The role of highly conserved non-coding DNA sequences in vertebrate development and evolution

Parker, Hugo

The role of highly conserved non-coding DNA sequences in vertebrate development and evolution

Hugo Parker

School of Biological and Chemical Sciences
Queen Mary, University of London

September 2010

A dissertation submitted for the degree of
Doctor of Philosophy
at the University of London

Declaration

This dissertation is submitted for the degree of Doctor of Philosophy at the University of London.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated.

No part of this dissertation is being submitted for any other qualification or at any other university.

Hugo Parker
September 2010

Acknowledgements

Thanks to my supervisor Greg Elgar for support and guidance. Thanks to members of the Elgar and Bronner-Fraser labs for friendship and advice. Thanks to Paul for the bio-informatics help and to Stefan, Debs, Emma and Heather for molecular biology and embryology advice. Thanks to Heather for maintaining the fishes. Thanks to Tatjana, Marianne, Natalya, Marcos, Benji, Jeremiah and Melinda for lamprey expertise. Thanks to QMUL SBCS for funding this thesis. Thanks to my friends. Special thanks to my parents for love and support.

Abbreviations

2R	two rounds of whole genome duplication
A-P	anterior-posterior
BBR	Boehringers blocking reagent
bHLH	basic helix-loop-helix
bp	base pairs
CNE	conserved non-coding element
CNS	central nervous system
CRM	*cis*-regulatory module
dCNE	duplicated conserved non-coding element
DDW	double distilled water
DEPC	diethylpyrocharbonate
EB	Enhancer Browser
EM	embryo medium
EM-PTU	embryo medium with PTU
F0	parental generation
F1	first filial generation
GFP	green fluorescent protein
GRN	gene regulatory network
HD	homeo-domain
hpf	hours post fertilisation
kb	kilo base pairs
MABT	maleic acid buffer with tween
MMR	Marc's modified Ringer's solution
MYA	million years ago
PBS	phosphate buffered saline
PFM	position frequency matrix
PG	paralogy group (*hox* genes)
PTU	phenolthiourea
PTW	phosphate buffered saline with tween
PWM	position weight matrix
r	rhombomere
RA	retinoic acid
RFP	red fluorescent protein

RT	room temperature
TF	transcription factor
TFBS	transcription factor binding site
UCE	ultra-conserved element
WGS	whole genome shotgun

Summary

Comparisons between vertebrate genome sequences, from mammals to fishes, have revealed thousands of conserved non-coding elements (CNEs) that are associated with developmental genes. Interestingly, the vast majority of these CNEs cannot be found in invertebrate genomes by sequence homology. As many CNEs have been demonstrated to act as enhancers *in-vivo*, it has been postulated that CNEs represent gene regulatory elements with crucial roles in aspects of development that are shared between vertebrates.

To trace the evolution of CNE sequences in vertebrates, a preliminary search for CNEs in the lamprey genome was conducted using the draft lamprey genome sequence. This thesis documents how the CNEs identified in lamprey have been used as a guide to ask questions about the function and evolution of CNEs in the vertebrate lineage. Through the combined use of comparative genomics and developmental biology techniques, including a newly developed reporter assay for sea lamprey embryos, crucial first steps have been taken toward systematically de-coding these ancient gene regulatory elements. Special attention is paid toward utilising the low sequence identity of lamprey CNEs for 'phylogenetic footprinting', an approach which uncovers striking enrichment of CNEs for a set of motifs that are characteristic of Hox-regulated elements. These findings help to establish CNEs within a developmental and evolutionary context.

Contents

List of tables

List of Figures

List of figures continued

Chapter contributions

Specific contributions of collaborators are described in detail in the text within each chapter where necessary.

Chapter 3

The research described in this chapter was published in:

McEwen GK, Goode DK, Parker HJ, Woolfe A, Callaway H, Elgar G. (2009). Early evolution of conserved regulatory sequences associated with development in vertebrates. PLoS Genet. 5: e1000762.

McEwen GK, Woolfe A, Elgar G and Piccinelli P contributed to the bio-informatic analyses and Goode DK to the functional analyses. Goode DK contributed to the figures.

Chapter 5

Piccinelli P and Elgar G contributed to the bio-informatic analyses.

Chapter 7

Piccinelli P contributed to the bio-informatic analyses.

1 Introduction

Evolutionary developmental biology

Animal taxa with diverse morphologies have arisen from a common ancestor by the process of evolution. Any explanation of how morphological evolution has occurred (and continues to occur) must address the changes in development that underlie heritable morphological changes. This is the aim of evolutionary developmental biology (Evo-Devo). In order to elucidate the developmental changes that led to evolutionary transitions, the mechanism of development must be understood at a molecular level. This mechanism must explain how a single cell can give rise to a multi-cellular organism composed of heterogeneous tissues. It is clear that simple asymmetries of the expression of early developmental genes can give rise to complex expression patterns of genes later in development, directing different tissues to be produced in different positions. Early developmental genes achieve this by regulating, via the proteins they encode, the transcription of other developmental genes, which are themselves regulators of other genes (Carroll, 2000; Levine & Tjian, 2003). This regulation ultimately depends upon transcription factors (TFs), which are proteins that bind to DNA in a sequence-specific manner at transcription factor binding sites (TFBSs) in the genomic vicinity of their target gene. In this model, TFs associated with binding sites in the genes promoter region directly influence the formation of the transcription initiation complex. Transcription of the gene is also strongly influenced by TFs binding to more distal regulatory elements, which also regulate the formation of the transcription initiation complex, either enhancing its formation, repressing it, or insulating the promoter from enhancers associated with other nearby genes (Kadonaga, 2004) (Figure 1.1). The proximity of distal regulatory elements to their genes promoter can vary widely between different elements, with some having a range of influence of several megabases (Vavouri *et al*., 2006).

The combinatorial regulation of a gene by many TFs makes that gene tightly regulated, only being expressed in areas of the developing organism where the right combination of TFs is present, enabling that gene to have a complex expression pattern. Clusters of TFBSs have been termed '*cis*-regulatory modules' (CRMs) (Howard & Davidson, 2004) (Figure 1.1), and are a feature of all genes. Indeed, a gene can be regulated through the action of multiple CRMs, each able to act independently to drive expression

of their target gene in complementary domains of the developing embryo (Arnone & Davidson, 1997).

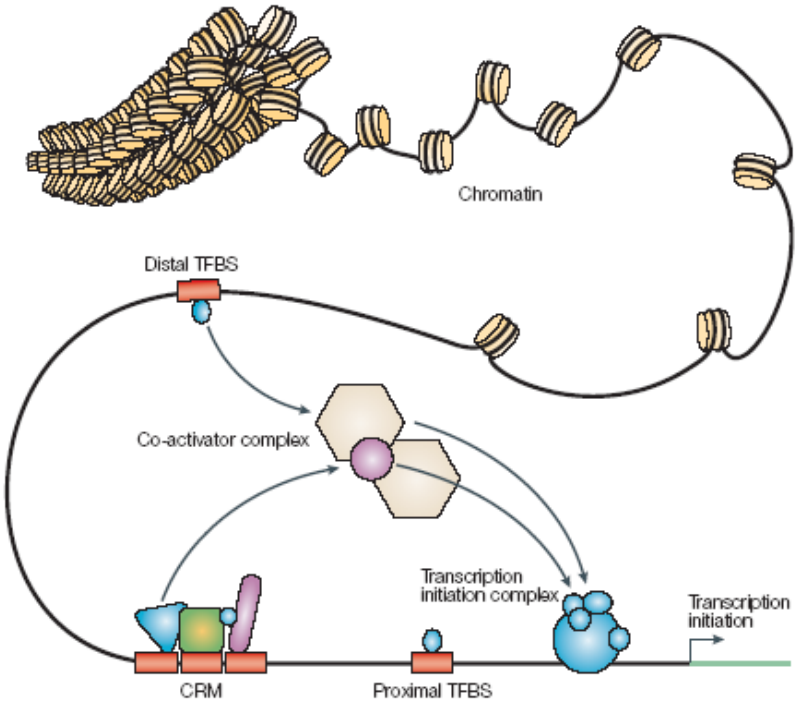


Figure 1.1. A model of transcriptional regulation by transcription factors. Transcription factors, represented by blue, green or purple shapes associating with TFBSs regulate the formation of the transcription initiation complex (blue circles). CRM: *cis* -regulatory module. Figure adapted from Wasserman & Sandelin (2004).

Complex sets of expression cascades resulting from genes regulating the expression of other genes can be represented in the form of hierarchical gene regulatory networks (GRNs) (Levine & Davidson, 2005), several of which have been well characterised for particular developmental pathways in invertebrates through large scale gene perturbation analyses (e.g. Davidson *et al.*, 2002; Shi *et al.*, 2005). An increasing number of GRNs are now being characterised in vertebrates (e.g. Sauka-Spengler & Bronner-Fraser, 2008; Alexander *et al.*, 2009; Morley *et al.*, 2009). The characterisation of GRNs is a powerful approach for investigating developmental processes (Davidson, 2006). Further, GRNs can provide insights into evolutionary mechanisms, as the evolution of body plans can be viewed as being the result of inherited changes in the architecture of GRNs (Davidson & Erwin, 2006; Hinman & Davidson, 2007; Erwin & Davidson, 2009).

It has been posited that the GRNs of early metazoa were relatively simple and plastic, with changes in these networks giving rise to different animal lineages. The diverging

core networks then became impervious to major changes as they increased in complexity, resulting in taxa of animals that have distinct body plans between groups but conserved body plans within groups (e.g. phyla) (Davidson, 2006). Thus, the GRNs of animals from different phyla are expected to be similar in architectural principles, but different in terms of the specific components making up that architecture, whereas the GRNs of different species from the same phyla are expected to be mostly identical in their core regions, with changes predominantly at the external network levels (Davidson & Erwin, 2006). In order for us to construct GRNs, the developmental genes in the network, and their interactions with each other, must be characterised. The identification and characterisation of CRMs that regulate these genes is a crucial aspect of inferring GRNs, as these modules represent direct regulatory links between the components of the network.

Great advances have been made in annotating gene sequences within genomes; unfortunately the identification of regulatory elements is less easy. The importance of identifying these regulatory elements is highlighted by the prediction that the human genome contains 25-30,000 genes, which is only modestly more than the genome of the morphologically 'simple' nematode *C. elegans* (~19,000) (Hahn & Wray, 2002). Not only do metazoan genomes contain similar numbers of genes, they also share many gene families. There are some cases in which pairs of distant orthologous genes have been demonstrated to be functionally equivalent when substituted between hosts, suggesting that the biochemical properties of their proteins and their interactions with other factors have changed little between distantly related species (e.g. Hox factors: Malicki *et al*., 1990; McGinnis *et al*., 1990; Pax6: Halder *et al*., 1995). This has led to the hypothesis that changes in *regulatory* complexity underlie the evolution of more complex body plans (Carroll, 2008). However, it is not clear to what extent inter-clade functional equivalence is a common feature of ancient orthologous transcription factors. Furthermore, clade-specific expansions of transcription factor repertoires have occurred during metazoan divergence, such as KRAB-associated zinc finger genes in tetrapods (Huntley *et al*., 2006), so the developmental gene 'toolkits' of animal clades are similar but not identical. There is empirical evidence for the contribution of mutations both in genes and *cis*-regulatory elements to morphological evolution. Interestingly, the examples of mutations in genes predominantly concern those with single roles in development, such as those influencing colouration (e.g. Protas et al., 2006). Whilst the list of examples is not extensive enough to be conclusive, this is in line with the

prediction that mutations giving rise to negative pleiotropic effects, such as those in genes with multiple developmental roles, would reduce organismal fitness and thus be unable to contribute to morphological evolution. Conversely, mutations in *cis*-regulatory elements and in genes with single roles in development would be less likely to have negative pleiotropic effects and could therefore provide variation that could be positively selected. Taken together, there is a strong theoretical and empirical basis underlying the notion that *cis*-regulatory changes have played an important role in morphological evolution, whilst the relative significance of genic versus regulatory changes is still a matter for debate. Thus, identifying the regulatory elements of developmental genes in vertebrates is a crucial step toward elucidating the genetic changes underlying the evolution of the vertebrate body plan and characterising vertebrate GRNs.

Identifying *cis*-regulatory elements through genomics approaches

Traditionally, *cis*-regulatory elements have been identified by deletion analysis of genomic sequences near the gene of interest, followed by testing fragments by reporter assay (Pennacchio & Rubin, 2001). A wide range of computational tools for TFBS and CRM prediction have been developed for searching genomic sequences (e.g. Berman *et al.*, 2002; Ho Sui *et al.*, 2007), yet the success of these methods is limited by the fact that many binding sites are short sequences (5-10bp) so performing searches on large vertebrate genomes is likely to falsely identify many binding sites by chance (Wasserman & Sandelin, 2004). Furthermore, these approaches often require prior knowledge of the TFBSs to be searched for, restricting the discovery of CRMs composed of novel TFBSs. The availability of genomic sequences of many different species enables searching for homologous regions between species that may harbour regulatory elements; a technique termed 'phylogenetic footprinting' (Wasserman *et al.*, 2000). The assumption is that essential CRMs will be conserved by negative selection, so will be identifiable as non-coding sequences conserved between divergent species.

Alignments of whole vertebrate genomes have revealed numerous highly conserved non-coding regions of considerable length (>100 bp). For example, whole-genome human-mouse alignments identified more than 300,000 conserved non-coding elements of 70% identity over at least 100 bp, which are uniformly distributed throughout the genome (Dermitzakis *et al.*, 2003). Many conserved elements show evidence of

sequence constraint through purifying selection, rather than a low mutation rate, which suggests that they have functional roles (Drake *et al.*, 2005; Lunter *et al*., 2006). Putative *cis*-regulatory elements can be validated by *in-vivo* experiments using reporter constructs (e.g. Muller *et al*, 2002), but current methods of testing elements for regulatory function are relatively slow, making it unfeasible to test such a high number of sequences. In some cases, deletion of large regions containing many conserved non-coding elements had little phenotypic effect (Nobrega *et al*., 2004). Thus, it is unclear what the functions of many of these human-mouse conserved sequences are. Approaches to filter out smaller sets of sequences with high regulatory potential include searching for elements with more strict conservation parameters (e.g. Bejerano *et al.*, 2004) and searching for conservation between multiple and more divergent species (e.g. Woolfe *et al*. 2005, Pennacchio *et al*., 2006).

Conserved non-coding elements between mammal and teleost genomes

A whole-genome comparison between human and the Japanese pufferfish, *Fugu rubripes* (Fugu), identified nearly 1,400 highly conserved sequences of at least 100 bp in length that had little or no evidence of transcription (Woolfe *et al.*, 2005). The mean length of the conserved sequences was 199 bp with a mean identity of 84%; considerably higher than the mean level of coding-sequence identity between the two organisms. The sequences were called 'conserved non-coding elements' (CNEs), a term which has subsequently been used to refer to all non-coding sequences conserved between distant organisms, not just to this set of sequences. These human-Fugu CNEs have been retained in their host genomes more-or-less unchanged since the divergence of lobe-finned and ray-finned fish roughly 450 MYA. The majority of CNEs were also found to be conserved in other vertebrate genomes, namely rat, chicken and zebrafish, indicating that they are probably common to all bony vertebrates. More sensitive searches involving multiple species alignments have increased the number of identified CNEs conserved between human and Fugu to about 6,000 (Woolfe *et al*., 2007). Searching the invertebrate whole-genome sequences of a urochordate: *Ciona intestinalis,* fly: *Drosophila melanogaster,* and a nematode worm: *Caenorhabditis elegans,* for sequence identity with the vertebrate CNEs revealed no significant matches, suggesting the majority of these sequences to be conserved only within the vertebrate lineage. The distribution of CNEs was found to be highly clustered around genes involved in transcriptional regulation and development (termed 'trans-dev'

genes). This association of mammal-fish CNEs with trans-dev genes has been confirmed by a number of other studies (Sandelin *et al.*, 2004; Sironi *et al.*, 2005; Ovcharenko *et al.*, 2005).

Although most of the CNEs within the human genome appear to be unrelated to each other, 124 families of two to five duplicated CNEs (dCNEs) were identified, which were proposed to have arisen through at least one ancient whole-genome duplication early in vertebrate evolution (McEwen *et al.*, 2006). For the majority of these families, a set of paralogous genes could be assigned, removing much of the ambiguity involved in identifying the specific genes associated with CNEs. This also enabled confident measurements of the enhancer range of these dCNEs to be made, revealing that half of the dCNEs were situated more than 250 kb upstream of their target gene's promoter (Vavouri *et al.*, 2006).

The gene-regulatory function of CNEs

The gene-regulatory ability of an increasingly high number of CNEs has been confirmed through testing them for enhancer activity by reporter assay in zebrafish (Woolfe *et al.*, 2005; Kikuta *et al.*, 2007), frog (de la Calle-Mustienes *et al.*, 2005) and mouse embryos (Pennacchio *et al.*, 2006). CNEs often show reproducible enhancer activity, with spacio-temporal expression patterns generally reflecting the endogenous expression domains of their nearby trans-dev gene. The significance of CNEs for normal development is not always clear (Ahituv *et al.*, 2007), but crucial roles for many CNEs in development have been inferred through identification of genetic diseases arising from mutation or deletion of CNEs (Lettice *et al.*, 2003; Visel *et al.*, 2009; Ragvin *et al.*, 2010).

These data support the proposal that CNEs represent highly conserved CRMs. However, the conventional wisdom regarding TFs is that their interactions with binding sites show high levels of degeneracy (Sandelin & Wasserman, 2004), leading to the suggestion that CNEs are composed of multiple, tightly arranged TFBSs, with small sequence changes having deleterious effects on the binding of TF complexes to the CNE sequence (Elgar & Vavouri, 2008). It is likely that many vertebrate CRMs with conserved and important roles in development will not necessarily be found as CNEs, due to their operation through less sequence-restrictive mechanisms. Thus, CNEs may represent only a subset

of functionally conserved developmental *cis*-regulatory elements. A key issue is to what extent this set of elements operates through mechanisms that have been characterised for other, less well conserved CRMs.

The binding specificities of TFs can be inferred through *in-vitro* binding assays, which produce a set of similar sequences to which a given TF can strongly bind. These sequences can be aligned to calculate a position frequency matrix (PFM) (Stormo & Fields, 1998), which tabulates the frequency at which each nucleotide (A T G or C) is found for each position of the alignment. PFMs are converted into position weight matrices (PWMs) by weighting each base according to its average frequency in a background sequence set (Hertz & Stormo, 1999). These PWMs are usually depicted as a logo, in which each nucleotide at each position is represented as a letter with a size proportional to its weighted frequency. Publicly available databases of TF PWMs have been created (e.g. JASPAR (Bryne *et al*., 2008)), enabling genomic sequences to be scanned for the presence of putative TFBSs. However, only modest progress has been made in systematically identifying enriched TFBSs within CNEs, using either targeted or *de-novo* motif discovery approaches (Bailey *et al*., 2006; Pennacchio *et al.*, 2006; *Li et al*., 2010). Furthermore, although many CNEs have now been shown to exhibit enhancer activities in developing embryos, relatively few have been dissected to elucidate the specific sequence components responsible for their enhancer functions (e.g. Pöpperl *et al*., 1995; Tümpel *et al*., 2006). Thus, the regulatory language of CNEs remains somewhat of a mystery.

The evolution of CNEs

Whilst homologs of vertebrate CNEs are largely untraceable in the genomes of invertebrates, these invertebrate phyla each have their own characteristic sets of CNEs. Over 20,000 conserved non-coding sequences have been identified between closely related *Drosophila* species, with a proportion being traceable in the more distantly related mosquito genome, showing a bias in their genomic distribution towards the loci of developmental regulatory genes (Glazov *et al*, 2005). A comparison of the genomes of two nematodes, *Caenorhabditis elegans* and *Caenorhabditis briggsae*, which show a similar level of divergence to that between human and Fugu genomes, revealed worm-specific CNEs with similar properties to vertebrate CNEs (Vavouri *et al*., 2007). The smaller worm CNEs are found near trans/dev genes and share the same base

composition signals as vertebrate CNEs. Their regulatory function is supported by many of the worm CNEs containing previously identified transcriptional regulatory sites. Interestingly, many of the genes associated with CNEs in invertebrates are orthologs of vertebrate CNE-associated genes. Of 190 *C. elegans* genes with CNE-associated orthologs in humans, 60 are associated with worm CNEs in *C. elegans*, with 40 also having orthologs in *Drosophila* that are associated with conserved fly elements (Vavouri *et al*., 2007). This is consistent with the evolutionary model outlined above, in which developmental networks composed of CRMs associated with key developmental genes in the metazoan common ancestor were initially evolutionarily plastic, their divergent evolution and the subsequent fixation of different regulatory sequences in different lineages giving rise to groups of animals characterised by vastly different body plans.

To trace vertebrate CNEs deeper into the vertebrate phylogeny, the genome sequence of a cartilaginous fish, the elephant shark (*Callorhinchus milii*) was searched for non-coding sequence conservation with the human genome (Venkatesh et al, 2006). Cartilaginous fishes (*Chondrichthyes*) represent an extant group of jawed vertebrates that diverged from the common ancestor of the bony vertebrates about 530 MYA (Figure 1.2) (Kumar & Hedges, 1998). A total of 4782 human-shark CNEs were identified, almost all of them being vertebrate-specific sequences, and many having representatives in the human-Fugu CNE set. This suggests that a large cohort of CNEs evolved prior to the divergence of bony and cartilaginous fishes and were retained in both lineages for ~530 million years. In order to ascertain when these sequences first arose, the genome of a more distantly related vertebrate must be investigated.

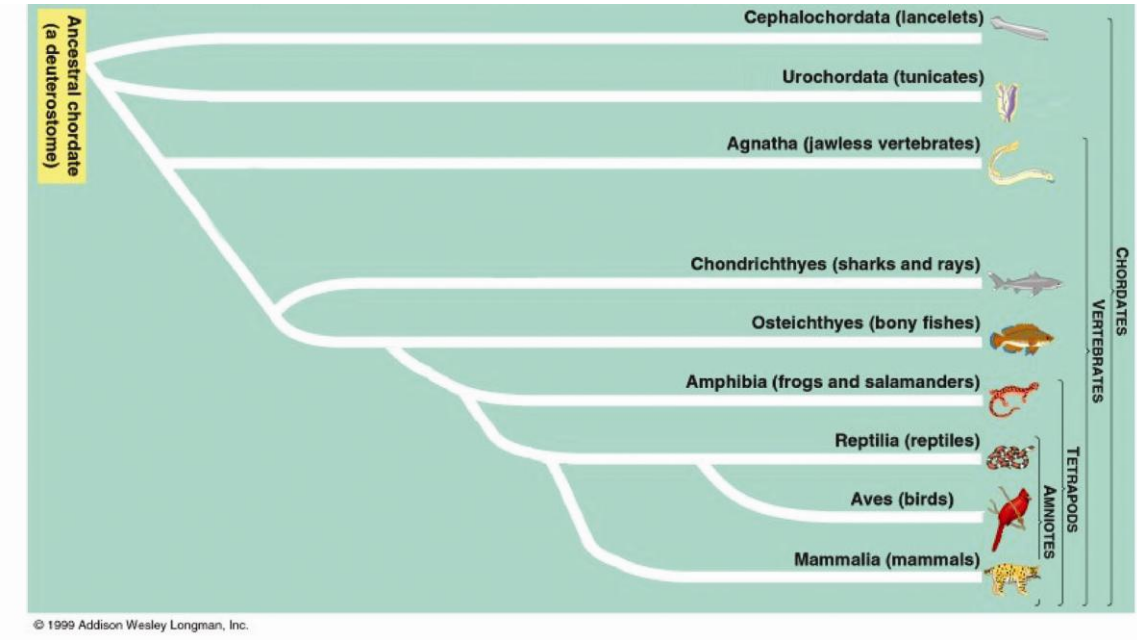


Figure 1.2. A phylogeny of the major vertebrate groups. The sea lamprey is a member of the agnathan lineage. Figure adapted from Campbell *et al*., 1999.

The sea lamprey as a model for investigating the evolution of vertebrate CNEs

The phylogenetic position of the sea lamprey (*Petromyzon marinus*), an extant jawless fish (agnathan), makes it ideally suited to answering questions regarding the early evolution of vertebrates (Figure 1.2). Given the inferred monophyly of cyclostomes (hagfish and lamprey) from molecular phylogenetic analyses (Kuraku & Kuratani, 2006), characteristics common to lamprey and jawed vertebrates (gnathostomes) can be assumed to have been present in the common ancestor of all extant vertebrates. Morphologically, the lamprey lacks certain characters present only in jawed vertebrates, including paired appendages, hinged jaws, an adaptive immune system, and specialisation of the axial skeleton along the anterior-posterior axis, which were acquired by the gnathostome lineage (Shimeld & Holland, 2000). The availability of large numbers of lamprey embryos during their summer mating season makes lamprey a useful model organism for evo-devo (Nikitina *et al*., 2009). Studies into lamprey development have revealed insights into the evolution of vertebrate characteristics such as the jaw (Shigetani *et al*., 2002), paired fins (Freitas *et al*., 2006) and neural crest (Sauka-Spengler *et al*., 2007). With the establishment of further molecular biology and histochemistry techniques for use on lamprey embryos (Kusakabe *et al*., 2003; McCauley & Bronner-Fraser, 2006), and a project to sequence the lamprey genome underway, the lamprey is poised to become a crucial evo-devo model. The sequencing

of the lamprey genome presents a fantastic opportunity to address questions about the timing of fixation of vertebrate CNEs, the functional evolution of CNEs in vertebrate lineages, and the evolution of vertebrates in general.

Thesis overview

This thesis traces the use of the sea lamprey as a model for investigating the emergence of CNEs in vertebrate genomes and the significance of these elements to vertebrate development and evolution. In Chapter 3 the pattern of CNE conservation across vertebrates is elucidated by searching for CNEs in the lamprey genome. The main findings from this search were that a relatively small, but significant, proportion of CNEs were found in the lamprey genome, and these elements were able to function as developmental enhancers in a zebrafish assay. An investigation into the functional significance of the lack of many CNEs in the lamprey genome is described in Chapter 4. Chapter 5 details the use of the lamprey CNE sequences for phylogenetic footprinting. This approach led to the identification of enriched Pbx-Hox TFBS motifs within lamprey and gnathostome CNEs, which correlate with hindbrain and pharyngeal arch enhancer function. In Chapter 6, the development of a reporter assay in lamprey embryos, and its use to address whether CNE gene-regulatory functions are conserved across vertebrates, is described. An evolutionary model predicting a role for many CNEs in the evolution of the vertebrate head is proposed. The topic of the identification of TFBS motifs in CNEs is returned to in Chapter 7, where the use of a *de-novo* motif discovery approach on CNEs is detailed.

2 Materials and Methods

Materials

Fish embryo reagents

20x Embryo medium (1L):

NaCl	17.5g
KCl	0.75g
$CaCl\text{-}2H_2O$	2.9g
add DDW to 800ml	
KH_2PO_4	0.41g
Na_2HPO_4(anhydrous)	0.142g
$MgSO_4\text{-}7H_2O$	4.9g
Vacuum filter sterilise	

Embryo medium (1L):

20x Embryo medium	50ml
500x $NaHCO_3$	2ml
Fill to 1L with DDW	

100x PTU (50ml):

PTU	150mg
Embryo medium	50ml
Heat to 65˚C, aliquot and freeze	

Embryo medium with PTU (700ml):

1x embryo medium	693ml
100x PTU	7ml

10x MMR (1L):

NaCl	58.44 g
KCl	1.491 g
$MgSO_4$	1.204 g
$CaCl_2.2H_2O$	2.94 g
HEPES (pH7.8)	11.915 g
adjust pH to 7.4 using 10M NaOH. Autoclave.	

***In-situ* reagents**

10X PBS (IL) pH 7.2:

$Na_2HPO_4.7H_20$	11.5 g
NaCl	80 g
KH_2PO_4	2 g
Kcl	2g

Make up to 1L with DDW

Hybridization mix (500 mL) **DNAse/RNAse free**:			final conc.
Formamide	250 ml		50%
20X SSC $_{DEPC}$(pH 5 w/citric acid)	32.5 ml		1.3X
0.5 M EDTA (pH 8)	5 ml		5mM
tRNA (20mg/mL in H_20_{DEPC})	5 ml	(or 100 mg)	200µg/ml
Tween-20	1ml		0.2%
10% CHAPS in H_20_{DEPC}	25 ml	(or 2.5 g)	0.5%
Heparin	50 mg		100µg/ml

Fill up to 500 mL with DEPC DDW

5X MABT 1L:

Maleic Acid	58g
DDW	500 ml
pH to 7.5 with Tris-base	100-150g
NaCl	43.5 g
20% Tween-20	5 ml

Fill up to 1l with DDW

10% Boehringer blocking reagent in MAB (no Tween-20) (200 ml):

BBR (Roche)	20g
1X MAB	180 ml

heat to 70°C to dissolve

autoclave or microwave to a boil

aliquot in 1 ml

freeze at -20°C

PTW (make fresh):

PBS with 0.1% Tween-20

20X SSC (200 mL) **RNAse/DNAse free**:

NaCl	35g
NaCitrate	17.6g

pH with 1M citric acid to 5.0

Fill up to 200 ml with DEPC DDW

NTMT (50 mL):		Final conc.
5 M NaCl	1 ml	100mM
1 M Tris.Hcl pH 9.5	5 ml	100mM
2 M $MgCl_2$	1.25 ml	50mM
10%Tween-20	0.5 ml	0.1%

Fill to final volume with DDW

Glycine:

0.1g in 50 mL PTW

sterile filter, make fresh on day of use

MEMFA:

16% formaldehyde	2.5ml
10x MEM salts	1ml
DEPC-DDW	6.5ml

make fresh on day of use

10X MEM salts 500ml:

MOPS (pH7.4)	104.65g
EGTA	3.804g
MgSO4	0.602g

DDW to 500ml

Methods

Molecular biology protocols:

PCR

General PCR mix using Taq polymerase:

5 µl	10x buffer
5µl	10x dNTPs (2mM)
2.5µl	forward primer (10µM)
2.5µl	reverse primer (10µM)
0.5-2µl	template DNA (25-50ng)
0.5µl	Taq polymerase
to 50µl	DDW
Basic cycle:	
95˚C	2 mins
95˚C	30 secs (denature)
55˚C	30 secs (anneal)
72˚C	30 secs (elongation)
repeat final 3 steps 25-35 times	
72˚C	5 mins

Adjust elongation temperature according to the instructions included with the enzyme being used (for instance, the optimal elongation temperature for Accuzyme (Bioline) is 68˚C whilst for BioTaq (Bioline) it is 72˚C).

