes. Development 125, 1113-1122.

Wada, S., Katsuyama, Y., Yasugi, S., Saiga, H., 1995. Spatially and temporally regulated expression of the LIM class homeobox gene HRLIM suggests multiple distinct functions in development of the ascidian, Halocynthia roretzi. Mechanisms of Development 51, 115-126.

Wallace, J.A., Felsenfeld, G., 2007. We gather together: insulators and genome organization. Current Opinion in Genetics & Development 17, 400-407.

Walther, C., Gruss, P., 1991. Pax-6, a murine paired box gene, is expressed in the developing CNS. Development 113, 1435-1449.

Wan, L.-B., Pan, H., Hannenhalli, S., Cheng, Y., Ma, J., Fedoriw, A., Lobanenkov, V., Latham, K.E., Schultz, R.M., Bartolomei, M.S., 2008. Maternal depletion of CTCF reveals multiple functions during oocyte and preimplantation embryo development. Development 135, 2729-2738.

Wang, P.J., 2004. X chromosomes, retrogenes and their role in male reproduction. Trends Endocrinol Metab 15, 79-83.

Wang, R.H., Xu, X.L., Kim, H.S., Xiao, Z., Deng, C.X., 2013. SIRT1 Deacetylates FOXA2 and Is Critical for Pdx1 Transcription and beta-Cell Formation. International Journal of Biological Sciences 9, 934-946.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., Barton, G.J., 2009. Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189-1191.

Wedeen, C.J., Shankland, M., 1997. Mesoderm Is Required for the Formation of a Segmented Endodermal Cell Layer in the LeechHelobdella. Developmental Biology 191, 202-214.

Weiner, A.M., Deininger, P.L., Efstratiadis, A., 1986. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem 55, 631-661.

Weiss, J.B., Von Ohlen, T., Mellerick, D.M., Dressler, G., Doe, C.Q., Scott, M.P., 1998. Dorsoventral patterning in the Drosophila central nervous system: the intermediate neuroblasts defective homeobox gene specifies intermediate column identity. Genes & Development 12, 3591-3602.

West, A.G., Gaszner, M., Felsenfeld, G., 2002. Insulators: many functions, many mechanisms. Genes and Development.

Wheeler, S.R., Carrico, M.L., Wilson, B.A., Skeath, J.B., 2005. The Tribolium columnar genes reveal conservation and plasticity in neural precursor patterning along the embryonic dorsal-ventral axis. Developmental Biology 279, 491-500.

Whelan, N.V., Kocot, K.M., Moroz, L.L., Halanych, K.M., 2015. Error, signal, and the placement of Ctenophora sister to all other animals. Proceedings of the National Academy of Sciences 112, 5773-5778.

White, R.A.H., Wilcox, M., 1985. Regulation of the distribution of Ultrabithorax proteins in Drosophila. Nature 318, 563-567.

Whittaker, J.R., 1997. Cephalochordates, the lancelets., in: S.F., G., A.M., R. (Eds.), Embryology: constructing the organism. Sinauer Associates Inc, Sunderland, MA, pp. 365-381.

Wiebe, P.O., Kormish, J.D., Roper, V.T., Fujitani, Y., Alston, N.I., Zaret, K.S., Wright, C.V.E., Stein, R.W., Gannon, M., 2007. Ptf1a binds to and activates area III, a highly conserved region of the Pdx1 promoter that mediates early pancreas-wide Pdx1 expression. Molecular and Cellular Biology 27, 4093-4104.

Wiens, M., Batel, R., Korzhev, M., Muller, W.E.G., 2003. Retinoid X receptor and retinoic acid response in the marine sponge Suberites domuncula. Journal of Experimental Biology 206, 3261-3271.

Wilt, F.H., 2002. Biomineralization of the spicules of sea urchin embryos. Zoological Science 19, 253-261.

Winterbottom, E.F., Illes, J.C., Faas, L., Isaacs, H.V., 2010. Conserved and novel roles for the Gsh2 transcription factor in primary neurogenesis. Development 137, 2623-2631.

Woltering, J.M., Durston, A.J., 2008. MiR-10 Represses HoxB1a and HoxB3a in Zebrafish. Plos One 3.

Woolfe, A., Goode, D.K., Cooke, J., Callaway, H., Smith, S., Snell, P., McEwen, G.K., Elgar, G., 2007. CONDOR: a database resource of developmentally associated conserved non-coding elements. Bmc Developmental Biology 7, 11.

Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., Gilks, W., Edwards, Y.J.K., Cooke, J.E., Elgar, G., 2005. Highly conserved non-coding sequences are associated with vertebrate development. Plos Biology 3, 116-130.

Wray, G.A., 2003. Transcriptional regulation and the evolution of development. International Journal of Developmental Biology 47, 675-684.

Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V., Romano, L.A., 2003. The evolution of transcriptional regulation in eukaryotes. Molecular Biology and Evolution 20, 1377-1419.

Wright, C.V., Schnegelsberg, P., De Robertis, E.M., 1989. XlHbox 8: a novel Xenopus homeo protein restricted to a narrow band of endoderm. Development 105, 787-794.

Wu, C.-H., Chen, S., Shortreed, M.R., Kreitinger, G.M., Yuan, Y., Frey, B.L., Zhang, Y., Mirza, S., Cirillo, L.A., Olivier, M., Smith, L.M., 2011a. Sequence-Specific Capture of Protein-DNA Complexes for Mass Spectrometric Protein Identification. PLoS ONE 6, e26217.

Wu, H.-R., Chen, Y.-T., Su, Y.-H., Luo, Y.-J., Holland, L.Z., Yu, J.-K., 2011b. Asymmetric localization of germline markers Vasa and Nanos during early development in the amphioxus Branchiostoma floridae. Developmental Biology 353, 147-159.

Wu, K.L., Gannon, M., Peshavaria, M., Offield, M.F., Henderson, E., Ray, M., Marks, A., Gamer, L.W., Wright, C.V.E., Stein, R., 1997. Hepatocyte nuclear factor 3 beta is involved in pancreatic beta-cell-specific transcription of the pdx-1 gene. Molecular and Cellular Biology 17, 6002-6013.

Wu, L.H., Lengyel, J.A., 1998. Role of caudal in hindgut specification and gastrulation suggests homology between Drosophila amnioproctodeal invagination and vertebrate blastopore. Development 125, 2433-2442.

Wysocka-Diller, J., Aisemberg, G.O., Macagno, E.R., 1995. A Novel Homeobox Cluster Expressed in Repeated Structures of the Midgut. Developmental Biology 171, 439-447.

Xie, X., Mikkelsen, T.S., Gnirke, A., Lindblad-Toh, K., Kellis, M., Lander, E.S., 2007. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proceedings of the National Academy of Sciences of the United States of America 104, 7145-7150.

Xu, X., Xu, P.X., Suzuki, Y., 1994. A maternal homeobox gene, Bombyx caudal, forms both mRNA and protein concentration gradients spanning anteroposterior axis during gastrulation. Development 120, 277-285.

Yajima, M., Kiyomoto, M., 2006. Study of larval and adult skeletogenic cells in developing sea urchin larvae. Biological Bulletin 211, 183-192.

Yajima, M., Suglia, E., Gustafson, E.A., Wessel, G.M., 2013. Meiotic gene expression initiates during larval development in the sea urchin. Developmental Dynamics 242, 155-163.

Yano, Y., Saito, R., Yoshida, N., Yoshiki, A., Wynshaw-Boris, A., Tomita, M., Hirotsune, S., 2004. A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene. Journal of Molecular Medicine-Jmm 82, 414-422.

Yanze, N., Spring, J., Schmidli, C., Schmid, V., 2001. Conservation of Hox/ParaHox-related genes in the early development of a cnidarian. Developmental Biology 236, 89-98.

Yasui, K., Urata, M., Yamaguchi, N., Ueda, H., Henmi, Y., 2007. Laboratory culture of the oriental lancelet Branchiostoma belcheri. Zoological Science 24, 514-520.

Yee, N.S., Yusuff, S., Pack, M., 2001. Zebrafish pdx1 morphant displays defects in pancreas development and digestive organ chirality, and potentially identifies a multipotent pancreas progenitor cell. Genesis 30, 137-140.

Yekta, S., Shih, I.H., Bartel, D.P., 2004. MicroRNA-directed cleavage of HOXB8 mRNA. Science 304, 594-596.

Yekta, S., Tabin, C.J., Bartel, D.P., 2008. MicroRNAs in the Hox network: an apparent link to posterior prevalence. Nature Reviews Genetics 9, 789-796.

Yoshihama, M., Uechi, T., Asakawa, S., Kawasaki, K., Kato, S., Higa, S., Maeda, N., Minoshima, S., Tanaka, T., Shimizu, N., Kenmochi, N., 2002. The human ribosomal protein genes: Sequencing and comparative analysis of 73 genes. Genome Research 12, 379-390.

Young, R.M., Reyes, A.E., Allende, M.L., 2002. Expression and splice variant analysis of the zebrafish tcf4 transcription factor. Mechanisms of Development 117, 269-273.

Young, T., Deschamps, J., 2009. Hox, Cdx, and anteroposterior patterning in the mouse embryo, in: Pourquie, O. (Ed.), Hox Genes. Elsevier Academic Press Inc, San Diego, pp. 235-+.

Yu, J.-K., Satou, Y., Holland, N.D., Shin, T.I., Kohara, Y., Satoh, N., Bronner-Fraser, M., Holland, L.Z., 2007a. Axial patterning in cephalochordates and the evolution of the organizer. Nature 445, 613-617.

Yu, J.-K., Wang, M.-C., Shin, T.I., Kohara, Y., Holland, L.Z., Satoh, N., Satou, Y., 2008. A cDNA resource for the cephalochordate amphioxus Branchiostoma floridae. Development Genes and Evolution 218, 723-727.

Yu, J.K., Holland, N.D., Holland, L.Z., 2004. Tissue-specific expression of FoxD reporter constructs in amphioxus embryos. Developmental Biology 274, 452-461.

Yu, Z., Morais, D., Ivanga, M., Harrison, P.M., 2007b. Analysis of the role of retrotransposition in gene evolution in vertebrates. Bmc Bioinformatics 8.

Yuan, L., Pelttari, J., Brundell, E., Bjorkroth, B., Zhao, J., Liu, J.G., Brismar, H., Daneholt, B., Hoog, C., 1998. The synaptonemal complex protein SCP3 can form multistranded, cross-striated fibers in vivo. Journal of Cell Biology 142, 331-339.

Yue, J.X., Yu, J.K., Putnam, N.H., Holland, L.Z., 2014. The Transcriptome of an Amphioxus, Asymmetron lucayanum, from the Bahamas: A Window into Chordate Evolution. Genome Biology and Evolution 6, 2681-2696.

Yun, K., Potter, S., Rubenstein, J.L.R., 2001. Gsh2 and Pax6 play complementary roles in dorsoventral patterning of the mammalian telencephalon. Development 128, 193-205.

Yusufzai, T.M., Tagami, H., Nakatani, Y., Felsenfeld, G., 2004. CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. Molecular Cell 13, 291-298.

Zamudio, N., Bourc'his, D., 2010. Transposable elements in the mammalian germline: a comfortable niche or a deadly trap? Heredity 105, 92-104.

Zeller, R.W., Virata, M.J., Cone, A.C., 2006. Predictable mosaic transgene expression in ascidian embryos produced with a simple electroporation device. Developmental Dynamics 235, 1921-1932.

Zhang, T.J., Guo, X.G., Chen, Y.L., 2013. Retinoic Acid-Activated Ndrg1a Represses Wnt/beta-catenin Signaling to Allow Xenopus Pancreas, Oesophagus, Stomach, and Duodenum Specification. Plos One 8, 14.

Zhang, Z.L., Harrison, P., Gerstein, M., 2002. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Research 12, 1466-1482.

Zhang, Z.L., Harrison, P.M., Liu, Y., Gerstein, M., 2003. Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome. Genome Research 13, 2541-2558.

Zheng, Y.H., Rengaraj, D., Choi, J.W., Park, K.J., Lee, S.I., Han, J.Y., 2009. Expression pattern of meiosis associated SYCP family members during germline development in chickens. Reproduction 138, 483-492.

Zickler, D., Kleckner, N., 1999. Meiotic chromosomes: Integrating structure and function. Annual Review of Genetics 33, 603-754.

Zlatanova, J., Caiafa, P., 2009. CCCTC-binding factor: to loop or to bridge. Cellular and Molecular Life Sciences 66, 1647-1660.

Appendices

7.1. Appendix 1. Description and location of genes annotated upon the B.floridae ParaHox reassembly.

Table 7.1. Details of *B.floridae* ParaHox Reassembly annotated genes

Annotated Gene	**Strand**	**Number of Exons**	**Start (bp)**	**End (bp)**		**EST Support (+/-)**
Aminophosphoribosyltransferase	+	8	1300715	1320358		-
BRAFLDRAFT_119133	+	13	1310418	1320385		+
BRAFLDRAFT_119132	-	4	1320381	1322795		+
Transmembrane protein 185	-	7	1324085	1328951		-
Leucine-rich repeat and Ig containing protein	-	1	1332447	1335309		-
EST13 (Unknown)	-	2	1346086	1347020		+
EST12 (Unknown)	+	1	1357492	1357769		+
Transmembrane protein 45B	+	6	1358642	1364877		+
Kelch-like protein 31/36	+	4	1365524	1370270		+
BRAFLDRAFT_69541/2/3	+	2	1372251	1376582		+
BRAFLDRAFT_119117	+	2	1378172	1379861		+
Unknown_predicted (Hemicentrin?)	-	8	1382583	1406777		-
Predicted_(MS4A-like)	-	6	1382583	1447598		-
Histamine H3 receptor_(predicted)	+	1	1454813	1456117		-
Histamine H3 receptor_(predicted)	-	1	1458370	1459695		-
Alkaline Phosphotase(AP7)(Predicted)	-	10	1478053	1482640		-
Alkaline Phosphotase(AP5)(Predicted)	-	11	1489077	1494129		-
Alkaline Phosphotase (2)	-	7	1494522	1499439		+
RNA directed DNA Pol (Jockey)(predicted)	+	1	1502584	1503381		-
Alkaline Phosphotase (AP6)(predicted)	-	11	1504113	1509873		-
Alkaline Phosphotase (AP4)(predicted)	-	10	1510871	1516940		-
Alkaline Phosphotase (AP3) (Predicted)	-	10	1518817	1525513		-
Alkaline Phosphotase (1)	-	11	1526760	1536342		+
MFS-type transporter SLC18B1-like	-	6	1536849	1547046		+
MFS-type transporter SLC18B1-like (2)	+	6	1577477	1579588		+
MFS-type transporter SLC18B1-like (3) extended	-	14	1577477	1583077		+

MFS-type transporter SLC18B1-like (3)	-	8	1580066	1583077		+
Caspase-8 (Predicted)	-	2	1589598	1590746		-
CHIC	-	6	1592348	1603610		+
SCP1	+	4	1603681	1610748		+
Gsx	+	2	1617045	1620762		+
Xlox	+	2	1646363	1620762		+
Cdx	-	2	1661196	1673862		+
PRHOXNB	-	4	1683625	1689249		+
MFS-type transporter SLC18B1-like (4)	-	13	1692465	1701805		+
Ribosomal RNA processing protein 8-like	+	6	1703365	1707152		+
Ribosomal RNA processing protein 8-like (w/5'UTR)	+	6	1704949	1708266		+
SDK3/CLPB	-	13	1707533	1715585		+
GNPDA	-	7	1723882	1730109		+
Reverse transcriptase	-	3	1754722	1759557		-
EST2 (Unknown)	+	1	1761753	1762475		+
EST1 (Unknown)	+	1	1769150	1769897		+
Transmembrane protein 56-B-like	+	5	1801895	1804586		+
Carbohydrate Sulfotransferase 14	+	4	1805516	1811600		+
EST17 (Unknown)	-	4	1812178	1818278		+
EST16 (unknown)	+	1	1824954	1826141		+
EST15 (Unknown)	-	2	1830202	1831692		+

7.2. Appendix 2. EST sequences used in amphioxus Scaffold Annotation

BRAFLDRAFT_119133 (EST CLSTR 12291)

>gi|66316176|gb|BW729564.1|BW729564 BW729564 Amphioxus Branchiostoma floridae unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad052l03 3', mRNA sequence
TTTTGTCAATGAGACTTTTCGATCGTTCATTCAAAGTCATCTCCAGAAGAGGCCAAATTCTGCCATCCAG
CCCTGGTGACGTCAGTCAGCCGAGATGTCGTGTTTAGGCAGTCTGGTCAGGCCGGGCTGGCATGAAAGGT
GCCACCATGGTGCTGAAGGTGCTATCCATCTCCGCCTTCATGGTGGTGAACAGAGTCTCGCCCTGTCCGG
CAAACGCCCTCAGGACTTCCCTCTCCTGGGGGTTGGCGAGCGGCACGATACGCTCCACGAAGAAGCGGCG
AGCCTTACGCCACGAGTCCACCAGCTTGTTGTCCACAGCGCTCAGCCCCATCACAAGTTCCTCATCGTTG
TCCTTGTGAACCTTGAAGAACTCAGTCCACCTCTTGAAGGTAGACATGCTGCTCTCCTTGATCTTATCAT
TCATGGTCTTGACCTTTTCACGAACGTCGCGGCTCAATTGGTCCTTGATCTTGTCGTCATACTCGACGCG
CAGCTGGAAGGTGACCTCACGCAGCTTCTCTCCCATCAGGATGGAATACTGCTTCACGGAGTCGGCAAAC
GGAGACAGGCCGGTTTGGGCATCGGACAGGGTCTGTGCTTCTGCCTGCAGCTGCATCATGGTCGGCTTCA
GGTGGCCGTCAAAGAAGGTGGCGAAGGAAGCCTGCATGGTGGCGGCCTGGGCGGACAGGGTGTCCCAGTA
GCTCCTCATGACGCGGCGCTCATGGGCAGTCAGCAGCGGCTTGATCTGGTTCTGGTANTANTTCTGGGCC
TCTCCATGANCGCGGGCCATCTTGGCACTGTATCCAACCA

>gi|66297193|gb|BW710619.1|BW710619 BW710619 Amphioxus Branchiostoma floridae unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad052l03 5', mRNA sequence
ACATCGTTCGCTCCGAGGCTCACTGTGCAGTTCAGAATGAAGGTGTTCTTGTTGGTGGTGCTCGCTGCCG
TCTATGTGCAGGCGGAGCCGACCCCGTTCGCGCTCGAGGTTCGGGCCTACACGGAGCTGGTCGCTACATC
CATGCGGGGCTTCCGCGTCATGGCTGCCGGCGAGTTTACGGACAAGATCAAGGATGAACTGCGTCCCGAG
ATCCGTCAGAAGCTGAGGAACTCTGCCGACCGCATCGACGCCAAGATGGCCGTTCTGTCCGACAAGTGGA
AGGCGACTTACCAGGCGAACAGGGAGAACGACAGGGGACTGGCACGTGCTCTGGTTGGATACAGTGCCAA
GATGGCCCGCGCTCATGGAGAGGCCCAGAACTACTACCAGAACCAGATCAAGCCGCTGCTGACTGCCCAT
GAGCGCCGCGTCATGAGGAGCTACTGGGACACCCTGTCCGCCCAGGCCGCCACCATGCAGGCTTCCTTCG
CCACCTTCTTTGACGGCCACCTGAAGCCGACCATGATGCAGCTGCAGGCAGAAGCACAGACCCTGTCCGA
TGCCCAAACCGGCCTGTCTCCGTTTGCCGACTCCGTGAAGCAGTATTCCATCCTGATGGGAGAGAAGCTG
CGTGAGGTCACCTTCCAGCTGCNCGTCGAGTATGACG

>gi|66299474|gb|BW712900.1|BW712900 BW712900 Amphioxus Branchiostoma floridae unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad013d10 3', mRNA sequence
ATCGTTATTCAAGTCATCGCCAGAAGAGGCCAAATTCTGCCATCCAGCCCTGGTGACGTCAGTCAGCCGA
GATGTCGTGTTTAGGCAGTCTGGTCAGGCCGGGCTGGCATGAAAGGTGCCACCATGGTGCTGAAGGTGCT
ATCCATCTCCGCCTTCATGGTGGTGAACAGAGCCTCGCCCTGTCCGGCAAACGCCCTCAGGACTTCCCTC
TCCTGGGGGTTGGCGAGCGGGACGATACGCTCCACGAAGAAGCGACGAGCCTTACGCCACGAGTCCACCA
GCTTGTTGTCCACAGCGCTCAGCCCCATCACAAGCTCCTCATCGCTGTCCTTGTGAACCTTGAAGAACTC
AGTCCACCTCTTGAAGGTAGACATGCTGCTCTCCTTGATCTTATCATTCATGGTCTTGACCTTTTCACGA
ACGTCGCGGCTCAATTGGTCCTTGATCTTGTCGTCATACTCGGCGCGCAGCTGGAAGGTGACCTCACGCA
GCTTCTGGCCCATCAGGATGGAATACTGCTTCACGGAGTCGGCAAAAGGAGACAGGCCGGTTTGGGCATC
GGACAGGGTCTGTGCCTCCGCCTGCAACTGCATCATGGTCGGCTTCAGGTTGGCGTCGAAGAAGGTGGCG
AAGGAAGCCTGCATGGTGGCGGCCTGGGCAGACAGGGTGTCCCAGTAGCTCCTCATGACGCGGCGCTCAT
GGGCAGTCAGCAGCGGCTTGATCTGGTTCTGGTANT

>gi|66280644|gb|BW694073.1|BW694073 BW694073 Amphioxus Branchiostoma floridae unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad013d10 5', mRNA sequence
GTGTTCTTGTTGGTGGTGCTCGCTGCCGTCTATGTGCAGGCGGAGCCGACCCCGTTCGCGCTCGAGGTTC
GGGCCTACACGGAGCTGGTCGCTACATCCATGCGGGGCTTCCGCGTCATGGCTGCCGGCGAGTTTACGGA
CAAGATCAAGGATGAACTGCGTCCCGAGATCCGTCAGAAGCTGAGGAACTCTGCCGACCGCATCGACGCC
AAGATGGCCGTTCTGTCTGACAAGTGGAAGGCCACTTACCAGGCGAACAGGGAGAACGACAGGGGACTGG
CACGTGCTCTGGTTGGATACAGTGCCAAGATGGCCCGCGCTCACGGAGAGGCCCAGAACTACTACCAGAA
CCAGATCAAGCCGCTGCTGACTGCCCATGAGCGCCGCGTCATGAGGAGCTACTGGGACACCCTGTCTGCC
CAGGCCGCCACCATGCAGGCTTCCTTCGCCACCTTCTTCGACGCCAACCTGAAGCCGACCATGATGCAGT
TGCAGGCGGAGGCACAGACCCTGTCCGATGCCCAAACCGGCCTGTCTCCTTTTGCCGACTCCGTGAAGCA
GTATTCCATCCTGATGGGCCAGAAGCTGCGTGAGGTCACCTTCCAGCTGCGCGCCGAGTATGACGACAA

BRAFLDRAFT_119132

>gi|169561523|gb|FE575959.1|FE575959 CAXF9373.rev Amphioxus Branchiostoma floridae unpublished cDNA library CAXF, gastrula whole animal Branchiostoma floridae cDNA clone CAXF9373 3', mRNA sequence
GTTCAGTTCTTTTGTCTTTTATTTTCCACTGCACCAACGTATACAGAGTAAATCATTCACGCTGTTTAAA
CAACAAAACCAAGATGTGGTACTGATATGAGATCTAGTGGAAGAAGTACAATAATACAAAAAGAGTTACA
ATGTACTGATGGTACAATGTACTGAGAAACATATCACACAGAAAGAATGACGCAGATCCCTTACTAAGTT

ATGCACCCTTTCGACGACTTTTAGAGTTAATTATCAACAAAGCTCGCAGACTTCGGCCTCTCTTTTCATC
GTATATCTAT

>gi|169561524|gb|FE575960.1|FE575960 CAXF9373.fwd Amphioxus Branchiostoma floridae unpublished cDNA library CAXF, gastrula whole animal Branchiostoma floridae cDNA clone CAXF9373 5', mRNA sequence
TGGGAAGCCAGCACCGACCTCTCCACTTTATGCAGTGTTAGACCTCGGAGGACGGCTCTACCCGGTTGAC
TTAAAAAAAGTCATCGGTTGGAGAACGGTCAAAGTTGGTGACGTAGTGTGCGGAAAGTGGTCAAGACGGA
TCATAAAGGGCACATTGGTGAAAATTGGAACTATACATGAGGTGGCGCCTCTCCTCAGCGCTGACGCCAA
CAGGCACTACTATGACGACACCGTGCACGACCTCGATGCCTACACCGTCAAGAAATTCAAAGGGGCGTGC
ACGATCTTGACACGCTCTTCGCCAGCCCATTCGCCGGCGATTCAGCGCCAGTGGTCGGCCCCGCCATCAC
AGCCAGCCGAGACGACACAGGTACACACCGGAATGGAGGAGCGGATGGACAACCTGACGATATTATATAT
AGTGCTACTGTATGTAACAGGTTTCTTCATTGGTTAATGACAGGCTCCAAACTTGTGTTAGAGATATTTC
GATTGAATATCATATTCGATACTGACAAATGTTGTTGAATAACCATAGATATACGATGAAAAGAGAGGCC
GAAGTCTGCGAGCTTTGTTGATAATTAACTCTAAAAGTCGTCGAAAGGGTGCATAACTTAGTAAGGGATC
TGCGTCATTCTTTCTGTGTGATATGTTTCTCAGTACATTGTACCATCAGTACATTGTAACTCTTTTTGTA
TTATTGTACTTCTTCCACTAGATCTCATATCAGTACCACATCTTGGTTTTGTTGTTTAAACAGCGTGAAT
GATTTACTCTGT

>gi|169554449|gb|FE568550.1|FE568550 CAXF17025.fwd Amphioxus Branchiostoma floridae unpublished cDNA library CAXF, gastrula whole animal Branchiostoma floridae cDNA clone CAXF17025 5', mRNA sequence
GTCTAGCTGTACGTCGTTACGTTCACCCAAGTTATCGGCAGTTATGTCGGTTCAGCTCCAGCCAGGGACG
GCTTCAGAAGACACCACCACAACTATGGATCATGACCATACGCAACAACCACAACCCGCTGCAACATCGC
AGAAGGCAACTAAGAAAAGGAAGAAGAGGAGAGCGAAATATGAGAAGCCCAGAGTTCTAGCCGTTGGGAA
GCCAGCACCGACCTCGCCACTTTATGCAGTACTAGACCTTGGAGGACGGCTCTACCCGGTTGACTTTAAA
AAAGTCATCGGTTGGAGAACGGTCAAAGTTGGTGACGTAGTCTGCGGAAAGTGGTCAAGGCGGATCATAA
AGGGCACACTGGTGAAAATTGGAACTATACATGAGGTGGCGCCTCTCCTCAACGCTGAAGCCGATAGGCA
CTACAATGACGACACCGTGCACGACCTCGATGCCTACACCGTCAAGAAATTCAAAGGGGCGTGCACGATC
TTGGCACGCTCTTCGCCAGCCTAGTCGCCGGCGATGCAGCGCCAGTTGCCGACCCCGCCATCACATTTAG
CCGAGACGACACAAGTACACACCGGAATGGAGGAGTGCATGGACAACCTTACAATATCATAGTGCTACTG
TAACAGGTTTTTTCATTGGTTAATGACAGGCTCCAAACTTGTGTTAGAGATATTTCGATTGAAGATCATA
TGGTAGGTCGATACTGACAAATGTTGTTGAATAACCATAGATATACGATGAAAAGAGAGGCCGAAGTCTG
CGAGCTTTGTTGATAATTAACTCT

EST 13 (Unknown)