DNA purification

DNA can be purified from enzymatic reactions (such as PCR and restriction digests) and agarose gels using the illustra GFX PCR DNA and Gel Band Purification kit (GE Healthcare).

Ethanol precipitation of DNA

Add one tenth volume of 3M sodium acetate (pH5.2) to DNA solution. Add two and a half volumes of cold 100% ethanol, place at -20˚C for 30 mins. Centrifuge sample for 30 minutes at high speed at 4˚C. Decant supernatant. Add cold 70% ethanol. Centrifuge for 10 minutes at high speed at 4˚C. Decant/remove supernatant. Dry on bench or in a vacuum. Resuspend pellet in DDW or TE buffer (10mM Tris-Cl, pH 7.5, 1mM EDTA).

Agarose gels

In general, use 1% agarose in 1x TBE buffer. Add 1µl ethidium bromide (10mgµl^{-1}) per 100ml agarose.

Megaprimer PCR

This technique for introducing mutations into DNA is based on that of Barik (2002).

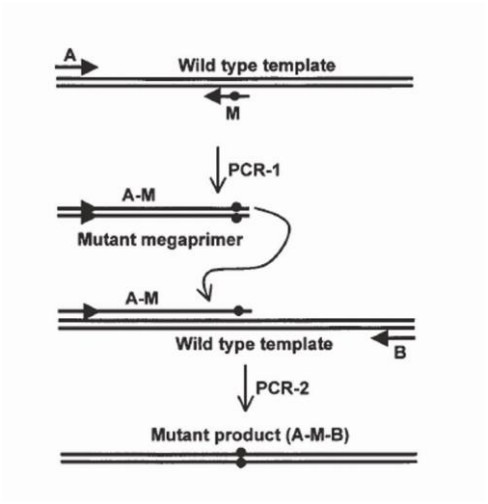


Figure 2.1. The megaprimer PCR method for site-directed mutagenesis. A and B are wild type primers, whilst M is a primer containing site-specific mutations.

PCR1: megaprimer creation – normal PCR settings, gel extract product

PCR2: megaprimer PCR – use product from PCR1 as primer, ~65˚C annealing temperature, 2 min annealing step, 10 min final extension, add primer B at cycle 6 annealing step, add enzyme at cycle 1 annealing step. Gel extract product.

PCR3: optional amplification PCR – use product of PCR2 as template, amplify element using internal primers. Gel extract product.

DNA extraction from animal tissue

Low-scale genomic DNA extraction can be performed using the DNeasy Blood and Tissue Kit (Qiagen).

Lamprey sperm DNA extraction

Obtaining lamprey sperm:

Dissect gonads from adult male lamprey, anaesthetised by Tricaine. Wash briefly with PBS, before mincing tissue with a razor and extensively triturating in PBS in a 50 ml falcon tube. Filter with a 40 µm filter before centrifugation at 500 x g for 10 mins. Resuspend the pellet in PBS. Check the colour of the pellet

– if it is white/cream then it contains pure sperm; blood appears as a red pellet at the bottom – this can be removed by resuspending only the white part of the pellet. Resuspend in 5ml PBS.

Counting cells on a haemocytometer:

Clean the haemocytometer with 70% ethanol, moisten the shoulders and affix the coverslip. Mix sperm cells by agitation. Quickly transfer 1 ml with a pipette into a separate eppendorf tube. Mix cells in this tube and transfer 100 µl into a new eppendorf tube. Add 100 µl of trypan blue (0.4% in PBS) and mix gently. Fill the haemocytometer gently with 10 µl of this mix. Count the cells in the 4 corner squares (not the blue cells!). Average the four counts and multiply by 2 x 10^6 to obtain the cell count for the original sample.

Sperm DNA extraction:

This method is based on that of Hossain et al (1997). Add 5ml of guanidinium lysis buffer (6M guanidinium thiocyanate, 30mM sodium citrate, 0.5% sarkosyl, 0.3M β-mercaptoethanol 0.2 $mgml^{-1}$ proteinase K (added fresh)) per 1 x 10^7 cells. Mix and incubate at 55˚C for 3-4 hours. Add two volumes of isopropanol to the lysate and gently invert the tube until DNA fibres clump together to form a 'cotton ball', which should be recovered with a glass sheppard's crook and briefly washed in 70% EtOH. Dissolve DNA in TE buffer.

Determining the concentration and purity of DNA

Use a NanoDrop spectrophotometer according to the users manual. Expect a 260:280 absorbance ratio of ~1.80 for pure DNA, ~2.00 for RNA. A low 260:280 ratio could indicate protein contamination and influences concentration determination.

Plasmid transformation

Add 1µl ligated DNA to 15µl competent cells (such as One Shot TOP10 (Invitrogen)) in a pop-top tube. Leave on ice for 10 mins. Heat shock cells at 42˚C for 47 secs and place on ice. Add 500µl of 37˚C SOC and incubate with shaking for 1 h. Plate desired volume (can try a range) onto pre-warmed 1% LB-agar plates containing the required antibiotic. Incubate plates overnight at 37˚C.

Plasmid preps

Inoculate the appropriate volume of LB medium containing the required antibiotic and incubate with shaking at 37˚C overnight (minimum 8 hours). Harvest cells by centrifugation and proceed with prep using either QIAprep Spin MiniPrep Kit (Qiagen) or QIAGEN Plasmid Maxi Kit.

Restriction digestion

Standard digest mix:

1µl	10x buffer
6.5µl	DDW
2µl	DNA (choose concentration appropriate to the size of the band to be visualised – the smaller the band the higher the concentration)
0.5µl	enzyme (5-10 units)

Mix well and incubate for 1h at 37˚C.

DNA ligation

Standard ligation mix (sticky ends):

1µl	10x T4 ligase buffer
6:1 molar ratio of insert to vector (~10ng vector)	
make up to 9.5µl	DDW
0.5µl	T4 ligase (200 units)

Leave for 10mins at RT. Proceed with transformation.

Fish embryo protocols:

Zebrafish strains

QMWT

Tubingen WT

rh3/5:KalTA4 (aka: r3r5 RFP) (Distel *et al*., 2009)

Zebrafish transgenesis –

Co-injection reporter assay

This assay is adapted from a protocol presented in Müller *et al*. (1999).

Preparation:

PCR-amplify βglobinGFP promoter-reporter cassette, ethanol precipitate and column purify to make a stock of 200ngµl^{-1}. PCR-amplify putative enhancer, ethanol precipitate and column purify to make a stock of 500ngµl^{-1}.

Injection:

Prepare the injection mix –

25ngµl^{-1}	βglobinGFP promoter-reporter cassette
75-125ngµl^{-1}	putative enhancer
0.5 µl	phenol red (0.5% Sigma)
Make up to 5µl with DDW	

Inject –

Pre-warm agarose injection dishes by placing them at room temperature. Load injection needle with injection mix and insert into micro-injector. Inject embryos at early 1-cell to 2-cell stages, either in the cell or into the yolk adjacent to the cell. Inject a volume equivalent to 1/5 the volume of the cytoplasm.

Post-injection –

Transfer embryos into a petri dish containing EM, incubate at 28˚C. In evening, remove unfertilised embryos and transfer embryos to EM-PTU, incubate at 28˚C. Change EM-PTU daily.

Embryo screening –

Dechorionate embryos manually using fine forceps. Anaesthetise embryos with 6-8 drops (glass Pasteur pipette) of Tricaine stock per small petri dish. Screen under fluorescent microscope. Note GFP expressing cells on embryo schematic data sheet. Compile composite expression using Adobe Photoshop.

Tol2-mediated transgenesis

This method is based on that of Fisher et al. (2006). The destination vector into which elements are to be cloned is pGW_cfosEGFP (Fisher et al, 2006).

PCR-amplify the element from genomic DNA with a high-fidelity enzyme using a 5 mins final extension step. For a successful TOPO reaction, column purify the PCR product.

Insertion of element into the entry vector:

Use the pCR8/GW/TOPO TA vector to clone the PCR product

50-100ng (4 µl max volume)	fresh, purified PCR product
0.3 µl (5ngul^{-1})	entry vector (TOPO)

1 µl saline buffer (from TOPO kit)

make up to 6 µl with DDW

Leave for 5 mins at room temperature.

Mix 3 µl of the mix with 50µl competent cells (DH5α, TOP10 or Match1) and transform. Plate all the volume on pre-warmed 1% LB-agar plates containing spectinomicine (100mgl^{-1}). Leave overnight at 37˚C.

Pick three colonies and inoculate them into 2ml of LB with spectinomicine (100mgl^{-1}). Incubate overnight at 37˚C with shaking.

Purify the plasmids by mini-prep and check for the presence of the insert by PCR. Select one plasmid for recombination.

Recombination with pGW_cfosEGFP destination vector:

Use the Gateway LR Clonase II enzyme (Invitrogen) to transfer the TOPO-cloned element into the pGW_cfosEGFP vector-

1 µl TOPO vector with insert (100ng)

1 µl pGW_cfosEGFP (100ng)

0.5 µl Clonase enzyme

Leave at 25˚C for 1h minimum. Add 0.25 µl of Proteinase K (2µgµl^{-1}). Transform 1.25 µl in Match1 or TOP10 competent cells. Plate all the volume on pre-warmed 1% LB-agar with ampiciline (100mgl^{-1}). Leave at 37˚C overnight.

Pick 3 colonies and inoculate them into 2ml of LB with ampicilin (100mgl^{-1}). Incubate overnight at 37˚C with shaking.

Purify the DNA by mini-prep, followed by column purification.

Injection:

Prepare the injection mix –

1 µl DNA (purified plasmid: 125ngµl^{-1})

1 µl transposase RNAenzyme (175ngµl^{-1})

0.5 µl phenol red stock

Load an injection needle of 1mm diameter with the mix. Inject with a micro-injector 1-3nl per embryo, aiming for the yolk just adjacent to the cell. Inject into embryos at the 1-2 cell stage.

Lamprey transgenesis –

Lamprey husbandry and embryo care, as described in the transgenesis protocols below, is detailed in Nikitina *et al*. (2009).

Linearised construct injection

Preparation of vector:

The vector used is pGW_cfosEGFP. Clone the element into the vector as described for the zebrafish tol2 assay. Linearise vector by restriction digestion with KpnI or XhoI. Digest enzyme with proteinase K (100-200 $\mu g \mu l^{-1}$). Column purify, elute in DDW to make stock of 100ngµl.

Injection:

Obtain lamprey eggs by massaging a gravid female lamprey in a large glass dish containing 18˚C Sparkletts water (mineral water with a suitable ionic concentration). Fertilise eggs by massaging a mature male lamprey, mix and leave eggs to fertilise for 10 mins. Check activation under a dissecting microscope – chorions can be seen to expand. Wash thrice with distilled water (18˚C) and replace with Sparkletts water. Incubate at 18˚C. The first cell division occurs at 5-6 hpf. Check the viability of the batch by the proportion of embryos showing cell cleavage - <50% cleavage is bad, >80% is good. Inject 2-3µl linearised plasmid at $100ng\mu l^{-1}$ into embryos at 1-2 cell stage. Store embryos in 0.1x MMR at 18˚C.

Lamprey embryo care:

Keep embryos at 18˚C in an incubator. Change the 0.1x MMR daily. Remove the dead embryos daily. At e4 (3dpf), spread embryos apart in the dish and leave them to gastrulate until late e5/e6. Remove dead embryos and screen survivors for reporter expression.

I-sceI meganuclease-mediated transgenesis

This method is based on that of Ogino *et al.* (2006), which was used on frogs.

Cloning element into the vector:

The cfos-IsceI-EGFP vector was created from the β-globin EGFP construct by cloning the mouse cfos promoter in place of the β-globin promoter, upstream of the EGFP coding sequence (Figure 2.2). I-SceI sites flank the promoter-GFP cassette. Enhancers to be tested can be cloned upstream of the promoter using eg 5’ XhoI and 3’ HindIII sites, which can be added to the primers used to PCR amplify the elements. After cloning the element into the plasmid, extract the plasmid using an EndoFree Plasmid Maxi Kit (Qiagen) and elute with water through QIAQuick columns (Qiagen). Dissolve the DNA in DDW to 1µgµl-1 and make a working stock of 100ngµl-1. Store both at -20˚C.

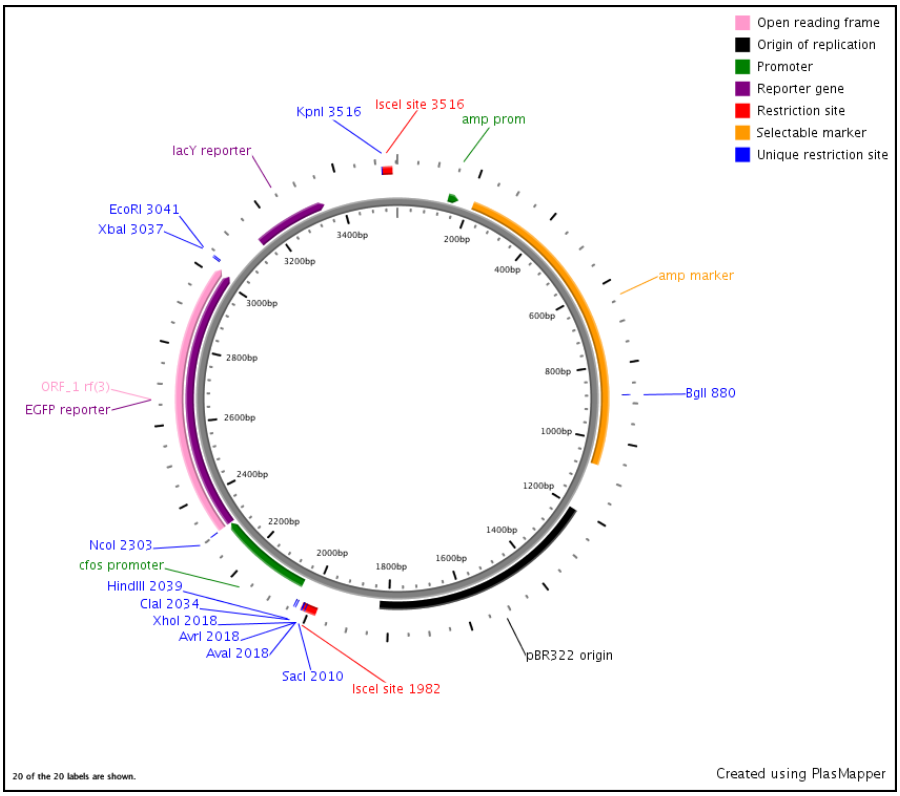


Figure 2.2. Map of the cfos-IsceI-EGFP plasmid. Key features are highlighted and explained by the colour-coded key.

Preparation of injection solution:

To maximize the number of full transgenics it is good to inject the DNA whilst the embryos are still at the single cell stage. However, it was found that injection at 2-3 hpf resulted in abnormal division followed by death during gastrulation, whereas injection at 5-6hpf was fine. At this stage the embryos start to show cleavage furrows.

The standard reaction is

2µl	10 X I-sceI buffer + BSA (pre-mixed)
4µl	cfos-IsceI-EGFP plasmid (100ng μl^{-1})
3µl	I-SceI enzyme (5 units μl^{-1})
11µl	Water

Digest at 37˚C for 40 mins. The I-SceI enzyme should be aliquoted and kept in a freezer at -80˚C to prevent degradation. This reaction mix results in a plasmid DNA concentration of 20ngμl^{-1}.

Injection:

After 40 mins of digestion, take the mix out of incubation. This is the injection mix. Inject roughly 2-3nl per embryo (quite a small drop) using the standard

injection procedure. Inject immediately - freshness is the key. Transfer injected embryos to 0.1x MMR and incubate at 18˚C.

In-situ hybridisation on lamprey embryos

Fixing Embryos:

Collect embryos and leave in MEMFA at RT FOR 1h on nutator.

Add 1 X DEPC-PTW solution. Shake on nutator for at least 15 mins. Discard solution. Repeat this wash three times.

Dehydrate embryos: 15 mins on nutator each

25% MeOH + 1X PTW/DEPC

50% MeOH + 1X PTW/DEPC

75% MeOH + 1X PTW/DEPC

Add MeOH 100%. Shake for 15 mins. Discard solution. Do this step twice

Add MeOH. Store at –20 °C

Pretreatments and Hybridization:

Re-hydrate embryos - 5-10 mins each, on nutator

75% MeOH + PTW/DEPC

50% MeOH + PTW/DEPC

25% MeOH + PTW/DEPC

Wash twice with PTW/DEPC for 5 mins each

Bleaching step: Replace PTW/DEPC with freshly made bleaching solution, which is 0.5% SSC, 10% H_2O_2, 5% formamide (For 10ml add in the following order: 500µl Formamide, 6.45ml H_2O, mix together then add 250µl 20X SSC and finally, 2.8ml H_2O_2). Place on the light box for 10 mins (this time can be increased to 15 mins if stronger signal is desired). Dilute the bleaching solution with an equal amount of water.

Wash thrice with PTW/DEPC for 5 mins each

Treat with 14-22µg/ml of proteinase K in PTW/DEPC (1:1000 dilution of the Roche PK stock). Incubation time depends on the level of the penetration desired. It may vary between 10 – 15 mins. Do not use the nutator after this step until the Fixing step!

Wash in 2 mg/ml of glycine in PTW/DEPC. Incubate for 10mins

Wash twice with PTW/DEPC for 5 mins

Post-fix with 4% PFA + 0.2% glutaraldehyde for 20 min @ RT

Rinse 4X with PTW/DEPC for 5 mins

PTW_{DEPC}: Hybe Mix (1:1) 10 min @ RT

Hybe Mix 10 min @ RT, add another HybeMix wash if necessary to get as close to 100% HybeMix as possible.

At this point the embryos can be stored indefinitely in HybeMix at -20°C. The procedure can then be continued at any convenience.

Change to new hybe mix and pre-hybridize:

Incubate at 70°C for approximately 3 hours

Add new pre-warmed (70°C) hybe mix + RNA probe-Dig (1-10 µl/ml hybe)

Incubate at 70°C O/N (at least 16 hours). Position tubes horizontally if using an oven.

Post-hybridization washes (Most important for specificity):

Remove the probe in Hybe solution and save it @ -20°C (to be re-used)

Wash 2X with pre-warmed hybridization solution (50 ml tube ON in the oven) at 70°C for 15 mins each

Wash 4X with warmed hybridization solution at 70°C for 30 –45 mins each

Wash once with HybeMix:MABT (1:1) @ 70°C for 10 mins

Wash once with HybeMix:MABT (1:1) @70° C for 30 mins with agitation

Wash 4X with MABT for 30 mins at RT on nutator

Change MABT to MABT + 20%Sheep Serum + 2%BBR for 15 mins at RT (blocking solution) on nutator

Replace with fresh MABT + SSC + BBR for 3-4 hours at RT on nutator

Change to Anti-Dig-AP AB (1:2000) in MABT + sheep serum + BBR and leave O/N at 4°C on the nutator

Post-Ab washes and Histochemistry:

All these steps are on the nutator-

Rinse 2X with MABT for 5 mins each wash at RT

Wash 2X with MABT for 30 mins each at RT

Wash 6X with MABT for 1 hour each wash at RT

Wash at 4°C O/N

Wash 4X with NTMT, 15 min at RT

Change NTMT solution to BCIP in NTMT (filtered). Be sure to cover vials with foil. Change the substrate after 1h. Closely follow the development of colour. If necessary leave at 4°C O/N.

After the desired color has developed, proceed with washing steps.

Wash 3X with PTW for 5 mins each at RT (in obscurity-foil)

Change to 4% PFA, leave at 4°C O/N OR 2 hours RT (in obscurity-foil)

Wash 3X with PTW for 10 mins at RT (in obscurity-foil)

Change solution to MeOH in PTW - 5 mins each

25% MeOH + PTW

50% MeOH + PTW

75% MeOH + PTW

Wash twice with 100% MeOH for 5 mins each

Keep @ 4°C or -20°C ON

Re-hydrate embryos before photographing:

Change solution to MeOH + PTW - 5 mins each

75% MeOH + PTW

50% MeOH + PTW

25% MeOH + PTW

Wash 3X with PTW for 5 mins each

Embryos can be further equilibrated to 75% Glycerol in PTW (gradual steps of 25%, 50% and finally 75% Glycerol/PTW)

3 CNEs in the Sea Lamprey Genome

Abstract

CNEs show an intriguing pattern of conservation across chordates – whilst they are incredibly highly conserved amongst jawed vertebrates, there are barely any traces of them in invertebrate chordate genomes. This is in contrast to the conservation patterns of the genes with which they are associated, the majority of which have invertebrate homologues. As an increasing number of CNEs have been shown to have *cis*-regulatory capabilities, they have been proposed to represent regulatory elements that arose early in the vertebrate lineage and are crucial for specifying the development of the vertebrate body plan. Whilst a wealth of genomic data is available for jawed vertebrates, there is another vertebrate lineage containing species whose genomes have not been so well characterised – the agnathans. This lineage occupies a unique phylogenetic position, having diverged from the other vertebrates very early in vertebrate evolution. In this chapter, emerging genomic sequence data from an agnathan – the sea lamprey – is used to trace the evolution of CNEs deep into the vertebrate phylogeny. In doing so, a set of ancient, pan-vertebrate CNEs is defined. This set of elements provides a useful resource for investigation into the crucial roles that CNEs are predicted to play in vertebrate development.

Introduction

The high sequence conservation of gnathostome CNEs can be traced back as far as the Chondrichthyes (Venkatesh et al, 2006), showing that these sequences evolved prior to the divergence of cartilaginous and bony fish more than 500 MYA (Blair & Hedges, 2005). However, traces of only a very small number of vertebrate CNEs are visible in the genome of an invertebrate chordate, amphioxus, illustrating that the majority of CNEs represent a defining characteristic of the ancestral vertebrate genome (Putnam *et al*., 2008; Holland *et al.*, 2008). Invertebrate groups have also been found to posses their own sets of CNEs (Glazov *et al.*, 2005; Vavouri *et al*., 2007) and there is a correlation between the families of genes around which both vertebrate and invertebrate CNEs are clustered, suggesting parallel evolution of the GRNs in which these CNEs act (Vavouri *et al*., 2007). Thus, despite the fact that the evolution of coding sequences can be tracked across the invertebrate/vertebrate boundary, the evolution of CNE sequences

seems to have followed a different pattern, with their comparatively rapid emergence very early the vertebrate lineage. As a starting point to investigate the origin and evolution of CNEs in early vertebrates, I have focused on the sea lamprey, *Petromyzon marinus*. It can be assumed that the common ancestor of gnathostomes and angnathans possessed the developmental mechanisms and morphological characteristics that are shared by both modern groups. Comparative studies of the lamprey and its genome can therefore provide insights into the ancestral vertebrate state, and the common regulatory sequences that determined it.

It is likely that the early vertebrate genome was shaped by two whole genome duplication events, with gnathostomes possessing paralogous copies of genes that exist in single copy in invertebrates such as amphioxus (Putnam *et al*., 2008). There is evidence supporting the hypothesis that both rounds of duplication preceded the agnathan-gnathostome divergence (Kuraku *et al*., 2009). However, this is difficult to prove definitively without an assembled lamprey genome. Whilst the majority of CNEs appear to be unique in the human genome, a significant sub-set exist as duplicates and are found in the vicinity of paralogous genes (McEwen *et al*., 2006). These duplicated CNEs (dCNEs) indicate that certain CNEs were already in existence prior to at least one of the whole genome duplication events in vertebrates. If the agnathan and gnathostome lineages diverged after these duplications, then we might expect to find some CNEs to also be conserved in the lamprey genome.

The genome of the sea lamprey, *Petromyzon marinus,* has been targeted for a high quality draft and assembly. Over 18 million whole genome shotgun (WGS) reads, equivalent to a 6-fold coverage, have been made available. By utilising this lamprey sequence data, it is possible to investigate an ancient era between the emergence of vertebrates and the divergence of the agnathans and gnathostomes, in order to identify those CNEs that are common to all extant vertebrates.

Results

Identification of CNEs from the lamprey whole genome shotgun sequence

(this data was generated and analysed by G. McEwen)

Thousands of CNEs have previously been identified by sensitive multiple sequence alignment between human, mouse, rat and Fugu genomes (Woolfe *et al*., 2007). From a

database of these CNEs (http://condor.nimr.mrc.ac.uk/) 13 gene regions were selected based on the high occurrence of dCNEs within them. Smaller CNEs from these regions were removed by using an "LPC" score of >= 50, (the score is based on the sequence length and identity across the four species (Woolfe et al., 2007)). The 13 gene regions comprise a total of 27 Mb of sequence, containing 1205 elements (Table 1), including 108 duplicated CNEs (dCNEs) (McEwen *et al*., 2006) and 46 ultra-conserved elements (Bejerano *et al.*, 2004) (UCEs) - sequences with 100% identity over at least 200bp between mouse, rat and human genomes. The 1205 gnathostome CNEs were identified using a multiple alignment approach - MLAGAN (Brudno *et al.*, 2003) - however, this approach is not possible for the lamprey trace sequences due to their short length. To verify the efficacy of using BLAST (Altschul *et al*., 1997) as a tool for identifying CNEs, we searched the Fugu genome for homologous sequences to the 1205 human CNEs using sensitive parameters (word size: 8, mismatch penalty: -1, e-value cut-off: $5e^{-4}$), with the majority of the CNEs being identified (1035/1205 = 85.9%) (Table 1). The 1205 CNEs were then searched against the lamprey reads, which were downloaded from the NCBI trace server (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi), using BLAST (parameters as above for the Fugu genome search).

Some of the lamprey reads were predicted to represent contamination, as a number of hits were almost identical to chicken when compared to all vertebrates in Ensembl using BLAST. The full length lamprey reads corresponding to these hits were then compared to the chicken genome using BLAST (with default parameters) and those matching to chicken with >90% identity across most of their length were removed (a total of 34 lamprey hits). Due to the unassembled lamprey genome sequence, many lamprey hits were to multiple redundant reads. Consensus sequences were generated for each hit if the sequences were more similar than 95% identical, with overlapping hits then being joined to make a contiguous hit.

From this search, 74 lamprey CNEs were identified, including 38 dCNEs and 8 UCEs, with matches to gnathostome CNEs in all but one of the gene regions (Table 1). This signifies a widespread distribution of CNEs across trans-dev genes in lamprey. The proportion of lamprey hits to dCNEs (38/74 = 51.3%) was found to be greater than would be expected based on their proportion in the gnathostome CNE set (dCNEs only constitute 3.7% of the total CNEs across the 13 regions). This demonstrates a considerable enrichment for this set of ancient elements. Furthermore, whilst 17.4% of

the UCEs were detectable, more than twice as many (35.2%) of the dCNEs were identified across the regions. Despite lamprey CNEs being found for each of the 13 gene regions except *Dach1*, the total number that were identified is low relative to the numbers conserved across jawed vertebrates – only 74 of 1205 were found. This could be an underestimate due to the incomplete coverage of the lamprey draft genome sequence. Alternatively this pattern may represent significant divergence of these sequences between lamprey and gnathostome lineages or the rapid emergence of many CNE sequences in the gnathostome lineage (see Discussion).

			All CNEs			**dCNEs**			**UCEs**		
	Human genome		**Multi-LAGAN**	**BLAST hits**		**Multi-LAGAN**	**BLAST hits**		**Multi-LAGAN**	**BLAST hits**	
Gene Region	**Chr**	**Length (Mb)**	**H/F/M/R**	**H/F**	**H/L**	**H/F/M/R**	**H/F**	**H/L**	**H/F/M/R**	**H/F**	**H/L**
BARHL2	1	0.789	55	49	1	4	4	0	0	0	0
BCL11A	2	2.634	72	55	4	6	6	3	8	5	1
DACH1	13	1.236	56	50	0	4	3	0	7	4	0
EBF3	10	1.753	138	118	9	13	13	3	2	1	1
FOXB1*	15	1.979	45	40	2	5	5	2	0	0	0
FOXP2	7	1.682	95	78	8	7	7	2	6	3	0
IRX5	16	4.421	192	163	10	20	20	6	3	1	0
MEIS2	15	3.079	118	103	10	8	8	4	4	3	0
NR2F1	5	3.412	117	98	3	5	5	2	1	1	0
PAX2	10	0.264	51	45	9	6	6	5	9	8	4
TSHZ3	19	2.194	109	101	10	15	15	8	2	2	2
ZIC2	13	0.835	36	29	3	4	4	1	0	0	0
ZNF503	10	2.759	121	106	5	11	11	2	4	4	0
Totals:		**27.039**	**1205**	**1035**	**74**	**108**	**107**	**38**	**46**	**32**	**8**

Table 3.1. CNEs from 13 human gene regions identified in the Fugu and lamprey genomes. Chr: chromosome, H: human, F: Fugu, M: mouse, R: rat, L: lamprey. H/F/M/R: four way multiple alignment (MLAGAN) used to identify CNEs.

The lengths of the lamprey hits were found to be on average considerably lower than those defined by gnathostome conservation. Lamprey sequences match on average only 47% of the length of CNEs defined through alignments between mammals and fish using the same BLAST parameters. Nevertheless, sequence conservation is high across these core regions, with an average identity of 80%, compared with approximately 90% between teleosts and mammals.

Analysis of a contiguous region of the lamprey genome that contains CNEs

(This search was performed by G. McEwen, with the alignment created by myself)

The pre-Ensembl lamprey draft assembly (PMAL3) was searched for assembled contigs that contain multiple lamprey CNEs. The longest of these (contig 1709) encompasses 33.7kb of contiguous lamprey sequence with just three short unresolved regions. This contig contains a number of CNEs as well as an uncharacterised gene, *C15orf41*, which lies immediately downstream of the *meis2* gene in gnathostome genomes (Figure 3.1). The identified lamprey CNEs reside directly adjacent to, or within the introns of, the *C15orf41* gene, and form part of a much larger genomic regulatory block covering nearly 3.5Mb of the *meis2* locus in the human genome, which contains over 200 CNEs.

The organisation of this region, with conserved coding exons acting as landmarks, allowed us to identify which gnathostome CNEs are detectable in the lamprey genome using multiple alignment approaches. Given the conserved positional relationship of CNEs in all other vertebrates, we have assumed that if lamprey CNEs are present, they will also be co-linear. From Figure 3.1 it is apparent that whilst some CNEs are clearly detectable in the lamprey genome, others are not found using sequence similarity. Furthermore, BLAST searches of the WGS reads do not identify these CNE sequences elsewhere in the lamprey genome.

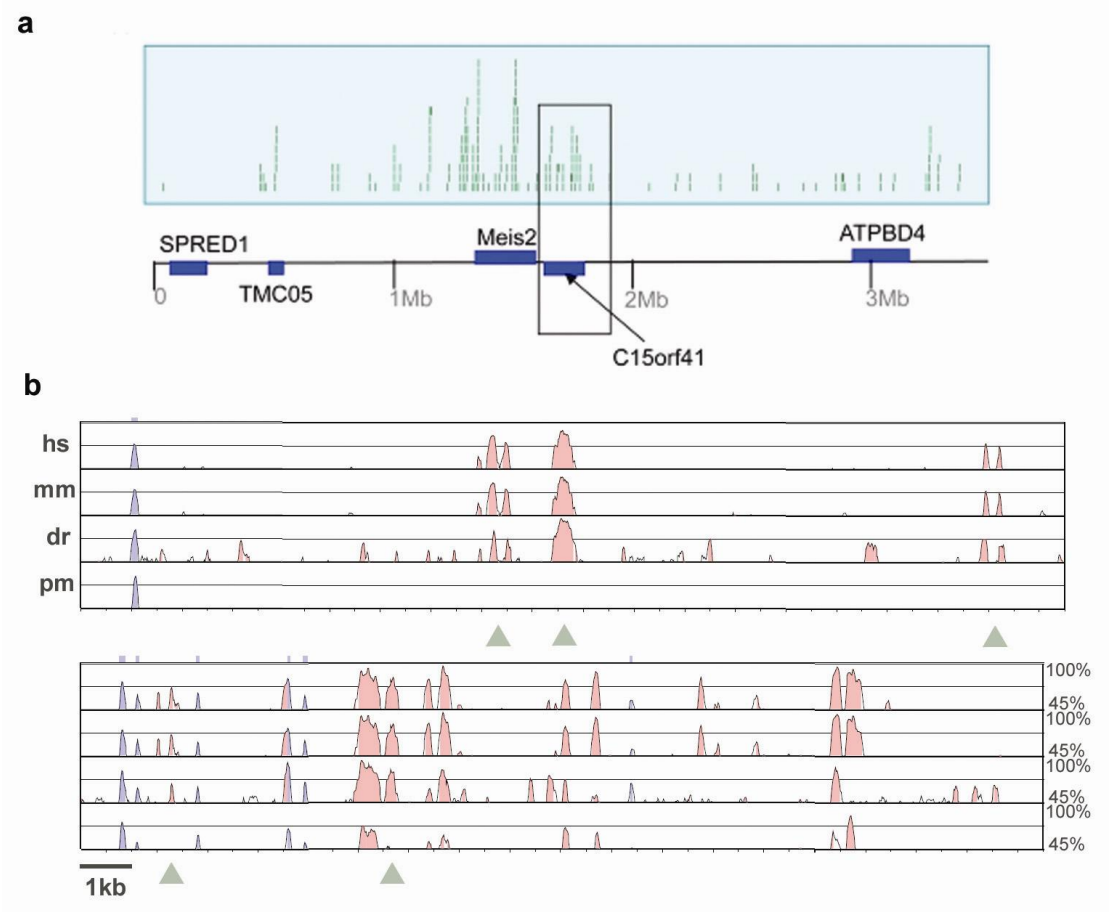


Figure 3.1. Conservation of non-coding sequences across the *meis2/c15orf41* locus in vertebrates. **a** Plot of non-coding sequence conservation between mammals and fish across a 3.5 Mb region of human chromosome 15q14, encompassing the *meis2* and *C15orf41* genes. Each vertical bar within the blue panel represents a CNE. **b** MLAGAN alignment of the *C15orf41* locus highlighted in **a**. Human (hs), mouse (mm), zebrafish (dr) and lamprey (pm) genomic regions are aligned with the orthologous region in the Fugu genome. Exons are represented by blue peaks and are detectable in all species. Pink peaks represent non-coding conservation. A number of these are conserved in lamprey but many are also absent (grey arrowheads). This figure is based on that of McEwen *et al*. (2009).

Functional conservation of lamprey CNEs

(This functional data was obtained by D. K. Goode and myself)

Gnathostome CNEs have evolved extremely slowly, given their high identity between sharks and mammals (Venkatesh *et al.*, 2006), which diverged over 500 MYA (Blair & Hedges, 2005). Lamprey CNEs appear considerably shorter (half as long on average) and less well conserved in sequence than their gnathostome counterparts. One of the most highly conserved CNEs in our data set is found within the *Ebf3* gene region and

extends to 491 bp at greater than 90% identity between Fugu and human. The corresponding element identifiable in the lamprey genome is only 211 bp long (with 79% identity). A second representative CNE, associated with the *Pax2* gene, is 85% identical across 425 bp between Fugu and human, but only 123 bp is conserved in lamprey (73% identity). We predicted that these 'core' regions of sequence conservation might comprise critical *cis*-regulatory modules common to all vertebrates.

In order to verify that lamprey CNEs represent functionally conserved developmental enhancers, we used a functional assay (co-injection: see Materials and methods) to test the ability of the core regions from the Ebf3 and Pax2 CNEs to up-regulate GFP reporter expression in zebrafish embryos. Testing the orthologous core regions of both CNEs from human and lamprey genomes, we found that in all four cases the core elements up-regulated GFP expression in a temporal and tissue-specific manner consistent with the endogenous pattern of expression of the associated gene. Strikingly, the patterns of expression were seen to be very similar between lamprey and human elements.

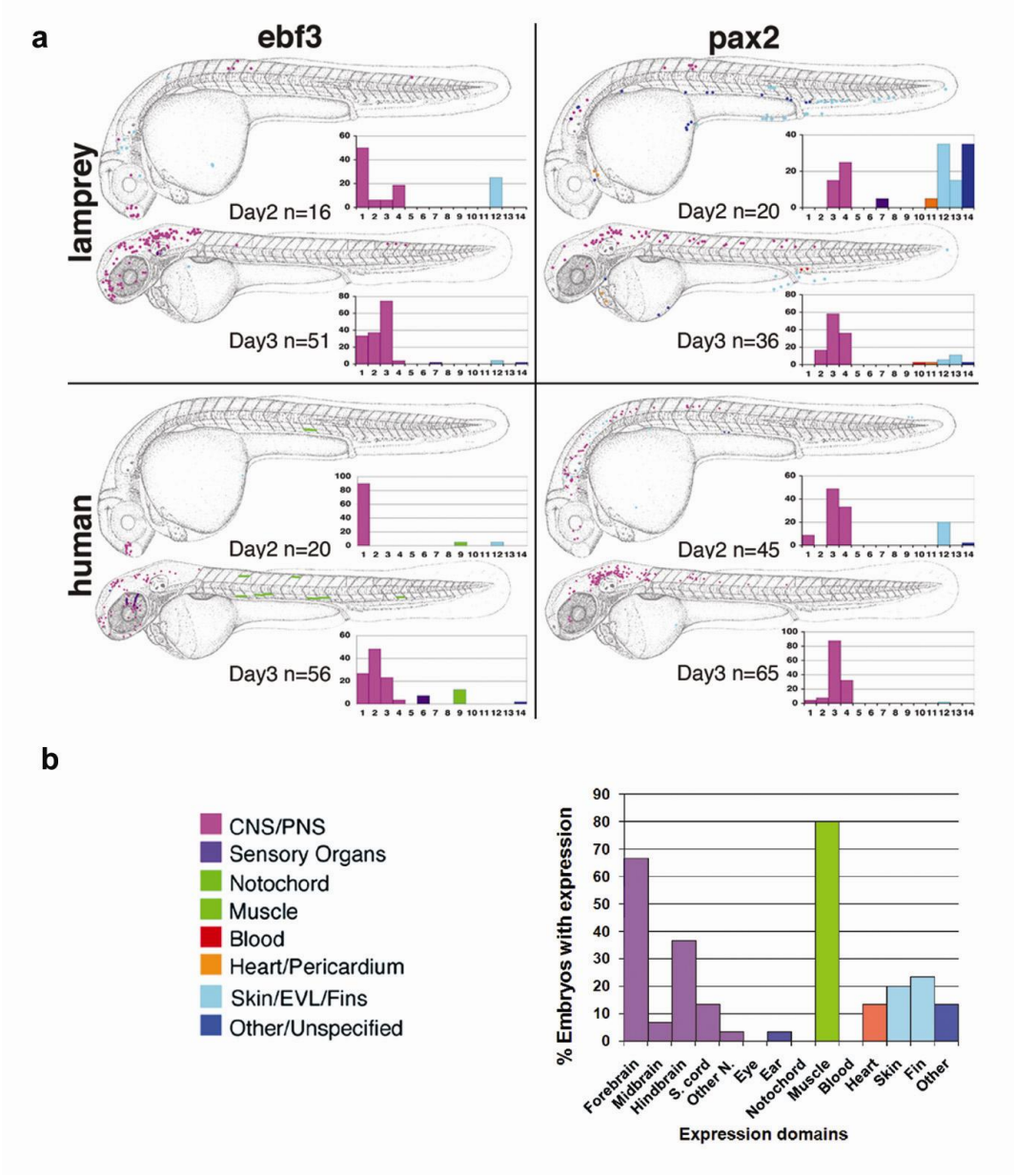


Figure 3.2. Schematic representations of GFP expression patterns driven by core CNEs. **a** GFP-positive cells are marked onto camera lucida drawings of a zebrafish embryo on day 2 (24-30hpf) and day 3 (48-54hpf) of development. The results from all embryos with expression are overlaid, giving a composite depiction of the GFP expression pattern. The number of GFP-positive embryos are noted beneath each schematic (n=). The charts show the percentage (y axis) of GFP positive embryos with expression in each domain. **b** Key for the schematics and charts shown in **a**. In both the charts and the schematics, broad domain categories are colour-coded as shown. An example chart identifies the different expression domains represented by the x-axis. This figure is based on that of McEwen et al. (2009).

Ebf3 is a member of the COE (*Col-Olf-Ebf*) gene family, which consists of the vertebrate orthologs of the *Drosophila collier* gene (Crozatier *et al.*, 1996) and *C. elegans unc-3* (Prasad et al, 1998). It is expressed in the developing central nervous system (Garel *et al.*, 1997) and appears to be a key regulator of neurogenesis, associated with the maturation of specific neuronal cell types in the spinal cord and brain (Crozatier *et al.*, 1996). On day two of zebrafish embryo development, at 24-30 hours post-fertilisation (hpf), both the lamprey and human elements directed expression of a GFP reporter gene predominantly in the forebrain (Figure 3.2). Levels of expression were much higher on day three, 48-54 hpf, with a more widespread pattern encompassing the spinal cord as well as the fore-, mid- and hindbrain regions. Both the lamprey and the human *Ebf3* elements appeared to up-regulate GFP expression specifically in a particular set of neurons in the zebrafish embryo (Figure 3.3), demonstrating strong functional conservation between the lamprey element, and the equivalent core region of the human CNE.

Pax2 is a member of the vertebrate *Pax2/5/8* family of transcription factors that is likely to have arisen from early duplications in the vertebrate lineage (Pfeffer *et al.*, 1998). It contributes to the development of the eye, ear, pronephros and midbrain-hindbrain boundary. In lamprey, it has been demonstrated that the expression pattern of *pax2* in each region is similar to that of gnathostomes (McCauley & Bronner-Fraser, 2002). Injection of *Pax2* elements derived from both lamprey and human resulted in GFP expression in the CNS and skin at day two, with a more specific pattern of neuronal expression, particularly in the hindbrain, on day three (Figure 3.2). Together, these reporter-assay data indicate striking functional conservation of non-coding enhancer elements that are separated by more than a billion years of vertebrate evolution.

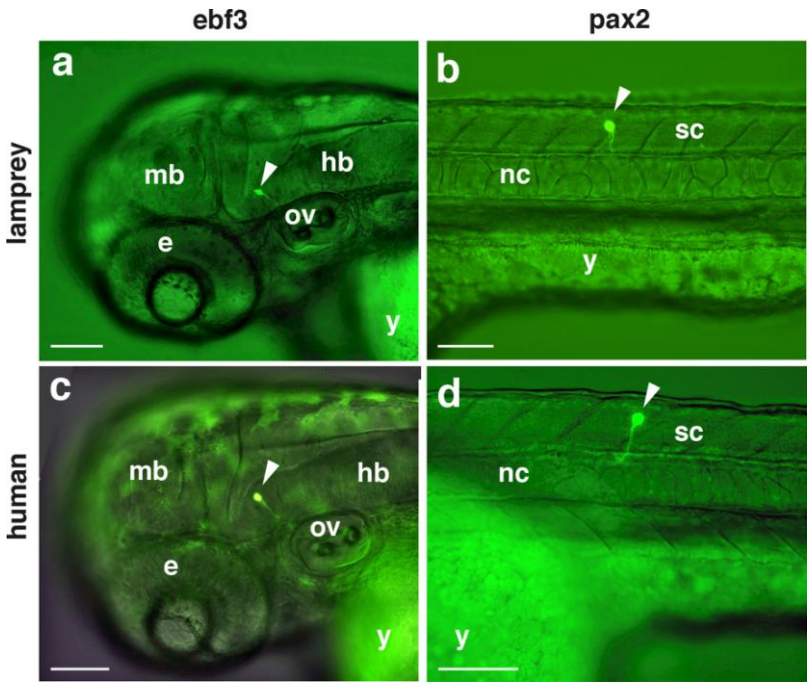


Figure 3.3. Up-regulation of GFP by orthologous lamprey and human CNEs. Images of live zebrafish embryos 48-54 hpf (**a-c**) and 24-30 hpf (**d**), lateral views, anterior to left. **a** and **c**, GFP expression in the hindbrain, driven by an EBF3 CNE derived from lamprey (**a**) and human (**c**). **b** and **d**, GFP expression in the spinal cord driven by a pax2 CNE derived from lamprey (**b**) and human (**d**). e, eye; hb, hindbrain; mb, midbrain; nc, notochord; ov, otic vesicle; sc, spinal cord; y, yolk. Scale bars represent 100μm. This figure is based on that of McEwen *et al*. (2009).

Comparison of sequence divergence between ancient orthologous CNEs and between dCNEs

The period of time during which the 2R duplications occurred, early in the vertebrate lineage, is likely to have been roughly contemporaneous with the divergence of the agnathan and gnathostome lineages. Thus, the divergence times between pairs of orthologous gnathostome-lamprey CNEs and between pairs of paralogous gnathostome dCNEs are more or less equivalent. Because some lamprey CNEs exist as dCNEs in gnathostomes, it is possible to compare the patterns of sequence divergence (as indicated by the length of BLAST alignments, using the paramaters defined on p27) between these different pairs of elements. I selected four dCNE families for such comparisons, with the main aim of addressing whether the 'core' CNE regions conserved between gnathostome-lamprey CNEs are also conserved between the corresponding dCNEs from the human genome (Figure 3.4). The dCNE families were selected on the basis of having a large size difference between gnathostome and

lamprey CNEs, exhibiting extreme cases of the phenomenon of a ‘core’ conserved region defined by lamprey homology.

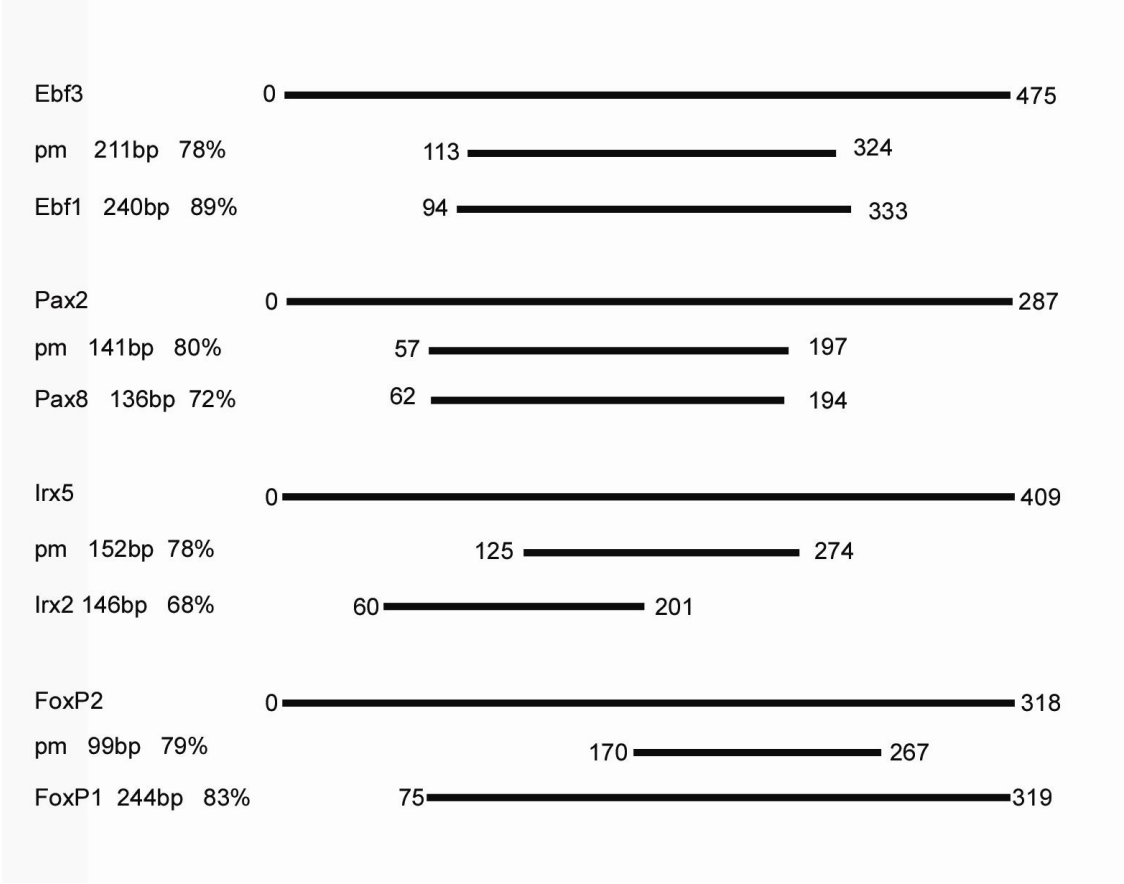


Figure 3.4. Sequence overlap between gnathostome and lamprey CNEs and dCNEs. Four gnathostome CNEs, associated with Ebf3, Pax2, Irx5 and FoxP2, are depicted as lines. Their lengths are indicated in bp (not to scale). Beneath each CNE, the aligned ‘core’ conserved region defined by the orthologous CNE in lamprey (pm: *Petromyzon marinus*) is shown, with positions of overlap between the aligned sequences given in bp and the length and percentage identity between human and lamprey elements detailed. Beneath this, the region of overlap between the human CNE and its paralogous dCNE from the human genome is depicted.

From Figure 3.4 it is evident that, for the four CNEs analysed, the ‘core’ regions defined by lamprey CNE conservation are approximately reproduced when the corresponding dCNEs are aligned. In the cases of the Ebf3 and Pax2 elements, this overlap is strikingly clear, with the lamprey conserved regions and dCNE conserved regions overlapping almost entirely. In the case of the Irx2 element, the core regions defined by lamprey conservation and dCNE conservation are of almost equal length but

only overlap with each other over half of this length. For the FoxP2 element, the core region of lamprey conservation is considerably shorter than that between the FoxP1/2 dCNEs. For these elements, there is no consensus pattern for the level of sequence identity within alignable regions between human-lamprey CNEs and dCNEs. Thus, the human Ebf3 element shows higher sequence identity with its dCNE than it does with its lamprey homolog, over equivalent stretches of its sequence. However, this relationship is reversed in the case of the Pax2 element, which shows greater identity with the lamprey homolog than with its paralog in the human genome.

Further identification of CNEs in lamprey

(This search was performed by P. Piccinelli and myself)

Our initial search for gnathostome CNEs in the lamprey genome was restricted to 13 gene regions that contained high numbers of dCNEs. We subsequently performed a further search for a set comprising all human-Fugu CNEs represented in the Condor CNE database, using BLAST with sensitive parameters. 6,693 non-redundant human CNEs (average length 116bp) were retrieved from the CONDOR database at. We used these to search lamprey sequence reads with sensitive parameters (word size: 7, mismatch penalty: -1, e-value cutoff: $5e^{-4}$). As before, lamprey sequences satisfying this initial parametric threshold were further analyzed for contamination, and those showing >90% homology to human or chicken across the whole read (i.e. extending outside the evolutionarily conserved region in other vertebrates) being removed. This search resulted in the identification of 246 lamprey CNEs. In agreement with the findings of the previous search, the average length of lamprey CNE hits was considerably smaller than that of their human-Fugu counterparts (115.4bp vs 230.9bp = 50%), and they had a relatively low sequence identity to the human sequences (79%).

Discussion

Identifying ancient vertebrate CNEs using the sea lamprey genomic sequence

This study presents evidence that all extant vertebrates, including the lamprey, possess a substantial repertoire of CNEs associated with genes that regulate development. The pattern of conservation in lamprey of dCNEs relative to CNEs and UCEs is in keeping with the hypothesis that these duplicated elements are evolutionarily ancient, predating the divergence of gnathostome and lamprey lineages, whilst a larger proportion of

CNEs may have evolved in the gnathostome lineage after this divergence. This suggests that the lamprey lineage diverged from the rest of the vertebrates at a time when the vertebrate body plan was taking shape, with large cohorts of *cis*-regulatory elements rapidly evolving and becoming fixed in both sequence and function, generating CNEs.

The identification of the *C15orf41* contig provides an insight into the CNE landscape of the lamprey genome. At one end of the gene region, a majority of gnathostome CNEs are detectable in lamprey and show conserved synteny, yet in the other half of this region, there are no lamprey CNEs present. This region presents a unique opportunity to investigate, by reporter assay, the functional implications of the low CNE quota in the lamprey genome.

The gene-regulatory role of ancient vertebrate CNEs

We chose two very highly conserved CNEs for functional analysis. The first, a CNE associated with the *Ebf3* gene, is over 90% identical across almost 500bp in jawed vertebrates, yet the lamprey identity extends to just over 200bp across the centre of this CNE. The human core sequence and the lamprey element drove similar patterns of GFP expression in the developing zebrafish brain, confirming that the shorter lamprey region of reduced conservation still retains the basic instructions for this enhancer function. A similar result was obtained for a CNE from the *Pax2* region, which shows an even greater reduction in length in lamprey, being less than 30% of the length of the gnathostome CNE. The long length and high sequence identity of CNEs has made them recalcitrant to analyses that aim to identify a regulatory language encoded within them. The lamprey sequence, combined with functional assays, provides a new angle to this approach and may facilitate the identification of important functional motifs within CNEs.

The striking overlap of core conserved regions defined by either human-lamprey or human-human dCNE alignments suggests that these core regions may have a functional and evolutionary significance, rather than being an artefact of sequence divergence. These core regions may be more evolutionarily ancient or may be more recalcitrant to sequence divergence than their flanking sequences. It is interesting that, in some cases, the same core regions are conserved between ancient orthologs as well as paralogs, as these two pairs of elements might be expected to have faced different selective forces –

for instance, paralogous regulatory elements might reasonably be predicted to have functionally diversified upon duplication otherwise they would be redundant. In which case, the core regions may represent essential components of cis-regulatory modules, whose regulatory function may be modifiable by flanking sequences – the fact that some of these dCNEs also exhibit inter-species conservation in their flanking regions that is not conserved between paralogs could support this notion. It would be of interest to investigate the patterns of sequence conservation of lamprey CNEs and dCNEs more systematically and to address the functional significance of these patterns.

Investigation of the role and function of these ancient CNEs could provide a critical starting point for characterising ancient vertebrate developmental GRNs. Gene-regulatory interactions can be inferred by the characterisation of gene expression patterns and by gene perturbation (e.g. by morpholino injection). These experiments are a crucial source of information for investigating the conservation of GRNs across vertebrates, and even across metazoans. Nevertheless, CNEs represent a complementary, genomic source of information regarding GRN conservation. Due to their high sequence conservation and length, it is likely that these elements are regulated by complex combinations of factors. If this is the case, then lamprey CNEs may represent the pan-vertebrate conservation of not just a few gene-regulatory links, but of whole circuits. Therefore, lamprey CNEs represent a fantastic resource for the detailed investigation into the GRNs operating during the development of all vertebrates.

Conclusion

CNEs show extraordinary sequence conservation across jawed vertebrates but the vast majority show no traces in the genomes of invertebrate chordates. This is in stark contrast to the conservation patterns of developmental genes, many of which can be traced across the invertebrate-vertebrate boundary. We sought to further characterise the pattern of conservation of CNEs across the vertebrate phylogeny by capitalising on the emerging genomic sequence data for the sea lamprey, the phylogenetic position of which makes it an ideal model for characterising the early vertebrate genome. We identified a significant number of CNEs in the lamprey genome, which we predict to represent ancient regulatory instructions for the ancestral vertebrate body plan. Indeed, testing the regulatory functions of two of these CNEs in zebrafish embryos shows that they can act as developmental enhancers, sharing tight functional conservation across

highly divergent species. Having elucidated the pattern of CNE sequence conservation across vertebrates, the next challenge is to interpret the developmental and evolutionary significance of this pattern. We predict that the *C15orf41* contig will be a useful subject for investigating the developmental significance of the lack of many CNEs in lamprey. Additionally, the low sequence conservation and small size of the CNE regions defined by lamprey sequence conservation may be a useful guide for identifying crucial sequence motifs within these elements, which could elucidate details of the GRNs in which they are presumed to operate.

4 Functional Conservation of Lamprey CNEs

Abstract

The hundreds of CNEs that are conserved between gnathostomes and lamprey point towards developmental GRN circuits that are shared across all vertebrates. On the other hand, it is interesting that the majority of gnathostome CNEs could not be traced back to the vertebrate common ancestor using the lamprey genome. If CNEs represent conserved aspects of GRNs, then the opposite may also be true – that lack of CNEs could be indicative of divergent GRN architecture. Alternatively, elements homologous to CNEs may have been present in the vertebrate common ancestor but diverged in sequence between lamprey and gnathostome lineages, whilst retaining their developmental roles. In this chapter I investigate whether this could be the case, by characterizing the enhancer function of sequences of the *C15orf41* genomic region, identified in the previous chapter, from which many lamprey CNEs are missing. I find CNEs of this region to function as developmental enhancers, particularly of the hindbrain. For the regions that are missing CNEs in the lamprey genome, I find little evidence for conservation of enhancer function between gnathostomes and lamprey. Importantly, this investigation highlights a number of issues that must be resolved in order to systematically infer the role of CNEs in vertebrate development and evolution.

Introduction

In the previous chapter we uncovered hundreds of CNEs that are conserved between gnathostomes and lamprey, suggesting that aspects of a developmental program are highly conserved across all vertebrates. Indeed, for certain developmental pathways, gene expression and knockdown studies in lamprey have revealed certain links within the relevant GRNs to be conserved (Sauka-Spengler *et al.*, 2007; Murakami *et al.*, 2001). However, the majority of gnathostome CNEs could not be traced back to the lamprey genome using sequence conservation. This may reflect crucial differences between the GRNs governing the development of gnathostomes and agnathans. Alternatively, many gnathostome CNEs may have functional homologs in the lamprey genome that have diverged in sequence, as has been suggested to have occurred for

elements between mammals and fish (Fisher *et al*. 2006), and documented between flies (Hare *et al*., 2008).

In order to assess the functional significance of the lower CNE quota in lamprey, I have characterised *cis*-regulatory elements associated with *meis2* by reporter assay in zebrafish embryos. As described in the previous chapter, this region, containing the bystander gene *C15orf41*, is missing many gnathostome CNEs in lamprey. I can address the functional significance of this by testing the enhancer capabilities of lamprey regions from which CNEs are missing and comparing these to the enhancer activities of CNEs from the equivalent gnathostome regions.