>gi|66467126|gb|BW858910.1|BW858910 BW858910 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne057e03 5', mRNA sequence
TAATTTATGCATGGTATTGTTCATCATTGACTAACGTACACATGTCACCATCATAAAATTCCCATTATTA
ATCATAAAGCGTTTTTGCAATTTTTACATAAATTATGCAAATAAGTTCCTCATTACCATATTTAGTATCT
GCTTATATTCCACCTATCATAGTTAGCATGTGTTACATTTATTGAAGTCCAGTTATTGAAAACAATGGAA
TTATACAATTTCCTCATTAATCATGCAAATTAAGTCCTCATTTGCATAAAATGTATATCATTATGAACAT
ATTTGCCTAAGGTACCCGCATGCCTAGTATGATGCCAATCCATCAATCCTTTCTGCAGTTATCCTCTTTA
GAATGTCTTGACAAAAACGCCCCTGCAGTTCC

>gi|66543391|gb|BW916863.1|BW916863 BW916863 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne057e03 3', mRNA sequence
ATGAGTTCAAAACAAATTAAAGAATTACAATTAATATCATTTGATGGTTTCTTCGAACACCAATAAAGAC

AATGCCTACCGTTATATACAAACCCGAGATAGAACTGGCATGTTTGCATAATTACCTCCATGAAAAAATG
GAGGTATAGTTTTGAGTGTGTCTGTGTGTGTGTGTGTTTGTGTGTGTGTGTGTGTCTGTGTGTCCGCA
TATTTGTGGTCATCATAACTTGAGAACCTCTTGATGGACTACGATGATATTTGGTATGTAGGTAGGGGTT
GGGAAGACGAAGGTCAAGGTCAATTTTGGGCCCCCTGGTGTGTGGCCTTGGTACTGCAACGCANCTTCCG
GTTTTGCTATCTCGGTGTTCTGAACATGCTATGGTC

EST 12 (Unknown)

>gi|66449905|gb|BW841689.1|BW841689 BW841689 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne015f24 5', mRNA sequence
ATTAATTTTGACCAAGAGTAGAACAATCAATTGGTAGATGATTTAAGTACTTAATTTATCACAAAAGTTA
GAAGTAACATGGGCAAGATAATAGACATATTCAGAACTGTTCTATTGTTTCCATGTGTACTTCATCTGGT
CTACATTTCTTTTCATGCACATGCCAGAATACCTTGCTAATATTACATCATTACATAAGTGTAGAAGTGG
TCACCTTTAAGTGCACAAAGAATATTGCCAATTATTAAAGCTCGTGTAAAACAAAATAAATTTCATTTAC
TACGTCGCAAAAAAAAAA

>gi|66519593|gb|BW899393.1|BW899393 BW899393 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne015f24 3', mRNA sequence
GAGCTTTAATAATTGGCAATATTCTTTGTGCACTTAAAGGTGACCACTTCTACACTTATGTAATGATGTA
ATATTAGCAAGGTATTCTGGCATGTGCATGAAAAGAAATGTAGACCAGATGAAGTACACATGGAAACAAT
AGAACAGTTCTGAATATGTCTATTATCTTGCCCATGTTACTTCTAACTTTTGTGATAAATTAAGTACTTA
AATCATCTACCAATTGATTGTTCTACTCTTGGTCAAAATTAAT

Transmembrane Protein 45B

>gi|66510871|gb|BW893194.1|BW893194 BW893194 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne161b20 5', mRNA sequence
ACACACACAGACTGGTGCCGTCTCTGGGATATTTTCCTTCTGTTGTACCGATTTCCGTGGACAAGTAGAC
GAGCTGTAACCATGGATATGGACCACGGGCACCACGACCATTCACACGGCGCGGATGCTATACAAGACCA
TCCGGGATCGGGGACGTTTGGAAGCCACGCCGTCCTGGGGACGTTCTTCTTCGTGTTCGGGCTGTGGTAC
GCCGTGAAGACCTGCTTCTTTACGCTGGAGAGGCTTCACACACAGGGACAGGGAAAACAGCCTGCCAGAA
ACAAGACATGGAGAGATCACATAGGCTGTGCCAAACGTACGTTGGGATTCCTCCTCTACTCCATGGACCC
CATGTTCAAGATCATCTCTTGTACGATAGGAATGTTAAGCCAGATGTCTCTTGGTGCCCACTGGAGGCTT
CGTGACCCGGTCACAGGAGAGTTTGTGGAACAAGCCGACTGGCAGCTGGTCACCATGTTCTCCTTCTTCT
TCTTCTCGGGGCTTGTGGACATCTTGGTCAGAGTGAAGTCACCGATTCCACCAAAACAGCGACAAGTTCT
TCATGAGTCTGGCGTTGTTCGTCGAGTCCTACTTTTTCTTCTATCATGAA

>gi|66666354|gb|BW951085.1|BW951085 BW951085 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne161b20 3', mRNA sequence
ATTAATTTGTTACTAACAAGTAGATTGGATTCTTAAAAACTTTTAAAAAACAACGTGTGGTAACAAAGGT
AGTTCGTCAACACTTGCTACAGTCACAATACAAATACACAATATTCAAACAACAGTCAACACAGTAACTC
ATCTTCAACAGGATATAACGTAAAGCATTAACATGCCATGTTCTAGGTTCTAAGTCATTGCATTGTTGCC
AATATTCTATCAGACGAGCGACGAACAGAACTCTGTATTGCCATCTATGTCATATAAAATTGCTACACAT
GTAGTTATGTAGGAATGACTTTAAACTTTATATTAAAGATACCAACAGGCCTGTAGTCGTATAACGATTT
TGCTATAATATCATTGCATAGTAACAGGAATCTTGCAGAAATAAAATGTATGATACAATTTATATCATAG
TGAATAAGAGACNCATATCTAGCTATGACATGGGCTCCCNGGTANTT

>gi|66282993|gb|BW696422.1|BW696422 BW696422 Amphioxus Branchiostoma floridae unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad001c24 5', mRNA sequence
ACACAGTCTGGCGTTGTCTCTGGGATATTTTCCTTCTGTTGTGCCGATTCCCGCGGACAAGTAGACAAGC
TGTAACCATGGATATGGACCACGGGCACCACGACCATTCACACGGTGCGGATGCTATACAAGACCATTCG
GGATCGGGGACGTTTGGAAGCCACGCCGTCCTGGGGACGTTCTTCTTCGTGTTCGGGCTGTGGTACGCCG
TGAAGACCTGCTTCTTCACGCTGGAGAGGCTCCACACACAGGGACAGGGGAAACAGCCTGCCAGAAACAA
GACATGGAGAGATCACATGGGCTGTGCCAAACGTACGCTGGGATTCCTCCTCTACTCCATGGACCCCATG
TTCAAGATCATCTCTTGTACCATAGGAATGTTAAGTCAGATGTCTCTTGGCGCCCACTGGAGGCTTCGTG
ACCCGGTCACAGGAGAGTTTGTGGAGCAAGCCGACTGGCAGCTCGTCACCATGTTCTCCTTCTTCTTCTT
CTCGGGGCTTGTGGACATCTTAGTCAGAGTGAAGTCACCGATTCCACCAAACAGCGACAAGTTCTTCATG
AGTCTGGCGCTGTTCGTCGAGTCCTACTTTTTCTTCTATCATGAA

>gi|66301723|gb|BW715149.1|BW715149 BW715149 Amphioxus Branchiostoma floridae unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad001c24 3', mRNA sequence
CAACAATTAATTTGTTACTAACAAGGTAGATTCGACAAATTGGATACTTAAAAACTTTTAAAAAACAACG
CGTGATAACAAAGGTAGCTTGCTACAGTCACAATACAAATACACAATATTCAAACAACAGTCAACACAGT
AACTCATCTTCAACAGGACACAACGTAAAGCATTAACATGCCATGTTCCAGGTTCTAAGTCATTGCATTG
TTGCCAATATTCTATCAGACGAGCGACGAACAGAACTCTGTATTGCCATCTATGTCATATAAAATTGCTA
CACATGTAGTTATGTAGGAAGGACTTTAAACTTTATATTAAAGATACCAACAGGCCTGTAGTCGTATAAC
GATTTTGCTATACTATCATTGCATAGTAACAGGAATCTTGCAGAAATAAAATGTATGATACAATTTATAT
CATAGTGAATAAGAGACACATATCTAGCTATGACATGGGCTCCACGGTAATTTCTGAAGACAGATCCTCT
TATCTGAAATCTGAACGGCCTGTCAAACCTTTGTAATCTAGCTGCAAGTGCTGTTTAAAGCCATCTCATG
AGGATGCCATAAATGTATTTGATGGAAACGAGTATGGAAACTAGTACAACACCGTGATGGTTCTAAGAAA
AANCAATCATAGGTTGATAACTCTGTTCTAA

Kelch-like protein 31/36

>gi|66283828|gb|BW697257.1|BW697257 BW697257 Amphioxus Branchiostoma floridae unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad021i19 5', mRNA sequence
TGGCTGCCCAACCGCCGGTTGACCTCTTTCTCTGCAGACAAACCGTCAAAAGATGTTTCGTATGAAAACG
TTGAAAAATCCACGATTGTGAAACAACACATGCGCAGGTGAAGATAAGACAGGATGGAGAAAATAAACGA
ACGCAAGTTTTCCCTGGCGTCCCATGGGGCCTCTGTTCTCGCCGGATTTAGGGAACTTTACCAGACTGAG
CTGCTGAGTGACGTGTGTTTGGTCGCAGAAAAACGGGAATTCAGATCACACAAGACACTCCTAGCCGCTT
GCTGCCCCTACTTCAGGTCAATGTTTTCAATCGACTTGAGAGAAAAAGAGGAGACGACGGTTGAGATGCA
CGGCACGACCGCTAGAGGGCTCTCCGCCATATTGGATTTTCTGTACAGCGGAGATCTCACGTTGAACGAC
GAGAATAAGGAAGACGTGTTGTCGACTGCTTGTTATCTCCAGGTAGACGCAGTGATTGACATGTGCTGCT
CTTATCTAAGAGAGAACATCCACATGAACAACTGCATCGGAATTTGGAACTTGGCCTGCGCGCTCAACCT
GCACGAATTGAAGGATTTCGCGGAGAACCACGTGACCAACAACCTGATCGAAGCTTCG

>gi|66486113|gb|BW871436.1|BW871436 BW871436 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne105e22 5', mRNA sequence
TTCGGTATCCCAGCGACCAGACCACTCGGAGGCGAACATGTTGCTACGCGACACGATTCAGCGGACACTT
CCCGAGAATCCACGATCTGCAGAGCTTCAAGGGAAGTAGTTCTGGCGGTGGGAGGAAGGCTGATGTATAA
CGAACGGCCCGCGGTCCAGAGGACTTGCGTCAGTTTCTGCGACGTCAGGTCGCCTGACGGAGACAAGCCG
TGGTACGAGATGACCCAGATTCCCATCCGGAGGAGGAACTACTGCGCAGCCGTGCTGGACGATGAGATTT
ACGTTGTCGGGGGGAGGGAGTGGGACAAGGAGGCCCGCGGGTACGACCGGTGGTCCGCTGCCGCCTTCTG
CTACAACCTCCGGACAGCGAAATGGCGGGAAGTCTCCAGCTTGTCCACGAAAAGGAGCTGCTTCTCTATG

GACGCCATCGAAGGGAACCTCTTCGCGGTTGGAGGGGACGAGGATCACGAGGACACTAGCATCCTGTCCT
CAGTGGAACGGTTTGATCCCATACAAAACCTTTGGTGGCCCTGCTCGGAAA

>gi|66562246|gb|BW929634.1|BW929634 BW929634 Amphioxus Branchiostoma floridae
unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne105e22 3',
mRNA sequence
AGTTAGGGCATACAGACAATGGTTGCATCGCACACATTTGGTTGCATAGTTTTGTATTGAATAAATTACC
ATCCACAAGTTTATCCTCTAGGATAACAAACCACATTCATATACTCTTGGTGTTGTTAGATGTTCAGTAG
GCGTAAAATATTTGGCATGGCGTGTTGCACAGCAATGTTCGTGTCAAAGGTCAACGTCCGAGCAGATGTT
GGGGATAGAAAATCAAGAAGCAAACATCTGCAGTTATTTTGTATCAATGAACTTAGGAAAAGTCTGTCGT
CAAAACGTCCGACTGGCTCCTCCTATGGGCCCAACTTGGTCGGCTAGACGTTTCTCCCGTCCTGAATGGC
TCTATGAAAGAGCCGGGGACCATCATCTTGT

>gi|66434806|gb|BW826590.1|BW826590 BW826590 Amphioxus Branchiostoma floridae
unpublished cDNA library, larva whole animal Branchiostoma floridae cDNA clone bflv030f13 3',
mRNA sequence
GTTAGGGCATACAGACAATGTTGCATCGCACACATTTGGTTGCATAGTTTTGTATTGAATGAATTACCAT
CCACAAGTTTATCCTCTAGGATAACAAACCACATTCATATACTCTTGGCGTTGTTAGATGTTCAGTAGGC
GTAAAATATTTGGCATGGCGTGTTGCACAGCAATGTTCGTGTCAAAGGTCAAAGTCCGAGCAGATGTTGG
GGATAGAAAATCAAGAAGCAAACATCTGCAGTTATTTTGTATCAATGAACTTAGGAAAAGTCTGTCGTCA
AAACGTCCGACTGGCTCCTCCTATGGGCCCGACTTGGTCGGCTAAATGTTTCTCCGTCCTGAACGGCTCT
ATGAAGGAGCCGGGGACCATCATCTTGTAGGCAGGTACGGGCCCGGGGTACCGCAGGGTGTTACTGACCG
TCCACTCACGGGCTGACTCGTCATACACTTGAACCTTGCTGACGTAGTTCATCTCACCGGAAAATGTGCT
CTGGTCACCCCCGAGAACCGTCATGTAGCCCTCCACGACCACGGCGCTGCACAGCCCGGCCGGTACTCTG
AGCGGCTCTATGAAGCTCCACTGGTCCCGCTCGGGACTGTAGCACTCCACCACGTCTAACGGCTCATTAT
GAGTCCG

BRAFLDRAFT_6954[1/2/3]

>gi|169579254|gb|FE595968.1|FE595968 CAXG9204.fwd Amphioxus Branchiostoma floridae
unpublished cDNA library CAXG, larva whole animal Branchiostoma floridae cDNA clone CAXG9204
5', mRNA sequence
AAATGTGCCTCAACTGTCGTCGGCGTGTACTGTTGTCGCCATGTCGGATGCTCCAACCATCGTTTGTGTC
GTCTGGAAAGCGGCGGCTCTGATCTACTGGCTCAGGGACTCCTGGGAGAACATTTTCACCCCTCCCATCC
CCGAGCTGGTCGTAGGGACCTTTGGTATAAGTGCACTGTATGACTGGGTGGCAACTTTAGTTCCACTGGA
CTTGCCCTACGCAGTGCTAGTGAAGGTTCTCAATCAGTGTGTTTTGATTGGATGTCTGCTCGTTGCACAC
AAAACGCTCGAGAACTTAGAGAAGCAGGTTTCCCACCAAATGTCCCTCTGTCACCGCGCGTGTTACGCCC
TGCTGGTTGAGAACTGTGTGGCCTTCAACCTGACCTGGAACTACGTCCAGACCGCGGCGCTGGTCAGCGA
GCTGCTCGTCCAGGACTTCCATATCGTGCCGGACACCGTGGTGACCCTTCACCTGGTCCTACTGGCGATC
GGCGTCCTGGTTACTGCTACAGTTGAGATAGTCTTCCGTGATAAATTCAGGTGGACAGTAGCGTCTTTTC
CTCCCCTCCTGATATGGGCGTTTTGTCTCCGTCGTGCTAGCGGTGTGGATGACACGTTTCTGTACGGCCT
ACTCGTCCTAACCGCGCTCGTCATGTTAGCGAAGGTTGCTCCGGTAACACGAAGTCGTCCCGACTTTTCC
CAAGAGATAACAAGAGAAATTGGTTGAAGATCAGCTGTCAGAAACTGTTGACACTGACTCAGATCTTGTA
AC

>gi|169579253|gb|FE595967.1|FE595967 CAXG9204.rev Amphioxus Branchiostoma floridae
unpublished cDNA library CAXG, larva whole animal Branchiostoma floridae cDNA clone CAXG9204
3', mRNA sequence
GGGNGAAATTAGTAAAGGCGTATTACACCGTTTATGATACAATGTAGTGGACAACGAGTGAATCGCATGC
TGCATAAAATCACAATACACTGACTATTTGTTTACAACACAGTAAAAGCCCGGTTCAAATATCCTATATA
TATTATCATGTCGGGTTTCCTTGCATAAAATGTCTCTCACTGAAACCTTAATTGAACCATATTTTCACGT
AGTTTTACTCAAAGTGTACAGTGTTTTGCCATGTACTGCTGGTAGGGTTTTTTAAAATAACTCCATTCTC

AGAAGAGTCTTGAAAGTTGTTTACAAGATCTGAGTCAGTGTCAACAGTTTCTGACAGCTGATCTTCAACC
AATTTCTCTTGCTTATCTCCTTGGGAAAAGTCGGGACGACTTCGTGTTACCGGAGCAACCTTCGCTAACA
TGACGAGCGCGGTTAGGACGAGTAGGCCGTACAGAAACGTGTCATCCACACCGCTAGCACGACGGAGACA
AAACGCCCATATCAGGAGGGGAGGAAAAGACGCTACTGTCCACCTGAATTTATCACGGAAGACTATCTCA
ACTGTAGCAGTAACCAGGACGCCGATCGCCAGTAGGACCAGGTGAAGGGTCACCACGGTGTCCGGCACGA
TATGGAAGTCCTGGA

BRAFLDRAFT_119117

>gi|66466520|gb|BW858304.1|BW858304 BW858304 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne055j17 5', mRNA sequence
TCGTCTACAAAACTTCCTGGAACGACTACATCCGGGAACTGAGCCCAGTGATCACCAGCTTCAGTGTGAG
CATCCTCTGTAACGCTGCCTCCACCATTTTGGATTTTTATGGTTTCCCCCTGTCCGCCGGTGTGGCCAAA
ATGCTCCCGCCGTTCTTCCTGTGCGCATGCCTGTTCCTGACGCTGAACATTGCGGAGAACCACATGGCTG
AGATGTCCACAGCACACCGCTGGTGTTACCTGCTGCTGGTCGAGAACTGCGAGGGTTTCAGCCTCACGTG
GAGTCACGCGGAAACTACCTCCACAGTCAGCACCATTCTGGTCNGACATTTCAACGTG

>gi|66542562|gb|BW916252.1|BW916252 BW916252 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne055j17 3', mRNA sequence
ATTGTAACTGTCGCGTGGTTTTGTACTCATTTCTACTCAAGGCCTGGCCACATGCGTTGTGCGATAGTAG
GACGATCAGAAACTGTGAGCATTCATGAGAAAGTCACGAACTGATACAGAAGTATTGTTACGGAAAAGGT
TGTGTTAGTTCATTGTCAGCTGTTGGTCCTGCCACATCTTGAAAATTATACGACATCAAAATCGCACGAC
GGGTGAGACCATACATCAGGCTGTGTGTGTCTGTGACGAACTTTGAGACATTGAGCATTGTTTATATTAT
GGCAAGATGATCGAACCAAAATGAAGATAGATCTAGATACTACTCGTCAACTTATAAGATATCCAAAACC
TAGATCTCTCATTTCATGATCATTCTATTTATTTAT

Alkaline Phosphotase 2

>gi|66493081|gb|BW878404.1|BW878404 BW878404 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne131d17 5', mRNA sequence
TTTGAGTGCATATACTATCTGTAGGTCTGCACAACTAAAGTTAGATAGATATTAACAACTTGGTACCACG
TGGTACTTTTTGTTGCAAACATCTGGAACCGCTAGATGGCACTTCTGCAACTAAGCCGTCTGGTTACAGC
TCTTTTTGGCCTAGTTGTTACTGCTGTATGTAAATGAGCAGATGTCGTTAATGCGTAACTTACTAAAGCT
AAGTCGCAACTCACACCGTCTGAACTGTCTATATTGTGAGATGATGAACCACGGCAATATTTTACATGAG
AAATAAAGGCATGTTGTTAAGCACACGGTACTAAAAATAAAAATAAATAAATCGTATGTGAACGTGGCTT
CTGGAGGTCATATAAAGGTCAACGGACTTTACACAGGAGGTAGAATCTACACTACTTCAATAAGAGTAGT
TTTCGTCAAGCCGAGTTATTCATATGTTAGAT

>gi|66569964|gb|BW936068.1|BW936068 BW936068 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne131d17 3', mRNA sequence
GACGATCTACATATGAATAACTCGGCTTGACGAAAACTACTCTTATTGAAGTAGTGTAGATTCTACCTCC
TGTGTAAAGTCCGTTGACCTTTATATGACCTCCAGAAGCCACGTTCACATACGATTTATTTATTTTTATT
TTTAGTACCGTGTGCTTAACAACATGCCTTTATTTCTCATGTAAAATATTGCCGTGGTTCATCATCTCAC
AATATAGACAGTTCAGACGGTGTGAGTTGCGACTTAGCTTTAGTAAGTTACGCATTAACGACATCTGCTC
ATTTACATACAGCAGTAACAACTAGGCCAAAAAGAGCTGTAACCAGACGGCTTAGTTGCAGAAGTGCCAT
CTAGCGGTTCCAGATGTTTGCAACAAAAAGTACCACGTGGTACCAAGTTGTTAATATCTATCTAACTTTA
GTTGTGCAGACCTACAGATAGTATATGCACTCAAA

>gi|169536376|gb|FE550389.1|FE550389 CAXC1779.fwd Amphioxus Branchiostoma floridae unpublished cDNA library CAXC, neurula whole animal Branchiostoma floridae cDNA clone CAXC1779 5', mRNA sequence
AGAATGTGGTGCTGTTCCTAGGCGATGGGATGGGCGTGGTCACTGTGACGTCAGCACGGATCCTGAAGGG
ACAAAAGGCGGGGAACCCGGGAGAGGAGACCGTGTTGAACATGGAGACACTGCCGCATGTGGCACTGTCC
AAGACCTACAGCATAGATGCCCAGACGCCGGACTCGGCTTCGACGGCCACCGCGTACCTGTGTGGCGTGA
AGGCTCCTTACGAAACGGTCGGGTTAGACGGCAGGGCTCGGCATATAAACTGTAGCTCTTCAAAAGGCAC
AGAAGTGCTGTCAGTTTTAGATTGGGCAGAATCTGCAGGAAAGTCCACTGGTATCGTGACGACAGCCCGC
GTGTCTCATGCCACCCCAGCGGCAGCCTACGCTCACTCAGCCTACCGTGGGTGGGAGGTGGACAGCGTTC
TTACGCCTGAAGCTGTCCAGAACGGATGTAAGGACATCTCTGCTCAACTGGTGGACGACAATCCAGGCAT
TGAGGTGATCCTAGGAGGAGGACGTGCGACGTTCCATGCCGGGGCCGATCCGGAATATCCGGACGATCCT
AGATTTAACGGTGTCCGGAGTGACGGCAGAGACCTGGTGCAGGACTGGCTGGACGGGAAGACATCAGCGC
GTTACGTCTGGAACGGGACGGACTTCCGGACCATCAATCCACAAACAACGGACTATCTTCTGGGTCTTTT
TGAGTTCAGTCATATGAAATACATGACAGACAGAGAGGATTCTCCGTCAGAAGACCTACCCTTGCGGGAA
TGACGCGAACTGCCATC

>gi|169536375|gb|FE550388.1|FE550388 CAXC1779.rev Amphioxus Branchiostoma floridae unpublished cDNA library CAXC, neurula whole animal Branchiostoma floridae cDNA clone CAXC1779 3', mRNA sequence
AAGAGGTGAAAATAGTTTATTTCAAAGACCGATAGAACATATGAATGACTCGGCTTGACAAAAACTACTC
TTATTGAAGGAGGGGAGATTCCACCCCCGGGGGAAAGGCCGTTGACCTTTATATGACCCCCAGAAGCCCC
GTTCACAAACGATTTATTTATTTTTATTTTTAGGACCGGGGGCTTAACGACATGCCTTTATTTCCCATGG
AAAAAATTGCCGGGGGTCATCATCCCACAATAAAGACAGTTCAAACGGGGGGAGTTGCGACTTAGCTTTA
GGAAGTTACGCATTAACGACATCTGCTCATTTACATACGGGAGTAACAACTAGGCCAAAAAGAGCTGTAA
CCAGACGGCTTAGTTGCAGAAGGGCCATTTAGCGGTTCCAGATGTTTGCAACAAAAAGTACCACGGGGGA
CCAAGTTGTTAAAATCTATCTAACTTTAGGTGGGCAGACCTACAGATAGTATATGCACTCAAAGCCACCA
AACAAGCTGGA

Alkaline Phosphotase 1

>gi|169563656|gb|FE577252.1|FE577252 CAXG1044.fwd Amphioxus Branchiostoma floridae unpublished cDNA library CAXG, larva whole animal Branchiostoma floridae cDNA clone CAXG1044 5', mRNA sequence
ACAAGAAACATGGCTACAACCCGGACCGTCACACGTCTCTTTCTTTTGGTGGCGTTTACGGCTACAACGT
TCGCCAAACCAGCAGCCGATCGCGACGATGAGTCTTATGACGTGGACCACTCCGGGCAGCTGTGGCAGAC
CCGGGCCCGCAGAGCGCATGCGCAGTCCAGGACGGTACCCACAGACCAGGAGCGCACGCCCGACTACTGG
ACCAACATGGCGCGGGCCTCCATCGATGAAGCTCTCCGCCTGCAGACCCTCAACACGAACGTGGCCAAAA
ACGTGGTGCTGTTCCTGGGAGACGGGATGGGCGTTTCCACGGTAACCACGGCACGGATCCTGAAGGGACA
AAAGGCGGGAAACCCGGGAGAGGAGACCGTGCTGGCCATGGACTCATTACCTTACACCGCCATGTCTAAG
ACCTACAATATCGACGCCCAAGTCCCTGACTCGGCCGGTACTGCTACAGCGTTCTTGTGCGGGGTGAAGG
CGGAGGCCGGGGTCATCGGCGTGGACGGGAGAACACGGTACGGCAACTGCAGTTCCTCCAAGGGTCACGA
GGCGGAGTCCATCATTGTGCACGCGGAAAGAGCAGGGAAGTCGACTGGGATCGTCACCACCGCCCGAGTG
ACCCACGCTACGCCGGCGGCGGCATACGCTCACTCGGCCGCACGGGGCTGGGAGGCCGACAGCGATCTGA
CGGCGGA

>gi|169563655|gb|FE577251.1|FE577251 CAXG1044.rev Amphioxus Branchiostoma floridae unpublished cDNA library CAXG, larva whole animal Branchiostoma floridae cDNA clone CAXG1044 3', mRNA sequence
AACAGGATTGACTGAGATTTTATTTCTCGGTATTTTATTCTAGGTATTGTATACAACCTTGTCCACATAC
TTGTCATACATTTTCCAATTCATTCGAGCAACACGGTTAATTTTCATGGCGAATCCTAACATAAGACAGG

AACTAAAGAAACTCAAGTGACGACAAAAAAAGTAACTTGAATACATACAGAGAAATGCATTTGTTCCATG
TACCATAAAATCAGGATTGATGACAAACTGAACTGACTATAGCAGTAACAATGGCGATCAATGAAAAAAG
GTGAACATGAAACGTTAATTGTGCAGTGTGTGTAGTTGGTGACTTGTCTGACTCAGAAAAATCGAGTCAA
AGCTAATTTGGCGGTGGGGAGGGACCTGTGCGTAGGAATTAAATATTTCGATCCAGGACATACTCTGGAT
CTAATCTCCAAGCAGAAGTCTGGATGGAAGATTGTAAAGCCCTCTGACCACCGGACGCCCGTGTGTGTGT
TGCTATCAAAGACGCTCGATAGCCAGAGAACGTTACAATCGTCCACCCAACATCTGCTTGGAGATTCTCT
CTGTAACTCTCTCGCTCTCTCTCTCTTTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCGCACGGG
GNGGG