The vertebrate *meis* family of homeobox genes comprises at least three paralogs (Steelman *et al*., 1997), which encode transcription factors related to the Drosophila gene homothorax (Rieckhof *et al*., 1997). These proteins can act as co-factors to Pbx (Rieckhof *et al*., 1997; Berthelsen *et al*., 1999; Chang *et al*., 1997) and Hox proteins (Shanmugan *et al*., 1999; Shen *et al*., 1999; Jacobs *et al*., 1999), forming DNA-binding hetero-dimers and hetero-trimers. Meis proteins have diverse roles in vertebrate development, including morphogenesis of the lens (Zhang *et al*., 2002) and retina (Bessa *et al*., 2008; Heine *et al*., 2008), proximo-distal patterning of the limbs (Mercader *et al*., 2000; Capdevila *et al*., 1999), specification of telencephalic (Toresson *et al*., 2000) and hindbrain domains (Waskiewicz *et al*., 2001; Choe *et al*., 2002) and in neural crest development (Maeda *et al*., 2002). These roles are reflected by the widespread expression patterns of *meis* genes during development (Zerucha & Prince, 2001; Biemar *et al*., 2001; Cecconi *et al*., 1997) and may underlie the particularly high number of CNEs associated with these genes in vertebrates (Woolfe *et al*., 2004).

Results

Multiple alignment of the *c15orf41* genomic region from vertebrates

As introduced in the previous chapter, a genomic region containing *c15orf41*, a bystander gene of *Meis2* (*Meis2.2* in zebrafish), can be delineated in the lamprey genome through sequence conservation of the exons of *C15orf41*. Many CNEs reside within this genomic region in gnathostomes, a sub-set of which are also identifiable in the equivalent lamprey region through sequence conservation (Figure 4.1). However, there are also some CNEs that cannot be identified in the lamprey region through sequence comparison. Importantly, the regions from which CNEs appear to be missing in lamprey can be defined by the conserved genomic landmarks that abut them. These landmarks are either conserved exons of *C15orf41*, or the visible lamprey CNEs. Specifically, gnathostome CNEs 3271-80, all reside within an intron of *C15orf41* (intron 7-8). Whilst the equivalent intronic sequence can be defined in the lamprey genome, homologs of these CNEs cannot be identified within this intron. The same case is found for intron 5-6, which contains CNEs 3281-3 in gnathostomes but not in lamprey. Finally, CNE 3286 is conserved in sequence between gnathostomes, but the equivalent region in lamprey, between CNEs 3285 and 3287, appears not to contain CNE 3286. As described in the previous chapter, searches for homologs of these CNEs in lamprey using BLAST did not produce any hits, suggesting that these CNEs have not moved to other genomic positions by rearrangement in lamprey. The ability to identify equivalent genomic regions between gnathostomes and lampey enables me to address whether these CNEs are functionally conserved in lamprey.

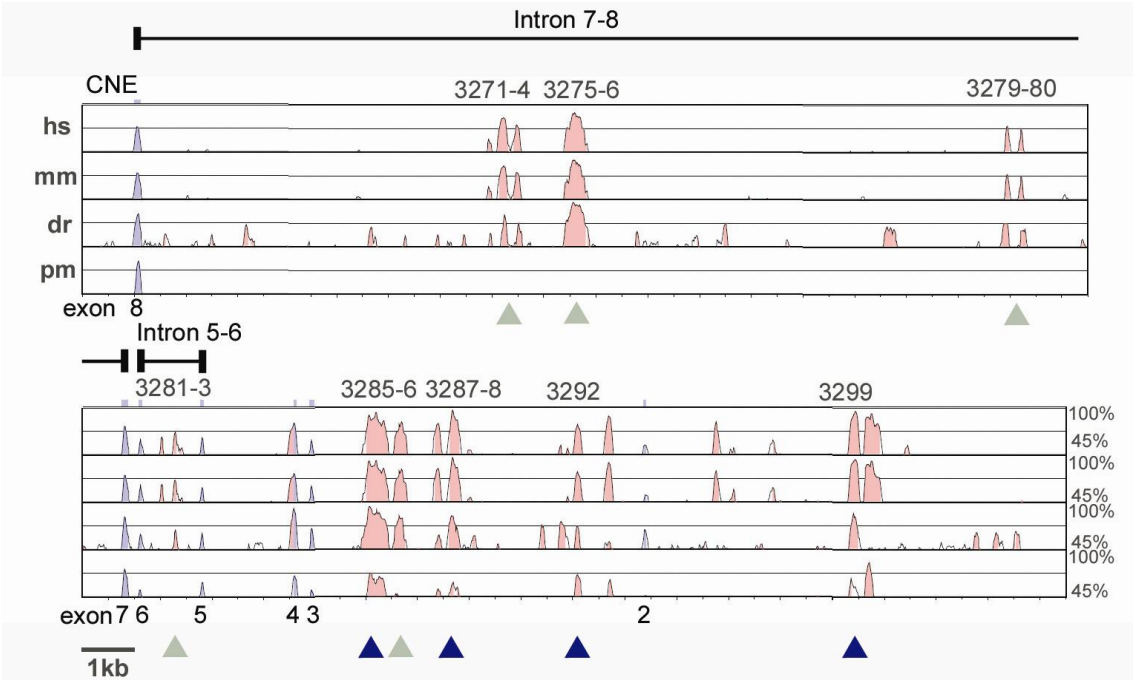


Figure 4.1. Multiple alignment of the genomic region containing the gene *C15orf41*. Genomic sequences from human (hs), mouse (mm), zebrafish (dr) and lamprey (pm) genomes are aligned with the fugu genomic region as a baseline. Exons of *C15orf41* are represented as blue peaks and are numbered below the alignment. Gnathostome and lamprey CNEs (pink peaks) are labeled above their positions in the alignment. The CNEs investigated in this chapter are highlighted, being conserved across gnathostomes and lamprey (dark blue arrowheads) or across gnathostomes but not lamprey (grey arrowheads). The introns tested for enhancer activity in this investigation (intron 5-6 and intron 7-8) are delineated and labeled above the alignment.

Functional conservation between zebrafish and lamprey CNEs

I first sought to verify that orthologous zebrafish and lamprey CNEs from the *C15orf41* contig represent functionally conserved *cis*-regulatory elements. I selected four pairs of zebrafish (dr) and lamprey (pm) CNEs for reporter assay – CNEs 3285, 3287-8, 3292 and 3299. Each of these elements was found to drive GFP expression in the hindbrain and spinal cord, consistent with the endogenous expression of *meis2.2* in zebrafish (Figure 4.2). Furthermore, orthologous zebrafish and lamprey elements drove highly similar expression patterns. Interestingly, some of these elements created quite restricted patterns, such as CNE 3299 in the anterior hindbrain. CNE 3299 is also noteworthy for the lamprey and zebrafish orthologs driving slightly different expression patterns, with

dr3299 driving expression in the branchial arches whilst that of pm3299 was restricted to the hindbrain.

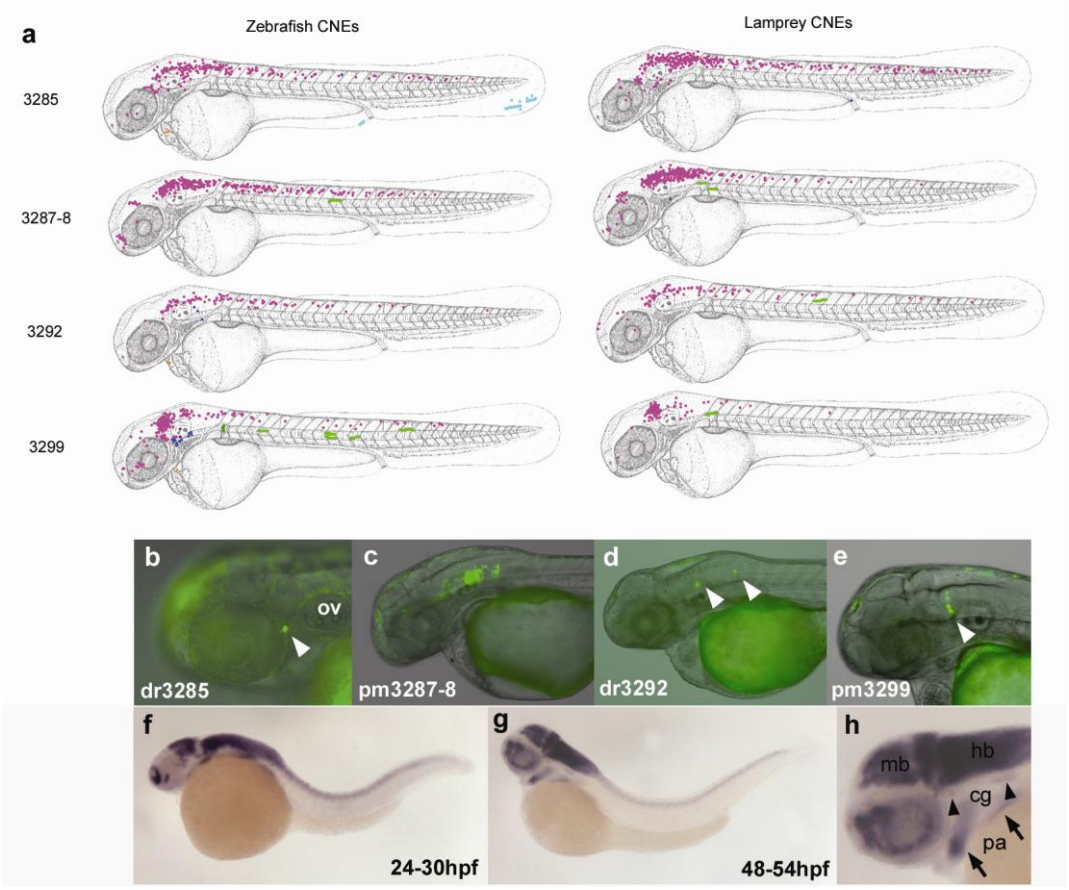


Figure 4.2. Patterns of GFP expression driven by orthologous zebrafish and lamprey *meis2* CNEs. Expression is shown for 48-54hpf zebrafish embryos. **a** Composite expression patterns for four pairs of homologous zebrafish (left) and lamprey (right) CNEs. **b** GFP expression in a neuron of the cranial ganglia, driven by element dr3285. **c-e** GFP expression in neurons of the hindbrain driven by lamprey (**c**, **d**) and zebrafish (**d**) meis2 CNEs. **f-h** Endogenous expression pattern of *meis2.2* in zebrafish embryos at 24-30hpf (**f**) and 48-54hpf (**g,h**) revealed by *in-situ* hybridization. The expression of *meis2.2* in the head at 48-54hpf is detailed in panel **h**, highlighting expression in the pharyngeal arches (pa – arrows) and cranial ganglia (cg – arrowheads) . mb: mid-brain, hb: hindbrain. The zebrafish *in-situ* data was obtained from Zebrafish Model Organism Database (ZFIN), University of Oregon, Eugene, OR 97403-5274; URL: http://zfin.org/)

Functional investigation of intron 5-6

The 1.2kb fugu intron (frIntron5-6), containing three elements conserved between mammals and fugu (CNEs 3281, 2 and 3), up-regulated GFP expression in the nervous system, particularly in the hindbrain and secondary neurons of the spinal cord, with greater activity at 48 hpf (Figure 4.3). Surprisingly, the 550 bp lamprey intron (pmIntron5-6) was also capable of driving GFP expression in a tissue specific manner in zebrafish; however, the expression domains have little similarity to those of the fugu intron, being the cranial ganglia, Rohon-Beard neurons of the spinal cord, and the pectoral fin, with expression also in the tail and skin, particularly at 24 hpf.

In order to assess the contribution of CNEs to the enhancer activity of the intron, the CNEs within intron 5-6 from fugu and zebrafish were tested for enhancer function. A 450 bp sequence containing the three Fugu CNEs (fr3281-3) drove GFP expression in a manner very similar to the pattern produced by the fugu intron (Figure 4.3), with a slightly higher proportion of embryos expressing in the fore- and mid-brain at 48 hpf and expression in the muscle and skin at 24hpf. The intronic flanking sequence either side of fr3281-3 was not capable of up-regulating GFP expression in this assay. From this I can infer that the CNEs within the fugu intron are largely responsible for the expression pattern of the intron.

To ascertain whether the CNEs of this intron have conserved enhancer function across gnathostomes I tested the orthologous zebrafish CNEs for enhancer activity. The MLAGAN alignment identified one CNE within the ~5kb zebrafish intron - dr3282. Close inspection of the zebrafish intronic sequence revealed fragments of CNE 3281 to be conserved beside dr3282. A sequence encompassing dr3281 and dr3282 (dr3281-2) drove GFP expression in domains that are highly consistent with those obtained for frIntron5-6 and fr3281-3, except for some additional muscle expression. This suggests that the *cis*-regulatory function of the CNE module is conserved between teleosts despite divergence in sequence (partial loss of dr3281 and loss of dr3283 in zebrafish), yet this function is not conserved in the divergent lamprey intron.

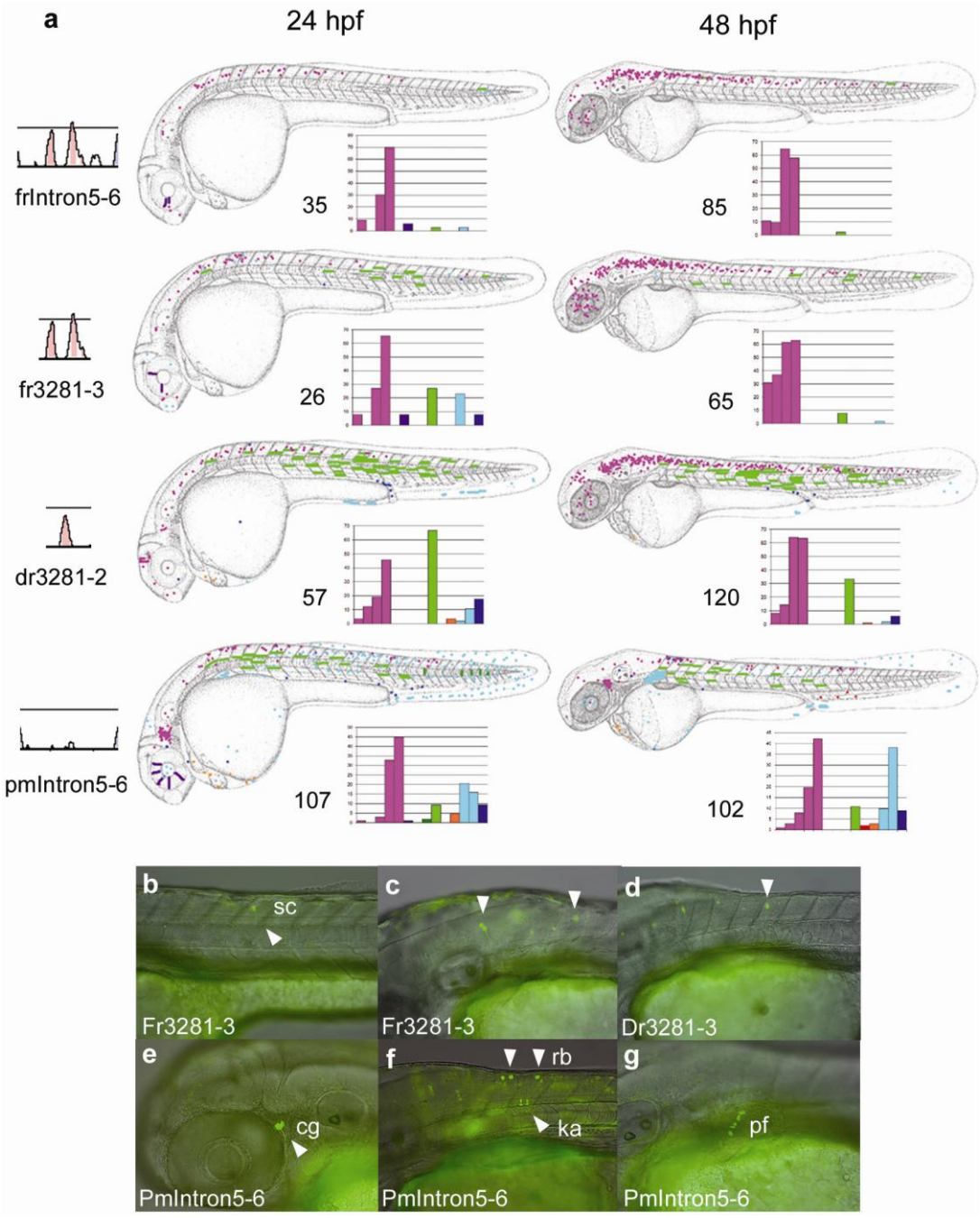


Figure 4.3. The enhancer function of *C15orf41* intron 5-6 is conserved between gnathostomes but not in lamprey. **a** Composite GFP expression patterns driven by homologous introns from fugu (frIntron5-6) and lamprey (pmIntron5-6) and by CNEs within these introns from fugu (fr3281-3) and zebrafish (dr3281-2) at 24 and 48 hpf. The charts on the left indicate CNEs within the regions injected and are in the MLAGAN output format, as in Figure 4.1. The number of GFP-expressing embryos from which the composites are compiled is shown for each element injected and for both timepoints. **b-g** GFP expression at approximately 48-54hpf. The element injected is identified on each panel. Arrow heads in **c** and **d** point to neurons in the spinal cord with ventrally projecting axons. Abbreviations: ka – Kolmer-Agduhr neuron, pf – pectoral fin, rb – Rohon-Beard neuron, sc – spinal cord, cg – cranial ganglion cell.

Functional investigation of CNE 3286

A lamprey homolog of CNE 3286 was unable to be identified using MLAGAN (Figure 4.4). I functionally investigated the regions between CNEs 3285 and 3287 from Fugu (fr3286+ length: 883bp) and lamprey (pm3286+ length: 1107bp) as well as CNE 3286 from fugu (fr3286 length 263bp) and zebrafish (dr3286 length: 219bp), by reporter assay. dr3286 and fr3286 drove similar expression in the hindbrain and spinal cord, mostly at 48 hpf (Figure 4.4). Expression was also evident in the developing gut for the zebrafish element. fr3286+, which abuts CNEs 3285 and 3287, up-regulated strong expression in the hindbrain and spinal cord at 48 hpf and weak expression in the gut, but also produced expression in the muscle and notochord. Intriguingly, the lamprey region, pm3286+, also drove expression in the hindbrain and spinal cord, with expression in the gut. This similarity of expression patterns driven by gnathostome CNEs and the lamprey region suggests that they may derive from an ancestral element which has been conserved in function despite sequence divergence between the two lineages. It is notable that short lengths of sequence conservation, which may represent conserved TFBSs, are visible upon close scrutiny of the aligned zebrafish, fugu and lamprey CNE 3286 sequences (see appendix).

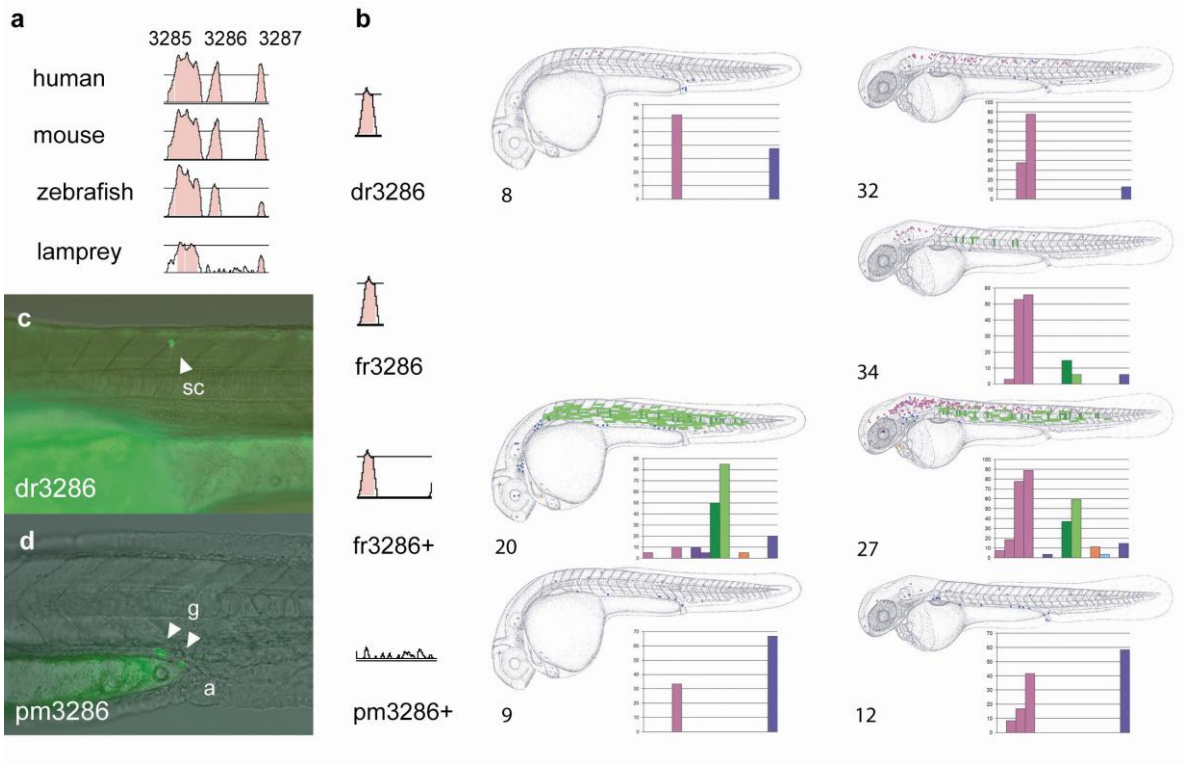


Figure 4.4. CNE 3286 provides evidence for functional conservation despite sequence divergence between gnathostomes and lamprey. **a** MLAGAN alignment of the genomic region containing CNEs 3285-7 in vertebrates, showing sequence homology to CNE 3286 to be absent in the lamprey region. **b** Composite GFP expression patterns driven by elements from zebrafish, fugu and lamprey. **c-d** GFP fluorescence in 48-54hpf zebrafish embryos. Abbreviations: a – anus, g – gut, sc – spinal cord.

Functional investigation of intron 7-8

I next tested the enhancer potential of lamprey intron 7-8 by dividing it into six abutting sequences of approximately 1.2 kb for functional assay. The same approach was not feasible for the equivalent intron in fugu or zebrafish due to the larger sizes of this intron in gnathostome genomes. Instead, three zebrafish sequences containing the CNEs within the intron were selected for reporter assay. Of these three elements, two acted as enhancers in our assay, with dr3271-4 producing no GFP expression. dr3275-6 drove expression in the telencephalon, hindbrain, muscle and tail, with dr3279-80 up-regulating GFP in the cranial ganglia, hindbrain and spinal cord (Figures 4.5 and 4.6).

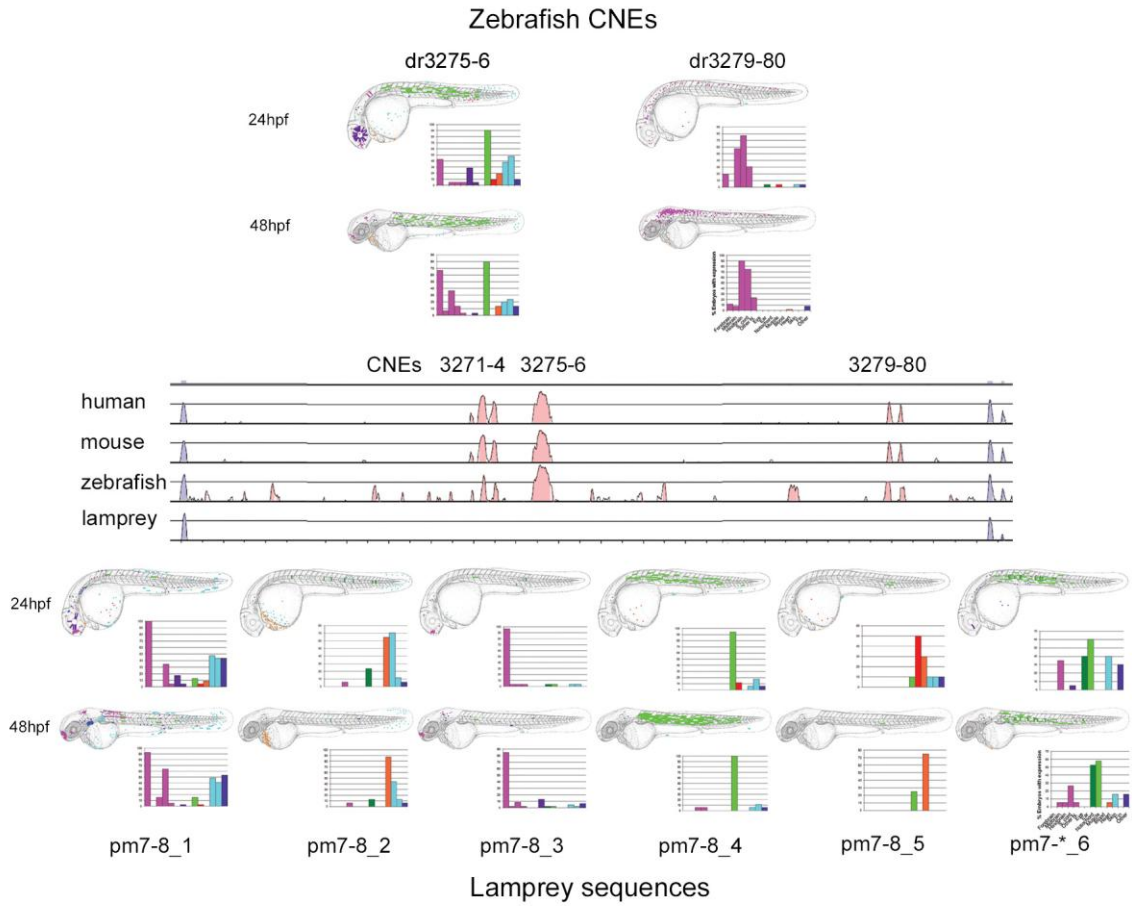


Figure 4.5. Lamprey sequences within intron7-8 act as tissue specific enhancers in zebrafish. Composite GFP expression patterns are shown for two zebrafish CNE modules within the intron (top) and six approximately 1.2 kb lamprey sequences comprising the lamprey intron (bottom). The MLAGAN alignment highlights the CNEs from zebrafish that were functionally tested.

Each of the lamprey sequences tested was capable of up-regulating GFP expression in zebrafish (Figure 4.6). pm7-8_1 drove highly consistent and strong expression in the olfactory organ, as well as in KA neurons of the spinal cord, in the pectoral fin, tail and posterior branchial arches (Figure 4.5 and Figure 4.6). pm7-8_3 drove GFP expression in the same domain as one of the zebrafish CNEs of intron 7-8, driving consistent expression in the telencephalon, in a similar fashion to dr3275-6 (Figures 4.5 & 4.6).

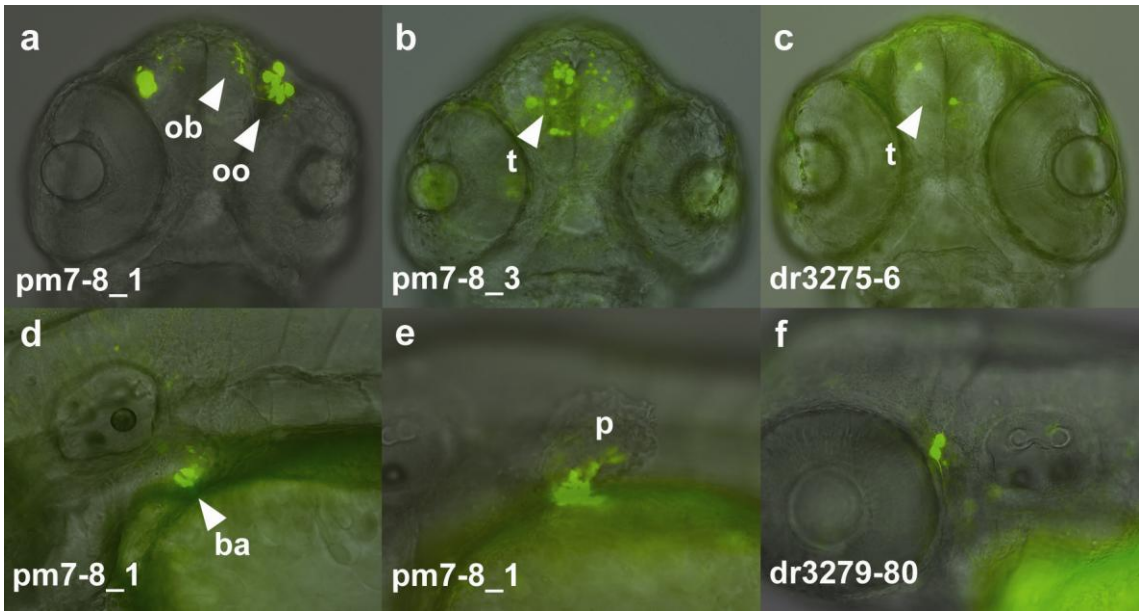


Figure 4.6. Lamprey enhancers and zebrafish CNEs of intron 7-8 drove tissue specific expression in zebrafish embryos. **a-f** GFP fluorescence of zebrafish embryos at 48-54 hpf. **a-c** Ventral views of the head, with anterior to the to the top. **d-f** Lateral views with anterior to the left. The element injected is identified at the bottom left of each panel. Abbreviations: ba – branchial arch, ob - olfactory bulb, oo – olfactory organ, p – pectoral fin, t – telencephalon.

Thus, lamprey genomic sequences that bear no overt homology to those of gnathostomes were able to drive a functional output that was, in some cases, very strong and tissue specific. Furthermore, one of the tissues in which expression was up-regulated was also seen to express GFP when gnathostome CNEs from the homologous intron were functionally tested, which could represent a degree of conservation of intronic function despite sequence divergence between lamprey and gnathostomes.