>gi|38190691|gb|CF919489.1|CF919489 Bflor531.000574 Amphioxus 26 hrs cDNA library (Name convention: BFL26 or MPMGp531) Branchiostoma floridae cDNA clone MPMGp531O1753;BFL26_53O17 5', mRNA sequence
CAGCGCGTCCGCTGGAACGGAACGGACTTCAGAGATGTCGACCCGGAATCGACGGATTATCTTCTTGGTC
TTTTTGAAAGGAGTCACATGAAGTACACTGCAGACCGCACGGATGACCCGGCCGAGGAGCCCACCATCGC
CGACATGACGAGGAAGGCGATAGAAATACTCCGCAAAAACGACAATGGTTTCTTCCTGCTAGTGGAAGGC
GGCCGAATAGACCACGGACATCACGCGTCCAAGGCGGTGAAAGCTTTAGAGGACACCGTGGCGTTTGATG
ACGCCGTGCAGGTGGCTAAAGACATGCTGGACACGTCTGACAGCCTCATAGTGGTGACGGCGGACCACTC
CCACACACTGACGTTTGCAGGGTACCCTGATAGGGGCCACCCCATATTTGGACAAAACGTGTACACGTCA
TCCACCCCTGATAATACCTGGGATGAGTTACCGTACACCACCCTGCTGTACGGGAACGGTCCGGGGTACG
CGCTGGTGGAGACTACAAACGGGAACGACACGCAGGTCACACGCCAAAACATCACGGACGTCGATACAGC
GGATAAGGAGTACGAACAGCATAGCGCCGCGCCGTTACGGAGCGAGACGCACGGAGGGGAGGACGTCATC
ATCATGGCGGACGGACCCATGGCTCACCTGTTCACGGCGTA

>gi|66505241|gb|BW889088.1|BW889088 BW889088 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne149m03 5', mRNA sequence
CTTCGTATTTGTTTCGACTACACTCCATATTTGGCATGATTTCAGCCGATTCTTTCTGTTGTGTTTTTTT
GTATGTGCTTGTACTTATGTCTATGTACTGGGGTGAAGTCTGCCAACTCTGATTGCAATGTTTCAGGACA
TAACATGTACACGTACTCTACCCCTGATTATGACTGGGACAAGTTACCGTACACCACCCTGCTGTACGGG
AACGGTCCGGGGTACGCGCTGGTGGAGACTACAAACGGGAACGACACGCAGGTCACACGCCAAAACATCA
CGGATATCAATACAGCGGATAAGGAGTACGAACAGCAGAGCGCCGCGCCGTTACGGAGTGAGACGCACGC
AGGGGAGGACGTCATCATCATGGCGGACGGGCCCATGGCTCACCTGTTCCACGGCGTACAGGAGCAACAC
TACATCCCACACGTTATGATGTACGCCGCCTGTCTGGGGGAGTACACAGAGCACTGTGACAAACCGGGAA
CACCCAAACCCGTCAGGGATGC

MFS-type-transporter_SLC18B1_like(1)

>gi|66412177|gb|BW803961.1|BW803961 BW803961 Amphioxus Branchiostoma floridae unpublished cDNA library, larva whole animal Branchiostoma floridae cDNA clone bflv029a20 5', mRNA sequence
TTATTATGTATGAGACGATTCGTTTCCTTTCCGTCATATGCACTGTGCAGCAGAAGATAACACAGCAGAC
AGGCGTGTGAGTTGTGACGTCATAAAGTACACACAGACCAACATTAGTTATATTCACCCACAAGAAGATT
ACAGAAATGAAGAAAGCCGAGAAAGACGACAGCGAGACCTCGCGCTTGCTTGATGCACAAAACGCCTCAA
GACATATACAGCTCGAGACAGACAGTGTCCCGGATGTTGTGACGTATGGCAGTGTGACGGACGAACACAC
GGAAGAAGTCGCTAATAACACAGAGGCTTCTACA

>gi|169566206|gb|FE581401.1|FE581401 CAXG13886.fwd Amphioxus Branchiostoma floridae unpublished cDNA library CAXG, larva whole animal Branchiostoma floridae cDNA clone CAXG13886 5', mRNA sequence
GATCCGTTTCCTTTCCGTCATATGCACTGTGCAGCAGAAGATAACATAGCAGACAGGCGTGTGAGTTGTG
ACGTCATAAAGTACACACGGGCCAACATTAGTTATATTCACCCACAAGAAGATTACAGAAATGAAGAAAG

CCGAGAAAGACGACAGCGAGACGTCGCGCTTGCTTGATGTACAAAACGCCTCAAGACATGTACAGCTCGA
CACAGACAGTGTCCCGGATGTTGTGACGTATGGCAGTGTGACAGACGAACACACGGAAGAAGTCGCTAAT
AACACAGAGACTTCTACAGAGGAGGTGGGCTTCAGTTTAAGGAGAGCATCAAAAAGGCAGATATTATCCT
TCGTCTCCATCGCCTTACTGAACTTTTCAGGATTCTGTTACTATTCTGTAATAGCCCCGTTCTTTCCGAA
CGAGGCTATAAAACGAGGGGTATCGCAGACCGTGGTGGGATTCATATTCGGATGTTTTGCTGTTGTCAAC
TTCTTTGCGAATCTAGTATTCGGAAAATACATCACGGCCATTGGGTCCAGGTTCCTGCTGACCAGTGGTG
TGTTTGTGGCGGGGAGTTGTTCTGTGTTGTTTGGGCTTTTGGAGTACATGGAAGGGACGACATTTATGGT
GTTTTGCTTCACGATCCGGTCTATAGAGGCCCTCGGTGTAGCTGGTTTCCAAACCGCTGGTACGGCCATT
CTCACCCATGCCTTCTCAAACAAGGTGGCAACAGTCA

>gi|169570559|gb|FE583237.1|FE583237 CAXG15355.fwd Amphioxus Branchiostoma floridae unpublished cDNA library CAXG, larva whole animal Branchiostoma floridae cDNA clone CAXG15355 5', mRNA sequence
ACTGTAGCAGAACTTCAGAACAGAGGAAAGCCGCTAGGGGGCCTAAGATAAGACCTAACTGCCTGCTGAA
TTCCACAATAACCCTTCCATGTACTGTAGCAGAACTTCAGAACTTTCCTATCTCTTGTTCTTGACCGTGA
TGCTTTGACCATATGCATGTTATGACTTGTTATAGACATTATTTTAGTGGTACATAGATCTCGTGGTCG
GGAAGCAGTGACATTTGATACCTAAGACCAGCGTGTTTGAATCCTGCATTGGGAAAGCTTCACCACTAAA
AAACATAACTTTATGCATACTGTATACTGTACATACAAACGCTACCAAAAACATAACTTTATGATTGGTT
TATTGAGTGAACTAACCTTCTAAACCTTCTTGGCGAAGGCAATAAGTAGGTCATTTAATATAAGCCCATG
CAGTTCCTTACGAAATCGCTAGATCAAGCCTCAGTCTGTACCCGAAAGAGACCGTTAGGGGTCACCAGAT
GTATTTGACACGCGAGCCGTCGCACAACAGCATTGTACTGTTAGCTAACGAAGAACGCACTGTGCTTCGT
CTCATTACAGGCCTAGGACAGTCTGCTTTGTCTTGATCCCTTTTATGTGCCATGTGGTCTTATAAATGAT
CAGTGTTCAAGCCACACCCTTTGTCTACTTCTGAATCAATAGTTTGATTGCGATCGCTCGTATTGTGGGG
AAATATTTGTGAGCACTTGGGCT

>gi|169570558|gb|FE583236.1|FE583236 CAXG15355.rev Amphioxus Branchiostoma floridae unpublished cDNA library CAXG, larva whole animal Branchiostoma floridae cDNA clone CAXG15355 3', mRNA sequence
AATAATTCATTTCATTTATTTACTTGCAAATCCGTATGCCTAATTGCAAGGGTACAGACAGTAAAAAAGT
AACATGTGTCTATTCTATACAGCTATCTTATGATTCTACTTACTTTACTGGGAGGTTTGACTTCTTCTTT
GGAGGCTGTGGCTGATGAAAAGTCCCACAATTTTAATGATTAGTGGGTTCTGTGATTTTAACAGAAATAA
TAGTGATAGGCCGAAAAGCAGCATTATCCATTCGCTAGCCGTAGCAACAGCCTGCCTTGGAGGCTAGGGT
AGTATTGAAAATCATTCACGTGATACACACACCTTACACACGTGCACACCGTAGTACGCGGGAGTGATTT
AGCTTGAATGCCATCACGTGCAGGACATGCATCGTCGTATTATAGGGTCACAGGAATGTACAAAGTACAA
ACTATATACATTGAATCTGTGTTGGTAATGGCAGTACATTCATCCTGAAGAAGGCGAAAATCTGATCCGT
GAACACCTTAGTGTTTAGCATTGTTTTTAACTATATTGTTTTCACACCTTAAACATTGGTTTTGTGGGCG
TAACTGGTTAGTATGTGAGTATGCCTCTGTACATACATACCCACCTGCCTAGTGTATGTACCCAGCCCAA
GTGCTCACAAATATTTCCCCACAATACGAGCGATCGCAATCAAACTATTGATTC

MFS-type-transporter_SLC18B1_like(2)

>gi|66294077|gb|BW707503.1|BW707503 BW707503 Amphioxus Branchiostoma floridae unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad043m13 5', mRNA sequence
ACAGAAGAAGTTGCCAATGACACAGAGACTCCTAAAGAGAAAGAAGGCTTCAGCTGTGGGAGGGCATCAA
AAAGACAGATCCTGTCCTTCTTCTGCGTTGCTTTTCTGAACTTTTCAGGGACGGCCTGCTCCACTATAAT
CGCTCCGTTCTTTCCAAATGAGGCTTTACGACGAGGGGCCTCGCAGACTACAGTAGGATTTGTATTCGGA
TGTTTCAGTGCAGTCCAGTTTCTAGGAGGGCTGGTCTTCGGCAAATTTATCACAACTATTGGGTCGAGGT
TCGTGATGATCAGTGGAGTGTTTGTGGCAGGGAGCTGTTCGCTGTTGTTTGGGTTCCTAGCGTACATGGA
AGGAACAACATTCATCGCCTTCTGCTTTGCCATTCGGTCTATGGCGGCCCTGGGTGTGTCTGCGTACATG
ACTGCAGCAACAACCATCATGGCCCACGAGTTCCCCAACGACATAGCGAAAGTCATGGGTACCCTGGAGA

TTTTCACCGGACTCGGCATGATGGCGGGTCCTCCCATTGGGGGTGTCCTGTACAACCTTGGCGGGTTCAA
ACTGCCGTTCTTCACGGTGGGGGGTCTGATGTTCTGCTGCTGCGCGGTGCTGGCTGTCCTGGTCCCGC

MFS-type-transporter_SLC18B1_like(3)

>gi|66312970|gb|BW726376.1|BW726376 BW726376 Amphioxus Branchiostoma floridae unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad043m13 3', mRNA sequence
GTATCCTTGAGCTTTCTTCTCAGTTGATGGTCCAGCATGGCATCTCTTGTACACGTTCTCACTGACGGTA
AAAGTGACCACCAGCAACATGGAAAATAAGATATACCCTGAGAATGCTGTTGAGGCCCATGGCAACCCAA
ACCTCTCTACCAGGGCACTGCTCACTGTCGGACCCAGAAATGACCCCATGCTCATAAATGCGGCAAAAGT
TCCAGATACCAGACCGTAAGTAGCAAAGTCGGTCTCCATACCTGCATCGCTGGCCGCCCAAAGCATCACA
TTGAAGAGGGGTGCTAGAACTGAGCCGATGGACAGTGCACTGACTACAACACCGACGATATTTATCCACA
GTACTTTAGGCAGGAGAGTGACGTAGTCTGTAAGGAGGGGTGACGGACCAATGAGGAGCGCCCCAGCTGA
AAGCACGAGCAGTCCTAATGTCATCATAAATCTGACACATTTCTTTTTGTCAGCCAGCCACCCCCATGCC
GGCGCGAAAAGGGCGTACACGCAGGCCAGGAGCAGGAATATCAGACCAACTTGTGGGGCTGTGACATCGA
ACTCTTCTGCTACGTATGGTTGTATTACAGGACTAAGATACTCAATAATAGAGTAAACCACAACGGTTAC
TCCACACGCCATAATGACGGTTGGGATGCTAAGGAAGAAAAGTAAAGACACGTCCTTTTTGCCT

MFS-type-transporter_SLC18B1_(3) extended

(-1 frame)
>gi|66312970|gb|BW726376.1|BW726376 BW726376 Amphioxus Branchiostoma floridae unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad043m13 3', mRNA sequence
GTATCCTTGAGCTTTCTTCTCAGTTGATGGTCCAGCATGGCATCTCTTGTACACGTTCTCACTGACGGTA
AAAGTGACCACCAGCAACATGGAAAATAAGATATACCCTGAGAATGCTGTTGAGGCCCATGGCAACCCAA
ACCTCTCTACCAGGGCACTGCTCACTGTCGGACCCAGAAATGACCCCATGCTCATAAATGCGGCAAAAGT
TCCAGATACCAGACCGTAAGTAGCAAAGTCGGTCTCCATACCTGCATCGCTGGCCGCCCAAAGCATCACA
TTGAAGAGGGGTGCTAGAACTGAGCCGATGGACAGTGCACTGACTACAACACCGACGATATTTATCCACA
GTACTTTAGGCAGGAGAGTGACGTAGTCTGTAAGGAGGGGTGACGGACCAATGAGGAGCGCCCCAGCTGA
AAGCACGAGCAGTCCTAATGTCATCATAAATCTGACACATTTCTTTTTGTCAGCCAGCCACCCCCATGCC
GGCGCGAAAAGGGCGTACACGCAGGCCAGGAGCAGGAATATCAGACCAACTTGTGGGGCTGTGACATCGA
ACTCTTCTGCTACGTATGGTTGTATTACAGGACTAAGATACTCAATAATAGAGTAAACCACAACGGTTAC
TCCACACGCCATAATGACGGTTGGGATGCTAAGGAAGAAAAGTAAAGACACGTCCTTTTTGCCT

(-3 frame)
>gi|66294077|gb|BW707503.1|BW707503 BW707503 Amphioxus Branchiostoma floridae unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad043m13 5', mRNA sequence
ACAGAAGAAGTTGCCAATGACACAGAGACTCCTAAAGAGAAAGAAGGCTTCAGCTGTGGGAGGGCATCAA
AAAGACAGATCCTGTCCTTCTTCTGCGTTGCTTTTCTGAACTTTTCAGGGACGGCCTGCTCCACTATAAT
CGCTCCGTTCTTTCCAAATGAGGCTTTACGACGAGGGGCCTCGCAGACTACAGTAGGATTTGTATTCGGA
TGTTTCAGTGCAGTCCAGTTTCTAGGAGGGCTGGTCTTCGGCAAATTTATCACAACTATTGGGTCGAGGT
TCGTGATGATCAGTGGAGTGTTTGTGGCAGGGAGCTGTTCGCTGTTGTTTGGGTTCCTAGCGTACATGGA
AGGAACAACATTCATCGCCTTCTGCTTTGCCATTCGGTCTATGGCGGCCCTGGGTGTGTCTGCGTACATG
ACTGCAGCAACAACCATCATGGCCCACGAGTTCCCCAACGACATAGCGAAAGTCATGGGTACCCTGGAGA
TTTTCACCGGACTCGGCATGATGGCGGGTCCTCCCATTGGGGGTGTCCTGTACAACCTTGGCGGGTTCAA
ACTGCCGTTCTTCACGGTGGGGGGTCTGATGTTCTGCTGCTGCGCGGTGCTGGCTGTCCTGGTCCCGC

CHIC (Bfl.19568 transcribed locus)

>gnl|UG|Bfl#S25136288 BFLG3_000088 Amphioxus 5-6 hrs cDNA library (Name convention: BFLG or MPMGp498) Branchiostoma floridae cDNA clone MPMGp498A0428 5', mRNA sequence

```
CACCACGCGTCTGGTTTATGGTTTGTGAAGATAGGAAAAGTCTCATCTTCTCTGAACTGC
TTAGGGGCTGAAGAAATAGTAGAAATAGGTGAATAATCTGCCAGAGAAGGAAAACTAACT
TCAAAACTATGTACCTAGCTTTAGTTGTTTCCAACAGCCGCTACCAAGCTTTTCTCCGCT
CCCAACATTTGCTAGCCAGCTTCTAATGGTTGCAACCCAGCTAAAGGCTGCCTTCTACAG
CCAGTACCCAGCTTTAGTCTGCTTCCAACGGTCTTAACCAAAGCTTCAGTTCACTTTCAA
TGGTCGCTGCCCATCTTTATTCCACTTCCAATGGTCGTTACCCAATTTTTTTTTCCATTT
CCAATGGTTTAGAAGGTGTTGGAATCTGCCTAAAGCTAACAAGACTAGATTCCATGCAAT
TCATCCCATTTGACCATTTCTGCTATTTCTTCAGCAATGAGTGTGATAGAGACTGGTGAA
AGTCCCAGAAATGCTGTCAGTTCAACACAGAGCTGCACTGTTTTTTTGACCAACCATTTC
TCCCAGTTAGGCACTTCTAAATTATTTCAGGGGATGTGGTATATATCTAGAAATCTTCCT
GTGAAGAGCATGCCTTATTTCCAGCTCTGTCCTCGGGATGAAGGGAGGACTGTAAGTAN
```

>gnl|UG|Bfl#S25239348 BW780548 Amphioxus Branchiostoma floridae unpublished cDNA library, gastrula whole animal Branchiostoma floridae cDNA clone bbga033i21 5', mRNA sequence

```
TATGTGTGTATGTATCTCAGGATTGGCGAGTTGTCGGAAACTCAGCGGGAGACGAACTTT
GTAGAGAAAACTTGAGACCAACCGATTTGGTGAGTGGACCTAACGTGCATTAATGACTAT
AGTTAATATAATGACCATGGTTGTGCCCATACAAGTATGTGATTCATCAACCTGTGTCTT
TCAGTACCAGCCAGCGAGAATGGTGGAAAGTACTTCCATGCAGTATATATATATATACAC
CACCTTTGTGAACGCTCTGATTAAGATCATAATGCAAGTAGGGGTAGACTTATAATGCTA
TAATGTATTATTATTTTCTACGAAACTTGTTGCTGTTTTACACATCGTTTTTTGATATAT
CATTGTATTATGTTGTTATGACTCANGCAGGTAGTTGAAATTGTTTTATTATTTAATACA
TGATTATTTGAATGCAGTGAAGATGTTGATGCCTTCCCTATGGATGAATACAAATTATTA
GATGATACAAATGTGTATAGAGTCAGAGATATGTTTATAAGATGTAGTATCAGAAGTTGC
CTCAGCTGTAGCGAGATGGCGGGTCACTTGTCAGAAGACTTCAGTCTTGTTTGTAGTTAA
ACACATGGTTTANTCACATTGGCACCAAGTATGTCACCAATGTACCAATGTTGCATTGTT
CCTTTCCCTCAACACAAATTGTTGTTTGGCTATGCCTCANGG
```

>gnl|UG|Bfl#S25252560 BW793760 Amphioxus Branchiostoma floridae unpublished cDNA library, gastrula whole animal Branchiostoma floridae cDNA clone bbga033i21 3', mRNA sequence

```
CTATCAGCATACCACATCTCCTCTACTTACAGTCCTTCCTTCATCCCGAGGACAGAGCTG
GAAATAAGGCATGCTCTTCACAGGAAGATTTCTATAGATATATACCACATCCCCTGAAAT
AATTTAGAAGTGCCTAACTGGGAGAAACGGTTAGTCAAAAACACCAGTGCAGCTCTGTGT
TGAACTGACAGCATTTTGGGGACTTTCATGAGTCTCTATCCCACTCAGTCCTGAAGAAAT
AGTAGAAATTTGTCAAATACGATGAATTGCATGGAATCAAGTCTTGTCAGCTTTAATCAG
ATTCCAACGCCTTCTAAACCATTGGAAATGGAAAAAAATATTGGGTAACGACCATTGGAA
GTGGAATAAAGATGGGCAGCGACCATTGAAAGCGAACTGAAGCTTTGGTTAGGACCGTTG
GAAGCAAACTAAAGCTGGTTACTGGCTGTAGAATGCAGGCTTTAGCTGGGTTGCAACCAT
TAGATGCTGGCTAGCAAATGTTGGGAGCGGAGAAAAGCTTGGTAGCGACCGTTGGAAGCG
AACTAAAGCTAGGTAGTTTTGAAGTTAGTTTTCCTTCTCTGACAGATTATTCACCAATTT
CTACTATTTCTTCAGCCACAAAGCCGTTCAGAGAAGATGAGACTTTTCCTATCTTCACAA
ATCATAAACAGTAACCAAATATAATGCTCCACCCTTGAGGCATAGCCAAACANCAATTTG
TGTTGAGGGAA
```

>gnl|UG|Bfl#S43180096 CAXC11026.rev Amphioxus Branchiostoma floridae unpublished cDNA library CAXC, neurula whole animal Branchiostoma floridae cDNA clone CAXC11026 3', mRNA sequence

```
GAACGAACAATAGTCTTTATTAACAACCTATCGGCATACCACATCTCCTCTACTTACAGT
CCTTCCCTTCATCCCGAGGACAGAGCTGGAAATAAGGCATGCTCTTCACAGGAAGATTTC
```

TATAGATATATACCACATCCCCTGAAATAATTTAAAAGTGCCTAACTGGGAGAAATGGTT
GGTCAAACAACCAGTGCAGCTCTGTGTTGAACTGACAGCATTTCTGGGACTTTCACCAGT
CTCTATCACACTCATTGCTGAAGAAATAGCAGAAATTGGTCAAATGGGATGAATTGCATG
GAATCTAGTCTTGTTAGCTTTAGTCAGATTCCAACACCTTCTAAACCGTTGGAAATGGAA
AAAATTTGGGTAACGACCATTGGAAGTGGAATATAAAGATGGGCAGCGACCATTGAAAGT
GAACTGAAGCTTTGGTTAAGACCGTTGGAAGCAGACTAAAGCTGGGTACTGGCTGTAGAA
TGCAGCCTTTAGCTGAGTAGCAACCATTAGATGCTGGCTAGCAAATGTTGAGAGCGGAGA
AAAGCTTGGTAGCGGCTGTTGGAAACAACTAAAGCTAGGTAGTTTTGAAGTTAGTTTTCC
TTCTCTGGCAGATTATTCACCTATTTCTACTATTTCTTCAGCCCCTAAGCAGTTC

>gnl|UG|Bfl#S43188518 CAXC17150.fwd Amphioxus Branchiostoma floridae unpublished cDNA library CAXC, neurula whole animal Branchiostoma floridae cDNA clone CAXC17150 5', mRNA sequence
AAAACTGCCTAAGAAGACCACTGATGGGACTGGCCAAATCTGGACTATGTACTAGTAGAC
AGGTGGTAACTATAGACATGATTCTCAATACTTGTGTCAATGGGAAAATTATCTATTGGG
ACCATCAAAAAGTGGTCACATTGTCCAGGGGGTCCTTTATGTAAAGGTGGTTACTTGAAC
AGCTGTGACTGGTAAACTTCTTCAGATCAGACTTTAGTACCAGCCGTGTAAACGTGTACA
CGTAGAGTGCACTCGCTCGGTGCTACCACGCAGATGTGAATCCTTTGTGCATCCTACTCC
AATGTGTGCATGTATGTGTGTATGTATCTCAGGATTGGCGTTGTCGGAAACTCAGCGGGA
GACGAACTTTGTAGAGAAAACTTGAGACCAACTGATTTGGTGAGTGGACCTAACGTGCAT
TAATGACTATAGTTAATATAAATGACCATGGTTGTGCCCATACAAGTATGTGATTCATCA
ACCTGTGTCTTTCAGTACCAGCCTGCGAGAATGGTGGAAAGTACTTCCATGCAGTATATA
TATATACACCACCTTTGTGAACGCTCTGATTAAAATCATAATGCAAGTAGGGGTAGACTT
ATAATGCTATAATGTATTATTATTTTCTACGAAACTTGTTGCTGTTTTACACATCGTTTT
TTAGATATATCATTGTATTATGTTGTTATGACTCATGCAGGTAGTTGAAATTGTTTTATT
ATCTAAGACATGATTATTTGAATGCAGTGAAGATGTTGATGCCTTCCCTATGGATGAATA
CAAATTATTAGATGATACAAATGTATAGAG

>gnl|UG|Bfl#S43190468 CAXC3027.rev Amphioxus Branchiostoma floridae unpublished cDNA library CAXC, neurula whole animal Branchiostoma floridae cDNA clone CAXC3027 3', mRNA sequence
ATTTGATCGAACAAAAGTCTTTATTAACAACCTATCGGCATACCACATCCCCCCTACTTA
CAGTCCTTCCCTTCATCCCGAGGACAAAGCTGGAAATAAGGCATGCTCTTCACAGGAAGA
TTTCTATAGATATATACCACATCCCCCGAAAAAATTTAAAAGTGCCTAACTGGGAAAAAT
GGTTGGTCAAACAACCAGTGCAGCTCTGTGTTGAACTGACAGCATTTCTGGGACTTTCAC
CAGTCTCTATCACACTCATTGCTGAAAAAATAGCAAAAATTGGTCAAATGGGATGAATTG
CATGGAATCTAGTCTTGTTAGCTTTAGTCAAATTCCAACACCTTCTAAACCGTTGGAAAT
GGAAAAAATTTGGGTAACGACCATTGGAAGTGGAATATAAAGATGGGCAGCGACCATTGA
AAGTGAACTGAAGCTTTGGTTAAAACCGTTGGAAGCAAACTAAAGCTGGGTACTGGCTGT
AAAATGCACCCTTTAGCTGAGTAGCAACCATTAAATGCTGGCTAGCAAATGTTGAGAGCG
GAGAAAAGCTTGGTACCGGCTGTTGGAAACAACTAAAGCTAGGTAGTTTTGAAGTTAGTT
TTC

>gnl|UG|Bfl#S43190469 CAXC3027.fwd Amphioxus Branchiostoma floridae unpublished cDNA library CAXC, neurula whole animal Branchiostoma floridae cDNA clone CAXC3027 5', mRNA sequence
TTGACCAGCTCTGCACATAGAAAAACACTTGTTAAAAGAGAAATACATTGCATTGAAAAC
ATCAGGTATAATATGAAGCTTAGCTGAAAGGACTTCCTAATTGTTAGAAATGTCAACATG
ATTGTGTTTTGTTAGTCTTGTCTGCACATGTACAGTCAAAACTGCCTAAGAAGACCACTG
ATGGGACTGACCAAATCTGGACTATGTACTAGTAGACAGGTGGTAACTATAGACATGATT
CTCAATGCTTGTGTCAATGGGAAAATTATCTATTGGGACCATCAAAAAGTGGTCACATTG

TCCAGAGGGTCCTTATGTAAAGGTGGTTACTTGAACAGCTGTGACTGGTAAACTTCTTCA
GATCAGACTTCAGTACCTGCAGTGTAAACGTGTACACGTAGAGTACGCTCGCTCGGTGCT
ACCACGCAGACGTGAATCCTTCGTGCATCCTACTCCAATGTGTGCATGTTTGTATGTATG
TATGTATCTCAGGATTGGCGAGTTGTCGGAAACTCCACGAGAGACGAACTTTGTAGAGAA
AACTTGAGACCAACCGATTTGGTGAGTGGACCTAACGTGCATTAATGACTATAGTTAATA
TAAATGACCATGGTTGTGCCCATACAAGTATGTGATTCATCAACCTGTGTCTTTCAGTAC
CAGCCAGCGAGAATGGTGGAAAGTACTTCCATGCAGTATATATATATATACACCACCTTT
GTGAACGCTCTGATTAAGATCATAATGCAAGTAGGGGTAGACTTATAATGCTATAATGTA
TTATTATTTTCTA

>gnl|UG|Bfl#S43213492 CAXF8591.rev Amphioxus Branchiostoma floridae unpublished cDNA library
CAXF, gastrula whole animal Branchiostoma floridae cDNA clone CAXF8591 3', mRNA sequence
GATCGAACAATAGTCTTTATTAACAACCTATCGGCATACCACATCTCCTCTACTTACAGT
CCTTCCCTTCATCCCGAGGACAGAGCTGGAAATAAGGCATGCTCTTCACAGGAAGATTTC
TATAGATATATACCACATCCCCTGAAATAATTTAGAAGTGCCTAACTGGGAGAAATGGTT
GGTCAAACAACCAGTGCAGCTCTGTGTTGAACTGACAGCATTTCTGAGACTTTCACCAGT
CTCTATCACACTCAGTGCTGAAGAAATAGCAGAAATTGGTCAAATGGGATGAATTGCATG
GAATCTAGTCTTGTTAGCTTTAGTCAGATTCCAACACCTTCTAAACCATTGCAAATGGAA
AAAAATATTGGGTAACGACCATTGGAAGTGGAATAAAGATGGGCAGCGACCATTGAAAGC
GAACTGAAGCTTTGGTTAAGACCGTTGGAAGCAAACTAAAGCTGGTTACTGGCTGTAGAA
TGCAGGCTTTAGCTGGGTTGCAACCATTAGATGCTGGCTAGCAAATGTTGGGAGCGGAGA
AAAGCTTGGTAGCGACCGTTGGAAGCGAACTAAAGCTAGGTAGTTTTGAAGTTAGTTTTC
CTTCTCTGACAGATTATTCACCAATTTCTACTATTTCTTCAGCCACAAAGCCGTTC

>gnl|UG|Bfl#S43213493 CAXF8591.fwd Amphioxus Branchiostoma floridae unpublished cDNA
library CAXF, gastrula whole animal Branchiostoma floridae cDNA clone CAXF8591 5', mRNA
sequence
CTCGCTGTACCTGTAGCACCATGGACTCCCCCCTCCAGTAACCATACACCGGTCACCTGG
GGCACACTTCTGAACAGTCTCAATACTCATTAAGTTATTATAATCCCTGTTACATACTCA
TAAAATCATACAGATTGGGCATACAGGCACATGATACTTAGCTACAGAATTTATACAACT
GCACAAACACAACAAATGCTCTGACAGACTTTTCCAGCAGTGCTCTGAGGGAAAAGTGTA
AACCGCTTTTCCAAATGTCTCGCTGTTTGTCAATCTTTTGTTCAGTTGTGTTGGTGCAGT
AGGTAGGAACATCTACTAGTATAATGTACCTGAATGCCTAATCTTTATCAATGGGCAGGT
ATTGTTTTCTCATGATTCAGTCAGTGTGGAAAATACTAAATTTTCCTGAATGTATGGTTT
TCAGGCTAAAGCCACTTTGGGGTCACCATATCCTTATTTCAGGACTAAGGAATGACAAAG
TAGAGGGACACATGATAAACTCTATTATATTTGTAGTACTAAAGGAGCTAGTCTAAGAGA
ATAATAACATATTATAACCTTTTTCATTTGAAATGACAAAGTCATATCAGGAAAAGTATT
TGCACTGTTTTGATAAACCAGCAATCATTTGATGAGATTTAAGACTTTCCTGTCATGTTC
AGAGACATGTTTTGACATGATTCCAGTGGCAATGTTTTGAGTACTTGTTTTGTACCTTTT
CTGTATGTCTTTGACCAGCTCTGCACAT

>gnl|UG|Bfl#S43234794 CAXG9514.rev Amphioxus Branchiostoma floridae unpublished cDNA
library CAXG, larva whole animal Branchiostoma floridae cDNA clone CAXG9514 3', mRNA sequence
GATCGAACAATAGTCTTTATTAACAACCTATCAGCATACCACATCTCCTCTACTTACAGT
CCTTCCCTTCATCCCGAGGACAGAGCTGGAAATAAGGCATGCTCTTTACAGGAAGATTTC
TATAGATATATTTACCACATCCCCTGAAATAATTTAGAAGTGCCTAACTGGGAGAAATGG
TTGGTCAAACAACCAGTGCAGCTCTGTGTTGAACTGACAGCATTTCTGAGACTTTCACCA
GTCCCTATCACACTCAGTGCTGAAGAAATAGCAGAAATTGGTCAAATGGGATGAATTGCA
TGGAATCTAGTCTTGTTAGCTTTAGTCAGATTCCAACACCTTCTAAACCATTGGAAATGG
AAAAAAATATTGGGTAACGACCATTGGAAGTGGAATAAAGATGGGCAGCGACCATTGAAA
GCGAACTGAAGCTTTGGTTAGGACCGTTAGAAGCAAACTAAAGCTGGTTACTGGCTGTAG

AATGCAGGCTTTAGCTGGGTTGCAACCATTAGATGCTGGCTAGCAAATGTTGGGAGCGGA
GAAAAGCTTGGTAGCGACCGTTGGAAGCGAACTAAAGCTAGGTAGTTTTGAAGTTAGTTT
TCCTTCTCTGACAGATTATTCACCAATTTCTACTATTTCTTCAGCCACAAAGCCGTTC

>gnl|UG|Bfl#S43234795 CAXG9514.fwd Amphioxus Branchiostoma floridae unpublished cDNA library CAXG, larva whole animal Branchiostoma floridae cDNA clone CAXG9514 5', mRNA sequence
CCAATGTGTCCATGTATGTGTGTATGTATCTCAGGATTGGCGAGTTGTCGGAAACTCCAC
GAGAGACGAACTTTGTAGAGAAAACTTGAGACCAACCGATTTGGTGAGTGGACCTAACGT
GCATTAATGACTATAGTTAATATAATGACCATGGTTGTGCCCATACAAGTATGTGATTCA
TCAACCTGTGTCTTTCAGTACCAGCCAGCGAGAATGGTGGAAAGTACTTCCATGCAGTAT
ATATATATATACACCACCTTTGTGAACGCTCTGATTAAAATCATAATGCAAGTAGGGGTA
GACTTATAATGCTATAATGTATTATTATTTTCTACGAAACTTGTTGCTGTTTTACACATC
GTTTTTTGATATATCATTGTACTATGTTGTTATGACTCATGCAGGTAGTTGAAATTGTTT
TATATTTAATACATAATTATTTGAATGCAGTGAAGATGTTGATGCCTTCCCTATGGATGA
ATACAAATTATTAGATGATACAAATGTGTATAGAGTCAGAGATATGTTTATAAGATGTAG
TATCAGAAGTTGCCTCATCTGTAGTGAGATGGCGGGTCACTTGTCAGAAGACTTCGTTCT
TGTTTGTAGTTAAACACCTGGTTTAGACACATTGGCACCAAGTATGTCACAAATGTACCA
ATGTTGCATTGTTCCTTT

B.la CHIC (with 3'UTR)
>gi|379310582|gb|JT881816.1| TSA: Branchiostoma lanceolatum Seq43418.bl mRNA sequence
GATACCGTGCGCAGATGTTTCCATCATGGCGGACTTCGATGCGATTTACGAGGAAGATGACGAAGATGAG
CGGCTTATGGAGGAACATCTACTCAGTACCGTGCCTGACCCCGTTATCGTTAGAGGATCAGGACATGTCA
CTGTGTTTGGTCTAAGCAATAAGTTCAACACAGAGTTTCCACAAGGCTTGGCTGCAAAGGTTGCGCCTGA
GGAGTACAAGGCGACGATCAGCCGGCTGAACGGTGTCCTGAGGAAGACGTTACCTGTGAACGTGAAGTGG
CTGCTGTGCGGCTGTCTGTGTTGCTGCTGTACCCTGGGCTGTTCACTCTGGCCAGTCATCTGTCTCAGCA
AAAGGACACGACATTCTATAGAGAAAGTACTGGACTGGGAAAACAGTCATTTGTATCACAAGCTGGGCTT
GCATTGGAGATTAGCGAAAAGGAAGTGTGAATCTAGCAACATGATGGAATATGTAATCTTAATAGAGTTC
ATTCCGAAGGTTCCCATCCACAGACCAGACTGACCCTGAAGTGTGCCCTGGCTGTACCGGTATATCTGTA
GCACCATGGACTCCCCCTCCCGCAACCCTGCACTCATCACATGAGAAACACTGAATACTAAGTTATTAAA
GCCCTGTTACAGAGTGATAAAATCATAAAGATTAGGCATTCAGGTAGATGATACTTAGTTTGGCCCTACT
TTCC

SCP1

>gi|66284246|gb|BW697675.1|BW697675 BW697675 Amphioxus Branchiostoma floridae unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad022l10 5', mRNA sequence
TTCTGACTTGGAGGCGCCATTTTTGGAAAAAGGGAGTTTCTTCCAACAAGAACTGACACGAACTCCCTTG
ACACTCATCAAATGCAGAGCAGTCCAGGTGCATGCACAGCATAGTTAAAATGCAGAAAAGCTAACTTAGC
TCCAAGAAAGCCCCACCCCCATGCAGTAGGAAACAGAAAACTGGAATCAGCCGTTGCCTTTCTACATCAA
AAACACAACAAACCTGTCTACCAGTGATGCAAGGTATACAGCAGGTGTATCATCAGCAAGAGCCATTCTT
CAAGCCTCTGTCACCCCAGCAACAAGAGCGCCAGCATAGTTTCTTCAAGATCGGCACAGACCACCAGGAA
GTGGAGATGGAATCGCTCTCACCCATGCGGCTTGGGCAACAGGTACACAGTGGTGAGCGCCTGACGAGTC
TGCATTCTCGCCTCCAGAAGGAGGCAGAAAAGATCAACAAGTGGAAGCATCAGACAGAGATGCAGATCCA
ACAGAGGGAAAAGAAGATACAGGACACTCAGCAAACCATTGATTCACAACGCAAGTCCATCCTTGAGCTA
CAGCTTCAGAATGAGAACCTCAGCTCCAAGCTGCAGGAG

>gi|66302869|gb|BW716295.1|BW716295 BW716295 Amphioxus Branchiostoma floridae unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad022l10 3', mRNA sequence

GCAGAAATTTTAACAATTTACAAGTCTTGAAATAACATTTACAAGTCTTGAAACAACATCTCGCCATAAT
GATACGTGCAACATAGACAGAGTAAAAGCACGTCAAATACTTAAGAAAAGTAATGGCAGACTTGGCGTTT
GGTTTGAATAATTACCAATCAGCTCAAATAGACAATGCTTTAGAAACAATTGCATTTGTCCTAATAATAT
TAGATAACATTCGTCATCACAGTCAACTTTACCATTCAATAAATAGCTTATCTAGAGTTAATAAATAATT
GATAAATTTCTATAGTTTCTTAAAACATTCGAACTCTTCTACTTAACGACGCAAACTCATTCAGGAATCT
TTGAATTTCAGATTCAAGACTTCGTAGCTATGACGTGCCCAAACATTTCATTTTTAAACACTGATATGAT
TAAGAATATGAAAATCATATCAAAATGGTTTACTTATACAAGTTATAATATTCAATATTATCGAATAGAA
TGAACGTTAGCTTAACATATCGCTATAACGTACACGAAAGAGCACTGAGAAACCGAGATAACCACACGTA
TCATATACGCTTTGCAACGGTATTTCATTTAAGATCGTTACTAAAACATGAACGCAAAACATTTGTTTCT
CAAAAGGTTAAAGGGAACTTGCAACTACAAACTGAAACCGTAGTTTGCTCATATGTGGAGCATGAAGGAT
GCGTCNTTTAAAGGTAT

Xlox 3'UTR (Bfl.23459 transcribed locus)

>gi|66663573|gb|BW948304.1|BW948304 BW948304 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne153i16 3', mRNA sequence
TACGTCTTTTCCTCTATCACATTTTCTGATACACTTTTACAACAGGCACTGTAGATACACATTCTCACGT
GTAACAAGGCAGATCTAACATACAAGCACCAACGATAGCGAATTATAGCTTTGTGTTTTCGTACATAGCT
GAGCAAATGACACTGATTCGATACTTCAACAAACGTTTCTATATTAACATTAAAACTTTCATTGCGTCTT
AATAAATAATCACAAGGAACATATTGAATAGATTTTACACTCTAAAATTTCTATATAAAGACAGCTTTGC
TACATTCATATCAGCCTGTGTTCTAAGGCTAGCTCTATTATACTTAGTATCTAGTACGGTTTCCACAGTG
GTCGACCCAGATGGCGGTCTATAAAATATGGGTCCATAAATCAGCGTGGTTTCCAATTACTATTGAGATG
GATATGTGAAATATATCTCCTGTGATATTCCTGAGT

>gi|66507123|gb|BW890457.1|BW890457 BW890457 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne153i16 5', mRNA sequence
ACGATAAAGAGCTCGGTACATCCCTAGGGTAAGCAAGAGGTGTTAGCATTGTGCCTTCTGCATATTGGGA
AAGGCTGACAATGGAACCCGGCGGCTCCTTTGGTGCTCTTGACATAAAAAAACGAAACCCTATATTATAT
ATCAGTGTATTTTTTTCAGTTTGAAGGAAGCAACGACTACGTGCGAACAGGACAGGGTGTGAACGAAACC
TTCCAGACAAAGCTATTGTTTCCCATCATTCATCAAACCAGACATGTAGCAGACCCGTGACAAATTGTTC
GTTAGGAAGCCAAGAAGAAACAATGTGTACGGAGACATGTGAATGAAACAATGTAGAAATATCAACTGCA
TTTTGAACGAGTCCGTTATGATAACGGATATAACGGATTTTCCAGTGACGTTTGTATTGAAACCGTTGCC
CTTGAAACGCCGGGGCCCTGTACAATAATTAGTGTGAAATCTGTACAAATTAGTCCAGCGTTTATGCTC

>gi|66282594|gb|BW696023.1|BW696023 BW696023 Amphioxus Branchiostoma floridae unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad018o19 5', mRNA sequence
TATATTATATATCAGTGTATTTTTTCCAGTTTGAAGGAAGCGACGACTACGTGCGAACAGGATAAGGTGT
GAACGAAACCTTCCAGACAAAGCTATTGTTTCCCATCATTCATCAAACCAGACATGTAGCAGACCCGTGA
CAAATTGTTCGTTAGGAAGCCAAGAAAAAACAATGTGTACGGAGACATGTGAATAAAACAATGTAGAAAT
ATCAACTACATTTTGAACGAGTCCGTTATGATAACGGATATAACGGATTTTCCAGTGACGTTTGTATTGG
AACCGTTGCCCTTGAAACGCCGAGGCCCTGTACAATAATTAGTGTGAAATCTGTACAAATTAGTCCGTTT
ATGCTCCCTAGGTTTTTATCAGTGACCTGGAAGCTCGCAGGTATTTGTCAACCGGTTGTTGTGGGCTTCG
TAGCGACGGGGTCCTCCTCTTGTAACCCCCGTAAAGATCAAATATTGGAGCCAAACTCAGGCCTGTTTAG
CGCCACAATGTTTACAAGAACATCAATGTTAGAGGTCCTGAGTTCACATCTGTAACAATGTTGCCCAAAT
AATCACATACAAGTCTGTTGATATGAAGTGAAATGCAAGCGAAAGACGATTAATATGAGTGATTTGCGTG
GCTTCACG

>gi|30911676|gb|BI376719.1|BI376719 BFLG3_000522 Amphioxus 5-6 hrs cDNA library (Name convention: BFLG or MPMGp498) Branchiostoma floridae cDNA clone MPMGp498J0612 5', mRNA sequence
AGGAGCATCCGTCGACAGCTGATATGAACGTAGCAAAGTTGTCTTTAAATAGAAATTTTAGAGTGTAAAA
TCTATTCAGTATGTTCCTTGTGATTATTTATTAAGACGCAATCAAAGTTTTAATGTTAATATAGAAACGC
TTGTTGAAGTATATAATCAGTGTCATTTGTTCAGCTATGTCCGAAAACGCAAAGCTATAATTCGCTATCG
TTGGTGCTTGATACTATGTTAGAGCTGCCTTGTTACACGTGAGGATGTGTATCTACAGTGCCTGTTGTAA
ATTGTATCAGAAAATGTGATAGAGGAAAAGGCGTTATAAATACATTAAACCGTTATTTTCCCGTTTAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGGGGGGGGGGCCCCAAAGGGGGCCCCG
TTTTAAAGGGGGGGCCCCTTTTAGGGGGGGATTTTATATGATATTGGGTAGTTTGTGCTTGGTATTTTCG
AGTTTTTCCTGGTGGTTTTGGTCATGTGGGTAGAAATCGGTATCACCTGTTCTCGTTACAGACTATGAGG
AAATTCAGGTCAGATAAAGCTTGGCTGCTAAAACGTAGAGGGAAGAGGGAGCG

B.la Xlox 3'UTR (UTR confirmation read)

>lcl|comp1023376_c0_seq3 len=3359
TGGTTTCAGAACCGCCGGATGAAGTGGAAAAAGGAGCAGGCCAAGCGGCGGCCGCTGCCCGAGACTGCCTC
CAGCACGACCCCCGGGGGCAGCAGCGGCGGGGCCGGCACCGCGGCGGGGGGCGCCGAGTCGACGGAGACC
AGCGGCACCGACCCCGAGACTTCACCGGTCAGTGAGCCGGTCTCGACGCCTCCCCCTTCCACGTCTTTACCGG
TGTCTCCACCTGTGAACTCAGGTGGGCAGGGGACCTCAGCACCTTCCCACACGGGCGGGGTTACCGTTCCCCC
CGTGCACCAAACACTGTCTCATAGCGTTACCGGACCGACAGAGCCCGCACTCCAACGGGAAAACCTCTCACAG
AGCCTTGCCTTTTCACGCTCCTGATTCCTGCAATAGATACGTCAACAACGGCTCTCCGGTCAATAGGACTAATT
TGTAAGGCATTCTTGTGGGATTTCTCCCAATCAAAGGCATCAACGCAGGTGGTTTTTCATCAGAAAAACACCA
GAAGAAAAGACTTGAAATCTATACCACAGATAAACCGATATTTGCAAGATAATCGTTCTAGTCAGTCGACTCG
CAAAGCAGCTACTAGTATGTATCACCAGGGGGAATGAAACAATCGTTTAGCGACGTTTTGCATAACCAAGACG
CAAGTGTCCTCCAGGAATACTTCCAAGTGTCGGCCGGACCCCGTCAAAGGCTATCCGGATTTGCATAATGAAA
GAGGTACAAATTAGTGGGAATCATTAGACTGAGGGGCGGGCGGGGCGCGTGGTCAAGTTGAGGGCACGGG
GTTACGGCTCATTCTCATGGAAAACTGGGAGTCGGCGGCGCCTTTTCTATGCATGTCCGTGCCGTTTGTAGTG
GGGCCAGGCCCGCCGGTAAAGACGGGCACGGCGGCGGCGGCTTGAAAGGGACTCTCGCGCCGTGATTAAAC
CGATGTACAAGATCGGAGACACACGATAAAGAGCTCGGCACATCCCTAGGGTAAGCAAGAGGCGTTAGCATT
GTGCCTTCTGCATATTGGGAAAAGCCGACAATGGAAGCCGGCGGCTCCTTTGGTGCTCGTGACATAAAAAAA
AAGAAACCCTACATGTTATATTCTTAATTTAGTTTGAAGAAAGCGTCGACTACGTGCGAACTGGACAAGGTGT
GAACGAAACCTTCCAGACAAAGCTATTGTTTCCCACCATTCATCACACCAGACATGTAGCAAATCCGTGACAAA
TTGTTCGTTAGGAAGCCAAGATGAAAAATGACTTTCAACGCATGGAAACATGTGAATAAAACAATGTAGAAG
CATCCACTACATTTTTGAATGAATCCGTTATGATAACGGATATAACGGATTTTCCAGTGACGTTTGTATTGGGT
CCGTTGCCCTTGAAACGCCGGAGCCGAGCGCAATAATTAGTGTGAAATCTGTACAAATTAGTCCAGCGTTTAT
GCTCCCTAGGTTTTTATCAGTGACCTGGAAGCTCGCAGGTATTTGTCAACCGGTTGTTGTGGGCTTCGTAGCG
ACGAGGTCCTCATCATGTAACCCCCGTAAAGATCAAATATTGGAGCCAAACTCGGCCCGGTTTAGCACCACGC
TGTTTACTAGAACATCGATGCTATAGGTTTTGAGTCCACTTAAAATAGAAATATTGGCAAAACAAGCACATACA
CGTCTGTTGATCTGAAGTGAAATGTAAGCGAAAGACGACTCAAATGAGTGATTTGCGTGGCTTCCCGACGCCG
GTCGTTGGCTTGACAGATCTGAGTGGTTTGGCTTAAATGAGCGTTATCAATTTGCACTCTGAATTGTTGACAAG
ATTTGCCGTAGTCAAGGGAAACGGGGTACAATAATGAGTGAAAGGGGCCCAAAAAATCCCGGCCGTTTCCGA
ACAAAAATTCAACATCAAACAGCCCGGATTAGTCGAGAACTGAAACGTTGGTTTCTGTTGTGCGGTTCAAAAA
CGACCCAACGTTGGGCCGGACGGTATTTTCTCGGCCGGTAGAGACCGACAAGACTTCCCTTTATTCCTGTGTA
CTCCAGCTGGGAATTAGGTACAAAGTCAGAGGAACGAAAGGTTTAACAAACGGTATGTACTTGGGCTAGAAT
ACATAGATATCCGGGGCAGCTTTTCTGTACTATATTATATGTGCATTCACGCCATCACGAAACGTCCGATCAGA
ATTCTGTAAACGTGCTAAAAGTTTAGCGACACAACTGTGGTTATCAGTTATTCCTTACAGGACGACTTGTTACA
TATAAACTTTAAGAAGCGTACACTGCCTTGTTTTTCCGGATAAATACAAACGTTGTAACGTTGAATGGAAATCG
TAATTTGATGATGCATTCTCGGGGCTTCTGTCAGACTTGAAGTTGGAATCATTCCAAAATATAGTCAGGATATG
TTGCTGAATTTGGCGGCGACTTAAACTGGTACAAGAACGGCTTTTCCTTCTTATTCAAGTGAAATATCTTGAAA
GCAACATTTCAACAATTCATGTTGAAAGCACATGAGCAGTAGAATATCTTTGCCACCAAAGAAAACGATGACG
GCCGTTCACAATTTTGTTTATGATTCATGGCGCTACCACAAACCGTGAGAGCATTCTAGGGATAAAGCTGTGC

ATGTGTCGTCCGAAATCCATCTGTTTTGTTACGGAAAAGGCTGTAATAGATTCTTGTCAGTGGTTTATTCCGTC
ACAGCCTTTTTAAAAACTCTCCAACATCAAAATCGCACGATGGTCTACAACGAATGTGCCCGCTCCGTTAGGAC
ACGTTGAAACGTCGCTGAACGATTGGCGGAGCTATCTAGAGCTTCTACCCCCTTGAAAGGACCAAATTAAAGC
AATAATCTAACACTCAGGAATATCACAGGAGATATATTTCACATATCCATCTCAATAGTAATTGGAAACCATGC
TGATTTATGGACCCATATTTTATAGACCGCCATCTGGGTCGACGACTGTGGAAAACGTACCAGATATAATACTT
AATAGGGCTAGCGTTAGAGCATTGGCTGCCATGGACGCAGCAAAGTCGTCTTTGTATAGAGACAAATTTTAGA
ATGTAAAATCTATTCAGAATATTCCCTTGTGATTATTTATGCCAGAAGAAATCAAAGTTTTAATGTTAATATAAG
AACGTTTGTTGAAATATGGAATCAGTTCATCGTTTAGCTGTGAAAACGCAAAGATATGATTTGCTATTGTTAAT
GCTTGTTACTATATCAGATCTAACTTTGTTATATTTGAGGACACAGATCTACAGTGCCTGTTGTTTAAGTGTATT
AGAAAATGTGATAGAAGAAAAGGCGTTATAAATACATTAAACCGTTACTTCCCCGTTTAAAAAAAAAAA

Cdx (Bfl.10986 transcribed locus)

>gi|66461877|gb|BW853661.1|BW853661 BW853661 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne042k05 5', mRNA sequence
TACGACTGGATGAGGAAAAGCAACTACTCCACAAGTCCTCCCCCAGGTAAGACGAGGACGAAGGATAAAT
ACCGGGTGGTTTATTCCGACCATCAGCGCCTGGAGCTGGAGAAGGAGTTCTACTCCAACAAGTACATCAC
CATCAAGAGGAAGGTTCAGCTGGCGAACGAACTGGGCCTGTCGGAGCGCCAGGTCAAGATCTGGTTCCAG
AACAGGCGCGCCAAGCAGCGCAAGATGGCCAAGCGGAAGGAGCTGCAGCATCCGGGCGGGCAGGGCGGGA
GTGACGATGGGGGAGGGGTGATGGGGAGAGGTGTCCACACTCACGGTAGGCCCCCACCCCACCAGCTCA
CCCTAAACCCCAGCGGCGTGGCGGCCTCCACCCTCAGCAACCCCGCTCTCCCCCCGTCCTCCTCCCCTCT
CATGACCAGCGCCATGACGCATGCAGTGACGTTGCCGTCGTGCGTTCCTTCCTCGTGACACTTGCANAAN
TTCCAGAGTGACAGCTGTTGATACGGACCAAAGGGAGCTGTTGTGTGGACACNAGGAGGAANAGACTTAT
CATTCCCCTCTAGCTGGA

>gi|66535887|gb|BW911521.1|BW911521 BW911521 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne042k05 3', mRNA sequence
AGGGGAGGAGAACGCCGCTAGGTGTACAAAGACTCCGTGTGTCTACGGGGGGAGGGGTGGTACACAGTAC
GTTGTCATGTTGGGAACTGTAAACGTTTCTACGGGCTATATCTGTACAGTATACGTCACATTTCCAACCG
AAAGAACGGGAACAACAACTGGGTTCTCAATAAGCCTACATACACTACAGGTATGATACGTGTCAAACGG
CTACACTACCACATGCTGAATAAAGAACAAAGATACATAATAATATATAATATAAAGTGAACGGATAAAA
TCTCCTTAATCATAGTTCGAAATTACACTTTAAGGTGAAAATATATACTCCAATATGCATAGCAAGAAAA
GACGTCATCTCATCATCGAATAATTATTATCTCAACAATACATACAATATGAGGTTCTTTGGATGTCAGC
TCGCGAAAAGAGTCGGCCTCGGTGCAAAACTATCCTTACAATGGACAAACGGTAGATACGAAGGAGCTGG
TGTAAACATAGCAGTCGATTCGTCACGATGGTTATAACTCCTAGCTACAAGCGTGCTTGATTGTACGAAG
AGAGATCGACAGCAGTTCGTCTTCTGGCCACAAACTCTCACCTAAACTATCTTTGGTTTGACCAAAGTAA
ACCGCACATTGCACCACTTCTGGGCCCGGTTTCAGACACACAAAAATATAAAGGGGACGAGAACAAGTAA
CGTTGTACAAACGTAGCCATTGTGCACGCTTTTCAT