Discussion

Conservation of sequence and function in lamprey CNEs

The prediction that lamprey CNEs represent conserved *cis*-regulatory elements (and therefore conserved GRN architectures) that are shared between all vertebrates assumes firstly that they are *cis*-regulatory elements and secondly that their sequence conservation correlates with functional conservation within their respective species. It was shown in the previous chapter that homologous elements from human and lamprey drive highly similar expression patterns when tested in a zebrafish reporter assay. I have

now shown, with a larger number of CNEs, that homologous zebrafish and lamprey elements associated with *meis2* drive consistent expression patterns when assayed in zebrafish. This confirms the assumption that these sequences can act as *cis*-regulatory elements and suggests that they are functionally conserved across vertebrates (functional testing by reporter assay in lamprey embryos would confirm this).

The expression patterns of the *meis2* CNEs are consistent with the endogenous pattern of expression of *meis2.2* in zebrafish (Figure 4.2) and *meis2* in mouse. CNE 3299 drives particularly interesting expression that appears to be restricted to the anterior hindbrain, making it a good candidate for functional dissection of potential TFBSs. A limitation of our assay is that detailed characterisation of expression domains is hindered by the high mosaicism of reporter expression. Further, during the creation of composite expression patterns, it is difficult to map positions of GFP-expressing cells with great accuracy. This makes it hard to interpret reporter expression patterns within the context of the expression of particular developmental genes. Thus, these expression patterns give a broad indication of tissue specificity but detailed analysis would require the use of a reporter assay that could provide less mosaic reporter expression.

Non-conserved lamprey sequences can function in zebrafish

A notable finding from this investigation is that lamprey sequences that are not conserved in gnathostomes are nevertheless able to upregulate GFP expression in a tissue-specific manner in zebrafish embryos. In fact, every lamprey sequence that was tested showed enhancer function. Controls in which non-conserved non-coding fugu sequences of comparable sizes to CNEs were tested for enhancer activity in zebrafish yielded no significant GFP expression using the co-injection assay (Woolfe *et al.*, 2004). In contrast, zebrafish non-conserved sequences have been demonstrated to have regulatory function in zebrafish embryos, although at a lower proportion compared to conserved elements around the same gene (McGaughey *et al*, 2008).

The lamprey elements frequently up-regulate GFP expression in domains that are consistent with *meis* gene expression in gnathostomes, for instance in the spinal cord, telencephalon, cranial ganglia, branchial arches and pectoral fin. Furthermore, some of the gnathostome CNEs drive expression in these domains too; for example in the telencephalon (dr3275-6) and cranial ganglia (dr 3285, dr3279-80). As for the

expression in the pectoral fin, which is found for two of the lamprey elements (pmIntron5-6 and pmIntron7-8_1), the role of MEIS proteins in specifying the proximo-distal axis during vertebrate limb development has been well documented (e.g. Capdevila *et al.*, 1999). Whilst lamprey does not possess paired appendages, there is evidence to suggest that aspects of the GRN governing vertebrate paired limb development originally functioned in the median fins of early vertebrates (Freitas *et al.*, 2006). Furthermore, ABD-B class HOX proteins (groups 9 and 10), which are capable of forming DNA-binding complexes with MEIS proteins (Shanmugam *et al*. 1999), are expressed in the median fin of lamprey (Freitas *et al.*, 2006). Thus, the functions of these lamprey elements in zebrafish may be consistent with ancestral developmental roles of Meis genes that have been retained in gnathostomes and agnathans, whilst the regulatory elements responsible for these roles may have diverged and changed position.

Although many of the expression domains of lamprey elements are consistent with endogenous *meis* expression in gnathostomes, some domains are not. For example, pm7-8_1 strongly and consistently expressed GFP in the zebrafish olfactory organ, however this is not a domain of endogenous *meis* expression in zebrafish. It is important to note that significant changes in trans-regulatory state may have occurred between agnathans and gnathostomes. If this is the case, the expression driven by lamprey elements in our zebrafish reporter assay would only represent half of the picture. In order to interpret these gene-regulatory elements within the context of cis- and trans-regulatory divergence, it is necessary to test their behavior in a lamprey reporter assay. This would inform us as to the relevance of their expression patterns in zebrafish and would illuminate the degree of similarity of the lamprey and gnathostome gene-regulatory architectures (this issue is focused upon, with regard to the conservation of CNE function between lamprey and gnathostomes, in chapter 6).

Conservation of function despite sequence divergence between gnathostome and lamprey enhancers

The key question that I posed regarding the CNE quota in the lamprey genome was whether *cis*-regulatory function is conserved between CNE-harboring regions from gnathostomes and the homologous lamprey regions. I addressed this by assaying equivalent genomic regions from gnathostomes and lamprey for reporter expression in

zebrafish. The data obtained provide evidence for both conservation and divergence of function. For CNE 3286, there is evidence for conservation of spinal cord and gut enhancer function. Within intron 7-8 the elements pm7-8_3 and dr3275-6 both drive GFP expression in the telencephalon despite no overt sequence similarity. However, similarity of function between homologous regions when tested in zebrafish does not necessarily mean *conservation* of function. Without knowledge of their regulatory potential in lamprey embryos, or of their mechanism of action, it is hard to interpret whether these are functionally conserved elements. The case for conservation of function is stronger for CNE 3286, as there are chunks of sequence similarity between the gnathostome and lamprey sequences. Further evidence for functional conservation could be obtained by characterizing whether these stretches of sequence conservation contain TFBSs that contribute to the enhancer function of these elements. In the case of intron5-6, there is no evidence for functional conservation between lamprey and gnathostomes, suggesting that CNEs either evolved in sequence and function within this region in the gnathostome lineage, or that they diverged in sequence and function in the lamprey lineage.

It should be noted that the approach toward characterizing the enhancer function of zebrafish intron 7-8 has limitations. I have assumed that the three CNE modules chosen contain the core functional elements within the intron and thus represent its transcriptional regulatory function. I have shown this to be valid for fugu intron 5-6, but that intron is relatively small compared to the size of the CNEs within it. It is not certain whether this rule holds up for larger introns such as intron 7-8. Recent investigations into the sequence conservation of functional elements show many functional elements not to be conserved (e.g. McGaughey *et al.*, 2008), meaning that the analysis of zebrafish intron7-8 function is not exhaustive and that there may be additional (or fewer) domains of expression regulated by the zebrafish intron.

Finally, the sea lamprey has recently been found to undergo extensive programmed genomic re-modelling during development, involving the loss of a significant proportion (>20%) of its germ-line DNA, including transcribed regions (Smith *et al.*, 2009). The lamprey tissue from which genomic DNA was extracted for the shotgun sequencing project was derived from the adult liver, meaning that it may represent a lamprey genome from which crucial developmental components, such as CNEs, have been lost. In this light, it is possible that the low CNE quota that we have found in the

lamprey somatic genome, and the functional ramifications that I show in this chapter, are partly the result of extensive loss of genomic DNA during lamprey development. Genomic sequence data from the lamprey germ-line would be required in order to thoroughly address this possibility.

Conclusion

I set out to investigate the gene-regulatory significance of the absence of a number of gnathostome CNEs from the lamprey genome. I addressed this by comparing the regulatory functions of equivalent gnathostome and lamprey regions in the vicinity of gnathostome *meis2* and its lamprey homolog. Of note, I have found that most, if not all, of the lamprey regions tested are able to drive GFP expression in a tissue-specific manner in zebrafish embryos, suggesting that these regions contain *cis*-regulatory elements in lamprey. Yet, for many of these, the significance of this expression is difficult to interpret without characterising their regulatory activity in lamprey embryos. I have found evidence hinting at the conservation of enhancer function upon sequence divergence between lamprey and gnathostomes. Yet, proof of conserved regulatory mechanisms would require detailed characterization of TFBSs within these elements, which is difficult given our lack of knowledge of their mechanism of action. I also find evidence against functional conservation of equivalent gnathostome and lamprey regions. This suggests that the absence of gnathostome CNEs from lamprey could, in some cases, indicate significant differences between developmental programs of lamprey and gnathostomes. However, the paucity of functional data regarding the regulatory roles of gnathostome CNEs makes it difficult to interpret which specific gene-regulatory interactions may differ between lamprey and gnathostomes. I have also gleaned further evidence for conserved regulatory functions between pan-vertebrate CNEs. Characterisation of these conserved elements in more detail could reveal deeper insights into the developmental mechanisms that are highly conserved across all vertebrates.

5 Pbx-Hox Motifs in CNEs

Abstract

Whilst CNEs hold a lot of promise for de-coding vertebrate gene regulation and inferring developmental GRNs, our knowledge of their functions and mechanisms of action remains poor. Without knowledge of how CNEs work and what their functions are, it is hard to interpret their significance in development and evolution. This chapter describes the characterisation of TFBS motifs within CNEs of the *c15orf41* gene region. Using sequence data from the sea lamprey genome to facilitate phylogenetic footprinting, I uncover Pbx-Hox motifs within these CNEs. I further discover a striking enrichment for Pbx-Hox motifs across the whole vertebrate CNE set. These motifs are found to be correlated with reporter expression in the hindbrain and pharyngeal arches during development. The implications that these findings have for the functions and mechanism of action of CNEs, as well as their evolutionary significance, are discussed.

Introduction

Identifying TFBSs within CNEs

The utility of inter-specific sequence conservation for the identification and characterisation of vertebrate *cis*-regulatory elements was recognised prior to the genome-wide identification of CNEs. A number of studies had already identified deeply conserved developmental enhancers and used phylogenetic footprinting as a guide to characterise crucial transcription factor binding sites within them (e.g. Pöpperl *et al*., 1995). However, these studies were mostly restricted to characterising individual elements, so it was not clear to what extent their findings were applicable to other *cis*-regulatory elements in the genome.

The identification of thousands of CNEs associated with developmental genes has provided genome biologists with a wealth of putative *cis*-regulatory sequences to investigate. These elements are predicted to represent crucial links in gene regulatory networks (GRNs) that underlie conserved aspects of vertebrate development. In order to place these links within the context of GRNs it is essential to uncover the factors

binding to CNEs, through identification of the TFBS motifs that they recognise. However, the majority of CNEs have not been functionally characterised by reporter assay, nor dissected at the level of TFBSs. Whilst early investigations into *cis*-regulatory mechanisms uncovered TFBS motifs for a range of factors, it has not been clear whether the information gleaned from these studies is applicable to CNEs – indeed, the notion that CNEs are composed of 'orthodox' TFBSs has been called into question. The main reason for this is that their high sequence conservation across distantly related species is hard to reconcile with the prevailing wisdom of how transcription factors bind to DNA, in which a given transcription factor is able to recognise short, degenerate sequence motifs, allowing enhancer sequences to diverge through mutation and shuffling of TFBSs, whilst maintaining the same function. Furthermore, attempts to systematically uncover enriched sequence motifs within mammal-fish CNEs, by 'top-down' *de-novo* motif discovery, have so far been relatively unsuccessful compared to similar approaches applied to mammalian promoter elements.

Thus, whilst CNEs are considered to hold a lot of promise for de-coding vertebrate gene regulation, our lack of knowledge of their mechanism of action makes it unclear whether the *cis*-regulatory language of well characterised enhancers is also used by CNEs, and whether any language that may be present in CNEs is applicable to less well conserved *cis*-regulatory elements. The low number of well characterised CNEs and the failure to systematically uncover enriched sequence signatures within CNEs have exacerbated the problem. Without detailed knowledge of how CNEs work and what their developmental functions are, it is difficult to place them within the context of developmental GRNs, and to interpret their roles in evolution.

The lack of knowledge of CNE function is being rectified by projects to systematically assay CNEs for enhancer activity in mouse (Pennacchio *et al.*, 2006) and zebrafish embryos (Li *et al.*, 2010), which will provide useful data enabling CNEs to be grouped according to their expression patterns, with the aim of scanning these groups for TFBS motifs. This 'bottom-up' approach - starting with a group of CNEs with similar function and identifying TFBS motifs that they have in common - is a complementary alternative to the 'top-down' method of *de-novo* identification of TFBS motifs across the whole CNE set (addressed in chapter 6).

Regulation by Hox factors through the Pbx-Hox TFBS motif

A few focused investigations have used phylogenetic footprinting to identify TFBSs within selected *cis*-regulatory elements that are conserved between distantly related vertebrates. A particularly well characterised TFBS motif is that recognised by the Pbx-Hox heterodimer (Popperl *et al.*, 1995; Tümpel *et al.*, 2006). Hox factors binding as monomers share similar binding preferences to each other, varying around a core TAAT motif (Noyes *et al.*, 2008; Berger *et al.*, 2008). This leads to the question of how different Hox factors are able to control the expression of different target genes. Part of the answer involves their interaction with co-factors belonging to the 'three amino-acid loop extension' (TALE) class of homeodomain proteins (including Pbx and Meis/Prep, and the drosophila homologs Exd and Hth) (Mann & Chan, 1996; Jacobs *et al.*, 1999; Moens & Selleri, 2006). Hox factors belonging to vertebrate paralogy groups 1-8 contain a conserved hexapeptide upstream of their homeodomain, which interacts with a hydrophobic pocket within the TALE domain of Pbx, leading to the formation of Pbx-Hox hetero-dimers (Shanmugan *et al.*, 1999; Joshi *et al.*, 2007). The interaction with Pbx increases the binding affinity and selectivity of the Hox factor to DNA, with Pbx-Hox heterodimers recognising an 8bp core DNA sequence motif: TGATNNAT (Chan & Mann, 1996). Pbx binds the TGAT half-site and Hox binds the NNAT half site, with different Hox factors having different binding preferences for the variable (NN) positions of the motif (Chan *et al.*, 1997). The two positions immediately 3' of the core Pbx-Hox site also influence binding of the heterodimer, with many characterised binding sites containing G/T and G/A at these two positions, although these positions are more variable than those of the core motif (Mann *et al.*, 2009).

Many *cis*-regulatory elements have been identified that are regulated by Pbx-Hox complexes, often in conjunction with Meis/Prep factors, which contribute to a ternary complex by interacting with PBX via n-terminal domains, and can be either DNA bound, to a typical TGACAR motif, or unbound (e.g. Ferretti *et al.*, 2000). As Pbx-Hox regulated elements have been characterised in *Drosophila* and *C. elegans*, these complexes constitute an ancient mechanism of gene regulation, conserved across bilaterians (Mann *et al.*, 2009).

In vertebrates, the majority of characterised Pbx-Hox TFBSs reside within elements that are involved in regulating the Hox genes themselves, mediating auto- and cross-

regulatory interactions between anterior hox factors and driving gene expression in the hindbrain and neural crest (reviewed by Tümpel *et al*., 2009). Nested patterns of Hox expression that are set up in the hindbrain contribute, via migratory neural crest cells, to anterior-posterior (A-P) patterning of the cranial ganglia, pharyngeal arches and their derivatives, and facilitate the connection of sensory and motor circuits between the hindbrain and tissues of the head and neck (Alexander *et al*., 2009).

Whilst Hox genes play crucial and evolutionarily conserved roles in A-P patterning of vertebrate embryos, our knowledge of their downstream targets is poor. Surprisingly, despite the identification of deeply conserved Pbx-Hox TFBSs in vertebrate enhancers, and our detailed knowledge of Pbx-Hox-DNA interactions, a systematic search for Pbx-Hox motifs in vertebrate conserved regulatory elements has not previously been carried out.

Results

CNEs from the *C15orf41* contig drive expression in the nervous system, especially in the hindbrain

As described in chapters 3 and 4, the genomic region downstream of *meis2* in jawed vertebrates, containing the bystander gene *C15orf41*, contains a number of lamprey CNEs (McEwen *et al*., 2009). In the human genome, these CNEs lie within an approximately 11kb stretch of sequence, roughly 450kb downstream of the *meis2* transcriptional start site (depicted in chapter 3, figure 3.1). In chapter 4, some of these elements from both zebrafish and lamprey were demonstrated to function as enhancers in zebrafish through our co-injection assay. These CNEs were chosen as a starting point for a bottom-up search for TFBSs in CNEs for two reasons. Firstly, they show clear and specific expression patterns in the hindbrain of zebrafish embryos in our co-injection assay, making them good subjects for functional dissection; and secondly, their conservation in lamprey facilitates phylogenetic footprinting, as alignments that include lamprey CNE sequences have lower overall sequence conservation, making crucial sequence motifs likely to stand out as highly conserved regions of the alignment (McEwen *et al*., 2009).

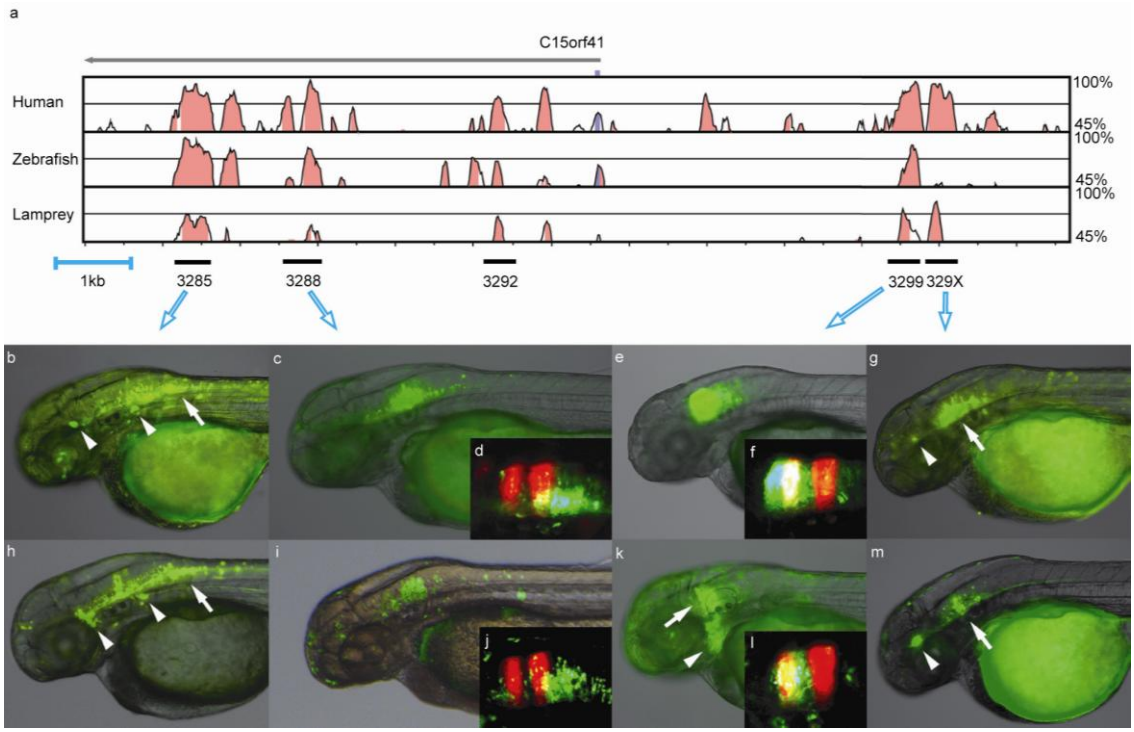


Figure 5.1. Reporter expression driven by CNEs of the *C15orf41* contig. The *C15orf41* contig contains a number of lamprey CNEs, which drive expression in the nervous system of zebrafish embryos, especially in the hindbrain. **a** multiple alignment of orthologous genomic regions containing the gene *C15orf41* (blue peak), downstream of *meis2*, revealing CNEs (red peaks). Human, zebrafish and lamprey sequences are aligned with the fugu sequence as a baseline. Zebrafish CNE 329X is translocated in the zebrafish genome assembly so does not appear in this alignment. **b-m,** orthologous elements from lamprey (**b**-**g**) and zebrafish (**h**-**m**) drive similar GFP expression patterns in the nervous system of zebrafish embryos at 54hpf: element 3285 in the cranial ganglia (arrowheads) and primary neurons of the hindbrain and spinal cord (arrows) (**b,h**); 3288 in the hindbrain posterior to rhombomere (r) 4 (**c,i**), as determined by comparison with r3r5-RFP expression (red) (**d,j**); 3299 in the anterior hindbrain – r2-4 for the lamprey homolog (**e,f**) and r3-4 plus the corresponding neural crest for the zebrafish homolog (**k,l**); 329X in the hindbrain (arrow) and neurons of the midbrain (arrowhead) (**g,m**).

In order to investigate their *cis*-regulatory function in greater detail, their enhancer activities were tested using the less mosaic tol2 reporter assay (Fisher *et al*. 2006). Using the tol2 system, injected embryos (F0) often show strong reporter expression with low mosaicism (see Materials and Methods). To obtain 'full transgenic' specimens, the F0 embryos can be grown to adulthood and mated with wild-type fish to create F1 progeny, a proportion of which will inherit the integrated reporter construct through the germline and would be expected to show clear, non-mosaic GFP expression. Whilst the

patterns of reporter expression observed for a given construct in F0 embryos often vary between individuals, probably due to position effects and multiple integrations, it has been shown that expression patterns shared across greater than 25% of F0 embryos are representative of the expression patterns found in F1 progeny (McGaughey *et al*, 2008). Thus, consistent patterns of reporter expression in F0 embryos are deemed to give a reliable indication of the expression that would be observed in full transgenics (and therefore the endogenous activity of the enhancer).

Five pairs of orthologous zebrafish and lamprey elements from the *c15orf41* contig were tested using the tol2 system. These included the element 329X, which, whilst not present in the Condor database, is conserved between gnathostomes and lamprey. In zebrafish it has translocated so does not align in the zebrafish trace of the MLAGAN alignment (Figure 5.1). Furthermore, it is transcribed in zebrafish so is not considered to be a non-coding element. However, it has a non-coding duplicate associated with *meis1* and has been predicted to be a *cis*-regulatory element overlapping with an exon (Dong *et al*., 2009), so was included in this functional analysis.

Four pairs of elements drive discreet and complementary patterns of reporter expression in the hindbrain of zebrafish embryos, which are consistent with the endogenous expression pattern of *meis2.2* in zebrafish (Figure 5 and compare to Figure 4.2). The expression patterns that they drive when tested using the tol2 system show close correspondence with those from the co-injection assay (Figure 4.2 – previous chapter), confirming the consistency of functional readout from these elements between the two different reporter assay approaches. Homologous zebrafish and lamprey elements drove similar expression patterns, yet for each pair there were clear differences between homologs. dr3285 and pm3285 both drove expression in the crania ganglia and primary neurons of the hindbrain and spinal cord, but pm3285 also frequently drove expression in non-neural domains such as muscle and eye, whilst these expression domains were not observed for dr3285. dr3288 and pm3288 both drove expression in the hindbrain, posterior to rhombomere (r) 4 (as shown by comparison with RFP expression in r3 and r5 (using the r3r5RFP transgenic line (see Materials and Methods)), but the expression driven by pm3288 was broader than that of dr3288. dr3299 and pm3299 both drove reporter expression in the anterior hindbrain. By comparison with r3r5 RFP expression, dr3299 was seen to direct expression to r3-4 and pm3299 to r2-4, with dr3299 also driving GFP expression in the neural crest populating the second pharyngeal arch,

whilst pm3299 drove no neural crest expression. Both dr329X and pm329X drove GFP expression in the ventral anterior hindbrain and to neurons of the midbrain, however the expression driven by dr329X in the hindbrain was more restricted.

CNEs from the *C15orf41* contig contain conserved Pbx-Hox and Meis TFBS motifs

The rhombomere-specific expression patterns of dr3299 and pm3299 are reminiscent of the expression patterns of anterior Hox factors. Searching for canonical Pbx-Hox motifs within CNE 3299 identified two motifs matching the TGATNNAT consensus, which are highly conserved between mammals, teleosts and lamprey (Figure 5.2). Furthermore, these two Pbx-Hox motifs are each flanked by a highly conserved Meis motif (TGACAG/A). Thus, these motifs were considered to be strong candidates for TFBSs necessary for the rhombomere-specific expression of dr3299 and pm3299. To test this, versions of the zebrafish element were created in which these motifs had been mutated. In dr3299_sub1, the 5' Meis motifs and the 5' Pbx-Hox motif were mutated, leaving the other cluster of meis and Pbx-Hox motifs unchanged. This element was found to drive a broader expression pattern in the hindbrain compared to the wild type element, extending anteriorly into r2, whilst the neural crest expression of the wild type element was absent in dr3299_sub1 (Figure 5.2). Mutating the second cluster of Pbx-Hox and Meis motifs abrogated reporter expression altogether (not shown).

Surprisingly, the other CNEs of the *C15orf41* contig, which act as enhancers in the zebrafish hindbrain, were also found to harbour Pbx-Hox and Meis motifs that are conserved between gnathostomes and lamprey. Of these elements, CNE 3285 was selected for functional dissection as it has an expression pattern dissimilar to that of CNE 3299 – driving GFP expression in the cranial ganglia and primary neurons of the hindbrain and spinal cord. This element contains a single conserved Pbx-Hox motif and a distal Meis motif (Figure 5.2). Interestingly, mutating the Pbx-Hox motif within dr3285 severely reduced its ability to drive reporter expression in each of the expression domains of the wild type enhancer. However, the mutations introduced in dr3285 sub1 extend beyond the Pbx-Hox motif and into a highly conserved 5' homeobox motif, so the loss of reporter expression may not be solely attributable to mutation of the Pbx-Hox motif.

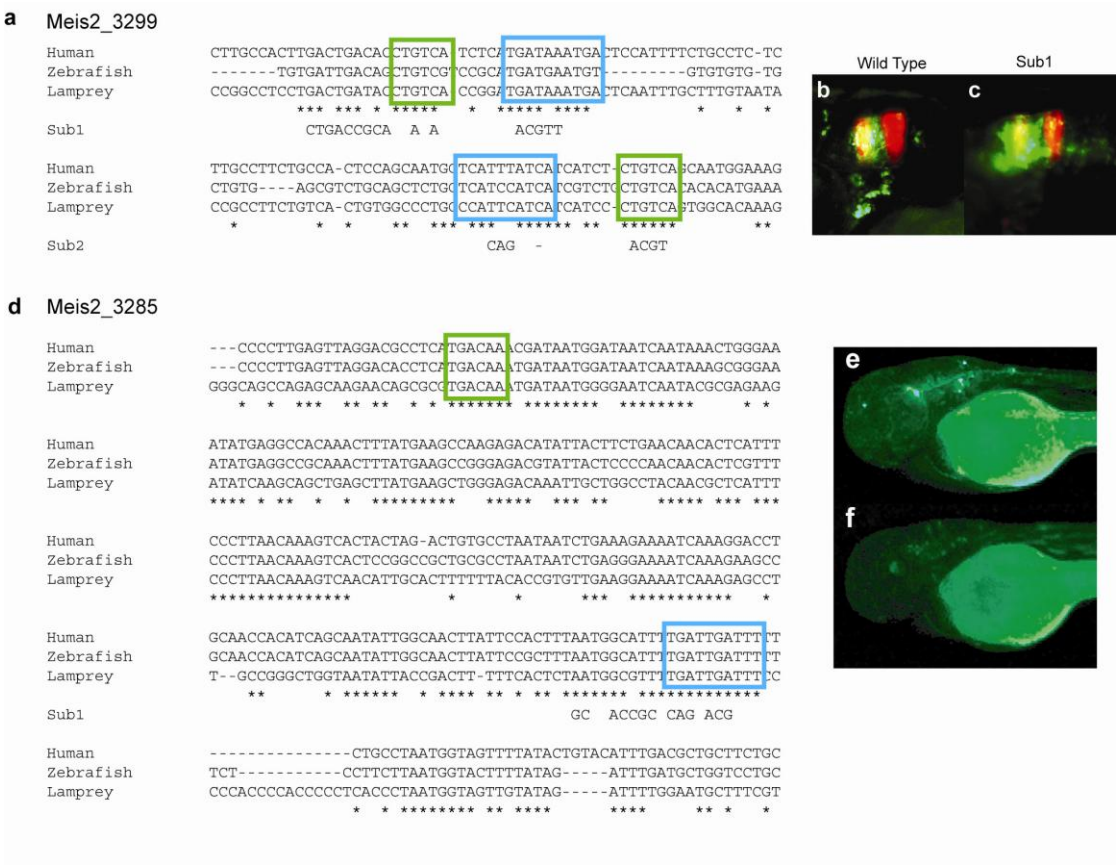


Figure 5.2. Pbx-Hox motifs are essential for enhancer activity of meis2_3299 and meis2_3285. **a** Multiple sequence alignment of meis2_3299 from human, zebrafish and lamprey genomes, highlighting conserved Pbx-Hox (blue boxes) and meis (green boxes) binding site motifs. The mutated bases in dr3299_Sub1 and _Sub2 are shown beneath the alignment. **b-c** dr3299_Sub1 (**c**) drove broader reporter expression in the hindbrain at 54hpf compared to dr3299 (wt) (**b**) and did not up-regulate GFP expression in the neural crest (r3r5 RFP expression shown in red). **d** Multiple sequence alignment of meis2_3285 showing conserved Pbx-Hox (blue box) and meis (green box) binding-site motifs. The mutated bases in dr3285_Sub1 are shown. **e-f** dr3285_Sub1 (**f**) drove severely diminished GFP expression at 54hpf compared to the wild type dr3285 element (**e**).

An *in-silico* search for conserved Pbx-Hox motifs in CNEs

(This search was devised and carried out by P. Piccinelli and myself)

The occurrence of many highly conserved Pbx-Hox motifs within the CNEs of the *C15orf41* contig suggested that they may constitute a widespread signature within CNEs. To address this, an *in-silico* search strategy was devised to identify Pbx-Hox motifs in two overlapping sets of CNEs – a lamprey set, consisting of 246 CNEs

conserved between human, fugu, zebrafish and lamprey, and a gnathostome set, consisting of 4259 CNEs conserved between human, fugu and zebrafish. The sequences of human, fugu and most zebrafish CNEs were retrieved from the CONDOR database (Woolfe *et al.*, 2007). Additional zebrafish CNEs were identified using BLAST against a more recent zebrafish genome assembly (Zv8 release 58). Sequences in each alignment were clipped to the same size to prevent unaligned edges. To align the sequences we used ClustalW version 1.83. As a control, for each CNE we also generated 1000 multiple alignments by randomly shuffling the columns of each alignment using the seqboot implementation in Phylip version 3.67.