>gi|66381723|gb|BW779229.1|BW779229 BW779229 Amphioxus Branchiostoma floridae unpublished cDNA library, gastrula whole animal Branchiostoma floridae cDNA clone bfga054d07 5', mRNA sequence
TACACATGACAAATTTAGGTTATCTAGATTTTTTCAAAGGCAACGCCAACGTGCACAATCATTTTACCGC
TCCCTTTTGACAAATCAGTGAACAGACAAAGTTGACGATGCAATTTGAATTTTCTTTTCAACTTCGACGA
ACTTCTGGCTTTGACGAAGCTCCTAGTTTTGGGGAAGACGCGAGCGAAGTCCTAGACACCTAGACACAAA
AAACTTTCTTCAGTGGTAGTAGCATCTTCGTCTCGACAGCCGTGCAATGAAAAGCGTGCACAATGGCTAC
GTTTTGTACAACGTTACT

>gi|66400635|gb|BW792419.1|BW792419 BW792419 Amphioxus Branchiostoma floridae unpublished cDNA library, gastrula whole animal Branchiostoma floridae cDNA clone bfga054d07 3', mRNA sequence
GAGAAGGGGAGGAGAACGCCGCTAGGTGTACAAAGACTCCGTGTGTCTACGGGGGGAGGGGTGGTACACA
GTACGTTGTCATGTTGGGAACTGTAGACGTTTCTACGGGCTATATCTGTACAGTATACGTCACATTTCCA
ACCGAAAGAACGGGAACAACAACTGGGTTCTCAATAAACCTACATACACTACAGGTATGATACGTGTCAA
ACGGCTACACTACCACATGCTGAATAAAGAACAAAGATACATAATAATATATAATATAAAGTGAACGGAT
AAAATCTCCTTAATCATAGTTCGAAATTACACTTTAAGGTGAAAATATATACTCCAATATGCATAGCAAG
AAAAGACGTCATCTCATCATCGAATAATTATTATCTCAACAATACATACAATATGAGGTTCTTTGGATGT
CAGCTCGCGAAAAGAGTCGGCCTCGGTGCAAAACTATCCTTACAATGGACAAACGGTAGATACGAAGGAG
CTGGTGTAAACATAGCAGACGATTCGTCACGATGGTGATAACTTCTAGCTACAAGCGTGCTTGATTGTAC
GAAGAGAGATCGACAGCAGTTCGTCTTCTGGCCACAAACTCTCGTCTAACCTATCTTGGNTTGACNAAGT
AACCGCACATNGCACCCTTCTGGG

>gi|38190506|gb|CF919304.1|CF919304 Bflor531.000389 Amphioxus 26 hrs cDNA library (Name convention: BFL26 or MPMGp531) Branchiostoma floridae cDNA clone MPMGp531O104;BFL26_4O10 5', mRNA sequence
GGAGTGCGCACGCGTCGGGCAGATGTGAGATTCTAGGCCCTGATGGTCAGACGAGGACGAAGGATAAGTA
CCGGGTGGTTTATTCCGACCATCAGCGCCTGGAGCTGGAGAAGGAGTTCTACTCCAACAAGTACATCACC
ATCAAGAGGAAGGTTCAGCTGGCGAACGAACTGGGCCTATCGGAGCGCCAGGTCAAGATCTGGTTCCAGA
ACAGGCGCGCCAAGCAGCGCAAGATGGCCAAGCGGAAGGAGCTGCAGCATCCGGGCGGGCAGGGCGGGAG
TGACGATCGGGGAGGGGTGATGGGAGAGGTGTCCACACTCACGGTAGGCCCCCCACCCCACCAGCTCACC
CTAAACCCCAGCGGCGTGGCGGCCTCCACCCTCAGCAACCCCGCTCTCCCCCCGTCCTCCTCCCCCCTCA
TGACCAGCGCCATGACGCATGCAGTGACGTTGCCGTCGTGCGTTCCTTCCTCGTGACACTTGCAGAAGTT
CCAGAGTGACAGCTGTTGATACGGACCAAAGGGAGCTGTTGTGTGGACAGAATGAGGAAGAGACTTATCA
TTCCCCTCTAGCTGGAAAACCGAGAGAAATACTACATGTTTGTGTCCTCCATGTGTGCATCGTCAGCGCA
GGAAGTTCGAGCTATCTTTTGTCCGAAGGAAACACAACAAGATCACACTTTAGGGGAAAAGTACGCATGA
CACATTTAGGT

PRHOXNB

>gi|66495870|gb|BW881193.1|BW881193 BW881193 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne095p03 5', mRNA sequence
AACATGGCGTCTTTGACGATTGCGGAAGTGAACAAACTCGACTCGGAAGAGTTTATAGAGATATTTGGGA
ATGTGGTAGAAAACTGCAAACTCGCTGCCGCCGCCGTGTGGTCCCACAAACCGTTCCAAGACGTCGATCA
CCTGAATCAGACGATAGCGGACTTTCTAGACGCCCTACCCCAGAAAGGTAAGGAAGGTGTTCTCCGCTGC
CACCCTGACCTGGCGGGACGACTGGCGCAGGCCGGACAGCTCACGGCGGAGTCTACACAGGAGCAGCGGT
CTGCCGGGCTGGACCAGCTCACTCAGGACGAGCTCACCACGCTCACGGACTTAAACCAACAGTACAAGGT
CAAGTTCGGCTTCCCGTTCGTCATCTGCGCCCGGCTGAACAAGAAGGCCGCCATCTTGAACGGATTGACG
GAGCGATTGAAGCACTCTTCGGAGGAGGAGACGCTAGCTGGGGTGGGGGAGGTGAAGAAGATCTGTCAGC
TGAGGATAGCGGATATCGTGACGTCAGANGCCAAGTTATAGTTCTCTAC

>gi|66654171|gb|BW938902.1|BW938902 BW938902 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne095p03 3', mRNA sequence
AGCGATACCTTTGATATGTGTTTGGTGGTAGTGTGTACACTGGGGACAATCTGGTACCTTGTCTGGACCC
GCGATCTGCCTGCAAAACCAAGGCTCATATATCATGCGGCATGATCTGAGGCGTTTGTTCTATAATGATT
TATGGTACATTATTAAGTGCTGTATTTCATGTGTCTGGAAATTCGGATTAGCCAAAATTGATTGAAGCAA
ATTCTTTTTCTAATTTTCGAATGGCAATGCTTCATGTAAGACATTTTCAAATTCTTTTATTAATGTGGAT
TTAAAAAATAACAATATCTATGATTAGCATTGATTGTTTAATTTTTGTACTTTGATCCGCTGAGTATCTC

TCCAAACGATCCATCTGATTATCCATCTCTTTATATATTTATGGATATGTAAAACGTAATTCATATCTAT
AACTTTTCCTGACATCATAATTGTGATTGTTGTTTTTTAACTCACATTCTCCGATGACTCAACTGATGAA
ACATTACAAATGATGAAAAATGCAAATGACAACATATTTTCAAAATTTTCGTGAATCTGTACACAACCTT
GATCTGTACATAATAATTTACAATAGCACGATCACAACCAATCAAAATACAACTAACATCCACTATCCAC
TGGGGCGACGACCGCGTTACGATCCAAACTGGATTTTATCCAAACCATCAACAAGCTCCCTGGTTAAAAC
CATTGACTTCAGAATTGGCATTTTAT

MFS-type-transporter_SLC18B1_like(4)

>gi|66406712|gb|BW798496.1|BW798496 BW798496 Amphioxus Branchiostoma floridae unpublished cDNA library, larva whole animal Branchiostoma floridae cDNA clone bflv012o20 5', mRNA sequence
ACCCTACCTAGCTGACAAGTACACATATCTGACCACAACTCAGATGGGTCTGATCTTTCTGTTGTTTGCG
TCAACTTACGCAATCCTCGCACCGCTGTGGGGCTGGATGGCAGACAAAAAGAAAGCGATGCGGTTCATGA
TCATTATAGGACTGATCATCTTGTCGGCAGCATTGCTGATGGTTGGACCGTCCCCTCTTCTGACAGACTA
CCTCAACGTTTTACCCAAGAAACAACTTTGGATTAACCTCGTTGGACTTGCTACTGTATCTATAGGAGGT
GGTATGGCCATAGCACCCATCTTTAATGAGATGCTTTATGCAGCCAGTGATGCTGGTCTGGAAGACAGCT
TTTCGACAAACGCCTTAGTAT

>gi|66428686|gb|BW820470.1|BW820470 BW820470 Amphioxus Branchiostoma floridae unpublished cDNA library, larva whole animal Branchiostoma floridae cDNA clone bflv012o20 3', mRNA sequence
ATGCACTGGATGTGATTATGTGTGTAACAATGTGTGCTNATGCTGAATGTTAAGTCTAATTGCATGTAAT
ACGGCACTAGTACTATTGAACATGAATAGTTGTAAAAAAAGTAGCTACATGATGCTAGATTAGCATTGTC
ATTGCCAGCTGGGACTACAAACTTAGTACCTAGAGTGTAAAAAGAAAATTTTACTGTTCTTTATGAATTT
TATCTATGTCAATAATTTACATG

>gi|66422318|gb|BW814102.1|BW814102 BW814102 Amphioxus Branchiostoma floridae unpublished cDNA library, larva whole animal Branchiostoma floridae cDNA clone bflv057o16 5', mRNA sequence
TACATGTAGTCACCACCTTATATCTACAGTCTACCTGTATTTACCTGAACCACAGAGGCTGTATCCCCGC
GGTCAGGAGTTAGGGTGGGTGTAGGAGCCGAGAGGACCGGGGTACCGTTTCTGTCCGCTTATAATTACGG
CCAGGCAGCCAGGAAAGAAACGTCGTGAACTAACGTACTGTCGCCATGATGGAGGAACAAGAAGGCCTCC
TCAGTACGTCACCCNTCACTGAGCAGCACAAATAAAAGTCTCAACAACCAGTACGGCTCC

>gi|66444707|gb|BW836491.1|BW836491 BW836491 Amphioxus Branchiostoma floridae unpublished cDNA library, larva whole animal Branchiostoma floridae cDNA clone bflv057o16 3', mRNA sequence
ATTAATCTAGTACAACCTACGCGTAGATTCTAAGCTATACAATACTGATAGAAGACACAACTATGTGTTT
TACATCTAGTACATTTGTATACAGTCATATTCTTTTACGTGTGTTTTTTTGTGTGCGTTGAGGTAGTAAT
ATACATGTAATGCCAAAACTAATGGCCTAAAATATCTATACAGATGTTGTGTCGCGACAGGCAAACATAG
ACATTCTCTCTATTTATCTTGGGTATAGTTGGACTTGTCTTACCATTCTATATCCGTTATAAGTAGGATA
TGTCATCGAGAAATAGTACTCATATTGAACACCTTTTTTTTTGCACGTAGGACATGCTTCAGTTGTTTTG
AAACAGCCTTATAATGGATTAACTCAAAAGCGCCATTTAG

Ribosomal-processing-protein-8-like (Bfl.3286 transcribed locus)

>gi|169566753|gb|FE577801.1|FE577801 CAXG1080.fwd Amphioxus Branchiostoma floridae unpublished cDNA library CAXG, larva whole animal Branchiostoma floridae cDNA clone CAXG1080 5', mRNA sequence
AGAAAATGGCCATGTTTGGAGCCTCCGACTGGGGAGATGATGATGATGCAGCACACCTGGAGGAGTCGCT

GTTTGCCTCTGGAGATGGTCTGTCTTCATTCACAAAGCACAAAAGGAAAAGTGTCAAATCACAAGAAGTT
ATTCCGCAAGGGGGAAAACAGGAAACTTTAGAACTGAAGAAAGAAGACCCCACAATTAACACTGAGCAGA
GCAATGGTTCTAGTACCGTGCCTCACAAGCCAAAACGGAAAAGAACCAGGAAGAAAAAGCCACACTTACA
GAGCGATTTGTCAGATGGCGGTACAGAAATATCCTTCAAGCCACAGCTAACACTGAAACAGAAGAGAAAA
CTGAAAAGAAAAAGACAGGAAGAGTCCAGTGTTGGGGAAGCCATATCAACAGCAACAAAGAAGGCAAGAA
AGTCAAGCACACAGCAAAGAGATGGTGTGGACCAGACAGACAACTACACAGATGCAGGACCTGTCAGGGT
GAAGAAACCAAGTTTCAAGAAAGAAACTGACAAAGACAGCAAGACATCTCCCTCGCATACAACTGAAGTT
CCAGAA

>gi|66416148|gb|BW807932.1|BW807932 BW807932 Amphioxus Branchiostoma floridae unpublished cDNA library, larva whole animal Branchiostoma floridae cDNA clone bflv040g15 5', mRNA sequence
AATAAGAACAAATTCAAATCAGCTGTTGGAGACAAGCCCACAGTACCAAGTGAGGAACACAGCAAGATTT
GTGTGTCAGTACAAAAAAGTACTGAGTCCTACCATGGTACAGAAAAGTCTAGGCCTTCCACTAAAGAAGA
TGTTCTGACTGGAAAAGCAGTTTCTGTGCCAAAAGAACAAGGAAAAGTGGCGATTCTTAAAGTTGGGAAC
AAAACAGGGATCGGAGGAAACACTCCGCAACAAGATGCAAAGGAGGGAAAGATAGAAACCTCTCCGAGTG
CCAAACCCCTAAAAGCAAAGAAAAATTCTCCCTTTGCAAAGCTGCAAAAGGTGTTGCAGTACATGCAGCC
NAAGCATAATGCTGAAAAGTCCCACCCCTCCTTTCCGGAGCAAAATATTTTACATTTGGTGAACATGTCG
GACGAGGA

Ribosomal-processing-protein-8-like w/5' UTR

>gi|169566752|gb|FE577800.1|FE577800 CAXG1080.rev Amphioxus Branchiostoma floridae unpublished cDNA library CAXG, larva whole animal Branchiostoma floridae cDNA clone CAXG1080 3', mRNA sequence
CCAATTTTTAGATTAATTTATTAAGCCACACCAAATCTACTCTTTTAGAACCTGCGAAGTATGCCAATCT
TCATGTCTGTGGGAAATTTGTGTACAAAAGGTTTGTATATGTGTATGTAGGTTTTACACCCCTTCTCCAC
AGGAGGGTGTGACTGCTTAGCGTCCAAAGATTTGTCTTTTTTAACTTGCAACATATTCAACTATAAAATC
TATCATGTTGCTATGCAGCGGCTCTGCAATTTCTCTCTAGTGGTAAAGGAGGCAAACAGCGTATTCAGGA
CCCAAATGACTGTACATTTTTTGGGGTTTTGAAGGTGTGTTGGGTCTTACATAAAATTATGGTTTTACAT
GTGAATAATGTTAGTCTGAACAGCTTGTACTACATCTCTGTGTATGGCTATGGACATGTCCTATCTCTTC
TTGTACAGACAGGGTCTGAGTTCCAGCCCAGCACTCCCCCTGCTGGTCTTGGGTTCTGAGATCTTTCTGA
ACTCAAACATGACAAAGTGGCTGTTGGACAGGTCCTTTGACACAATCTTGAATCCAAAGAGAGCAAGTCC
CCTGATGA

>gi|379310479|gb|JT881713.1| TSA: Branchiostoma lanceolatum Seq43315.bl mRNA sequence
AGAAAGCAGCAAGGTCAACACACAAAGTGGTGGTATTGCAGATTTAAGAACTAGACATCAACAGACTACT
TCTAAGACACCAGGAGAAAGTCTGAAAAGGATACCAAAGAAGAAGGAGTACTAGATCGCTCTTCTCTACT
GAGACAAAAGATGGAGGCCAGGTTGAAGTCAGCCAGGTTCAGACAGATCAACGAGATGCTCTACACAACT
ACAGGTGAAGAAGCCAGAAGGATGTTTCAGAAGGACCCAGGTGCATTCCAGGTGTACCACCAGGGCTTCT
CAGCACAGGTGCAGAAATGGCCAGTAAATCCTGTGGACAAGATCATCATCTGGCTTAAGAGAAGGCCGTC
TTCTGAAGTAGTGGCAGACTTCGGATGTGGAGATGCTAAAATAGCTCAGAGTGTGAAGAACCAGGTTCAC
TCCTTCGACCTGGTGGCTGTCAACAAACATGTCACTGTGTGTGACATCACAAAGGTTCCCCTGGAGGATG
AGGCTGTGGATGTGGCAGTGTTCTGCCTGGCCCTGATGGGAACCAACATCTCTGACTTCCTCAGGGAAGC
CAACAGGGTCCTCAAAATAGGCGGTGTTCTTAAAATTGCTGAAGTCGCCAGCAGATTCGAAAACATCAAT
GGCTTCATCAGAGGACTGGCACTCTTTGGATTCAAGCTCGCGTCAAAGGACCTGTCCAACAGCCACTTCG
TCATGTTTGACTTCACAAAGATCTCAGAACCAAGGACCAGCAGGGCCAGTGCTAGTCTGGAACTCAGACC
CTGTCTCTACAAGAAGAGATAGGACATGTACAGAGATGTAGTAAAGCCGCTAGAGCCTGCTTTGAAGGCT
AAGATTTTTAACATTTAAGACCCACCATCACCATAGACACAAACAGGCAGTTTGTTGATTAAAGCCTTTT
GTCGACTTATTTCCCATATACAAAGAGTAATGAAATTTGGCATACTAATTAGATACTGAAAGAGTACAAT
CTTGTGGATGGTCCTGATGGGAATAAATAATTCTAGAAATTGGACCAGGCTGTGTCTCTGTCATCAGGTA

CTCTTGGCAAGTTCAAGTCCACACTTTTGCTCCCCTCTTTAACAAGCTGTAGTTTGATGACGGGGAGTTT
CTCGGGCTCCTTGTTGTCATTCTTTGCCATGACGTCGTTAGGGTTGTCCACGGCGATGCGGAGCGTGCAA
CCTGGGTGTATCAGGCCTCTCTCATGTGCTACTGCAACCTGGTTAACTACTCTTCTTTCCACCTGTGATA
GTACAAACAACACAATCAGAAATCTCTTACTTCGGCT

SDK3/ClpB

>gi|66496006|gb|BW881329.1|BW881329 BW881329 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne096f05 5', mRNA sequence
TAATGACGTCACAGAAATCAAACGGCTGGCAGAGATTGGAGTGGATGTGAACCAGAGACATATTTTGGGC
TGGACTCCATTAATGGTCGCAGCTGTCAGTAGGAACTTAGGAGCAGTGAAGGCTCTACTGGAGGCTGGTG
CTGACCCCAACATGAGGGAAGAGTTTGTCAATGTTTACCAGACTGCACGAGAGAAAGGAATGCATTCATT
GGACGTTCTTGTGACCAGAGAAGACGAGTTCAGTAACCGGCTGAATAACCGAGCCAGTTTCCGGGGATGC
ACGGCTCTGCACTATGCAGTGCTGGCAGACGATGTTCACATTGTCAAGGCACTTCTGGAAGCTGGTGCAG
ATCCCACCATGGAGAATGACAGTGGCCATGCAGCTGGTCTGTATGCACACAACATGGAAGTCAAAAGACT
CCTGGAGGAATACAAAGACAAGTATGCAGAGCTGCAGAGACAAAAGCAAGTGGAGGAGAGGAGAAAGTTC
CCCCTGGAGGAGCGGCTACGGGAGCACATTATTGGACAGGAGGGAGCTATCACTACAGTCGCTGCAGCCA
TTA

>gi|66654315|gb|BW939046.1|BW939046 BW939046 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne096f05 3', mRNA sequence
AATGATTCTATATTCTTGTGTTTCTTTGCAGGACCTGTCTAACAGCCACTTTGTCATGTTTGAGTTCACA
AAGATCTCAGAACCCAAGACCAGCAGGGGGAGTGCTGGGCTGGAACTCAGACCCTGTCTGTACAAGAAGA
GATAGGACATGTCCATAGTCGTGCACAGAGATGTAGTACAGGCTAAGATTTTTCACATGTAAAACCATAA
TTTTATGTAAGACCCACCACACCTTCAAAACCAAAAAAATGTACAGTCATTTGGGTCCTGAATACGCTGT
TTGCCTCCTTTACCACTACAGAGCAATTGCAGAGCCGCTGCACAGCAACATTGATTTTATAGTTGGAATA
TGTTGCAAGTTAAAAAAGACAAATCTTTGGACGCTCAGCAGTCAGAGGGTGTGTAAAACCTTTTGTACAC
AAATTTCTCACAGACATGAAGCTTGGTATACTTATTAGGTACTGAAAGAGTTGGTGTGGCTTAATAAATG
AATCTAAAAATTGGACCAGGTCTGGTCTCTGTCATCAGGTACTCTCGGCAAGTTCAGGTCCACACTTTTG
CTCCCCTCTTTAACTAGCTGGAGTTTGATGATGGGGAGTTTCTCAGGCTCCTTGTTGTCATTCCTTGCCA
TGACATCATTAGGGTTGTCCACGGCGATGCGAAGTGTGCAACCTGGGTGTATCAGGCCTCTCTCGTGTGC
TACTGCAACCTGGTTAACCACTCTTCTT

GNPDA

>gi|66404687|gb|BW796471.1|BW796471 BW796471 Amphioxus Branchiostoma floridae unpublished cDNA library, larva whole animal Branchiostoma floridae cDNA clone bflv004h24 5', mRNA sequence
AGCTGGCTCACGTGTCGGGTAGGACCAGTTCTCCAGTCCCGCGACTTTATCTCCGCCTTATCTGCAAAAT
ACACAACAAACAAGCTTCTTAAAATTAAGCGCTAACAACAGTGGATCAAGATGCGTTTGGTTATCCTTGA
TGACTATGACAAGGCCAGTGACTGGGCAGCCAGGTATATCATGAACAGGATCCTACAGTTCAACCCTGGT
CCTGACAATTACTTTGTCATGGGGCTACCTACAGGAAGCACTCCAGTTGGAACATACAAGAAGCTGATAG
AATTCCACAAAGCTGGGCAGCTGTCTTTCAGATATGTCAAGACTTTCAATATGGACGAGTATGTGGGCAT
TGCCCGTGATCACCCTGAGAGCTACCACTCCTTCATGTGGACAAACTTCTTCAAGCACATCGACATCCTG
CCTGAGAACGCACACATTCTGGACGGCAATGCAGAGGACTTGGAGGAGGAGTGCAGACAGTACGAGGAGA
AAATCAAAGAAGCTGGCGGCGTGGAACTTTTCCTTGGTGGTATCGGTCCAGATGGTCACATTGCCTTCAA
CGAGCCTGGCTCTAGCCTTGTGTCGAGAACGCGAGTCAAGACGCTGGCGA

>gi|727949753|gb|JZ813352.1|JZ813352 Alin_C_10021 Amphioxus diverticulum cDNA library
Branchiostoma belcheri cDNA clone DIVERTICULUM_FL_3_5_46-F06-M13F_F06, mRNA sequence
CCGCCTTATCTGCAAAATACACAACAAACAAGCTTCTTACAATTAAGTTGTGTTGTAACTAAGTGAGTCA
AGATGCGTTTGGTTATCCTTGATGACTATGACAAGGCCAGTGACTGGGGCAGCCAGGTATATCATGAACA
GGATCCTACAGTTCAACCCTGGTCCTGACAAGTACTTTGTCATGGGGCTGCCTACAGGTAGCACTCCCGT
TGGAACATACAAGAAGCTGATAGAATTCCACAAAGCTGGGCAGCTGTCCTTCAGATATGTCAAGACTTTC
AATATGGATGAGTATGTTGCCATACCACGCGATCACCCCGAGAGCTACCACTCCTTCATGTGGACAAACT
TCTTCAAACACATCGACATCCTGCCGGAGAACGCCCACATTCTGGACGGCAATGCTGAGGACCTGGAGGA
GGAGTGCAGGTTGTACGAGGAGAAAATTAAAGAAGCTGGTGGTGTTGAACTTTTCCTTGGTGGTATTGGT
CCAGACGGCCACATTGCCTTCAATGAGCCGGGCTCCAGCCTTGTATCCAGAACGCGAGTCAAGACACTGG
CAAAGGAGACCATCATTGCTAACTCTCGTTTCTTTGGCGGCGACTTGGGCAAAGTGCCAACCATGGCACT
GACTGTGGGCGTCGGCACTGTCATGGATTCTAGGGAGGTGATGATTCTGATCACAGGAGCCCACAAGGCC
CTGGCCCTGTACAAGGCTATAGAGGAGGGTGTGAGCCACATGTGGACTGTCTCTGCATTCCAACAACACA
GGAAGGTCATCTTTGTTTGCGACGAGGATGCGACC

>gi|66342692|gb|BW756044.1|BW756044 BW756044 Amphioxus Branchiostoma floridae
unpublished cDNA library, egg whole animal Branchiostoma floridae cDNA clone bfeg002i08 3',
mRNA sequence
ATGACTTTTAGGCATACAAGGGTTAATGACAAGAAGACAACACCACACCATAGTACAACTTCTATATTAA
AGGTGTTCGCAGTCACATAGATATGTAAGTATTGTCCACCATGTTGGATGATTAAATATGGGAGTTGCCT
GGTCCTGTCAGGTTATTACCACATAGGTCCAGGTTTATCAATCCAGCGTAGCAGACAATACTGACATCTG
AGTCATTTAAGTCCTAACATCATATGTAGTATCTATGCTACTGAGAACATAGATGTCACATGCTGACACA
GTTTTGTTAACATACACTAGTTCTACCTGTACATGCTTGCATCCTTCCCTCATTTCATTTCTTTACCTTA
TTCTCTTCAGCATGCAACTCTACGCATGCACTTACCATATTTTTTTATGCACGATGTCAGGGTATTCATG
CAACATACCATTCAAAACCAAGCATTTAGTCCACTGCAATCCATGTTTTCTTTCCAACACTTTGTCATTT
CTCATTACCTTTGCATATTTCTGTAAAATTTATTTCTTCAAACTTGTGCAGGAGTGTCTTCATATGTTCT
TTTTCCAACTAATAGTACTACTGTAGTTCAGTGCAAAAATGTTTTTTTGTAAACAACAATACTTAAAATA
GTGCACATGGTATCTCTTAGCTTCTTTTCTTGACTTTTTTTCTCAAGGAAATGAGGGTCCATGTATCAAG
AAGCTACATCTAACACTTTGCATGAGTGAAATCAATTTTTGNAAATAATGNAACTGTTACATGGTGAAAN
AAATATGTTTGCAGGTA

>gi|66323299|gb|BW736669.1|BW736669 BW736669 Amphioxus Branchiostoma floridae
unpublished cDNA library, egg whole animal Branchiostoma floridae cDNA clone bfeg002i08 5',
mRNA sequence
TCAAGACAGTGAAGTACTTCAAGGGTCTGATGCACGTCCACAATAAGCTGATCGAGCCCATGGAGAATGG
ACCGGAGAAGAAGAAACGAAGAGTGGACGAAGGGTACAAGGATTGAGTCTTTTTGGTGGTGGCTGCATGT
CCTGTAGTGTACTAGTCTAACTTCCAACGTGTTTCAGAAGTGTAGCACTAACATATGCTGTGGTGTTCTA
ATTCTAAACCACTAGGAATATATACTAATCATACTTACAATGTCAAACTACATAACTGTATACATTGACT
GTTCTAACCAAAAATTCTTACATAGGAGGTAAATAGCTTAACAATAACTGCATCCAAACAACACAAAAAA
AAATTAATTGGAAAACTTTGATGATTATTTTACCTGCAAACATATTTCTTTCACAATGTAACAGTTACAT
TATTTACAAAAATTGATTTCACTCATGCAAAGTGTTAGATGTAGCTTCTTGATACATGGACCCTCATTTC
CTTGAGAAAAAAAGTCAAGAAAAGAAGCTAAGAGATACCATGTGCACTATTTTAAGTATTGTTGTTTACA
AAAAAACATTTTTGCACTGAACTACAGTAGTACTATTAGTTGGAAAAAGAACATATGAAGAC