To find evolutionarily conserved Pbx-Hox motifs (TGATNNAT) we employed the software CisFinder (Sharov & Ko, 2009) on our two alignment sets and their respective controls. A motif match was only considered if it matched all aligned species and occurred at the exact same aligned position. The lamprey set was found to contain 61 conserved Pbx-Hox motifs (within 47 CNEs), a 22-fold enrichment compared to shuffled CNE alignments. The gnathostome set contains 712 conserved Pbx-Hox motifs (in 591 CNEs), representing a 9-fold enrichment relative to shuffled alignments. Two alignments of CNEs with conserved Pbx-Hox motifs from the lamprey set are shown in Figure 5.3.

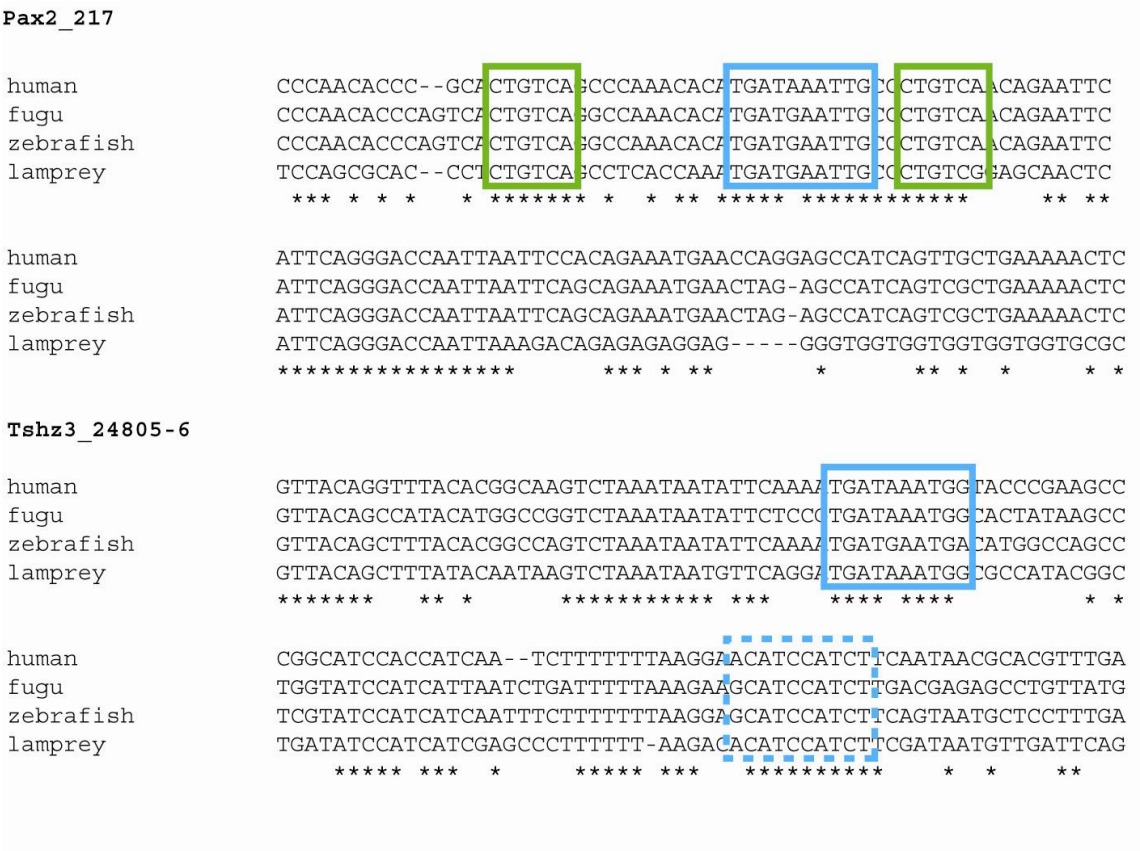


Figure 5.3. Conserved Pbx-Hox and Meis motifs within lamprey CNEs. Multiple alignments for two CNEs are shown with Pbx-Hox (blue boxes) and meis (green boxes) motifs highlighted. The blue dashed box in the Tshz2_24805-6 alignment indicates a conserved variant of the Pbx-Hox motif that does not conform to the TGATNNAT consensus.

Pbx-Hox motif hits identified in gnathostome CNEs strongly resemble Pbx-Hox binding sites identified in the literature

Further analysis of Pbx-Hox motif hits in the gnathostome set reveals a paucity of cytosines at variable positions 5 and 6 (Figure 5.4). This is a feature of characterised Pbx-Hox binding-sites, as cytosines at these positions destabilise binding of the Pbx-Hox complex (Joshi *et al*., 2007). Furthermore, positions 9 and 10, immediately 3' to the canonical Pbx-Hox motif, show strong bias towards G/T and A/G respectively, thereby defining a more stringent TGATNNATKR (KR) consensus motif that is also consistent with previously characterised Pbx-Hox binding-sites. This is revealed by comparison of the nucleotide frequency logo of the gnathostome CNE Pbx-Hox hits with that compiled from characterised Pbx-Hox sites from the literature (Figure 5.4).

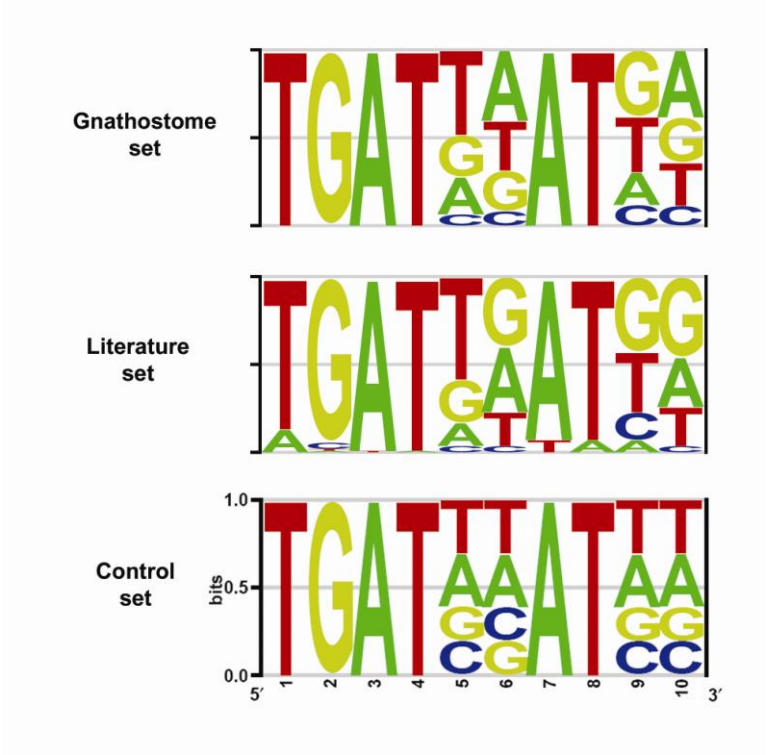


Figure 5.4. Frequency logos representing different sets of Pbx-Hox motifs. Those identified in the gnathostome alignments (gnathostome set) are compared with previously characterised Pbx-Hox motifs (literature set) (from Mann *et al.*, 2009 and Wassef *et al.*, 2008) and Pbx-Hox motifs from the gnathostome control set. The 10 positions of the Pbx-Hox motif are numbered on the x-axis, with nucleotide frequency for each position represented on the y-axis.

The Pbx-Hox hits from the gnathostome control sequence set, which has the same overall nucleotide frequency as the gnathostome CNE set, show nucleotide frequencies at the variable positions that are consistent with the overall nucleotide frequency of the set, indicating that the bias seen in the Pbx-Hox hits of gnathostome CNE is not due to an inherent nucleotide frequency bias of CNEs (Figure 5.4). Further analysis of the gnathostome CNE alignments resulted in strong support for the KR motif, with greater than 16-fold enrichment compared to shuffled alignment controls. The nucleotide frequencies at the variable positions suggest that the majority of the conserved motifs represent *bona-fide* Pbx-Hox binding sites.

Pbx-Hox motifs are enriched within other sets of vertebrate CNEs

(These motif searches were carried out by P. Piccinelli and G. Elgar)

We searched a set of 6,693 human CNEs, previously identified by human-fugu genome comparison (Woolfe *et al.*, 2007) for the stringent 10bp KR motif. The KR motif occurs 562 times in the human CNE set, representing a highly significant enrichment over

shuffled versions of the motif (p=0.000064), and when compared to control genomic regions and the entire human genome (Table 5.1). We also examined the distribution of Pbx-Hox motifs in other sets of evolutionarily-conserved developmental enhancers. The Enhancer Browser (EB) contains 1307 non-coding elements that vary in their degree of conservation across vertebrates, around half of which drive reporter gene expression in mouse embryos at day 11.5 (Pennacchio *et al.*, 2006). Across the entire dataset there is a significant enrichment for the KR motif (p=0.0033) compared with shuffled versions despite the fact that some of these are not deeply conserved. We then analysed a set of 4782 CNEs conserved between human and the cartilaginous chimera, *Callorhinchus milii* (shark CNEs) (Venkatesh *et al.*, 2006), and once again found significant enrichment for the KR motif (p=0.000064). It is important to note that whilst these sets show some overlap with our human CNE set, there are many CNEs that are unique to each set. 427/1307 EB sequences overlap 994 CNEs, covering a total of 146226 bases (7.4% of the EB sequence; 18.8% of CNE sequence). 1632 human sequences from the shark set overlap 2172 CNEs, covering a total of 271260 bases (26.5% of the shark set, 34.9% of CNEs). Thus, the enrichment for Pbx-Hox motifs is a phenomenon that is characteristic of all deeply conserved vertebrate non-coding elements, and not just of our CNE set.

motif	human CNEs	Shark CNEs	EB	EB HB/BA/ CN +ve	EB HB+ve	Zebrafish cne Browser	Zebrafish cne Browser HB+ve	Human genome
TGATNNATKR	**562**	**666**	**609**	**161**	**131**	**17**	**7**	**530509**
TGTANNATKR	171	188	388	65	52	12	3	543258
GTATNNATKR	152	168	279	54	39	9	2	442772
GTTANNATKR	150	178	325	79	65	8	2	475051
TTGANNATKR	201	245	447	80	64	9	1	626755
TTAGNNATKR	167	238	398	74	55	7	0	545811
ATGTNNATKR	259	297	452	86	72	20	1	765597
ATTGNNATKR	233	297	436	74	61	20	2	550250
AGTTNNATKR	215	254	431	85	68	9	0	547730
TAGTNNATKR	147	154	297	54	42	10	4	364553
TATGNNATKR	177	198	365	74	60	10	3	522892
GATTNNATKR	274	315	419	97	74	11	0	521158
TGATNNTAKR	106	143	314	65	50	6	1	489556
TGTANNTAKR	142	151	421	82	60	14	1	790068
GTATNNTAKR	59	73	195	41	34	1	0	372793
GTTANNTAKR	105	108	253	50	33	5	0	419041
TTGANNTAKR	163	205	385	72	62	10	0	548745
TTAGNNTAKR	73	97	235	41	31	0	0	371159
ATGTNNTAKR	103	124	376	64	55	3	0	615961
ATTGNNTAKR	137	158	305	57	42	6	1	474064
mean	165.38	196.583	353.54	70.5833	55.458	8.3333	1.25	521299
S.D	102.75	122.058	93.732	24.7402	20.915	5.522	1.67462	112360
z-score for pbxhox	**3.86**	**3.84**	**2.72**	**3.65**	**3.61**	1.57	**3.43**	0.082
p-value	**0.000064**	**0.000064**	**0.0033**	**0.0001**	**0.0002**		**0.0003**	

Table 5.1. Enrichment for Pbx-Hox motifs in different CNE sets and control sequence sets. The frequency of strict consensus Pbx-Hox (-KR) motifs is compared to those of shuffled variants across a number of different CNE sets, from which z-scores for enrichment of Pbx-Hox motifs are derived, with corresponding p-values. HB: hindbrain, BA: branchial arch, CN: cranial nerve.

CNEs with Pbx-Hox motifs frequently drive reporter expression in the hindbrain and pharyngeal arches

As the elements of the C15orf41 contig contained conserved Pbx-Hox motifs and drove segment-specific reporter expression in the hindbrain and pharyngeal arches, consistent with many other previously characterised Hox enhancers that contain Pbx-Hox motifs, I wished to test whether Pbx-Hox motifs predict CNEs with these domains of reporter expression. To do this, I assayed 21 CNEs containing Pbx-Hox motifs for reporter expression in zebrafish, comprising 11 CNEs that are conserved in lamprey and 10 that are conserved across gnathostomes but not in lamprey. These elements were selected according to two criteria: firstly, that they represented a number of different gene regions, and secondly, that they had multiple Pbx-Hox and Meis sites - similar to the elements from the C15orf41 contig. A list of CNEs assayed is found in the Appendix.

12 of these 21 elements up-regulated consistent patterns of reporter expression, consisting of 8 from the lamprey set and 4 from the gnathostome set. Remarkably, 11 of the 12 GFP-expressing elements (91.6%) drove expression either in the hindbrain, pharyngeal arches or both, with one element expressing in the trunk musculature (Figure 5.5). The expression patterns driven by these elements are in keeping with the endogenous expression patterns of the genes with which they are associated. In support of the hypothesis that these elements are directly regulated by specific Hox proteins, which have segmentally-restricted expression patterns, the majority of the elements expressing in the hindbrain do so in particular rhombomeres, as shown by comparison with r3r5 RFP expression (Figure 5.5). Interestingly, whilst in some cases the rhombomere-specific expression is seen across the whole rhombomeric domain, such as for Evi_40224 and Meis2_3299, in many cases reporter expression is limited dorso-ventrally (for instance, Tshz3_43509) or medio-laterally (for instance, Pax2_217).

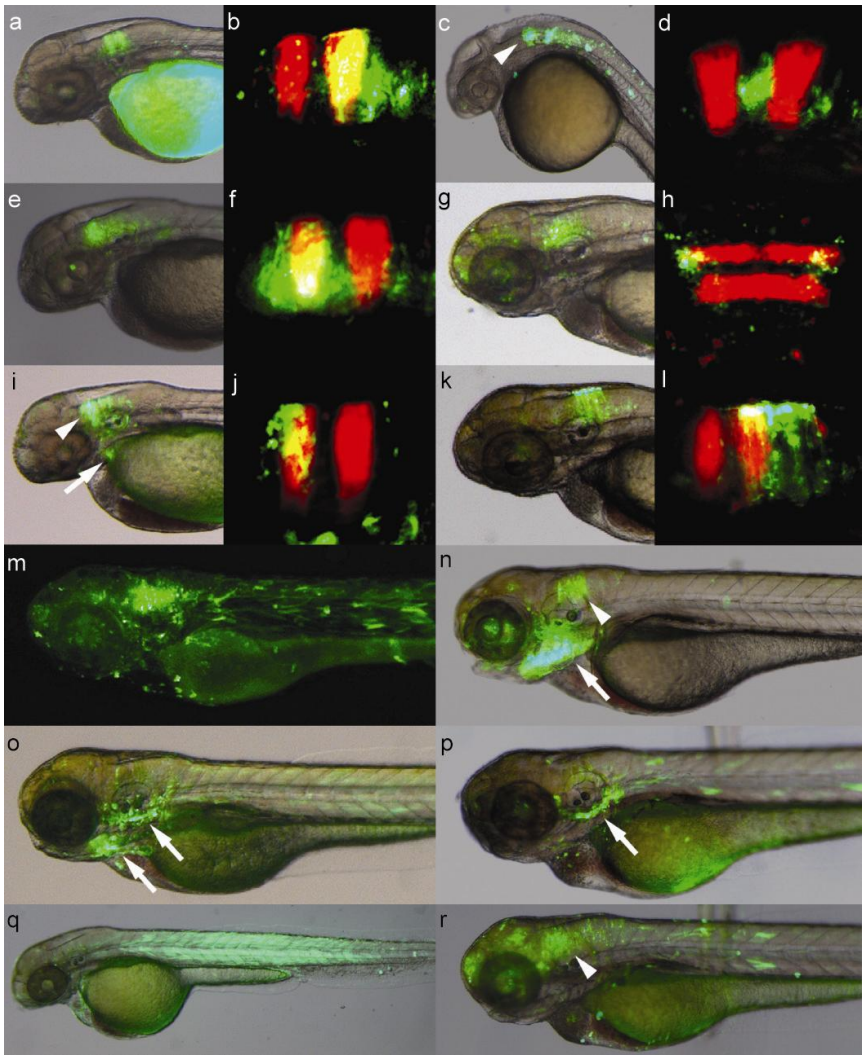


Figure 5.5. Pbx-Hox motifs predict segment-specific hindbrain and pharyngeal arch reporter expression. **a-r,** elements from the lamprey (**a-j,m,o,q**) and jawed vertebrate CNE sets (**k,l,n,p,r**) drive GFP expression either in the hindbrain – elements Evi1_40224 (**a,b**), Tshz3_43509 (**c,d)**, NR2F2_27254 (**e,f)**, Pax2_217 (**g,** dorsal view: **h**), Tshz3_24804 (**m**), Nkx6-1_4281 (**k,l**), BCL11A_2554 (**r**) – in the pharyngeal arches - TshZ3_24805-6 (**o**), FoxP1_886 (**p**) - or in both - ZNF503_32799 (**i,j**), Pax9_2099 (**n**). Expression in the hindbrain is often restricted to certain rhombomeres, as shown by comparison with r3r5 RFP expression (**b,d,f,h,j,l**). Tshz3_24807 drives expression in the trunk musculature (**q**). Arrowheads: hindbrain expression, arrows: pharyngeal arch expression. 24hpf embryos: **c,d;** 54hpf embryos: **a,b,e,f,i,j,q**; 78hpf embryos: **g,h,k,l,m,n,o,p,r**.

To address the association between Pbx-Hox motif presence and reporter expression in the hindbrain across a larger set of functionally characterised CNEs, we next examined functional data from the Enhancer Browser (EB) database (Pennacchio *et al*, 2006) (Bioinformatic analysis performed by Dr. P. Pichinnelli and Dr. G. Elgar). We found significant enrichment for the KR motif in those elements annotated as hindbrain

positive (64 motifs in 112kb) compared with those with no hindbrain annotation (85 in 238kb) (p=0.0042) (Table 5.1). We then looked only at those sub-regions within EB enhancers that align directly with CNEs. Within these conserved regions, there was more than two-fold enrichment for the stringent Pbx-Hox motif (30 occurrences in 24990bp of HB+ elements compared with 32 occurrences in 60341bp of HB- elements; p=0.001). We also analysed a smaller dataset from the cneBrowser set (http://bioinformatics.bc.edu/chuanglab/cneBrowser), which contains evolutionarily conserved enhancers associated with genes expressed in forebrain and hindbrain during zebrafish development. Although only 18 of 146 enhancers are annotated as hindbrain positive, 7 out of a total of 17 identified KR motifs reside in hindbrain positive enhancers (p=0.0003) (Table 5.1).

CNEs with Pbx-Hox motifs are associated with genes that are likely to be regulated by Hox factors

(The bioinformatic data was obtained by P. Piccinelli)

Whilst conserved Pbx-Hox motifs have a widespread distribution across the CNEs of many different genes, certain gene regions are more enriched for these motifs than others (Table 5.2). We analysed the distribution of Pbx-Hox motifs across CNEs of different genes within the CNE set, to ask whether genes with the greatest enrichment for Pbx-Hox motifs in their CNEs are known to interact with Hox factors during development or have roles in hindbrain or pharyngeal arch patterning.

Gene region	Number of Pbx-Hox motifs in CNEs	Total length of CNEs from gene region /kb	p-value
ZNF503 *	36	27.781	0.00E+00
TSHZ3	29	23.323	0.00E+00
TSHZ1	16	10.351	0.00E+00
ZNF703	8	2.991	0.00E+00
GLI3	8	3.432	9.10E-15
IRX2	21	23.981	1.08E-12
MAF	11	7.334	3.77E-11
ESRRB	7	4.537	2.47E-10
NKX6-1 *	10	6.853	3.11E-10
NR2F2	16	18.99	4.28E-10
POU3F2	6	3.297	4.90E-10
HOXD9	16	17.77	4.37E-09
IRX5	27	37.059	1.46E-08
PBX3	16	17.886	1.58E-08
SALL3	12	11.405	2.87E-08
EVI1 *	6	4.015	8.14E-07
SOX14	6	4.286	1.83E-06
MEIS1	9	9.298	3.60E-06
FOXP1	12	15.857	1.69E-05
PRDM16	4	3.736	2.21E-05
TSHZ2 *	7	7.221	2.47E-05
MEIS2 *	16	24.553	2.92E-05
POU6F2	4	3.103	6.94E-05
NR2F1	16	25.655	7.37E-05
PAX5	2	1.05	7.86E-05
SHOX2	8	7.615	9.07E-05
PAX9	4	3.217	1.19E-04
OTP	8	8.986	1.97E-04
MAB21L2 *	6	5.733	3.54E-04
ESRRG	8	8.743	3.86E-04
BCL11A	10	13.643	4.34E-04
ATBF1 *	6	7.514	9.66E-04
EMX2	7	8.736	1.10E-03
EBF1	4	4.032	2.14E-03
ZFHX1B	13	23.275	3.52E-03
PAX6	4	4.913	4.37E-03
ZIC1/4 *	6	8.332	4.76E-03
SATB1	5	5.558	5.87E-03
TCF7L2	7	11.045	6.52E-03

Table 5.2. Distribution of Pbx-Hox +ve CNEs across Condor gene regions. The 40 Condor gene regions from the human genome with the most significant enrichment for Pbx-Hox (TGATNNATKR) motifs within their CNEs are listed. For each gene region, the number of Pbx-Hox hits within the CNEs is compared with the average number of hits within 1000 equivalent sets of control CNEs generated by a zero-order Markov model, generating a p-value for enrichment for Pbx-Hox motifs in the CNEs compared to controls (using a z-test). This controls for the influence of the different numbers of CNEs between gene regions. Some gene regions contain clusters of genes, for

instance ‘IRX5’ contains the genes *irx3*, *irx5* and *irx6*, ‘IRX2’ contains *irx1*, *irx2* and *irx4*, and ‘HOXD9’ contains the HoxD cluster, including *Cdx2* and *Lunapark*, so CNEs within these regions may not necessarily regulate the gene after which the cluster is named. Gene names with asterisks beside them indicate that these genes were shown by micro-array experiments to be influenced by *hoxB1* in r4 of zebrafish or mouse embryos (Rohrschneider *et al*., 2007; Tvrdik & Capecchi, 2006).

In keeping with their well characterised roles as Hox co-factors, and their essential contributions to hindbrain patterning, we find *pbx3* and *meis2* to be amongst those genes with the highest enrichment Pbx-Hox motifs in their CNEs. Many of the other genes with high enrichment for Pbx-Hox motifs in their CNEs have characterised roles in anterior-posterior (A-P) head patterning and show segment specific patterns of expression during development. For instance, the *znf503/703* (*nlz1/nlz2*) zinc-finger proteins are essential for specification of rhombomere 4 (Hoyle *et al*., 2004; Runko & Sagerström, 2003). The *iroquois* (*irx*) genes, which exist in clusters in vertebrate genomes (Gómez-Skarmeta & Modolell, 2002), show rhombomere-specific expression in the hindbrain in mouse (Bosse *et al*., 1997) and zebrafish embryos (Lecaudey *et al*. 2005) with *irx1* and *irx7* interacting with Hox factors during the formation and specification of r1-4 in zebrafish (Stedman *et al*., 2009). The orphan nuclear receptors *NR2F1/2* (*COUP-TF1/2*) are negative transcriptional regulators that modulate the retinoic acid signalling pathway (Chung & Cooney, 2003), which has a key role in A-P patterning of the hindbrain and pharyngeal arches, partly through influencing the expression of Hox genes (Pereira *et al*., 2000). The members of the Teashirt protein family (*tshz1,2* and *3*) show segment-specific hindbrain expression (Santos *et al*., 2010), *tshz1* being essential for segmentally restricted gene expression in the hindbrain and pharyngeal arches of frog and mouse (Koebernick *et al*., 2006; Coré *et al*., 2007). In *Drosophila*, Teashirt has even been suggested to be a Hox co-factor (Robertson *et al*., 2004), however no instances of co-operative binding have been characterised (Taghli-Lamallem *et al*., 2007). Thus, the distribution of CNEs with Pbx-Hox motifs across different gene regions is consistent with them playing roles in Hox-dependent patterning mechanisms during development.

Genes with Pbx-Hox +ve CNEs overlap with characterised Hox targets in r4

There is good agreement between the genes highlighted by our in-silico binding-site search and by micro-array screens for downstream targets of Hoxb1 in rhombomere 4 of

zebrafish (Rohrschneider *et al*., 2007) and mouse (Tvrdik & Capecchi, 2006) (Table 5.2). Specifically, the expression levels of *znf503*, *tshz2*, *evi1*, *zic4* (within gene region '*zic1*' in Condor), *shox*, *meis2.1* and *foxd3* are decreased upon knock-down of *hoxB1* in zebrafish, with *znf503, nkx6-1, atbf1, mab21l2* and *phox2b* down-regulated in *hoxB1*-/- mouse embryos. Accordingly, the CNEs around each of these genes contain Pbx-Hox motifs and in the majority of cases show enrichment for these motifs relative to control sequence sets (Table 2, Appendix). However, not all the genes identified in the microarray screens are highlighted by our motif search. The differences arise because many of the genes in the microarray screens do not have CNEs associated with them; furthermore, the focus of the screens is restricted to *hoxB1* in r4, whilst our search is also likely to identify Pbx-Hox motifs that are regulated by other Hox factors in different embryonic domains. Additionally, microarray approaches are unable to differentiate between direct and indirect effects of Hox perturbation, whereas our Pbx-Hox motifs predict direct interactions. Nevertheless, both of the microarray datasets support our prediction that Pbx-Hox motifs in CNEs represent direct regulatory links between Hox genes and their targets during development.

CNEs with Pbx-Hox motifs contain other relevant TFBS motifs

The finding that genes with the highest enrichment for Pbx-Hox motifs in their CNEs frequently have roles in hindbrain or pharyngeal arch patterning, or are known to regulate or be regulated by Hox factors during development, hints that these genes may function in common GRNs and thus might also regulate each other. It is possible to address this by searching for TFBS motifs of these factors within Pbx-Hox +ve CNEs.

As an initial study into the feasibility of searching for other motifs within Pbx-Hox +ve CNEs, the 10 characterised lamprey CNEs that drove reporter expression in the hindbrain (Meis2_3285, Meis2_3288, Meis2_3299, Meis2_329X, Evi1_40224, Tshz3_43509, NR2F2_27254, ZNF503_32799, Tshz3_24800, Pax2_217) were scanned for NR2F1 motifs that are conserved across gnathostomes and lamprey, using the motif scan function of Jaspar (Bryne *et al*., 2008) with stringent parameters (relative profile score threshold of 80%). The NR2F1 motif was chosen for this search because the NR2F1 gene is expressed in the developing hindbrain, so is likely to play a role in the GRN for hindbrain patterning and therefore may regulate some of these elements. Further, this gene has a well characterised TFBS motif in the Jaspar core database

(Bryne *et al.*, 2008). Interestingly, two CNEs were identified that contained conserved 14bp NR2F1 motifs: NR2F2_27254 (shown in figure 5.6), which contained two conserved motifs, and Tshz3_43509, with one conserved motif.

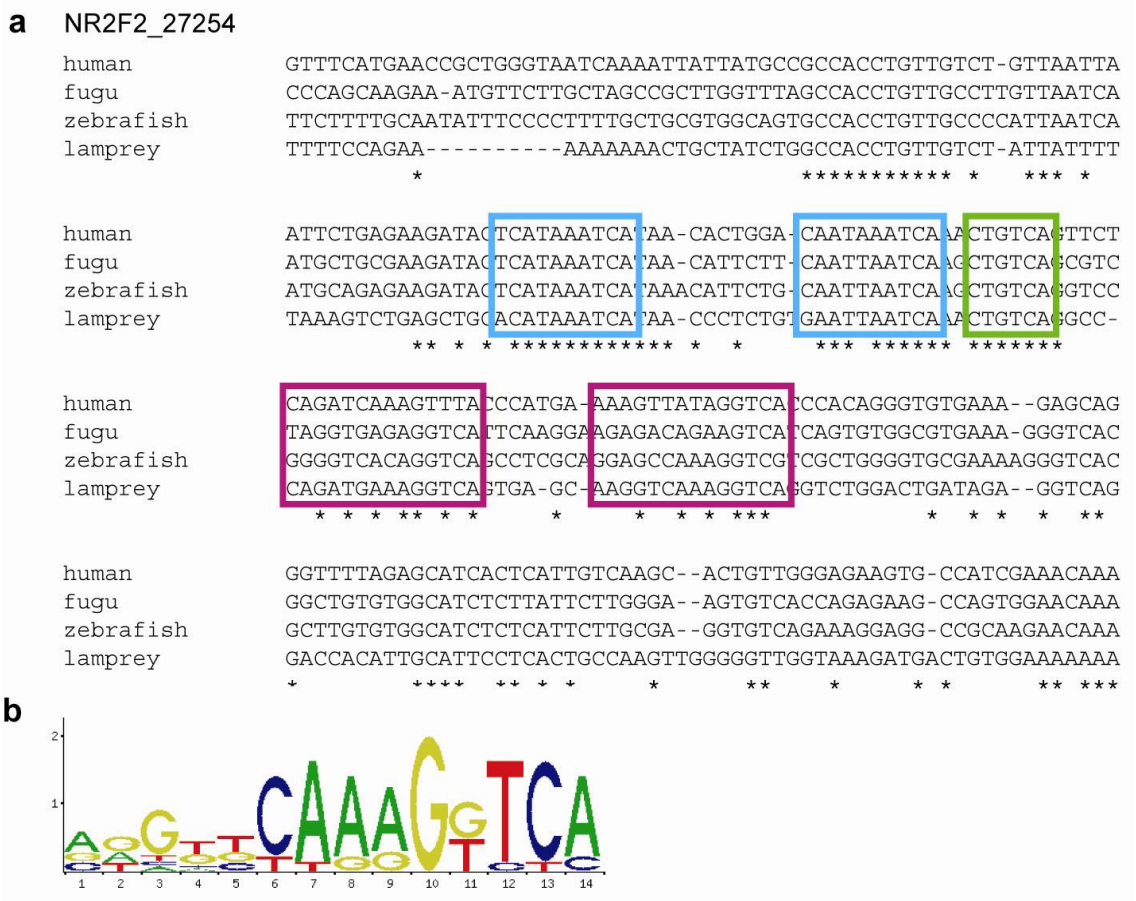


Figure 5.6. The lamprey CNE NR2F2_27254 contains conserved NR2F1 motifs. **a** multiple alignment of NR2F2_27254 orthologous sequences from gnathostomes and lamprey, with conserved Pbx-Hox (blue), meis (green) and NR2F1 (purple) motifs highlighted. **b** Sequence logo representing the binding preference of NR2F1 (Kimura *et al.*, 1993).