EST2 (Unknown)

>gi|66376778|gb|BW774562.1|BW774562 BW774562 Amphioxus Branchiostoma floridae
unpublished cDNA library, gastrula whole animal Branchiostoma floridae cDNA clone bfga023f01 5',
mRNA sequence
ACATTTCCGAAGCACTAACAAAAACACACGTACGTGAATAATAGTTTCTCATTATAAACACTATCTACGC
GCAAGTTGATATAGCTGTTGCTAAATGCTACACAAACACTGGTAAAGTACCAGTCTTAAGAAACACGGGT

AAATGCATCAGTTCCTAAATACTAGAAAAAACAGTCATGTTGGAGGTGAAGATATCCCTTGGCCAGGTGC
ATATACCAAAATGTGCGTTATTCTAATTAATACCTAATATTTGCATAATTAATAAAGACATTACATAGTT
CTTTTGTGGTCACTTCATGGTAGAGACTTCATATTGGAGATATATAGGTTGCTTGAGGACAGGTGAATAC
AATGAAATACATCTTATGTCAAGACTCAGGTATATGCAATAATGGGAATACAAGGGACCCAACCCTCTTT
GTCTGAAACAAGTTCAACTATCTTATTACCATGTAATATTGATACTATGAATGTATCAAACTCGTGTACT
TATATCATTCTGAAACTACATTTTGTGATTTATACCAATCAACTTCTAAAATGTCATGTAAACCCCCCCC
CAGGAGCAGCTGGTTACTACATATTACAGGTACGCTACCTGGTAACGTTACTATACTGCACCTGGGGAGT
AACGTATACGGTGTGTTACCTGGTACCTATATGGCACCTTATTACCGTACCG

EST1 (Unknown)

>gi|66515002|gb|BW896016.1|BW896016 BW896016 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne077c03 5', mRNA sequence
AAGTTGGTATCTCACTGCACTTTGGGTACTGGTGTGGCCCTGGGGGGTTAAGCCCCGGTCACAAAGCGCG
TACGATTTCTTGCGATGGTATATTCCGCATATCGTACGATGAGCGCAGTAAATCGCAGCAAAATCGTAAG
CAAAATCAGCCATCGCAACGCATCGGATGGTGTTCAAAATTTTTCCAGCGTCGTACGATTTTTCACTTGT
GTCTCTTTTGAATAGCCGCCTGGCCGCTTTTGTGCTTCTTTAACCAACAAAATGTGGACTTAGCTGTTTA
ATACCTTCCTATGATATATTCAACTGTAAAAATAGATTAAAAGAGACCATAACAACAAGAGCATTACAAG
AAAAGTTCCAATCAAGCGTGAGTACTGGTATGATTATCTCGGCCTGCTTGCCGGCTCTACGTGTCAGACT
CTTTCGAAGATCGTATCACGATATGCTATCAACAAAAAGATGCCATCATACAAAAATCATATTATAAGCA
TTATATCATATATATTTCCGACTCATAGAGCCTGTCTTTATGTTCGAGCCGGTCCCCATGGTTATTACGA
TTATGGTAAACTGACCCAAGACCGACGAGAGCTGAGATTTGATCTAATTATAAACTCACGTGCATGAAGC
GATCAAACAATTGTCTGCATCACCTGTGCGGGGCGCGCTGTTCTCTGGTTGCGACTGTGGCAGTTGCGAC
TGATATATTTTCCCTTTTCGTCAATATCCACAATATCAGGGGAAACTGAAACCATCACAACATGCTAACA
ATTCATAAATCTATAACCTCATCAGT

Transmembrane Protein 56B-like

>gi|66293573|gb|BW706999.1|BW706999 BW706999 Amphioxus Branchiostoma floridae unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad042f18 5', mRNA sequence
GACAACATGAGTTTCGAGGCGTTTCTTGTCGTCACGGCAGCGTCTTCGTGCGCCACTTGGATGTCCATCT
TCGCCTTCAGTCCCGTTCTCTTCACATGGCTGTCTTCGGCGTATAGGAGTCTCCCCAAGGACAGGCAACG
GCTGGTGGACAACCATTTCAAGACCGTGGTGCATGGAACGGCCGTCGCAGCACTCGCCTGGTATGCCTAC
ACCTGCACAGAGGTTCCGCCAGAGGGCGTTTGGCTCGATGCGCCACTTGTGAGGTTTGAGTCTGCGGTTT
ATTTCGGCTACTTGATATCAGACTTGATTCAAACAGCGATTTACCCGCACGTCAGCAACATAGAGTTCGT
CTCACATCACGTGTTCTCGTTGTATTCCTCCCTTATAGCAGCAAGCTATCCCGCTATGCCTTACTACGCC
AACATCTGCCACATGATGCAGCTTAGCAACCCCAGTGCGTTTTTCCGGTATGAACCGTTTAGTAACTAAC
AGTTTTATAGTGAGTACGTCATGTAATTTTACTTCTATATATTGGTACTTTTGTTCGAGGCACACTTGTA
ACGTTACTTTAGAATAAAGGTTGTTTTTTTTTGCAAAACTGTGACGGTAAAATCCTAATTGCGACATATC
ACAAG

>gi|66312465|gb|BW725871.1|BW725871 BW725871 Amphioxus Branchiostoma floridae unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad042f18 3', mRNA sequence
GTCGGCCCCACCAAGGGCAACTTCTATCTTTTACTGGTTAATAAAGGTCTTCATTCATTCGTTCATCAAA
AAAGAATGAACGACTACAATTAATGTAATGTTTAAAAACTTTTGTGACGCGTCTTTGCGGTCTGCGCAGG
TTTTGGCTCTGACTCTACAGCTTCTGCTTTGTTTCCTTTCCCCCAAAGTAGTCTACGACCCCTTTACATA
TCAAGCCAAACCAGTAGTAGTTCATGGCATTAAACAACAGTGAACCAAAGATATAACACGAGGAAACATG
CAGCGGCAGCTGACTGAAAGAGTCTTGGAAAATCATGATCTTGGCTAAGTTGACGGTCGCGATGCCGGTG

AACAGTATTCGGGAGATGAAACACGTCACCAGCAAGGTGATACCGTTCCACGTGTAATACTTGGAACCTT
TCAACCCCAGCTCCTCAAGGATGACCCTTGTATGGGAGGAAAATCCAAGAATTACGTATATCAATGTCTG
ATAATTGTTTTCAGTATCTAGCAATGATTGATCTTGCACTGCTAGATGTAGTTCATGCATTACATCATTC
TATTTCATGGTTTTTGTACCTAAAAAAGCGACGACCAAGGAAATAGAGAAAATCAGACTTTTAGTGAGTT
TTAACTTGAAGTGGCAATGAGATGACGCCCAGTACGAAATGTGATACTTGTG

Carbohydrate Sulfotransferase 14 like

>gi|66414109|gb|BW805893.1|BW805893 BW805893 Amphioxus Branchiostoma floridae unpublished cDNA library, larva whole animal Branchiostoma floridae cDNA clone bflv034k11 5', mRNA sequence
GGAACGGTACGAAGAGCAGGAAAACGTCATGATGTAGTCGTGTGCCTCCTCTTTTAACTCTGCGCTGAGA
CTTTCAGCAGTATTCAGTATGTTCAGCACGTTACGGAGAGCGTCGTGGTTGTTCCTGACTCTAATGGTGC
TGTCATGTGTGGGAATGGTCTTGTTTTGGAGGAAAATGTTTCTGACGGGGCAGGCGGGCGAGCCGCTCGG
CAGGAAGTCGCTCGGCACGCGGGTGTTATCTGAGAACCGTCCGCCTACAGATGGAACCGCAGACGACGTT
TTACTTAGGAGGTTTCTGGAGAAAGTCAACAACACCATAGACGCACGGATAGCGGCGAAAGTCAGGGACA
ACAAGCAGCCACAGTTACAGCTCCACCCC

>gi|66436347|gb|BW828131.1|BW828131 BW828131 Amphioxus Branchiostoma floridae unpublished cDNA library, larva whole animal Branchiostoma floridae cDNA clone bflv034k11 3', mRNA sequence
ACATTTGGCACAGCACAGCATAGCATAGATATAGTAGCAATCTTTGAATAAACAAAAAGTGGCATTACAA
TCACAGAGAACTCTAACATAATATAACAAACGTATAAAATAGCACACTTCATAAATTTCTGTGAATAAAG
GCAAGACTTTCAGCAAAAGTTCTGCTACCACTTGTTACTTAACCCTATTGCACTTGGTATACAAATTGTA
ATACATTTACATGTAAGTACATTGCCCCAGGAATGAAGTGGGTATAGACTTACTAGAAAATATAGAAAAA
CAGAAAATATTCATTGCCAAAATATAGATTTACCTCTGTATACCACATGATACATACAATGTACCAGCAA
ATGTCAGACAGTTTTTCTTATAAAATGGGGATAGTGGCATTTTCTTATGAACCTCATTTGCACATATTCT
GGTCGAAATTAAGGTATGTCAACCTAGATTCCTAGAAGTCCCAGACTGTAATTATGTTCTTCTGA

>gi|66503427|gb|BW887805.1|BW887805 BW887805 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne146d08 5', mRNA sequence
TTACAGTCTGGGACTTCTAGGTATCTAGGTTGCCATACCTTGATTTCTACCAGAATATGTGCATGCAAAG
AAGGTTCAAAGGAAAATGCCACTATCCCCATTTTATAAGAAAAACTGAATGTCTGACATTTGCTGGTACA
TTGTATGTATCATGAGGTATACAGAGGTAAATCTATATTTTGGCAATGAATATTTTCTCTGTTTTTCTAT
ATTTTCTAGGAAGGATTTTCTAGTAAGTCTATACCCACTTCATTCCTGGGGCGATGTACTTACATGTAAA
TGTATTACAGTTTGTATACCAAGTGCAATAGGGTTAAGTAACAAGTGGTAGTGGAACTTTTGCTGAAAGT
CTTGCCTTTATTCACAGAAATTTATGAAGTGTGCTATTTTATACGTTTGTTATATTATGTTAGAGTTCTC
TGTGATTGTAATGCCACNT

>gi|66660868|gb|BW945599.1|BW945599 BW945599 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne146d08 3', mRNA sequence
GGCACAGGCACAGGCATAGCATAGATATAGGTAGCAATCTTTGAATAAACAAAAAGTGGCATTACAATCA
CAGAGAACTCTAACATAATATAACAAACGTATAAAATAGCACACTTCATAAATTTCTGTGAATAAAGGCA
AGACTTTCAGCAAAAGTTCCACTACCACTTGTTACTTAACCCTATTGCACTTGGTATACAAACTGTAATA
CATTTACATGTAAGTACATCGCCCCAGGAATGAAGTGGGTATAGACTTACTAGAAAATCCTTCCTAGAAA
ATATAGAAAAACAGAGAAAATATTCATTGCCAAAATATAGATTTACCTCTGTATACCTCATGATACATAC
AATGTACCAGCAAATGTCAGACATTCAGTTTTTCTTATAAAATGGGGATAGTGGCATTTTCCTTTGAACC
TTCTTTGCATGCACATATTCTGGTAGAAATCAAGGTATGGCAACCTAGATACCTAGAAGTCCCAGACTGT
AA

EST17 (Unknown)

```
>gi|66296981|gb|BW710407.1|BW710407 BW710407 Amphioxus Branchiostoma floridae
unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad052b13 5',
mRNA sequence
TACATGTGATTGGGGAAGGCAGGGAGGCACACAAAAGTTTGTGATCTTAAAACTACCAGATATATCCATG
CTGGAAAGTAAACAGTAGAAGACATGTGTGACAGTGCTACAGCCACATTTGCTGGAATCCCTTGTCTACC
CAGTGAGGTACAGGGGGAGATTTTGTATCGCCTCCATGATGGGGTGGCATTGACAAATGCTCGACAGGTG
TGTCGCCTTTGGAAACAACTTGTGGACCAACCAGGCGACAAGCAGGTATGGTACACCATCTGCCGACACT
GCATCCCAAATGGAGTCCTCCAACAATTGACTCAGTTTGACAGGGACACCTTCTTCCAGACGTCTTCCAA
CAAGGTTGCAAAGAGCTCATGGTCCTGTGGAAAAAGACACCAGCTTGAGCAAAGGATATCTTTGTCATCT
CAAGATAGAAGACATTCTCAAAACACCATTCCTGACTGTATTGTCCACAAGCATATAAACTGCCAGTGTA
ACAGCCAGATACAGCCGTGTGTATCAGACCCTGTGGTGTACTGGAGGAGTGTGTATATGGAGTGGTACAG
AGGAAGGTTTGCAGGGAAGTGGGCCATGGTGA
```

```
>gi|66315962|gb|BW729350.1|BW729350 BW729350 Amphioxus Branchiostoma floridae
unpublished cDNA library, adult whole animal Branchiostoma floridae cDNA clone bfad052b13 3',
mRNA sequence
AAAAGGTTAATAAGTCTGCACATGAATGACTAGATACAGACATTAATCTGGCTAATGACGTCAATACTTT
TCCAACTAATGCACTGAAGCTTACGACAGCAAGTACACCATATGTACCAGTACACCATATGTACCATCTC
TAGATTTTTAACTGAAGTTAAAAATTGACTAAAGAAATGTGAAAATATTTGCATTGTTATCTTCAAACTA
AACGTAGGGTAACTGTTCATAGGAGCCTATGCAAGAAGATTATTTCAGTCTGGTGCCTGCAAACAATGAC
GAATCACAACAACAAAGTGTGTGACCACGTTTGCTGGCCAACTTCAGCCTGAGGGTCCGTGTGAAATGGG
GTAGAAAATACTGATCCCTTGACAACATGCAACAACAAAGCTAGTACGTGACTGATCTTCCTTCATGGTA
GGGAACGAAATATAAAACTCAGGTGTGTCCCTGTTGCTTCTTTGTCTTCAAGGTTGAACTTTCATTGAAC
CTTTAGGAAAGGTTGTGCCATGCCCTGAGTCTTTAGCATGAACAGAAGGTGCCTCTTGTCACATGTTCTA
CCCCTCTAGCAAATTCCACTATCTCAGACACACATGCTGTAACTCAACATCAATAGCAACAAGTGACAGC
GCTTAAGGCATGAGTTGAACTTTCATCCAAAAGGTAATTTGCCTCCTGATCACATTCTTCTCCATGAGTC
ATCAAGTTTGTATGTCAA
```

EST16 (Unknown)

```
>gi|66497852|gb|BW883175.1|BW883175 BW883175 Amphioxus Branchiostoma floridae
unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne133g12 5',
mRNA sequence
AGACTTTCTGTGCAACATCATGCCAAAGTTCTTCATCCACGTTTTGGTGCATACATACATACATACATAC
ATAACAATCTCTTGGATGTCATGTTAGGCAACTTACTTTACTTAGGCAACTGTTAAGTACAATGTATCAT
CAGCTCTCTAAGTCTCTTCCATTATTTTTAAGTTTCTCATTTAGTCATTGAAGGAAAGGAAATGCCTGTT
TTGAGAAACAAATTTGACAAACATGTAGCATTATCCTCCTATAATGTTTTTTGCTTTCTTGTTGAGAATA
CACATACATGTATCACGGCCTTCAGCTACAAAACCACTTGTACCCTGGGTACCATCCTAGTTTGNTTTCC
GCAC
```

```
>gi|66656181|gb|BW940912.1|BW940912 BW940912 Amphioxus Branchiostoma floridae
unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne133g12 3',
mRNA sequence
TATTTTCCGGGACAGATACATGATAGGAGAGGAATTGTCTGTACATGATGGAACTTGGATGATGTACAGT
TTGGGTACATAGAGACATGGCACACAATAAAAACAGCTAAACATTTGGGTCATTGCTTGAAGTAAACCTT
GTTTGCTTTACAAAAATATCCTCTAGTGGGGTAATACACCCTTTTCACACAGCAGATGTGATAGATTGAA
ATCGACGCCAATGTTGCCAACAATAGACTGGTCCTGACTATCCATCATAGTTAGCAGTTCAACCAATGAT
AGGATGGCAATCAGGAAATAGACTAGTTAGAGAATGGCTGTATGTGGGTTAAATATTTGCATTACTTAGG
CCACACCAATTTAATGTCTTGGTTCTCGGATTTCCCGATGGTCCCAACAAAAATAAGTAGGTTTTGTTAT
TGCAGACCTGTAAACAAAACCCATTCCATGAGTACATACTGCTTATTAACCAATCAGAGAGCTTCTTTGT
```

AGTCGCATGATAGAAAACAACCAATAAGCTTCTTGCATTTCTTCAACATACATAATTTCAGTATAAAAAC
AATTCCCTCTATTCAGAACAGTGGTATCAAATCA

>gi|66498387|gb|BW883710.1|BW883710 BW883710 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne134o10 5', mRNA sequence
AGACTTTCTGTGCAACATCANGCCAAAGTTCTTCATCCACGTTTTGGNGCATACATACATACAGTACATA
CATAACAATCTCTTGGATGTCATGTTAGGCAACTTACTTTACTTAGGCAACTGTTAAGTACAATGTATCA
TCAGCTCTCTAAGTCTCTTCCATTATTTTTAAGTTTCTCATTTAGTCATTGAAGGAAAGGAAATGCCTGT
TTTGAGAAACAAATTTGACAAACATGTAGCATTATCCTCCTANAANGTTTTTTGCTTTCTTGTTGAGAAT
ACACATACA

>gi|66656726|gb|BW941457.1|BW941457 BW941457 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne134o10 3', mRNA sequence
ATTTTCCGGGACAGATACATGATAGGAGAGGAATTGTCTGTACATGATGGAACTTGGATGATGTACAGTT
TGGGTACATAGAGACNTGGCACACAATAAAAACAGCTAAACATTTGGGTCATTGCTTGAAGTAAACCTTG
TTTGCTTTACAAAAATATCCTCTAGTGGGGTAATACACCCTTTTCACACAGCAGATGTGATAGATTGAAA
TCGACGCCTATGTTGCCAACAATAGACTGGTCCTGACTATCCNTCATAGTTAGCAGTTCAACCAATGATA
GGATGGCNATCAGGAAATAGACTAGTTAGAGAATGGCTGTATGTGGGTTAAATATTTGCATTACTTAGGC
CACACCAATTTAATGTCTTGGTTCTCGGATTTCCCGATGGTCCCAACAAAAATAAGTAGGTTTTGTTATT
GCAGACCTGTAAACAAAACCCATTCCATGAGTACATACTGCTTATTAACCAATCANAGAGCTTCTTTGTA
GTCGCATGATAGAAAACAACCAATAAGCTTCTTGCATTTCTTCAACATACATAATTTCAGTATNANAACA
ATTCCCTCTATTCAG

EST15 (Unknown)

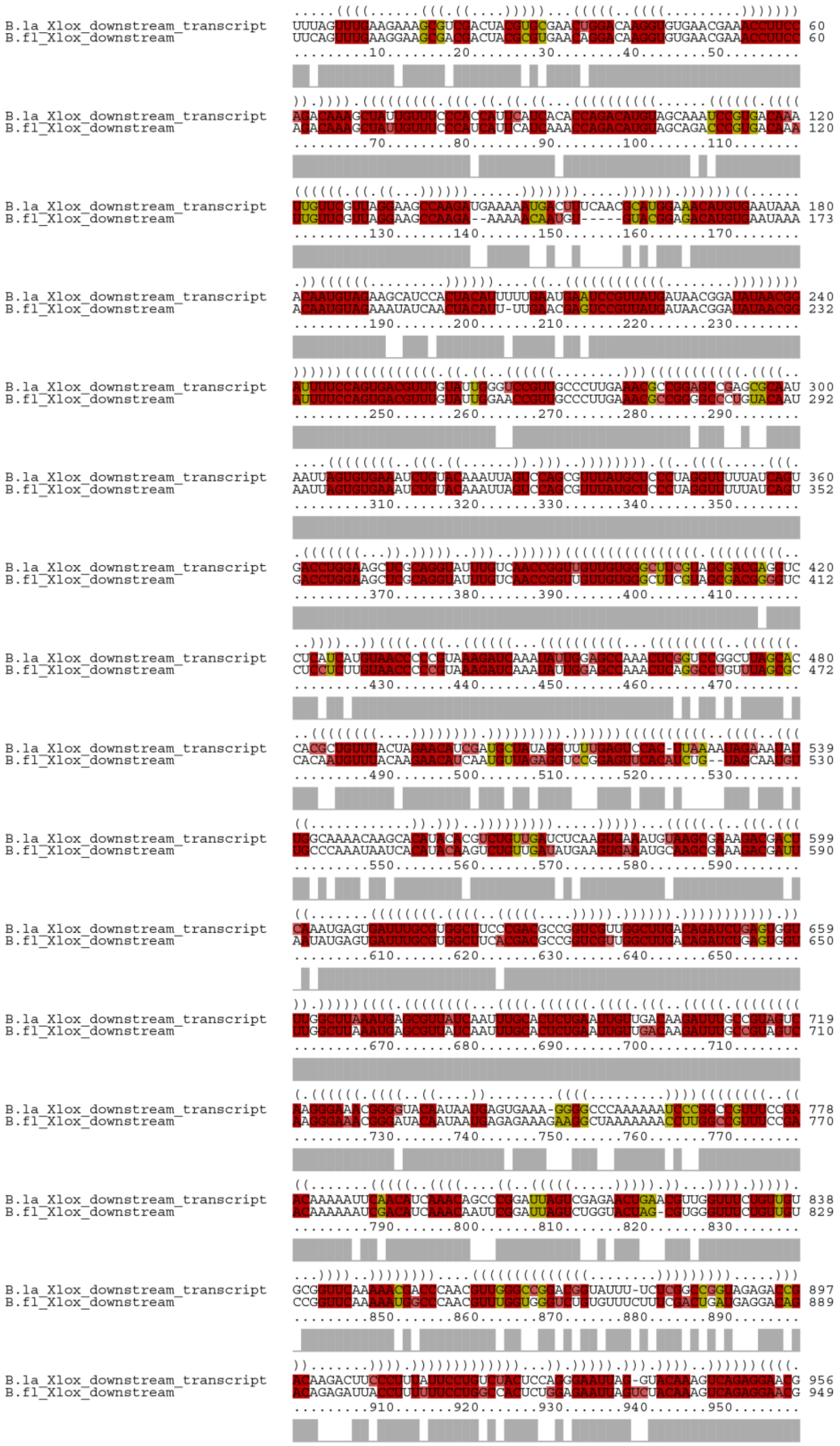
>gi|66506895|gb|BW890294.1|BW890294 BW890294 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne153b10 5', mRNA sequence
CCAGCTTTCATACATCTTGGTTAAATGCACAGGCAGGTGACTGAAGTCATATGCATGCATGTACTGGGAG
GAAAAATGTAAATTTTAAAGAAAATCTAAGACATCTATAACTCTAACCGTACAGGCAAGATAGTGGCTAC
CTGGAATTGTTGTAACAAGTTGATACATAAATGTAACAACATATGTAACAAGATTGCACATACGTAACAT
CATTCTCAGGTAGCTGGCCACTTGCAACATTCTTCTTTTAACTTGTCAATGCTGTGGACAATCTAAGCTT
GCAATATTGTCGACTGAAATTGAAATCATGACTAAAAGATTCTGCATCGCAGTGCGATTTATGATTATTG
TGATTTTAGCATAATGTTGTCATTATTAACTTGTGCTTTTTATGCTCAAGCCNATCTTTTTATCTTATAC
TGCTAATTGA

>gi|66663411|gb|BW948142.1|BW948142 BW948142 Amphioxus Branchiostoma floridae unpublished cDNA library, neurula whole animal Branchiostoma floridae cDNA clone bfne153b10 3', mRNA sequence
GTACAGACTAGGATGGAATATTGGGAAATCTGATTTGTAAAATGGTCGCAGTAAACCTTGATCTTGTGGC
TGACAACTGCACCCCACGTTGGCTCTGACTTGTTTGATTGAAGCTGTTTGTCCAAAATAAAAACAAGGAA
TTTCATAATCTCCAAAGTTGGGTTCCTAGAATCTTCAAGTTAGACAGTGACTTCCCCTGACTTTCACACT
GGCATAATGGTTCCTGTTACAGGTACCCCAAACTTGCTAACTTTTCTGTCAGCATGGGATGAGATAATAG
CATTTTCTTAGTAAAAATCATAACCTTATGACTTGTAAGAGAAGATCTTCAGAGAAATTTCGTGGTCCAC
AATGCTAATCCACAGTAGAGGAAAATGTCTGTGCAGGGAGTATCAACTACCCTTCCAATGTCAGTGTACA
TTATATGACTATAGCACCATAGGT

```
>gi|66416507|gb|BW808291.1|BW808291 BW808291 Amphioxus Branchiostoma floridae
unpublished cDNA library, larva whole animal Branchiostoma floridae cDNA clone bflv041g19 5',
mRNA sequence
TTATTGTGATTTTAGCATAATGTTGTCATTATTAACTTGTGCTTTTTATGCTCAAGCCAATCTTTTTATA
TCTTATACTGCTAATTGAATTTGTGTAGTATTCGCTTGAAATGTAGACTTCAATAATGGAATATTTGCAC
ACTCTGTGCATCTGTACTATAAAATCAATGCTGCTGTAATATTACCACAGTAAGAAAAATGTAGAGCTAG
AAACATCACTGTGAATAGGACGATATGCAACTTTGTGGGAGCAGGATTTAGTGTTTATTAAGCAGCAAAA
TCATCTCATCTTACTTATGATGTATCACATTTAGGTAGCAAATCTGCACAGAGTGTTTTATTTTGAACTC
AAAGTCTAGATCTAGTTTTG
```

```
>gi|66438796|gb|BW830580.1|BW830580 BW830580 Amphioxus Branchiostoma floridae
unpublished cDNA library, larva whole animal Branchiostoma floridae cDNA clone bflv041g19 3',
mRNA sequence
TACAGACTAGGATGGAATATTGGGAAATCTGATTTGTAAAATGGTCGCAGTAAACCTTGATCTTGTGGCT
GACAACTGCACCCCACGTTGGCTCTGACTTGTTTGATTGAAGCTGTTTGTCCAAAATAAAAACAAGGAAT
TTCATAATCTCCAAAGTTGGGTTCCTAGAATCTTCAAGTTAGACAGTGATTTCTCCTGACTTTCACACTG
GCATAATGGTTCCTGTTACAGGTACCCCAAACTTGCTAACTTTTCTGTCAGCATGGGATGAGATAATAGC
ATTTTCTTAGTAAAAATCATAACCTTATGACTTGTAAGAGAAGATCTTCAGAGAAATTTCGTGGTCCACA
ATGCTAATCCACAGTAGAGGAAAATGTCTGTGCAGGGAGTATCAACTACCCTTCCAATGTCAGTGTACAT
TATATGACTATAGCACCATAGGTTTGTTGATTACAAAAGGATTCAAACATTCCTGCGCCTCAAAGGCAGC
TAGTATCTATAGCGAGATCCACATAGTAGAATCTCAATCATATCTGCTGAAACATTTATCATACAAAATA
AACGACACATAACCAAACAAAACTACAAAACTAGATCTAGACTTTGAGT
```