These high scoring, conserved motifs suggest that Pbx-Hox +ve CNEs that drive expression in the hindbrain may also be regulated by other factors that are themselves regulated by Pbx, Hox and Meis. This would be consistent with a 'recursively wired' gene-regulatory network for hindbrain development. However, it should be noted that the NR2F1 motif used for this search also represents the DNA binding specificity of other nuclear hormone receptors, which often recognise sequences consisting of repeated AGGTCA motifs (Mangelsdorf *et al.*, 1995). Nevertheless, as many of these other receptors are also involved in retinoic acid signalling (some of which also have

CNEs with Pbx-Hox motifs – e.g. EsrrB, EsrrG (Table 5.2)), it is likely that the presence of these motifs signifies that these enhancers are regulated by retinoic acid signalling as part of a hindbrain GRN.

Pbx-Hox motifs are retained between duplicated CNEs

Due to the two rounds of whole genome duplication that took place early in the vertebrate lineage (2R) (Putnam *et al.*, 2008), as well as the teleost-specific whole genome duplication (Jaillon *et al*., 2004), there are a number of families of duplicated CNEs, which are a useful resource for investigating the evolutionary fate of duplicated *cis*-regulatory elements (McEwen *et al*., 2006; Woolfe & Elgar, 2007). Within our CNE set, the 2R duplicates are referred to as dCNEs, whilst the teleost-specific duplicates are referred to as co-orthologs. These duplicated CNEs have previously been the focus of studies characterising their sequence and functional divergence upon duplication, revealing patterns of both retention and divergence of function between dCNEs (McEwen *et al*., 2006) and patterns of sequence divergence consistent with sub-functionalisation in co-orthologs (Woolfe & Elgar, 2007). However, the insights from these studies are limited by their focus on CNEs that have not been characterised at the level of TFBS motifs.

The presence of Pbx-Hox motifs within a number of duplicated CNEs enables a more detailed, binding-site-specific, characterisation of CNE divergence upon duplication, with the potential to link sequence divergence in TFBS motifs with changes in reporter expression driven by duplicated enhancers. An example of this approach is a study focusing on *cis*-regulatory elements controlling expression of duplicated *hoxa2* genes in teleosts (Tümpel *et al.*, 2006). Whilst the authors found the majority of TFBSs that had been characterised in mouse to be alignable between both fugu gene regions, they were able to attribute differences in expression patterns between the paralogous fugu genes to a small number of nucleotide changes within these TFBSs, providing insight into *cis*-regulatory evolution. With a larger number of duplicated *cis*-regulatory elements at our disposal, it is possible to perform a similar study in a more systematic manner.

As a starting point, three CNE families were selected for *in-silico* analyses. These families were chosen as they each contain one of the CNEs for which there is clear functional data in zebrafish: Meis2_3299, Evi1_40224 and Tshz3_43509. For each

family the following questions were asked: are Pbx-Hox motifs (conforming to the TGATNNAT consensus) and Meis motifs (TGACAR) conserved between duplicated CNEs? Do different positions of the motif show different frequencies of conservation? Are some types of mutation more frequent than others?

The three CNE families are: 3299, including CNE 3299 from *meis2* and its dCNE associated with *meis1*; 40224, consisting of CNE 40224 from *evi1* and its dCNE associated with *prdm16*; and the large 43509 family, comprising CNE 43509 from *tshz3* and its dCNE associated with *tshz1*, as well as its co-orthologs in Fugu (from *tshz1.1*, *tshz1.2*) and zebrafish (from *tshz3a* and *tshz3b*). Multiple alignments for these elements are shown in Figures 5.7-5.9 and divergence within Pbx-Hox motifs is represented in Table 5.3.

In the majority of cases (10/12) Pbx-Hox motifs from duplicated CNEs can be aligned, with the core Pbx-Hox motif - TGATNNAT – being conserved in 6 out of these 10 cases. Inspection of the alignments reveals that the rest of the CNE has often diverged in sequence between duplicates, whilst the Pbx-Hox motifs have been retained, suggesting that they are of crucial importance for the functionality of their enhancers.

Table 5.3 documents the conservation of nucleotides at the different positions within the ten alignable, duplicated Pbx-Hox motifs. Positions 5, 9 and 10 exhibit the highest frequencies of mutation between duplicates, whilst positions 2-4 show no instances of mutation. This is in accordance with characterised Pbx-Hox binding sites, where 'variable' positions 5, 6, 9 and 10 have less stringent nucleotide specificity than the 'fixed' positions 1-4 and 7-8. This further supports the notion that these sites are indeed *bona fide* Pbx-Hox sites. Interestingly, there is relatively little divergence at position 6, either between paralogous elements or between orthologs, whilst position 5 has more frequent divergence (Table 5.3).

Type of duplicates	Motif comparison	1	2	3	4	5	6	7	8	9	10
		T	G	A	T	N	N	A	T	N	N
dCNEs	3299_PH1 Meis2 vs Meis1	C	C	C	C	D	C	C	D	C	D
dCNEs	3299_PH2 Meis2 vs Meis1	D	C	C	C	D	D	C	C	C	D
dCNEs	40224_PH1 Evi1 vs Prdm16	C	C	C	C	D	C	C	C	D	C
dCNEs	40224_PH2 Evi1 vs Prdm16	C	C	C	C	D	C	C	C	D	D
dCNEs	43509_PH2 Tshz3 vs Tshz1	C	C	C	C	D	C	C	C	C	C
dCNES	43509_PH3 Tshz3 vs Tshz1	C	C	C	C	C	C	C	C	D	C
Co-orths	43509_PH2 fugu Tshz1.1 vs fugu Tshz1.2	C	C	C	C	D	C	C	C	C	C
Co-orths	43509_PH3 fugu Tshz1.1 vs fugu Tshz1.2	C	C	C	C	C	C	C	D	C	C
Co-orths	43509_PH1 zfishTshz3a vs zfish Tshz3b	C	C	C	C	C	D	D	C	C	D
Co-orths	43509_PH2 zfishTshz3a vs zfish Tshz3b	D	C	C	C	C	C	C	C	C	D
Occurances of divergence		**2**	**0**	**0**	**0**	**6**	**2**	**1**	**2**	**3**	**5**

Table 5.3. Conservation of Pbx-Hox motifs between duplicated CNEs. The conservation (C) or divergence (D) between duplicated elements is tabulated for each position (1-10) of aligned Pbx-Hox (PH) motifs from duplicated (dCNEs) and co-orthologous (co-orths) CNEs shown in figures 5.7-5.9. Positions are considered conserved if they are conserved between each pair of paralogous CNEs from the alignment. Positions that were conserved between paralogous elements from one species but not between paralogous elements of another species were considered divergent.

There are some instances where mutations have occurred in the 'fixed' positions. Notably, there are two instances of nucleotide change at position 1 (PH2 of meis1 elements from the 3299 family and PH2 of zebrafish tshz3a from the 43509 family), both involving mutation from T to A, in which case the site is likely to retain its ability to be bound by a Pbx-Hox complex (examples of functional Pbx-Hox sites with A at position 1 include the mouse hoxB1 r4 enhancer repeat 1 (Pöpperl et al., 1995) and the fly fkh250 enhancer (Ryoo & Mann, 1999). There are also two separate instances of mutations at position 8 from T to C (PH1 of zebrafish and Fugu meis1 elements from the 3299 family and PH3 of the Fugu tshz1.1 element from the 43509 family). In both cases the rest of the site still conforms to the Pbx-Hox motif, suggesting these sites may still be constrained. In the case of the meis1 element, the mutation of position 8 from T to C occurred in the teleost lineage prior to the divergence of Fugu and zebrafish, as this same mutation is also found in medaka (not shown).

On the whole, Meis motifs are also conserved between duplicated CNEs, but, as is sometimes seen between orthologous CNEs from different species (for instance M1 in the zebrafish Meis2_3299 element), they can shuffle in position relative to the Pbx-Hox site. This shuffling is seen with M2 of CNE3299 between *meis1* and *meis2* dCNEs.

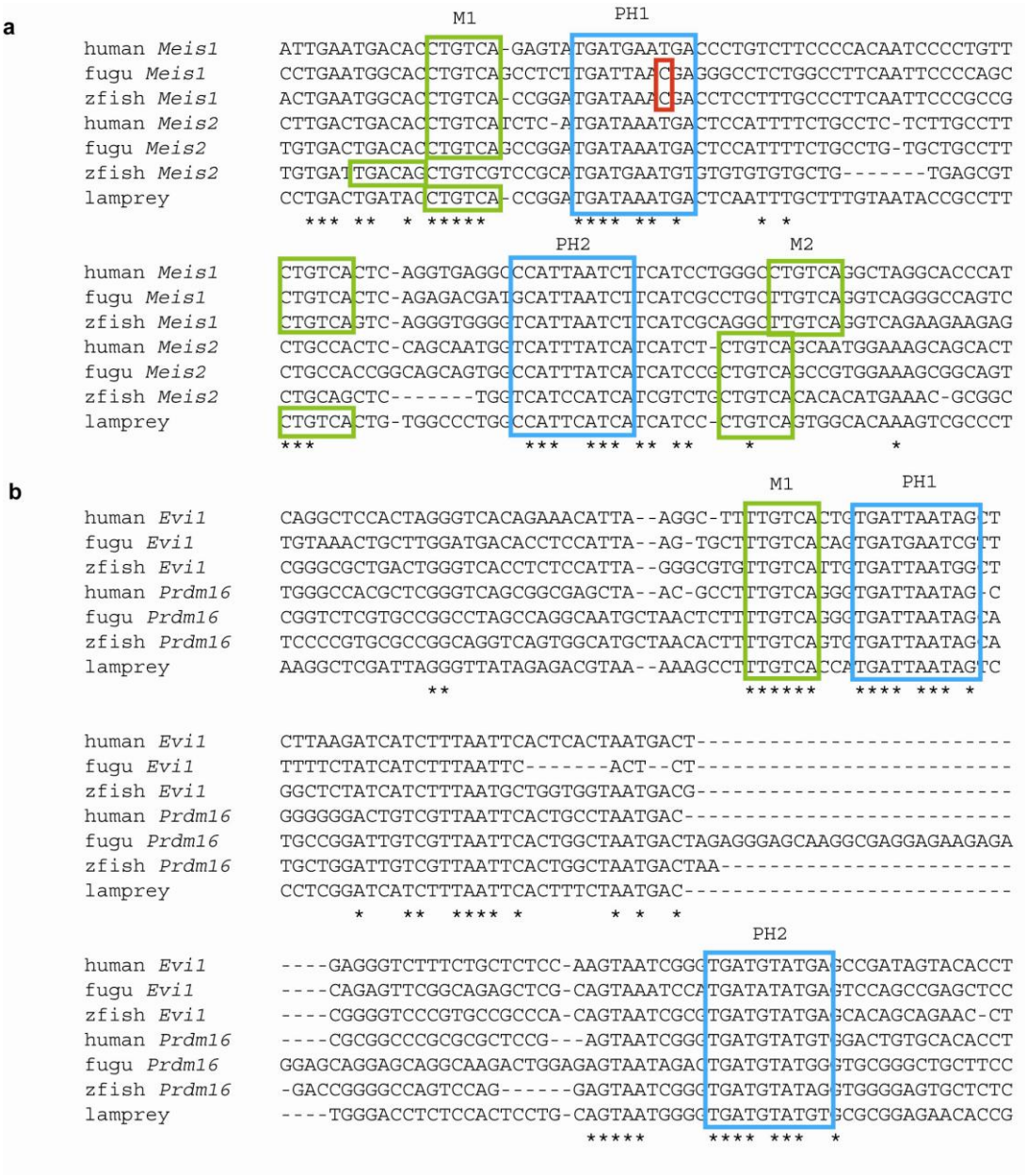


Figure 5.7. Patterns of retention of Pbx-Hox and Meis motifs between dCNEs I. **a-b** multiple sequence alignment of CNEs orthologous to Meis2_3299 and their paralogs associated with *meis1* (**a**), and CNEs orthologous to Evi1_40224 and their paralogs associated with *prdm16* (**b**) from gnathostomes, with the homologous lamprey elements included for comparison. Pbx-Hox (blue) and Meis (green) motifs are highlighted. Deviations from the canonical Pbx-Hox motif that are likely to reduce binding affinity are also highlighted (red).

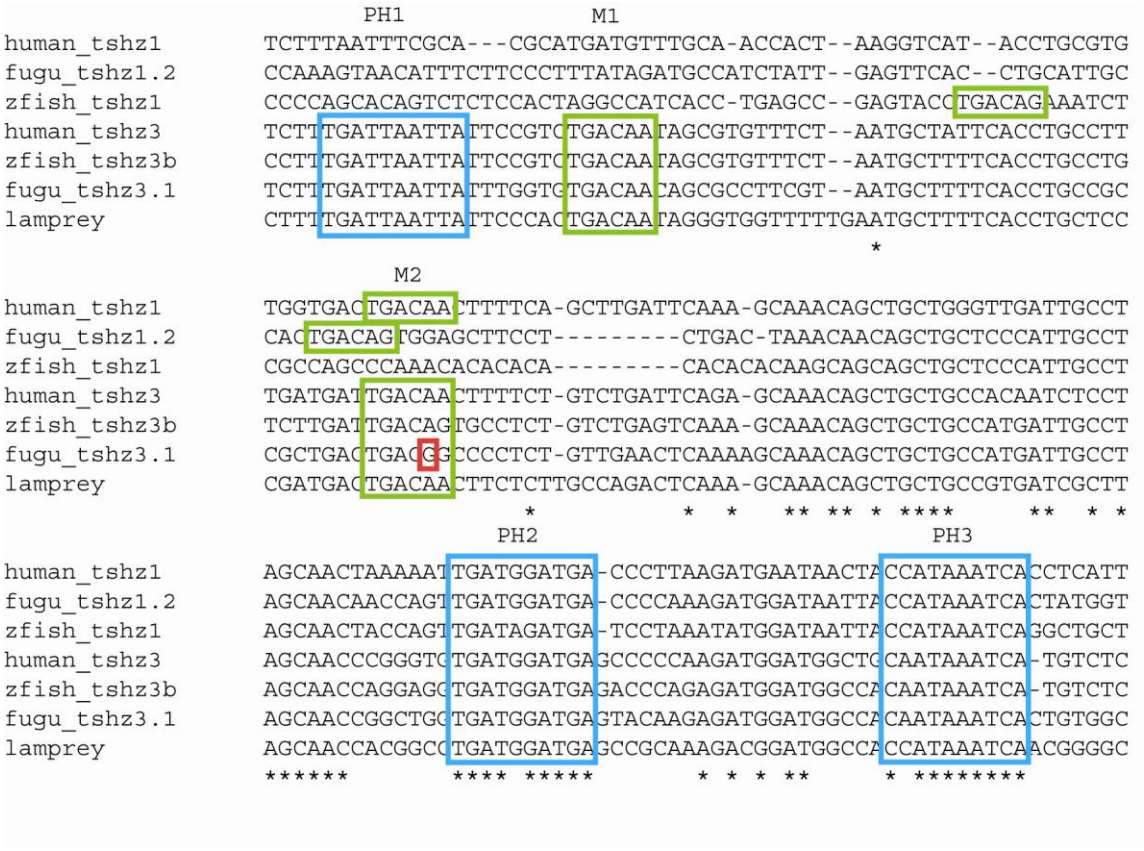


Figure 5.8. Patterns of retention of Pbx-Hox and Meis motifs between dCNEs II. Multiple sequence alignment between orthologous sequences of Tshz3_43509 and their paralogs associated with Tshz1 from gnathostomes, with a lamprey homolog included for comparison. Pbx-Hox (blue) and Meis (green) motifs are highlighted, as are deviations from consensus motifs likely to reduce binding affinity (red).

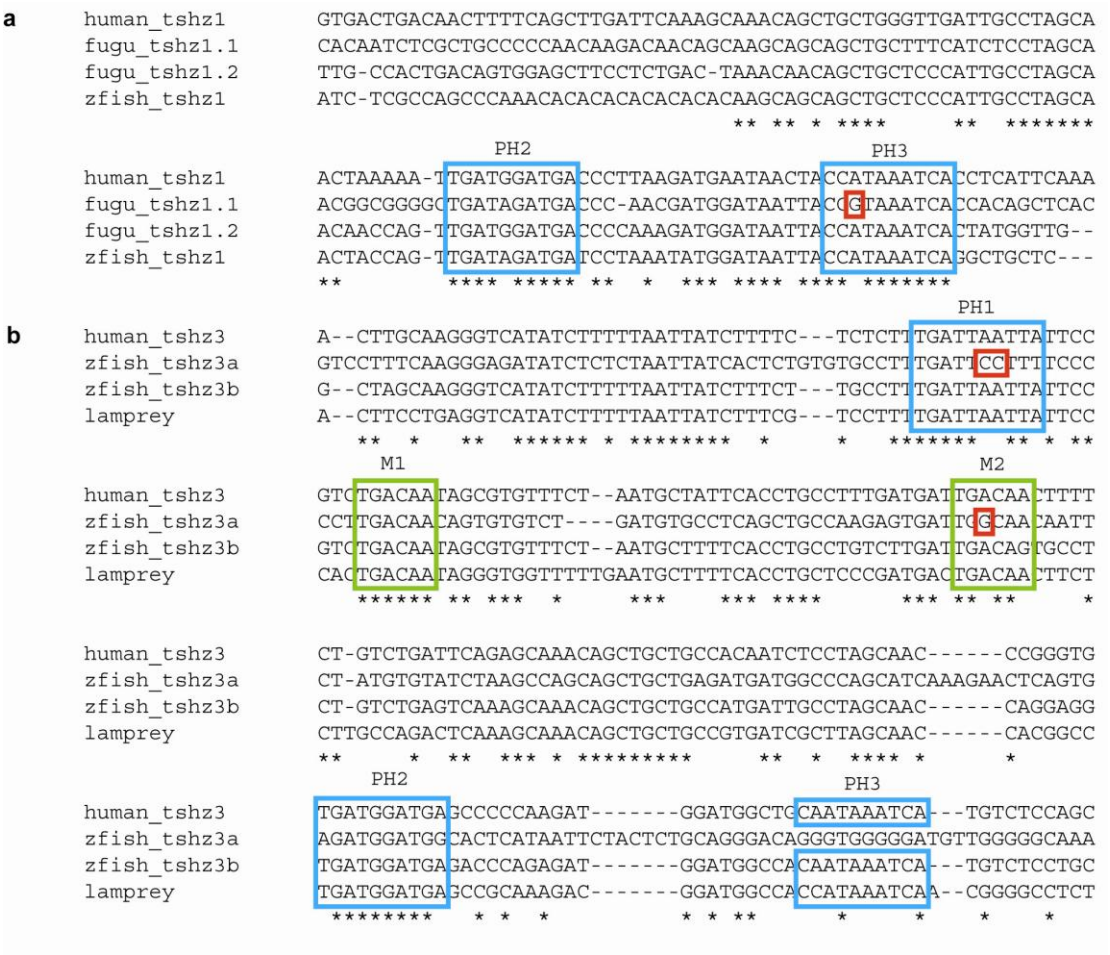


Figure 5.9. Patterns of retention of Pbx-Hox and Meis motifs between co-orthologous CNEs. **a** multiple sequence alignments of human and zebrafish tshz1_43509 sequences, with two co-orthologous fugu sequences. **b** multiple alignment of human tshz3_43509 with its homolog in lamprey and two co-orthologous zebrafish sequences. Pbx-Hox (blue) and Meis (green) motifs are highlighted as well as deviations from the motif consensus sequences that are predicted to reduce binding affinity (red).

Discussion

CNEs of the *C15orf41* contig

The CNEs of the *C15orf41* contig were chosen as a starting point for functional dissection due to their clear and specific expression patterns in the zebrafish co-injection assay. Their expression patterns in the tol2 assay are highly consistent with those from co-injection, despite these two assays using different minimal promoters, which can sometimes have an effect on enhancer expression (de la Calle-Mustienes *et al.*, 2005; Gehrig *et al.*, 2009). The generality of their expression patterns between different

reporter assays re-enforces the view that these CNEs represent tissue-specific developmental enhancers. Their expression domains are highly complementary and only partially overlapping, which is in keeping with the notion of *cis*-regulatory modules acting independently in specific domains to contribute to the overall expression pattern of their gene.

The patterns driven by homologous zebrafish and lamprey elements in zebrafish are very similar, but also show notable differences. For instance, dr3299 expresses GFP in the hindbrain and neural crest, whilst pm3299 is restricted to the hindbrain. To assess the evolutionary significance of this, it is necessary to functionally test these elements by reporter assay in lamprey embryos (this is addressed in more detail in chapter 6). Comparing zebrafish and lamprey elements, it appears that the sequences responsible for generating the restricted reporter expression of the zebrafish elements are mostly conserved within the lamprey elements, such that they are still recognisable to the zebrafish transcriptional machinery. However, it is notable that for three of these elements, 3288, 3299 and 329X, the lamprey elements up-regulate GFP at a greater intensity than their zebrafish homologues. This may reflect 'fine-tuning' in *cis*- and *trans*- within a lineage, such that elements drive less well regulated (more intense, broader) expression patterns when tested by heterologous reporter assay. This could also be further investigated by testing these elements in a lamprey reporter assay.

The functional dissection of dr3299 and dr3285 show that phylogenetic footprinting can be applied to CNEs to reveal motifs that are crucial for their enhancer function. The expression pattern driven by dr3299 in r3-4 of the hindbrain is highly consistent with data from other *cis*-regulatory elements that contain Pbx-Hox sites and drive rhombomere-specific reporter expression patterns. However, the expression pattern driven by dr3285, particularly the expression in cranial ganglia, is hard to explain by Hox regulation, as Hox factors are not expressed in the trigeminal ganglion. It could be that the Pbx-Hox motif is indeed bound by Pbx and Hox factors and contributes to expression in the more posterior cranial ganglia, the hindbrain or the spinal cord. It is also possible that this site is bound by a heterodimer consisting of Pbx and an orphan Hox factor that is expressed in the cranial ganglia (for instance, Tlx (Andermann & Weinberg, 2001). Alternatively, this site may resemble a Pbx-Hox site but is actually bound by other factors. Pbx-Meis heterodimers have been shown to recognise such sites, raising the possibility that these factors bind to this site, possibly facilitating the

binding of other factors to neighbouring sites. Thus, whilst the expression pattern and motif requirements of dr3299 are highly consistent with it being regulated by trimeric complexes consisting of Pbx, an anterior Hox factor, and Meis, the factors regulating the expression pattern of dr3285 are less clear.

The finding that all four of the elements of the *C15orf41* contig that drive GFP expression contain conserved Pbx-Hox and Meis sites is intriguing. It is possible that this has a functional significance, with the proximity of these elements to each other contributing to or being the product of a more global regulatory mechanism involving these motifs. Alternatively, the clustering of these elements with similar TFBS motifs may be an outcome of how they arose in the genome (e.g. Cameron & Davidson, 2009). The hypothesis that these elements 'interact' with each other in some way during gene regulation could be addressed by functionally testing the whole contiguous genomic region by reporter assay.

Identification of motifs in CNEs

The strong enrichment for Pbx-Hox motifs within CNEs suggests that many of these elements operate through TFBSs of the type that have been characterised previously, rather than through some hitherto unknown mechanism of action. Furthermore, the identification of a common motif within CNEs hints that a regulatory language within CNEs may be interpretable and that it could also be applicable to other, less well conserved *cis*-regulatory elements. Recent 'bottom-up' efforts to identify motifs within groups of tissue-specific enhancers have succeeded in generating sets of motifs that can be used to predict the expression patterns of other CNEs that are conserved between mammals (Narlikar *et al*., 2010) or between gnathostomes (Li *et al*., 2010). The motifs identified in these studies often match characterised binding profiles of developmental transcription factors, strengthening the case for CNEs containing an interpretable *cis*-regulatory code composed of 'conventional' TFBS motifs. The success of these bottom-up approaches, particularly the finding that Pbx-Hox motifs are clearly enriched within multiple different sets of CNEs, raises the question of how they had not been identified previously by top-down, *de-novo* motif search strategies (this issue is confronted in chapter 7).

We have managed to show enrichment for Pbx-Hox motifs within CNEs through two complementary approaches – firstly, identifying conserved instances of the core Pbx-Hox motif within CNE multiple alignments; and secondly, scanning CNE sequences from one species, using the more stringent –KR motif. The footprinting (alignment) approach gains stringency by searching for evolutionarily conserved motif hits, which is appropriate for the Pbx-Hox core motif as positions 1-4 and 7-8 are relatively invariant across characterised binding sites. The strong signal that this approach can provide is demonstrated by the high degree of enrichment for Pbx-Hox motifs within the test sets, particularly using the more divergent lamprey sequences, compared to shuffled alignment control sets. However, the requirement for multiple alignments reduces the size of the search set (as not all CNEs are found to be conserved in all gnathostome genomes, particularly for the teleosts (Venkatesh *et al.*, 2006)). Additionally, the requirement for motif hits to be positionally conserved within alignments prohibits the identification of TFBSs that have shuffled within CNEs. Thus, the increased enrichment signal provided by conserved motifs may be offset by an overall reduction in the number of motifs identified. Searching within CNEs of only one species enables a larger sequence set to be used and identifies motifs that may have changed position within other species, however this may also increase the number of false-positive hits and decrease the strength of the enrichment signal. To counteract this, we used the more stringent –KR motif when searching CNE sets composed of sequences from one species. The motif hits that these two approaches identified are likely to overlap considerably, yet each approach is also likely to have identified some motifs that the other approach did not. Finally, it is probable that there are Pbx-Hox TFBSs within CNEs that these search approaches failed to predict, such as those that may conform to the AGATNNAT motif – this highlights the balancing-act in TFBS prediction between maximising positive hits whilst minimising false-negatives.

In testing for the enrichment of motifs within CNEs, the approach should be designed based on prior knowledge of the motifs in question. It is likely that Pbx-Hox TFBSs represent a class of TFBS with motifs that are amenable to identification using strict consensus searches – they seem to be highly invariant within the core region and they appear to shuffle within CNEs relatively infrequently (as seen by their conservation between duplicated CNEs compared to that of meis motifs). Other TFBSs may be enriched within CNEs but might be harder to identify using strict consensus motifs and

evolutionary conservation, as their motifs may be more variable and may shuffle more frequently within CNEs (these issues are elaborated upon in Chapter 7).

It is important to appreciate the role that phylogenetic footprinting played in the initial discovery of Pbx-Hox motifs within the *meis2* CNEs. The inclusion of lamprey CNEs in the multiple sequence alignments stripped away many of the conserved positions within these CNEs, making the conserved Pbx-Hox and Meis motifs stand out upon manual inspection. This was crucial for their identification, as these motifs are not represented in the JASPAR TFBS motif database, so they were not identifiable by a systematic search for known motifs using JASPAR. Thus, deep sequence conservation within CNEs enables the identification of potentially important motifs without relying upon TFBS motif databases that represent only a modest proportion of factors.

Finally, in the introduction it was mentioned that a systematic search for Pbx-Hox motifs within vertebrate CNEs had not been performed before, despite these motifs being ideal for such a search. However, a study by Ebner *et al*. (2005) scanned the *Drosophila* genome for instances of the consensus Exd-Lab motif (TGATGGAT(T/G)G) accompanied within 40bp by a Hth motif (CTGTCA). Despite identifying 30 putative binding sites, only two were situated within 10kb of genes that had expression consistent with that of *lab*, one of these genes being *lab* itself. The expression of the novel putative *lab* target was subsequently shown to be regulated by a non-canonical Exd-Lab site, and not by the predicted site. This led the authors to conclude that the consensus Lab-Exd-Hth binding sequences are not sufficient to identify *lab* target genes, suggesting that *in-vivo* Hox binding-sites might be more divergent than anticipated and that the stringency of their motif search was probably too high. In light of our findings from vertebrate CNEs, it would be interesting to perform a search strategy similar to ours in *Drosophila* CNEs to investigate whether they contain the same degree of enrichment for canonical Pbx-Hox motifs.