7.3. Appendix 3. Alignment and secondary structure within the amphioxus Xlox 3'UTR



```
                                   .....((((......(((.((....)))))...(((((..((((.........)))))))
B.la_Xlox_downstream_transcript    UUUAGUUUGAAGAAAGCGUCGACUACGUGCGAACUGGACAAGGUGUGAACGAAACCUUCC 60
B.fl_Xlox_downstream               UUCAGUUUGAAGGAAGCGACGACUACGCGUGAACAGGACAAGGUGUGAACGAAACCUUCC 60
                                   .........10........20........30........40........50.........

                                   )).)))).((((((((.(((.((.((.((...((((((((((......((((((.((((
B.la_Xlox_downstream_transcript    AGACAAAGCUAUUGUUUCCCACCAUUCAUCACACCAGACAUGUAGCAAAUCCGUGACAAA 120
B.fl_Xlox_downstream               AGACAAAGCUAUUGUUUCCCAUCAUUCAUCAAACCAGACAUGUAGCAGACCCGUGACAAA 120
                                   .........70........80........90........100.......110........

                                   ((((((.((.((...)))))).......))))))).....)))))).))))))((......
B.la_Xlox_downstream_transcript    UUGUUCGUUAGGAAGCCAAGAUGAAAAAUGACUUUCAACGCAUGGAAACAUGUGAAUAAA 180
B.fl_Xlox_downstream               UUGUUCGUUAGGAAGCCAAGA--AAAAACAAUGU-----GUACGGAGACAUGUGAAUAAA 173
                                   .........130.......140.......150.......160.......170........

                                   .))((((((.........))))))....((..((((((((((((........))))))))
B.la_Xlox_downstream_transcript    ACAAUGUAGAAGCAUCCACUACAUUUUUGAAUGAAUCCGUUAUGAUAACGGAUAUAACGG 240
B.fl_Xlox_downstream               ACAAUGUAGAAAUAUCAACUACAUU-UUGAACGAGUCCGUUAUGAUAACGGAUAUAACGG 232
                                   .........190.......200.......210.......220.......230........

                                   ))))))((((((((((((.((.((..((((((........)))(((((((((((.((((..
B.la_Xlox_downstream_transcript    AUUUUCCAGUGACGUUUGUAUUGGGUCCGUUGCCCUUGAAACGCCGGAGCCGAGCGCAAU 300
B.fl_Xlox_downstream               AUUUUCCAGUGACGUUUGUAUUGGAACCGUUGCCCUUGAAACGCCGGGGCCCUGUACAAU 292
                                   .........250.......260.......270.......280.......290........

                                   ....((((((((..(((.((......)).)))..)))))))).((..((((.....(((.
B.la_Xlox_downstream_transcript    AAUUAGUGUGAAAUCUGUACAAAUUAGUCCAGCGUUUAUGCUCCCUAGGUUUUUAUCAGU 360
B.fl_Xlox_downstream               AAUUAGUGUGAAAUCUGUACAAAUUAGUCCAGCGUUUAUGCUCCCUAGGUUUUUAUCAGU 352
                                   .........310.......320.......330.......340.......350........

                                   .(((((((...)).)))))..)))..))))))(((((((((((((((((.(((((((((..
B.la_Xlox_downstream_transcript    GACCUGGAAGCUCGCAGGUAUUUGUCAACCGGUUGUUGUGGGCUUCGUAGCGACGAGGUC 420
B.fl_Xlox_downstream               GACCUGGAAGCUCGCAGGUAUUUGUCAACCGGUUGUUGUGGGCUUCGUAGCGACGGGGUC 412
                                   .........370.......380.......390.......400.......410........

                                   ..))))..))((((.(((..((((((...((((((((((...(((((((((..((((((.
B.la_Xlox_downstream_transcript    CUCAUCAUGUAACCCCCGUAAAGAUCAAAUAUUGGAGCCAAACUCGGUCCGGCUUAGCAC 480
B.fl_Xlox_downstream               CUCCUCUUGUAACCCCCGUAAAGAUCAAAUAUUGGAGCCAAACUCAGGCCUGUUUAGCGC 472
                                   .........430.......440.......450.......460.......470........

                                   ..((((((((....)))))))).))))))))).))))))((((((((((..((((..(((
B.la_Xlox_downstream_transcript    CACGCUGUUUACUAGAACAUCGAUGCUAUAGGUUUUGAGUCCAC-UUAAAAUAGAAAUAU 539
B.fl_Xlox_downstream               CACAAUGUUUACAAGAACAUCAAUGUUAGAGGUCCGGAGUUCACAUCUG--UAGCAAUGU 530
                                   .........490.......500.......510.......520.......530........

                                   ((............)).)))..)))))))))).....)))))...(((((.(..(((.(((
B.la_Xlox_downstream_transcript    UGGCAAAACAAGCACAUACACGUCUGUUGAUCUCAAGUGAAAUGUAAGCGAAAGACGACU 599
B.fl_Xlox_downstream               UGCCCAAAUAAUCACAUACAAGUCUGUUGAUAUGAAGUGAAAUGCAAGCGAAAGACGAUU 590
                                   .........550.......560.......570.......580.......590........

                                   ((.......((((((((.((((..(((((....))))).))))))).)))))))))))).))
B.la_Xlox_downstream_transcript    CAAAUGAGUGAUUUGCGUGGCUUCCCGACGCCGGUCGUUGGCUUGACAGAUCUGAGUGGU 659
B.fl_Xlox_downstream               AAUAUGAGUGAUUUGCGUGGCUUCACGACGCCGGUCGUUGGCUUGACAGAUCUGAGUGGU 650
                                   .........610.......620.......630.......640.......650........

                                   )).))))))((((.((((((((...((((.((((((.((((.(((..(((((.((((((((
B.la_Xlox_downstream_transcript    UUGGCUUAAAUGAGCGUUAUCAAUUUGCACUCUGAAUUGUUGACAAGAUUUGCCGUAGUC 719
B.fl_Xlox_downstream               UUGGCUUAAAUGAGCGUUAUCAAUUUGCACUCUGAAUUGUUGACAAGAUUUGCCGUAGUC 710
                                   .........670.......680.......690.......700.......710........

                                   (.((((((.((((..((....)).......((((.........))))((((((((..((
B.la_Xlox_downstream_transcript    AAGGGAAACGGGGUACAAUAAUGAGUGAAA-GGGGCCCAAAAAAUCCCGGCCGUUUCCGA 778
B.fl_Xlox_downstream               AAGGGAAACGGGAUACAAUAAUGAGAGAAAGAAGGCUAAAAAAACCUUGGCCGUUUCCGA 770
                                   .........730.......740.......750.......760.......770........

                                   ((.......(((((..((((....((..(((((.....))))).))...)))).)))))).
B.la_Xlox_downstream_transcript    ACAAAAAUUCAACAUCAAACAGCCCGGAUUAGUCGAGAACUGAACGUUGGUUUCUGUUGU 838
B.fl_Xlox_downstream               ACAAAAAAUCGACAUCAAACAAUUCGGAUUAGUCUGGUACUAG-CGUGGGUUUCUGUUGU 829
                                   .........790.......800.......810.......820.......830........

                                   ...))))..)))))))))....((((((((((((((........)))))))))).....)))
B.la_Xlox_downstream_transcript    GCGGUUCAAAAACGACCCAACGUUGGGCCGGACGGUAUUU-UCUCGGCCGGUAGAGACCG 897
B.fl_Xlox_downstream               CCGGUUCAAAAAUGGCCCAACGUUGGUGGGUCUGUGUUUCUUUCGACUGAUGAGGACAG 889
                                   .........850.......860.......870.......880.......890........

                                   ))........)))).))))))))))))))...)).)))))).))).))))..))))))((((.
B.la_Xlox_downstream_transcript    ACAAGACUUCCCUUUAUUCCUGUCUACUCCAGGGAAUUAG-GUACAAAGUCAGAGGAACG 956
B.fl_Xlox_downstream               ACAGAGAUUACCUUUUUUCCUGGCCACUCUGGAGAAUUAGUCUACAAAGUCAGAGGAACG 949
                                   .........910.......920.......930.......940.......950........
```

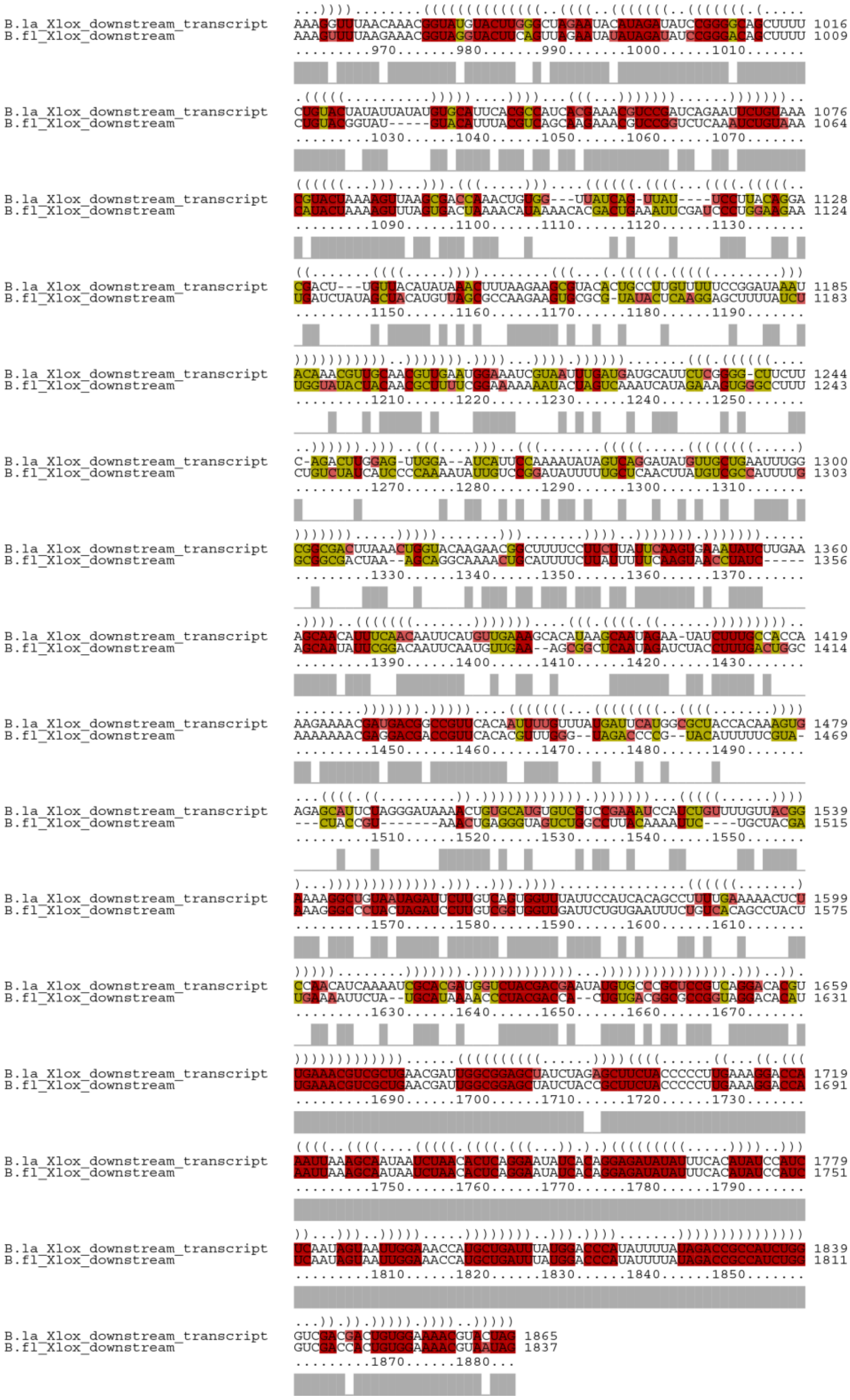


Figure 7.1. Alignment and conserved secondary structure within the amphioxus Xlox 3'UTR.

LocARNA alignment (Smith et al., 2010) of *B.lanceolatum Xlox* 3'UTR transcript and *B.floridae* genomic sequence. Brackets indicate the formation of pairing stem structures, whilst ellipses represent no folding constraint. This results in stem-loops in the form of (((…))). Red and yellow colouration represent different forms of base pairing. Grey boxes indicate conserved secondary structure.

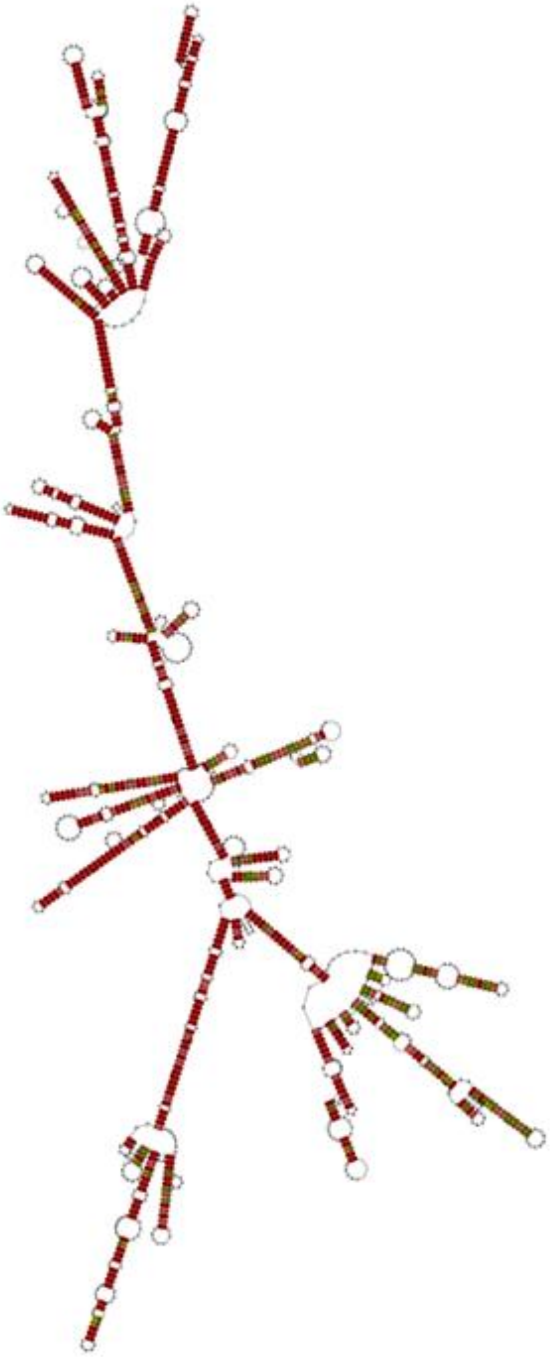


Figure 7.2. Conserved Secondary structure within the amphioxus Xlox 3’UTR.

LocARNA RNA secondary structure prediction (Smith et al., 2010) based on *B.lanceolatum Xlox* 3’UTR transcript and *B.floridae* genomic sequence. The position of Stem-loop structures is based upon the alignment within figure 7.1.

7.4. Appendix 4. Nucleotide and peptide sequences of sequenced B.lanceolatum *SCP1* clone.

>*B.lanceolatum_SCP1*_sequenced
GCAGGTGTATCATCAGCAAGAGCCGTTCTTCAAGGCTCTCTCACCTCAGCAACAGGAGCACCAGCATAGCTTC
TTTAAGATTGGTACAGAACAGCAAGAAGTGGAGATAAAAACCCTTTCACCCATGCGGCTTGGGCAGCAGATG
CACAGTGGGGAGCGCCTGACAAGCCTCCATTCTCGCCTTCAAAAGGAAGCAGAAAAGATCAACAAATGGAAA
CATCAGACAGAGATGCAAATCCAGCAGAGAGAAAGGAAGATCCAGGACACTCAGCAAACCATTGATTCACAA
CGCAAGTCCATTCTTGAGCTGCAGCTTCAGAATGAGAATCTCAGCTCCAAGCTGCAGGAGGAGATAGATGGC
CGTGTGGAGATCATGAAGAAGATCACTGCCACTCGAGACATGTGCTACCTGCTCAAGGATCATGCTGCTAATG
TTGAGGAGAGAATGGGGAAGTGTGAAGCCAACCGAGATGAGCTGCAATGTCTCCAACAAGACACGGTGTTC
CAACTCCAGGAGCTGACATCTAAATTCAACAACCTTCGCATTAATCATACTGAGGCAGAAAAAGTTCTCGGAA
ACAAGCTGAAGGAAAGTGTGAGTGAGCTCAACCAGGTTAAGTGTGATTACCAGAATGAGAAGGTCAACGTG
GAGAAAAGGCTTGAAGGCTTGATGCAGCAATGTTCTGAGAAAGAGATGGAAATCTCAAAGCTCACAGGTGCA
CTCACAGATAAGCAAGCTCAGCTCAGTGACCTGGAGCAACAGTGCAGTATGCTGGAGGAACATGTTGCAAAG
CTGGAAGATGAGTTCAAGTCTCTTCAAGATCAGCTGCAAGAGGCAAGTGACAAGATCTTCAGCAGGGATAAG
GAGATGGAGAAGATCTCGGGAGAGTTGACCAATACGGAGGCGCAGCTGCAAAAAGTGTCCGATGATTGTGA
GCAACTTGAGTCACACATTGGACTGTTGAAGGAACAGCATTCAATCGAAGTGACCAAGATTAATCATCAGCTG
GATACAGTAGAGGAAAAGTTGAATGTAGAGAAAACGAAGACAAAGGACATATCTTCCAAGCATCTCAGTGCT
GAGCAAAGAATTGGTGAGTTGCAAGCATCAATAAATGAGCAAGAGAAAAAGTATGAAACTCTGATAGAGGA
AAAGTCTAAAGTGGAAGGAGAAAAGGCAGCAGTTGAGGCTGAAGGAGTGCTGTTGAAGGAAAATATCATGG
CGCTTGAAACAGAGCAGAAGAGCATGAAAGATCAAGTGTCACAACTACGCACCACGGTGGCGGTCCTGACAG
ATGCAAAAATACACGCAGAGGAACAGTTGTCAATGCTTGAGAAAGAAAAGAAAGTAACAGGAGACGAAGTA
AAAGTCCTCACTAACACTTCNGCTGCAAGGGATAAGGAAATCAAGAAATTGCAAGATGACTTGAAGAAGGCA
GCTGAGCGCCAGAAAGAGGCAAAGAAGCAGGAGAAAGAATTGAAACAGCAACTGAAGACTAAAGATGCCG
AAGTTGACAAGACCTCTGCAAAGCTTAAGAAAGTGCAGGGTGATCTTGAGGAGATTCAGACTCTAATGGCTTC
CTTGCAGAGAGACCATGAGGAATTAAAAGACAAATTGAATGCAGCAGATCTACAACGATCCACGCTTCAGTC
GTCGCTTGATGAGGTCAACCAAGAGAAGGTATCTTTAAACGACACATTGAAACAGTTAGAAGAAAGTCTTAA
GGCACAAGATAAAGAGGCAGTGGGGAAGATACATCAACAGCAGGAGAGTACCAAAGCCCTTCAGACCGAAC
TAGACAGCAACAAGAAGTCCATGACTAAACTCGAAAACCGTGTTAAGTCACTAGAGAAACAAGTGGCTGAAA
AGACAAACAAAATAAAAGATCTTCAGCAAGACAACAAAACCGTGAAGAAAGAGCTTGGTATTCACCTTAAACT
GTCAAATGGATTTGAAGAAAAGGCCAAGTCCCTTGAGGAAGAGATCGCACAGGTGAAAAAGACAGCTGAAG
ACACAAAGATCGAATTGTCCACCGCAAAAGATGAAGTACAACGTGTGACGACGGATAAAAAGCAGGTTCAAG
CACATTGTGAGCAGCAGATTATGGAAATGACGGCCACACTGGAAAAGTACAAAGCCGAGAATCAGAAAATAG
TCAACCAGAAAGACAAGGAGATCGAGAAGATGAGGAAAGAGCACCAAACCTCCAAGAAACAGGATGTGCAA
GTAGCTGACTTGATGTCACAACTGCAAAGCGTGCAACAGCAGTTAGAAGACGTTAAGAAGGAGAAAGACGA
GATGCCAAAGACAAAGGAGTACCAACAGCAAGTCGACATGCTGAAGAAAACCATCGAAGATAAAGAAGGCC
AGTTGAAAGAACTTCGAACAGATCTTGAGAAAGCCAAAGCTGATGCTCTCTCTCCTACAACGCCAAAGACTTT
TTCAACGCCAAAGAATTATTCAACGCCGAAGACCAACGCTCCTCCATCCGCCTCGCTGCATCAGAGGCATGCC
GCGCGCAGAAACATTTCTCGCAAGGAGAATATGCCGAAAACAGATCCAATGGTTCCCCTGGCCGCGTCAACG
CCAATCCAGAACAAGACTCCACTACAACGGATCATCAAGCGCCCCGAGAGCGAACCGAAGAAACGCCGTGTC
GCGTTTGACATGACGGAGAAGACGGTGGAAATATCGATGGATGGTGGTGACTCGGAGGCGAACTCGTCCAC
ATCCGAACTCATGGAGTTGGATCCGGAAGACCTGCTGTCCGGAAAACCTCGTGGCGATCAACAAGTCCCCGG
GCAAGCATCACTCCAGGTTCACAAGTCTCCCTCCCATGGAATTCTCAAGTCGCCCGCCTTTGTCCGCAAGTCAC
CCGCTGCAAGGTTCGGAGCCCAAGCTCGAGCGTCTCCTGCAACGAAGACGCCAACACCACGTGGAAGTAAAT
CCTACAAAGTCCACATCTCCGACTCCTGGGGTAAGAAGAACAAGAACGTGAAGCAAGGGCCGAACAAGACAC
CCAAGTCTAAGGAGCAAGTAAAGAAAAGAAACAAGTCGTCCGAGAGAGGAGAGGAGCTGTCCTGGTTCGAG
TCTGATACTGTTTTTGGCTTCTTCGAGTAA

>*B.lanceolatum_SCP1*_sequenced_peptide

QVYHQQEPFFKALSPQQQEHQHSFFKIGTEQQEVEIKTLSPMRLGQQMHSGERLTSLHSRLQKEAEKINKWKHQT
EMQIQQRERKIQDTQQTIDSQRKSILELQLQNENLSSKLQEEIDGRVEIMKKITATRDMCYLLKDHAANVEERMGK
CEANRDELQCLQQDTVFQLQELTSKFNNLRINHTEAEKVLGNKLKESVSELNQVKCDYQNEKVNVEKRLEGLMQQ
CSEKEMEISKLTGALTDKQAQLSDLEQQCSMLEEHVAKLEDEFKSLQDQLQEASDKIFSRDKEMEKISGELTNTEAQ
LQKVSDDCEQLESHIGLLKEQHSIEVTKINHQLDTVEEKLNVEKTKTKDISSKHLSAEQRIGELQASINEQEKKYETLIE
EKSKVEGEKAAVEAEGVLLKENIMALETEQKSMKDQVSQLRTTVAVLTDAKIHAEEQLSMLEKEKKVTGDEVKVLT
NTSAARDKEIKKLQDDLKKAAERQKEAKKQEKELKQQLKTKDAEVDKTSAKLKKVQGDLEEIQTLMASLQRDHEEL
KDKLNAADLQRSTLQSSLDEVNQEKVSLNDTLKQLEESLKAQDKEAVGKIHQQQESTKALQTELDSNKKSMTKLEN
RVKSLEKQVAEKTNKIKDLQQDNKTVKKELGIHLKLSNGFEEKAKSLEEEIAQVKKTAEDTKIELSTAKDEVQRVTTDK
KQVQAHCEQQIMEMTATLEKYKAENQKIVNQKDKEIEKMRKEHQTSKKQDVQVADLMSQLQSVQQQLEDVKKE
KDEMPKTKEYQQQVDMLKKTIEDKEGQLKELRTDLEKAKADALSPTTPKTFSTPKNYSTPKTNAPPSASLHQRHAA
RRNISRKENMPKTDPMVPLAASTPIQNKTPLQRIIKRPESEPKKRRVAFDMTEKTVEISMDGGDSEANSSTSELMEL
DPEDLLSGKPRGDQQVPGQASLQVHKSPSHGILKSPAFVRKSPAARFGAQARASPATKTPTPRGSKSYKVHISDSW
GKKNKNVKQGPNKTPKSKEQVKKRNKSSERGEELSWFESDTVFGFFE*

7.5. Appendix 5. *AmphiSCP1* promoter predicted sequences

Table 7.2. Sequences of promoters predicted between *AmphiCHIC* and *AmphiSCP1*

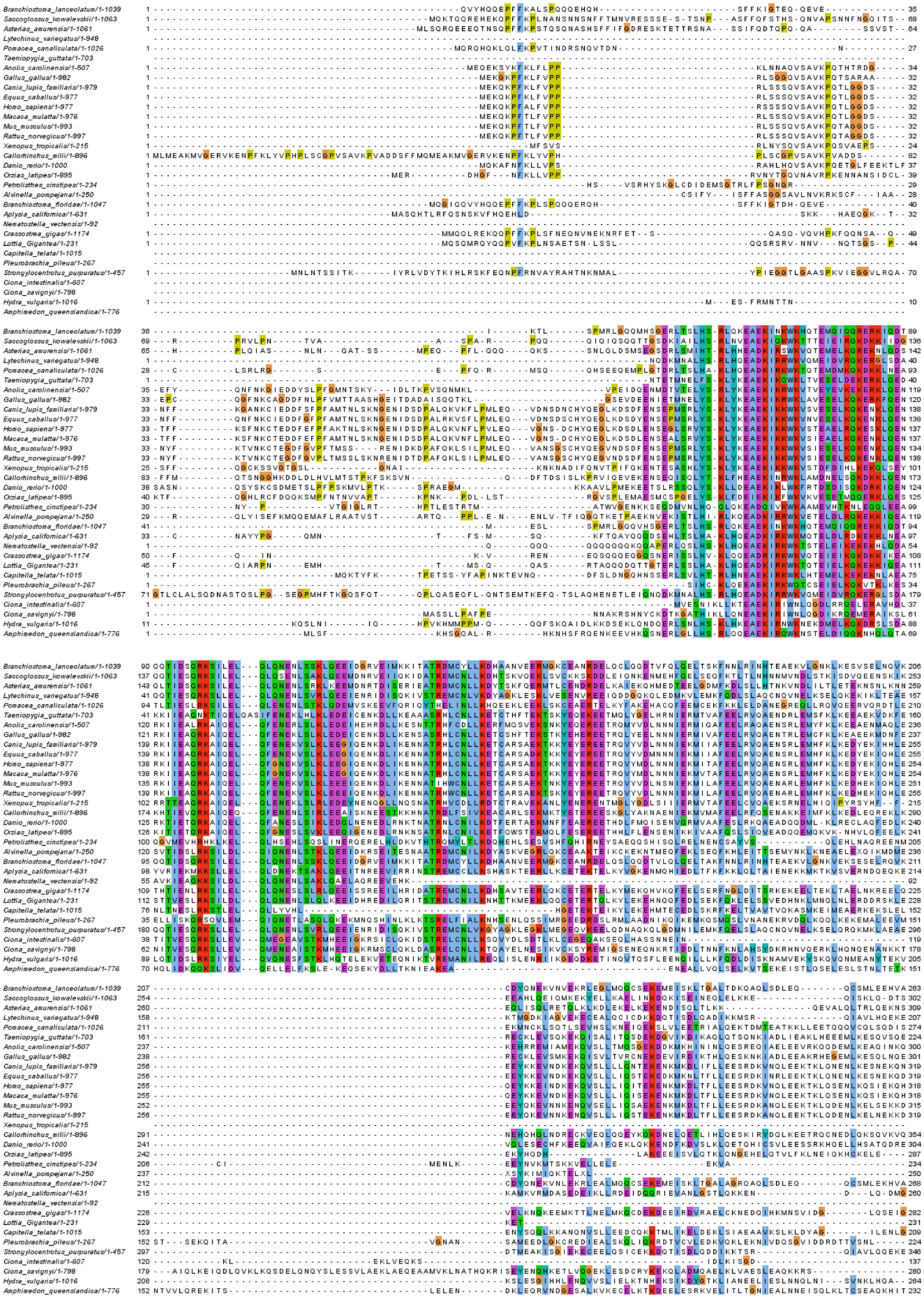
Identifier	Sequence (Bold large font indicates predicted TSS)
NNPP 1	TAGTACAGTGTATAAGAGTCTATTATAACGAAAAAAAATC**T**GACTTTCTG
NNPP 2	TACAAGTAGGAAAAAAAGCTGCAGAGAAGTCAGAAATCAG**A**AAGTCAGAT
NNPP 3	TCCCTTTTTCCAAAAATGGCGCCTCCAAGTCAGAATGTCG**T**CTGCAAGTG
NNPP 4	TATGTGTAGATATATGAATTGCGATCACGCCTAGCGTAGC**A**AAGTATTGT
NNPP 5	ATCTATTGTTCATAAAAGGAAACGAACCCCCTTGACACTC**A**CCAAAGGCA
TSSW Promoter TSS	TCACTTGCAGACGACATTCTGACTT**G**GAGGCGCCATTTTTGGAAAAAGGG
ProScan 1	n/a
Proscan 2	n/a

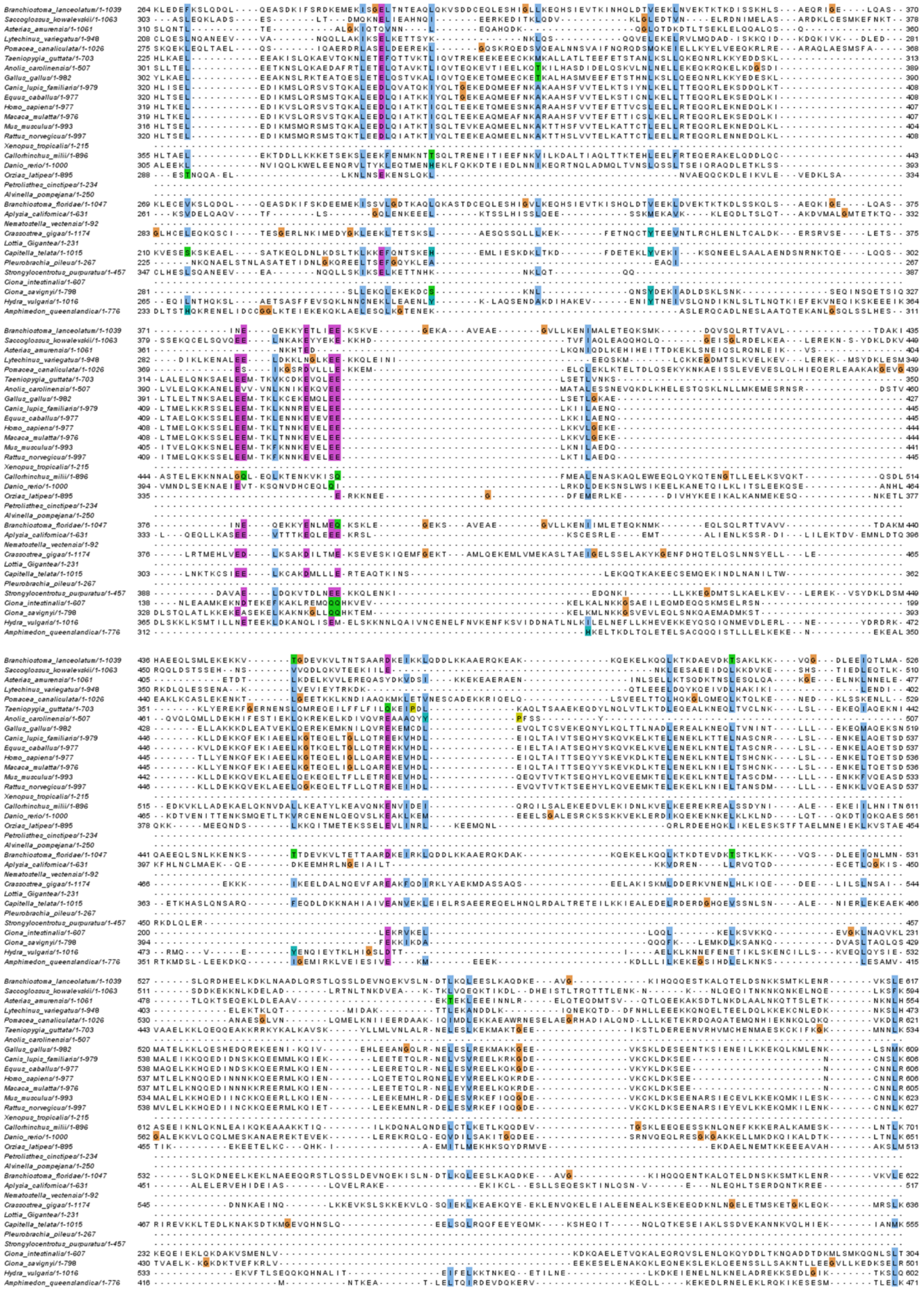
7.6. Appendix 6. Metazoan SCP1 protein alignment

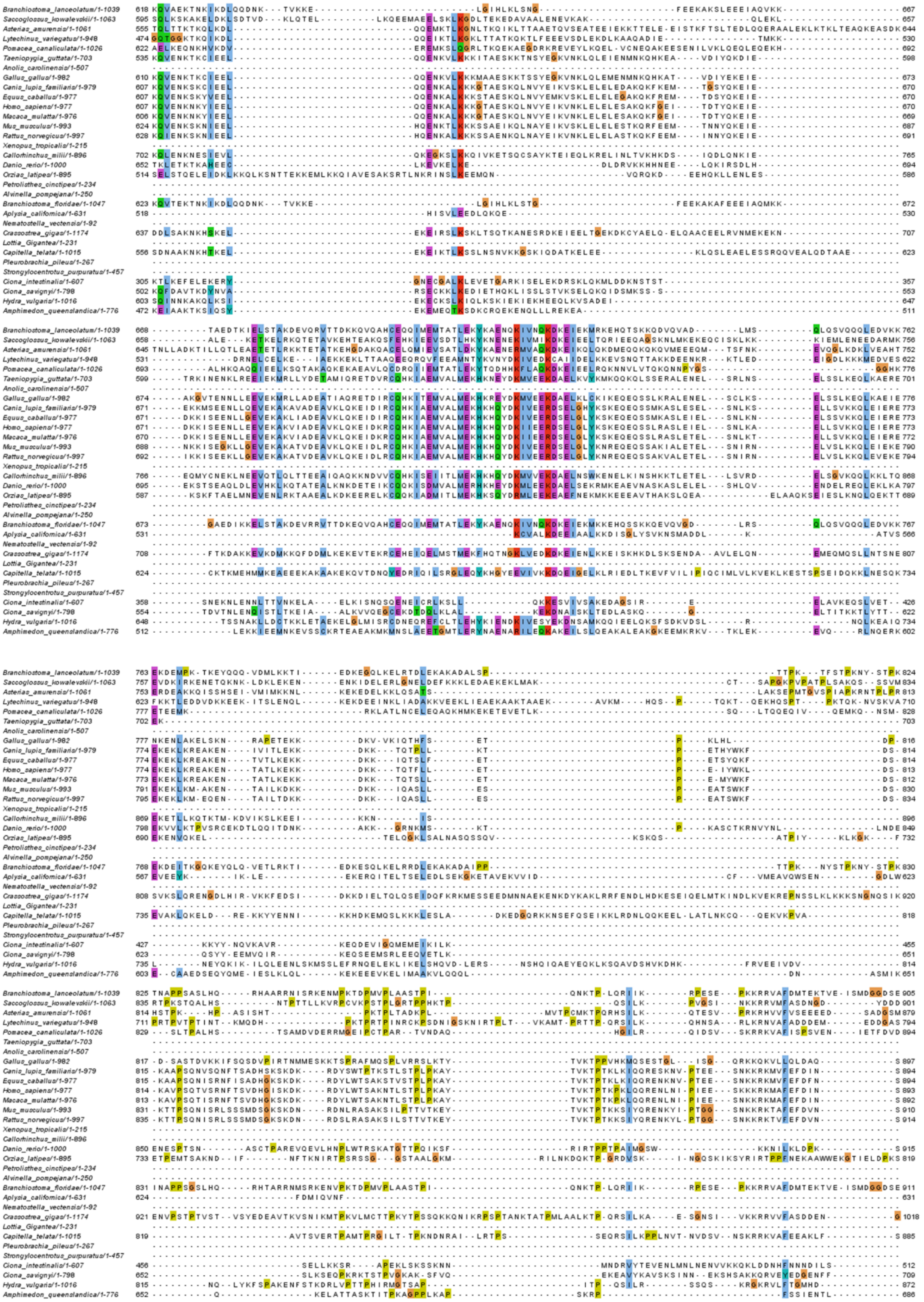
Table 7.3. Metazoan SCP1 sequences used for protein alignment

Species	Group	Accession number
Alvinella pompejana (bristleworm)	Lophotrochozoa-Polychaeta	GO222799.1
Amphimedon queenslandica (demosponge)	Porifera	ACUQ01003074.1
Anolis carolinensis (lizard)	Tetrapoda	XP_008108235.1
Aplysia californica (Sea hare)	Lophotrochozoa--Mollusca	XP_005108708.1

Asterias amurensis (northern pacific sea star)	Echinodermata-Asteroidia	GAVL01043045.1
Branchiostoma floridae (amphioxus)	Chordata-Cephalochordata	B.floridae ParaHox Reassembly
Branchiostoma lanceolatum (amphioxus)	Chordata-Cephalochordata	Unpublished cDNA
Callorhinchus milii (elephant shark)	Chordata-Chimaera	XP_007899851.1
Canis lupis familiaris (dog)	Chordata-Tetrapoda	XP_857086.1
Capitella telata (polychaete worm)	Lophotrochozoa--Polychaeta	http://genome.jgi.doe.gov/ scaffold_46000038
Ciona intestinalis (Sea Squirt)	Chordata-Tunicata	http://www.aniseed.cnrs.fr/ KH2012:KH.C8.516.v1.A.ND1-1
Ciona savignyi (Sea Squirt)	Chordata-Tunicata	AACT01000684.1
Crassostrea gigas (pacific oyster)	Lophotrochozoa--Mollusca	XP_011438578.1
Danio rerio (zebrafish)	Chordata-Actinopterigii	NP_001112366.1
Equus caballus (horse)	Chordata-Tetrapoda	XP_001496166.2
Gallus gallus (chicken)	Chordata-Tetrapoda	XP_004935063.1
Homo sapiens (human)	Chordata-Tetrapoda	EAW56621.1
Hydra vulgaris (freshwater hydroid)	Cnidaria-Hydrozoa	JQ906934.1
Lottia Gigantea (owl limpet)	Lophotrochozoa-Mollusca	FC693207.1
Lytechinus variegatus (green sea urchin)	Echinodermata-Echinoidea	GAUR01060760.1
Macaca mulatta (rhesus macaque)	Chordata-Tetrapoda	XP_001111808.1
Mus musculus (mouse)	Chordata-Tetrapoda	CAA86262.1
Nematostella vectensis (starlet sea anemone)	Cnidaria-Anthozoa	FC319267
Orzias latipes (medaka)	Chordata-Actinopterigii	JQ906936.1
Petrolisthes cinctipes (crab)	Ecdysozoa-Crustacea	FE795932.1
Pleurobrachia pileus (sea gooseberry)	Ctenophora	FP999277
Pomacea canaliculata (channelled applesnail)	Lophotrochozoa-Mollusca	GBZZ01052649.1
Rattus norvegicus (rat)	Chordata-Tetrapoda	NM_012810.1
Saccoglossus kowalevskii (acorn worm)	hemichordata	XM_006821456.1
Strongylocentrotus purpuratus (purple sea urchin)	Echinodermata-Echinoidea	XM_011679322.1
Taeniopygia guttata (zebra finch)	Chordata-Tetrapoda	UNIPROT: H0ZY72
Xenopus tropicalis (western clawed frog)	Chordata-Tetrapoda	XP_012811338.1
CCDC39-Drosophila Melanogaster	Ecdysozoa-Hexapoda	ACD81657.1
CCDC39-Homo Sapiens	Chordata-Tetrapoda	NP_852091.1
CCDC39-Strongylocentrotus purpuratus	Echinodermata-Echinoidea	XP_781717.3







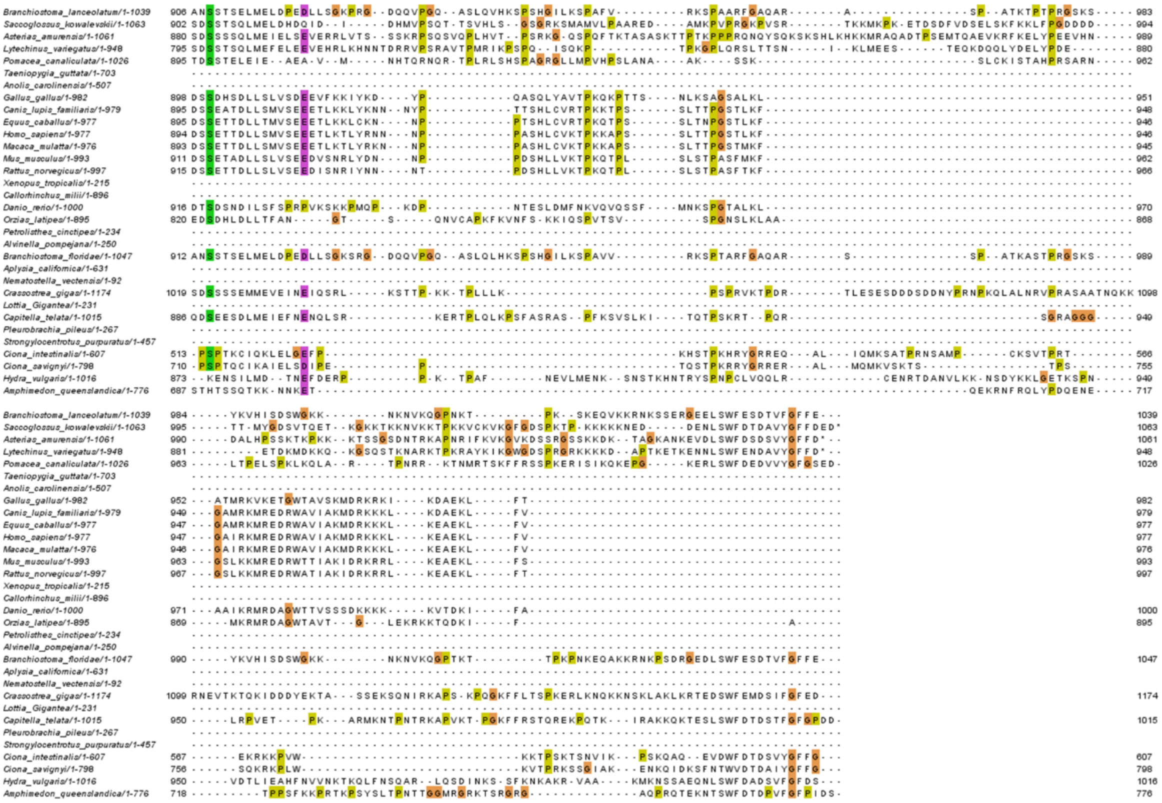


Figure 7.3. Alignment of metazoan SCP1 proteins

Alignment of metazoan SCP1 proteins with Clustal Omega (Sievers et al., 2011), using Jalview (Waterhouse et al., 2009). Colours indicate conservation of residues, whilst ellipses represent gaps where residues could not be aligned. ClustalW colour scheme for amino acid groups.

7.7. Appendix 7. Expression of the empty pCES reporter

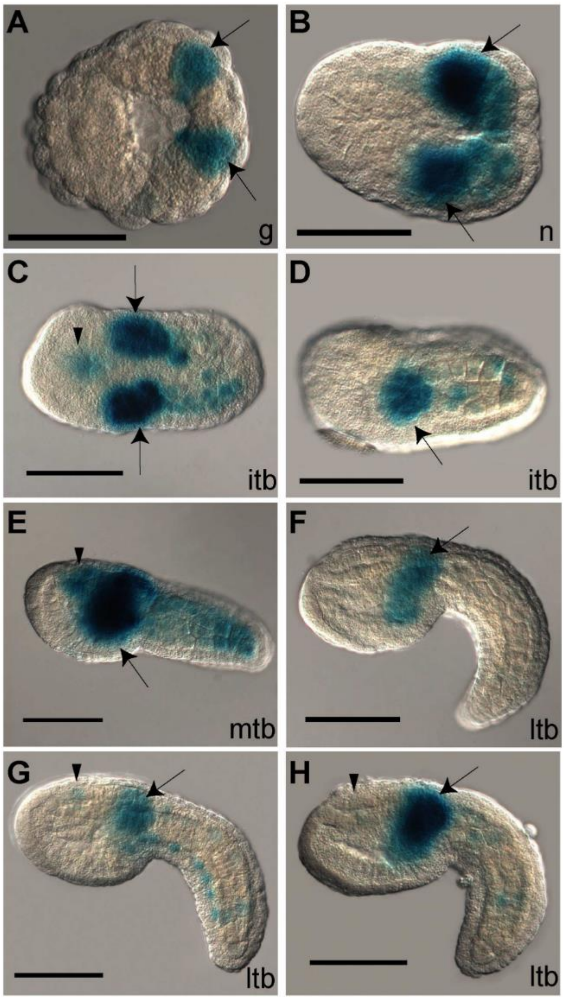


Figure 7.4. Ectopic expression of pCES vector in *Ciona intestinalis*.

Expression is almost always seen in mesenchymal tissue (A-H) (arrows). Expression is activated during gastrulation stages in mesenchymal cell lineages (A). At later stages, expression is visible posterior to the mesenchyme in variable numbers of tail muscle cells (C-H). Expression is also, though very rarely, observed in the centre of the sensory vesicle (black arrowheads) of tailbud stage embryos (A-C) show dorsal views, whilst (D-H) show lateral views. Lower case lettering refers to the stage of development; g, gastrula; n, neurula; itb, initial tailbud; mtb, mid tailbud; ltb, late tailbud. Scale bars represent 100 µm. (taken from Osborne 2009 Unpublished data) figure on next page. pCES vector from (Harafuji et al., 2002).

7.8. Appendix 8. Bf-Gsx-Up1c-repeat and Up1c-repeatreverse

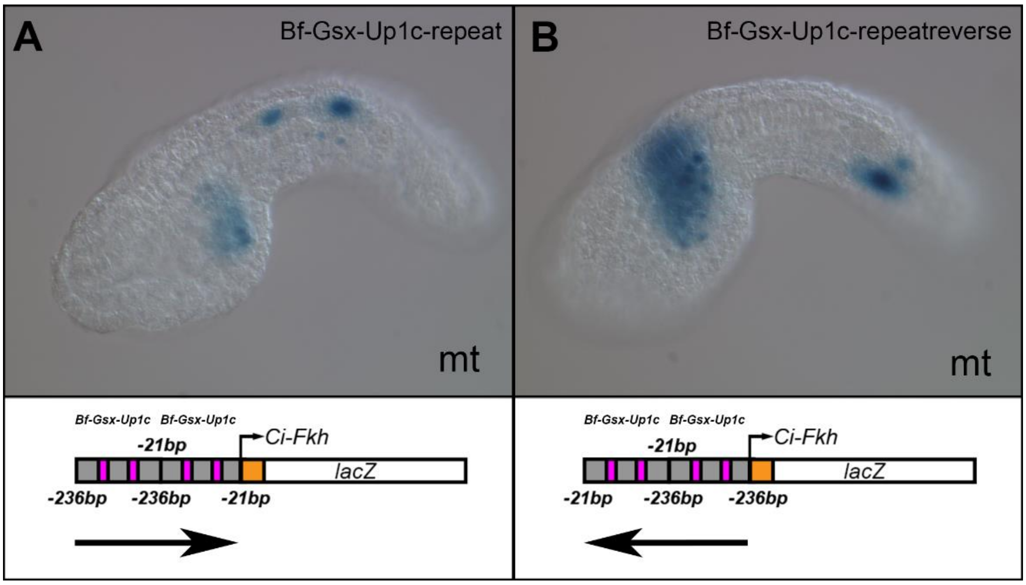


Figure 7.5. Typical Bf-Gsx-Up1c-repeat and Bf-Gsx-Up1c-repeatreverse transgenic Ciona embryos.

(A) Bf-Gsx-Up1c-repeat transgenic embryos display no nerve cord expression but high pCES background expression. Schematic shows orientation (black arrow) and structure of the Bf-Gsx-Up1c-repeat regulatory construct. (B) Bf-Gsx-Up1c-repeatreverse transgenic embryos display no nerve cord expression but high pCES background expression. Schematic shows orientation (black arrow) and structure of the Bf-Gsx-Up1c-repeatreverse regulatory construct. Pink boxes represent TCF/Lef binding sites. Orange represent the Ci-Forkhead promotor. Black arrows indicate orientation of the regulatory element in respect to the *Ci-Fkh* promoter. Scale bar represents 100µm.

7.8.1. Bf-Gsx-Up1c-repeat and Up1c-repeatreverse do not show cumulative TCF/Lef function.

As the Bf-Gsx-Up1 (figure 5.5) and Bf-Gsx-Up1+2b (figure 5.7) constructs indicated that TCF/Lef binding sites may be functioning collaboratively, perhaps cumulatively, it was hypothesised that a construct with many TCF/Lef binding sites, present within a functional context, may address if these binding sites do display cumulative function. The aim was to test whether increasing the number of TCF/Lef sites present would increase CNS LacZ expression efficiency. In order to achieve this, the Up1c 'minimal enhancer' region, which is known to produce nerve cord staining, was taken as a base to produce a repeated construct with two Up1c regions adjoining each other in front of the forkhead promoter. This could potentially double the amount of TCF/Lef sites present and increase LacZ expression efficiency above the 4.8% seen for Bf-Gsx-Up1c (figure 5.2). In addition, recent work had identified enhancers that only produce expression when placed in a specific orientation with a target gene (Hozumi et al., 2013), and so it was decided to test whether the Bf-Gsx-Up1c region may be orientation-sensitive. To this end, a reverse orientation construct of the Bf-Gsx-Up1c-repeat was created, or Bf-Gsx-Up1c-repeatreverse, as new restriction sites had already been introduced that would enable cloning of Up1c-repeat, and would also enable the cloning of Up1c-repeat-reverse. Two sets of primers were used to clone Up1c so that one copy had 5'-PstI, 3'BamHI sites and the other copy 5'BamHI, 3'PstI sites (Table 7.4). This enabled digestion and recovery of the resulting Up1c clones using the appropriate restriction enzymes and re-ligation, using the overhangs from both constructs, to produce both Up1c-repeat and Up1c reverse. Electroporation and analysis of these two constructs was not carried out as thoroughly as with previous constructs due to time constraints, but preliminary data is presented here.

Upon examining these two constructs, it was clear that expression was not as expected. Bf-Gsx-Up1c repeat transgenic embryos showed typical pCES background expression (figure 7.5 A), with no embryos showing CNS expression. The second construct, Bf-Gsx-Up1c-repeatreverse displayed similar expression patterns. Again, none of the embryos examined showed any CNS expression, and displayed only pCES background expression (figure 7.5 B). As stated, it was not possible to examine these two constructs with the same detail as previous constructs, and so numbers of embryos showing expression have not been characterised for Bf-Gsx-Up1c-repeat and Bf-Gsx-Up1c-repeatreverse as of yet. As such, further work needs to be carried out to properly characterise these two constructs.

Table 7.4 Primers and their modifications used for cloning of Up1c-repeat constructs.

Identifier	Sequence	Annealing Temp (°C)	Primer Modification
B.fl Gsx-up1c F	CTGCAGAAAGGGCCTCTATTGCTTTC	56	5' PstI
B.fl Gsx-up1c R	GGATCCAGCCCTTGCCAATGAAAAA		3' BamHI
Gsx-up 1c BamH1F	GGATCCAAAGGGCCTCTATTGCTTTC	56	5' BamHI
Gsx-up 1c Pst1R	CTGCAGAGCCCTTGCCAATGAAAAA		3' PstI

7.8.2. The Bf-Gsx-Up1c-repeat and Up1c-repeatreverse constructs highlight the complexity of regulatory interactions.

The Bf-Gsx-Up1c-repeat construct was designed to test if TCF/Lef sites would act in a simple additive fashion to increase expression efficiency within the context of this regulatory region. The Gsx-Up1c minimal enhancer region was chosen as the best region to observe increased expression, as it was known to show nerve cord expression, but any increases in expression efficiency would be more easily noted due to the low level of nerve cord expression seen in the wild-type Bf-Gsx-Up1c construct. In addition it was deemed best to avoid introducing sequence within this region, as this could potentially disrupt sequence and/or transcription factor binding sites that may also be important to the function of this regulatory element. As such, doubling up the regulatory region was chosen as the best solution to these problems, as it allowed the addition of known functional TCF/Lef sites present in a known functional context. The results, however, show that this construct ceases to produce nerve cord expression and no longer functions as a driver of CNS expression (Figure 7.5 A).

One explanation for the loss of nerve cord expression, as opposed to the gain of expression, could be that this approach is also increasing the amount of inhibitory sequence/repressive transcription binding sites present within the construct, or creating new repressive combinations across the join of the two Up1c elements. The second of these is perhaps more likely, as without an additional effect, such as new combinations of repressive factors being introduced, the balance of TCF/Lef activation to repression should have remain the same across single or doubled Up1c elements. Thus, repressive factors must be acting additively, or synergistically, with a greater degree of increased effect relative to the sum of activating TCF/Lef sites within the Bf-Gsx-Up1c-repeat constructs. The Bf-Gsx-Up1c-repeatreverse construct was designed as a way to test the sensitivity of the Up1c minimal enhancer to orientation, though this was based upon the expectation that Bf-Gsx-Up1c-repeat would produce stronger Up1c-like expression. However, results were similar to that of Bf-Gsx-Up1c-repeat (Figure 7.5 A,B) and this hypothesis could not be tested.

Further work might seek to address the questions posed when creating the Bf-Gsx-Up1c-repeat and Up1c-repeatreverse constructs. In order to examine whether TCF/Lef binding alone is sufficient to drive CNS expression, it could be possible to create a construct with a TCF/Lef binding site multimer (multiple successive, adjacent TCF/Lef binding sites) coupled to the Fkh-promoter in pCES. This method is established and used in other constructs as a reporter of Wnt/TCF (reviewed in Barolo (2006)). Alternatively, such a multimer could be used in conjunction with either Gsx-Up1c, Gsx-Up1 or Gsx-Up1+2b to examine if efficiency of CNS expression was increased in the presence of increased TCF/Lef binding. To address concerns with the addition of sites within the regulatory sequence, separate constructs containing TCF/Lef binding site multimers both adjacent to, and constructs containing multimers within, the regulatory sequence could be created and compared. With regards to examining the sensitivity of the Bf-Gsx-Up regulatory region to orientation, and examine if function was affected, a less complex solution would be utilised and constructs created that contain reverse copies of Gsx-Up1c, Gsx-Up1 and Gsx-Up1+2b and compare the function of these to each other and their forward orientation counterparts. The PiCh approaches referred to earlier could also be used to help understand these repressive functions and whether they are 'out-stripping' the TCF/Lef activation within Up1c multimers, by characterising the proteins involved in this repressive function.

Publications arising from this work

Garstang, M.G., Osborne, P.W. and Ferrier, D.E.K. Amphioxus SCP1: a case of retrogene replacement and co-option of regulatory elements adjacent to the ParaHox cluster (paper in prep).

Garstang, M.G., Osborne, P.W. and Ferrier, D.E.K. TCF/Lef regulates the Gsx ParaHox gene in central nervous system development in chordates. (*BMC Evol Biol*, In press).

Garstang, M. and Ferrier, D.E.K. Time is of the essence for ParaHox homeobox gene clustering. *BMC Biology* (2013) 11: 72.