Pbx-Hox motif association with hindbrain and pharyngeal arch expression

The aim of our bottom-up strategy for TFBS motif discovery was not solely to identify enriched motifs within CNEs, but to use these motifs as predictors of CNE function, thereby gaining broader insights into the roles of CNEs in vertbebrate development and evolution. From previous studies, the Pbx-Hox motif is strongly associated with

segment-specific reporter expression in the vertebrate hindbrain and pharyngeal arches. This association is also borne out in the *meis2* CNEs of the *C15orf41* contig. The expression patterns driven by Pbx-Hox +ve elements from the lamprey and gnathostome sets further support this correlation, as does reporter expression data from the EB and cneBrowser datasets. These patterns are consistent with the hypothesis that CNEs containing Pbx-Hox motifs are regulated by Hox factors during development, particularly by anterior Hox factors expressed in the developing vertebrate head. However, it is not clear whether this regulation is in the form of activation or repression, as it could conceivably be either – functional dissection of more CNEs would be informative in this regard. It is also not clear why these elements should only be regulated by anterior Hox factors in the head – other Hox factors bind as heterodimers with Pbx *in-vitro* (Chang *et al.*, 1996), although more posterior factors are less dependent upon Pbx and Meis for DNA binding (Uhl *et al*., 2010), and anterior Hox factors are also expressed in tissues outside the head, such as the somitic mesoderm and axial skeleton. This bias for expression in the head may represent an increased level of complexity within the head compared to other tissues, thus leading to a greater number of enhancer elements, or elements with more complex organisation (therefore more likely to be conserved in sequence between gnathostomes). There may also be an ascertainment bias due to the developmental time-points at which expression was sought – perhaps some Pbx-Hox +ve CNEs function in other domains at later time-points. Whilst our expression data suggest a correlation between Pbx-Hox motifs and hindbrain/pharyngeal arch expression, our results also show that not all CNEs with Pbx-Hox motifs express in these domains.

The association of Pbx-Hox motifs with genes that have characterised roles as Hox co-factors, or that play key roles in Hox-dependent patterning processes, particularly of the hindbrain, is consistent with the hypothesis that Pbx-Hox motifs in CNEs represent gene regulatory interactions between hox factors and their targets, and is in keeping with the observed reporter expression patterns of Pbx-Hox +ve CNEs. The finding that some of these genes have also been identified through microarray studies as being influenced by HoxB1 in r4 provides further support that Pbx-Hox motifs in CNEs can predict Hox-dependent enhancers. However, many of the genes with CNEs enriched in Pbx-Hox motifs also play crucial developmental roles in other embryonic domains, such as the *tshz* genes in patterning the axial skeleton - a role which is also likely to involve being regulated by Hox factors. Thus, whilst many of the Pbx-Hox motifs identified in CNEs

may well be involved in Hox-dependent head patterning, some may contribute to other Hox-dependent patterning networks.

A gene regulatory network for hindbrain patterning is conserved across vertebrates

The conservation of a significant number of Pbx-Hox +ve CNEs in sea lamprey, and the expression patterns that many of them drive in the hindbrain and pharyngeal arches, suggest that aspects of a GRN for hindbrain patterning are conserved across all vertebrates (this topic is dealt with further in chapter 6). These CNEs suggest essential – and, in some cases, previously unappreciated - roles for *tshz3*, *znf503*, *nr2f2*, *meis2* and *evi1* in vertebrate hindbrain patterning and highlight the important role that Pbx-Hox TFBS prediction can play in identifying new HOX target genes.

Importantly, whilst Pbx, Hox and Meis are predicted to play key roles in the regulation of these CNEs, their motifs do not explain all of the conserved regions within these CNEs, suggesting that other factors within the hindbrain GRN also play a role in regulating these elements. As GRNs are likely to be 'recursively wired', with kernels in the network consisting of a number of factors that all regulate each other (Davidson, 2006), it is probable that many of the genes with Pbx-Hox motifs in their CNEs interact with each other, with Pbx-Hox +ve CNEs representing many of the links within this network. Such '*cis*-regulatory embraces' have already been uncovered between Hox factors and other contributors to the hindbrain patterning network, such as *krox-20* and *rarb* (Serpente *et al*., 2005). More evidence for this model could be gained through systematic scanning of CNEs for relevant TFBS motifs, however this is limited by the paucity of characterised PWMs for these factors. It could be the case that binding-site motifs for some of these hindbrain patterning factors may significantly co-occur within certain CNEs. In which case, grouping CNEs according to function and to the presence of certain motifs (ie. Pbx-Hox, Meis), coupled with *de-novo* motif prediction, could be a useful approach to identify new motifs and possible patterns of co-occurance for factors involved in the hindbrain GRN.

The identification and characterisation of other motifs within Pbx-Hox +ve CNEs, coupled with testing them by reporter assay, could also refine predictions of CNE expression patterns. Furthermore, co-occurrence of multiple heterotypic motifs could

ultimately be used as a method to search genome-wide for Hox-regulated hindbrain enhancers without having to rely upon sequence conservation. This could facilitate the identification of new enhancers in gnathostomes, as well as providing a means for identifying homologous enhancers deeper in the chordate lineage (ie. in lamprey and amphioxus), through the identification of conserved signatures.

Mechanism of CNE action

Characterisation of enhancers has led to two different models of enhancer organisation, which have different expected evolutionary characteristics with regard to sequence conservation (Cameron & Davidson, 2009). 'Billboard'-type enhancers are relatively unrestricted with regard to the order and orientation of TFBSs, with functional output determined by the sum of transcription-factor input activities. Over large evolutionary time-spans the TFBSs within these enhancers can shuffle extensively, resulting in functional conservation despite sequence divergence. In contrast, 'enhanceosome'-type enhancers (based upon the interferon-β enhancer) serve as rigid assembly platforms for protein complexes, relying upon the precise composition, order, orientation and spacing of TFBSs, and resulting in high sequence conservation across large evolutionary distances (Merika & Thanos, 2001). Most enhancers are likely to lie between these two extremes. In this light, the deep evolutionary conservation of CNEs may reflect their action through the binding of large cohorts of transcription-factors in precisely ordered complexes and positions relative to one another.

With respect to hox-responsive enhancers, the term 'hoxasome' has been coined, describing protein complexes consisting of hox factors, their co-factors and other factors that contribute to the output of the enhancer (termed collaborators) (Mann *et al*., 2009). Naming these complexes specifically after hox factors reflects the crucial role for hox proteins in these complexes, as interfering with hox binding-sites frequently abrogates binding of the protein complex. Our data suggests that many CNEs could be considered to be hoxasomes.

The prediction that Pbx-Hox +ve CNEs are regulated by Hox factors enables us to interpret the reporter expression patterns of these CNEs in terms of the expression domains of Hox factors. By doing so we can speculate over the mechanisms by which these CNEs integrate positional information from various factors. Some Pbx-Hox +ve

CNEs show broad expression patterns within the hindbrain domains of hox expression, consistent with them being bound by a Hox factor across this whole domain and it being able to up-regulate reporter expression in all the cells in which it is bound. More frequently, however, reporter expression is limited to only certain subsets of cells within the expression domain of a Hox factor, showing dorso-ventral and medio-lateral restriction. In these cases, what is the role of the Pbx-Hox complex? It may only activate expression when in the presence of collaborators, which also have restricted expression patterns. If this is the case, how do these factors interact and do they bind co-operatively to the enhancer? Synthetic enhancers consisting of repeated Pbx-Hox binding-sites are able to drive robust and broad, but segment-specific, reporter expression (e.g. Pöpperl *et al*., 1995), suggesting that the role of the hox factors may be to provide relatively broad activation within particular A-P domains, which can then be sculpted by collaborating factors through repression. Many of the Hox-responsive enhancers that have been characterised so far drive reporter expression across whole rhombomeres, yet there are some examples of enhancers that integrate A-P Hox regulatory signals with D-V regulation (e.g. Samad *et al.*, 2004).

Another question relates to the tissue specificity of these enhancers – presumably other factors play roles in restricting or facilitating expression in certain tissues. Finally, it is conceivable that some Pbx-Hox binding-sites may act via repression. The set of Pbx-Hox +ve CNEs that we have identified will be a useful resource for investigating the mechanisms by which enhancers determine Hox-specificity and tissue specificity, whilst integrating the anterior-posterior patterning information from Hox factors with dorso-ventral and medio-lateral cues from other factors.

Patterns of evolution of duplicated CNEs

Based upon our selection of duplicated CNEs, Pbx-Hox motifs appear to be frequently conserved between duplicates. The conservation seen for the invariant positions of the Pbx-Hox motif is in keeping with the essential roles of these positions in binding of the Pbx-Hox heterodimer. Whilst nucleotides at ‘variable’ positions 5, 9 and 10 were seen to frequently differ between duplicates, as well as between orthologs, position 6 appeared to be more constrained. This is interesting, because it can contain A, T or G in different CNEs, but the identity seems to be constrained between homologous elements. Whilst mutating the variable positions of a Pbx-Hox site might not prevent the Pbx-Hox

complex from being able to bind, these positions may still be important in determining the specific Hox protein that binds, so they could be under evolutionary constraint. This might imply a greater role for position 6 than position 5 in determining Hox specificity. Whilst mutations at position 6 alone have been shown to be able to alter Hox specificity in-vitro (Phelan & Featherstone, 1997) as have combined mutations at positions 5 and 6 in-vivo (Chan *et al*., 1997), the relative contributions of positions 5 and 6 to Hox specificity have not been investigated.

Of the motifs in which the invariant positions are divergent between duplicates, those with mutations at position 1 - from T to A - are likely to still be functional, as Pbx-Hox binding-sites of this type have been characterised previously. However, there are no examples in the literature of Pbx-Hox sites with C at position 8. Duplicated motifs in which this mutation has occurred are found twice in our selection of elements. In one of these, associated with meis1, this mutation is conserved between teleosts, with the rest of the motif adhering to the strict Pbx-Hox consensus motif, suggesting that it is still under heavy evolutionary constraint. It is possible that these mutations could still result in functional Pbx-Hox sites, or they might create a new binding-site for a different protein or protein complex.

The study of Tümpel *et al.* (2006) found that duplicated Pbx-Hox TFBSs with mutations causing them to lose their segment specific expression functionality were nevertheless conserved across the rest of the motif, possibly mirroring the situation for some of the motifs described here. The authors suggested that those motifs may also have other, not previously characterised functions leading to their constraint. Assessing the enhancer function of the divergent elements described here could inform us whether this is a common phenomenon.

In comparison to the Pbx-Hox motifs, the Meis motifs appear to show a greater propensity for shuffling between orthologous and paralogous CNEs. This may be due to the short binding-site for the Meis protein making it more likely that Meis sites can evolve afresh within a DNA sequence. In-vitro studies have suggested that the position and orientation of the Meis site relative to the Pbx-Hox site does not strongly influence formation of the Pbx-Hox-Meis complex on DNA (Jacobs *et al*., 1999). Nevertheless, it could be possible that the changes in the positions of Meis sites observed for some of these elements may have a functional significance *in-vivo*.

Investigating the functional ramifications of these patterns of sequence retention and divergence by reporter assay could be enlightening. As well as enabling the impact of naturally occurring mutations upon the function of Pbx-Hox TFBSs to be inferred, with the potential of refining the Pbx-Hox TFBS consensus motif, comparisons of expression patterns between duplicate pairs could inform us on the role that other TFBSs play in the function of these enhancers. For instance, it has been demonstrated that TFBSs neighbouring a Pbx-Hox site can influence the hox specificity of this site (Li *et al.*, 1999). Our duplicated Pbx-Hox +ve CNEs often show retention of the Pbx-Hox motif but divergence of the neighbouring sequences – the effect that this has on reporter expression patterns would be of great interest, in terms of both the mechanism of enhancer action and of *cis*-regulatory evolution.

Another interesting question is whether the patterns of sequence and functional divergence between dCNEs are the same as those between the teleost co-orthologs. This could be informative as to whether the evolutionary processes acting upon duplicated CNEs were the same for the early genome duplications (2R) and later ones (e.g. teleost-specific), or whether the evolutionary ramifications of whole genome duplication can be different depending upon the context of these duplications. For instance, one might speculate that aspects of the GRNs underlying vertebrate development may have been more pliable at the time when the first rounds of genome duplication took place, at the base of the vertebrate lineage. In which case, dCNEs may show a greater tendency for neo-functionalisation, forging new links in GRNs, whereas duplicated CNEs arising from the teleost-specific genome duplication may have found themselves in more rigid, hard-wired GRNs and thus might show a greater tendency for sub-functionalisation.

The role of Pbx-Hox +ve CNEs in vertebrate evolution

A complex head patterned by Hox factors is a key vertebrate innovation. We ascertained in chapter 3 that hundreds of CNEs can be found conserved between lamprey and gnathostomes, yet a larger number of gnathostome CNEs are not identifiable in lamprey. Looking deeper in evolutionary time to invertebrate chordates, only a very small fraction of gnathostome CNEs have been found. When this pattern of conservation is viewed in the context of the findings in this chapter, regarding the multitude of CNEs that are likely to be regulated by Hox factors and associated with a

GRN for head patterning, it evokes an evolutionary scenario in which many CNEs evolved early in the vertebrate lineage to co-ordinate a Hox-dependent GRN for head patterning. This scenario is discussed in more detail, with reference to lamprey and amphioxus hindbrain patterning, in the next chapter.

Conclusion

The lack of knowledge relating to the functions and mechanism of action of CNEs makes it difficult to place these elements within the context of GRNs. Whilst these elements are assumed to operate through the binding of TFs, little progress has been made in characterising TFBSs within them. I sought to identify TFBS motifs in CNEs by performing phylogenetic footprinting, leveraging the increased sequence divergence of the lamprey elements as a guide. I have discovered that various sets of vertebrate CNEs, defined by conservation across different evolutionary distances, are enriched for Pbx-Hox motifs. The sequences of these motifs, the patterns of expression driven by the CNEs containing them, and their association with particular genes are all consistent with the hypothesis that they are regulated by Hox factors, with an apparent bias toward hindbrain and head development. These findings show that many CNEs contain TFBSs of the type that have been characterised for other, less well conserved elements. They are consistent with the notion that CNEs are composed of complex arrangements of TFBSs. Importantly, the identification of these TFBS motifs suggests that there may be other enriched motifs within CNEs. Furthermore, the identified Pbx-Hox +ve CNEs provide an opportunity to investigate the *cis*-regulatory evolution of CNEs, as they enable the functional divergence of CNEs to be interpreted within the framework of conservation/divergence of well characterised TFBS motifs. The expression patterns of Pbx-Hox +ve CNEs suggest that many of these elements represent links in a GRN for hindbrain and head development, aspects of which are likely to be conserved across all vertebrates. Key questions include whether there are other TFBS motifs enriched within CNEs, and whether the *cis*-regulatory functions of CNEs tested in zebrafish can be generalised to their homologs across other species.

6 Development of a Lamprey Reporter Assay

Abstract

Many CNEs have been shown to function as developmental enhancers by reporter assay within their host species, yet the degree to which these CNEs have conserved functions between species is investigated less frequently. From another angle, it is not clear whether their gene-regulatory function can evolve in a lineage specific manner. These are important questions, as they can inform us on the extent to which CNEs reflect conserved gene regulatory networks between species, whilst enlightening us on the mechanisms underlying *cis*-regulatory evolution. This chapter describes the development of a reporter assay in lamprey embryos and its use to investigate the functional evolution of CNEs. The CNEs of the *c15orf41* gene region are useful subjects for investigation in lamprey because I already have clues as to their mechanism of action. This enables me to correlate the reporter expression they drive in lamprey with the expression patterns of the factors that are predicted to be regulating them, thus placing them within the context of a gene regulatory network for hindbrain patterning. The patterns of functional divergence of these CNEs between lamprey and zebrafish are discussed, and a model linking Pbx-Hox +ve CNEs with the evolution of the gnathostome hindbrain is postulated.

Introduction

Do CNEs drive conserved expression patterns across lineages?

The degree to which CNEs show lineage-specific expression patterns is unclear. Despite CNEs having high sequence conservation, it is conceivable that the expression patterns driven by orthologous CNEs in their respective species could be significantly different. The divergence in expression patterns driven by duplicated CNEs when tested in the same species indicates how similar sequences can have significantly divergent *cis*-regulatory function (McEwen *et al*., 2006). As these duplicated CNEs were tested in the same species, the changes in expression pattern between them are due solely to differences in their sequences (changes in *cis*-). Between orthologous CNEs tested in their respective species these differences could be due to changes in both *cis*- and *trans*-

(the expression patterns or interactions between the transcription factors binding to CNEs, and the influence of the epigenetic machinery and cellular environment). There are three conceivable types of expression change that could happen between orthologous CNEs:

1. Change of expression domain. Developmental programs for different tissues or characters can utilise common gene-regulatory sub-circuits or 'plug-ins' (Davidson, 2006). A *cis*-regulatory element that is regulated by components of such a plug-in may have the potential to switch function to the development of a different character, through a relatively simple change in only one of its regulatory inputs.
2. Gain or loss of expression domains. Multi-functional elements could gain additional domains of expression or lose some domains, whilst their other domains stay conserved.
3. Modification of expression within a domain. In this case, orthologous CNEs would function in the development of the same morphological character, but would drive slightly different spatio-temporal expression.

Another regulatory change could result in conservation of function between orthologs, with changes in *cis*- compensating for changes in *trans*-. Importantly, changes in *cis*- or *trans*- could conceivably account for each of these types of expression divergence. For some regulatory interactions it has been possible to tease *cis*- and *trans*- effects apart – for instance the binding of certain transcription factors in hepatocytes has been compared between human and mouse, revealing conservation in *trans*- but significant divergence in *cis*- (Odom *et al*., 2007; Wilson *et al*., 2008). For CNEs, this separation of *cis*- and *trans*- effects can be achieved by testing pairs of orthologous elements by reporter assay within and across species.

An analysis of CNEs around the vertebrate *iroquois* gene clusters included assaying homologous CNEs from zebrafish and frog for reporter expression in their respective species (de la Calle-Mustienes *et al.*, 2005). Three pairs of CNEs were found that drove expression in both species, with the expression patterns in zebrafish and frog showing some agreement but also striking differences. For instance, one zebrafish element drove expression only in the forebrain in zebrafish embryos, whilst its counterpart from xenopus also drove expression in the midbrain and eye of frog embryos. The authors

suggested that differences in the assay methodologies (the zebrafish assay was very mosaic compared to the frog assay), in the embryos (the greater transparency of zebrafish embryos may reveal more reporter expression) and in the sequences of the CNEs (evolution in *cis*-) could each play a role. As they did not test any of the zebrafish elements in frog, nor the frog elements in zebrafish they were unable to gain insights into the relative roles of *cis*- and *trans*- evolution for these elements.

Navratilova *et al.*, (2009) have compared the expression patterns driven by human elements in zebrafish and mouse. Of six elements, associated with either *Sox3* or *Pax6*, four gave essentially the same patterns in transgenic mice and zebrafish. These comparisons are confounded by the data being derived from independent investigations, with different lengths of sequence being used for the assays in each species. As flanking sequences may influence the expression pattern of a conserved enhancer, the inclusion or exclusion of these sequences could alter the overall pattern driven by an element. Nevertheless, comparison of the broad expression domains, such as particular regions of the developing brain, showed the human elements to behave reasonably similarly in zebrafish and mouse assays – suggesting that little evolution in *trans*- had occurred between mouse and zebrafish for the factors regulating these elements. The orthologous zebrafish elements were also tested in zebrafish, generally showing a concordance of expression pattern with the human elements in zebrafish.

The expression patterns driven by orthologous CNEs from the human and amphioxus genomes were compared using reporter assays in mouse and amphioxus (Holland *et al.*, 2008). Orthologous CNE pairs were found to drive similar expression patterns in mouse embryos and were also able to up-regulate reporter expression in an amphioxus assay. However, the mosaicism of the amphioxus assay made interpretation of the expression patterns difficult. Further, the large flanking regions included in the injected sequences, which in human contained other CNEs not found in amphioxus, confounded the comparison.

From these examples, it is clear that differences between the transgenesis methodologies used for each species could have an influence on the expression patterns obtained. In order to infer any details regarding *cis*-regulatory evolution by cross-species reporter assay comparisons, experiments should be designed that eliminate such factors. Important factors to control are mosaicism, sequence length and the stages at which

reporter expression is characterised. If these are controlled, then divergence between expression patterns can be attributed to changes in *cis*- or in *trans*- by comparing the patterns obtained from an experiment in which pairs of orthologous elements are assayed both within and between species.

The utility of a lamprey reporter assay

The phylogenetic position and body plan of the sea lamprey make it a unique model for investigating the developmental changes responsible for the transition from ancestral invertebrate chordates to jawed vertebrates. Due to the availability of large numbers of embryos, lamprey can be used as a developmental model (Nikitina *et al*., 2009), but an efficient reporter assay has not previously been developed. This would be useful for investigating CNE functional conservation across large phylogenetic distances. On a broader scale, a lamprey reporter assay could be used to investigate many aspects of vertebrate gene-regulatory evolution.

Results

Kusukabi *et al*., (2003) conducted the first application of transgenesis in an agnathan, using a close relative of the sea lamprey - the Japanese lamprey, *Lampetra japonica*. In their study, circular plasmid constructs containing the GFP coding sequence downstream of either a viral promoter or 5' regulatory regions of medaka actin genes were injected into fertilised eggs before the first cleavage. Injections resulted in roughly 50% survival beyond day 2, with highly mosaic GFP expression in 20% to 40% of survivors, depending on the construct injected. Interestingly, muscle-specific actin promoters from medaka were able to drive muscle-specific GFP expression in lamprey embryos. Whilst hinting at the existence of a pan-vertebrate gene-regulatory mechanism for muscle development, this study also demonstrated the feasibility of lamprey transgenesis, serving as a useful starting point for the development of a more efficient reporter assay that would be suitable for testing CNEs in lamprey.

The two key hurdles to overcome in the development of a lamprey reporter assay are the identification of a suitable minimal promoter and the reduction of mosaicism of reporter expression. The tol2 plasmid utilised for zebrafish transgenesis - pGW_cfosEGFP (Fisher *et al.*, 2006) - has a mouse *cfos* minimal promoter. The cfos_pm3285 construct,

with the lamprey CNE 3285 (pm3285) cloned upstream of the promoter, was used to test whether the *cfos* minimal promoter could function in lamprey. pm3285 was selected as the first element with which to test the promoter as it drives strong and consistent expression in zebrafish embryos.

Whilst Kusukabi *et al*., (2003) injected circular plasmid DNA, linearised plasmid DNA would be expected to have a higher chance of genomic integration, so the cfos_pm3285 plasmid was linearised with KpnI (see Materials and methods). Injection of this construct during the first cleavage resulted in a high death rate immediately post-injection, as well as during gastrulation, such that the frequency of injected embryos surviving through gastrulation was less than 5% (Table 6.1) and they were often deformed. Of 18 survivors, 10 showed mosaic GFP expression in the ectoderm, with 5 of those also exhibiting mosaic expression in the nervous system and 2 expressing in the cranial ganglia (Figure 6.1, Table 6.1). The neural expression is in agreement with that driven by the same construct using the tol2 assay in zebrafish, suggesting that the *cfos* minimal promoter is capable of up-regulating CNE-dependent reporter expression in lamprey. Accordingly, the empty reporter construct, containing no enhancer, drives GFP expression only in the ectoderm. However, the survival rate was so low for this assay that the exact nature of any background expression driven by the *cfos* promoter could not be reliably inferred.

To decrease mosaicism and increase embryo survival, more sophisticated transgenesis methods were used. I reasoned that by increasing the probability of early genomic integration of the injected construct, the amount of DNA injected could also be lowered, lessening the toxic effect of exogenous DNA whilst decreasing mosaicism. The tol2 assay is highly efficient in zebrafish (Fisher *et al*., 2006) however, using the same method in lamprey embryos resulted in an extremely low survival frequency (data not shown) and no reporter expression.

The I-SceI meganuclease-mediated transgenesis method utilises the rare-cutting I-sceI meganuclease (Ogino *et al*., 2006). This technique requires a construct in which the DNA to be integrated is flanked by I-sceI recognition sites. This construct is digested with the meganuclease enzyme *in-vitro* and the reaction mix is injected immediately into fertilised eggs. Whilst the integration mechanism is unclear, it is considered unlikely that the enzyme cuts genomic DNA, as its 18bp recognition site is predicted to

occur once in every 7 X 10^{10} bp of genomic sequence (compared to the medaka genome size of 7 x10^{8} and lamprey 2.3x10^{9}) (Thermes *et al*., 2002). Rather, it has been suggested that the continued association of the enzyme with the digested construct prevents its degradation or concatamerisation, thus increasing the probability of genomic integration (Thermes *et al*., 2002).

The β-globin_egfp construct utilised in the zebrafish co-injection assay consists of GFP under the control of a mouse β-*globin* minimal promoter, with the promoter-reporter cassette flanked by *I-sceI* recognition sites. When pm3285 was cloned upstream of the promoter in this construct and injected using the meganuclease-mediated method, the construct was cleaved as expected and embryo survival was high, but no GFP expression was observed. The cfos-IsceI-EGFP construct was created by replacing the β-*globin* promoter with that of mouse *cfos* (see Materials and methods for plasmid map). Injecting this construct with no enhancer cloned into it, using the meganuclease method, resulted in a high rate of embryo survival (66% post-gastrulation), with a large proportion (87.5%) of survivors showing mosaic GFP expression in the ectoderm (Table 6.1, Figure 6.1). As this mosaic ectodermal expression is driven by the construct in the absence of an enhancer it can be considered 'background' expression.

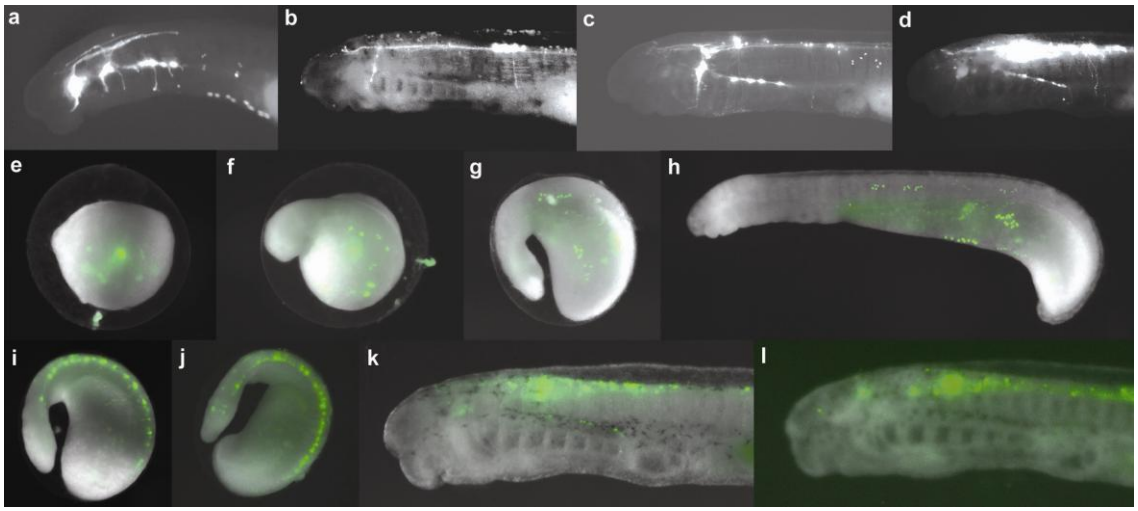


Figure 6.1. Development of a reporter assay in lamprey embryos. **a-b** Stage 26 transgenic lamprey generated through injection of the linearised cfos_pm3285 plasmid show mosaic GFP expression in the cranial ganglia (**a**) and neurons of the spinal cord (**b**), with low survival and frequent deformity. **c-d** Stage 26 lamprey embryos with GFP expression in the cranial ganglia (**c**) and spinal cord (**d**), obtained using the meganuclease assay with lamprey CNE 3285. **e-h** GFP expression in the ectoderm of stage 19 (**e**), 21 (**f**), 23 (**g**) and 25 (**h**) embryos, driven by the *cfos* minimal promoter in the absence of an enhancer (embryos shown are different individuals). **i-l** The lamprey homologs of CNE 3285 and CNE 3299, despite directing enhancer-specific expression (shown in **fig 2** and **3**), also both up-regulate GFP expression in neurons of the spinal cord in a proportion of transgenic embryos (see **table 1**). GFP expression in primary neurons of the spinal cord is shown in stage 24 (**i,j**) and stage 27 (**k,l**) embryos, driven by CNEs 3285 (**i,k**) and 3299 (**j,l**).

Method	Construct	Plasmid conc. /ngul	Embryos injected (approx.)	Survivors	Ectodermal background expression	Enhancer - dependent neuronal expression	Enhancer-specific expression
Linearised plasmid	cfos_pm3285	100	500	18	10	5	2
Meganuclease	cfos_IsceI_pm3285	20	600	220	not counted	35	10
Meganuclease	cfos_IsceI	20	350	232	203	1	NA
Meganuclease	cfos_IsceI_pm3299	20	700	302	not counted	56	8

Table 6.1. A comparison of lamprey transgenesis methods. Results are shown for linearised plasmid injection and meganuclease-mediated transgenesis. For the meganuclease method, results from three different constructs are shown, including the cfos_IsceI construct in the absence of an enhancer. Skin background expression was not counted for the injections of pm3285 and pm3299 with the meganuclease method, but in both cases the proportion of embryos with this background expression was roughly in keeping with that found for the cfos_IsceI construct.

Two lamprey CNEs, pm3285 and pm3299, were tested for enhancer activity in lamprey using the meganuclease assay. As mentioned above and described in the previous chapter, in zebrafish embryos pm3285 drives GFP expression in the cranial ganglia and primary neurons of the hindbrain and spinal cord, with pm3299 driving expression in the anterior hindbrain. Using the meganuclease assay, the pattern of GFP expression driven by pm3285 in lamprey is similar to that obtained by the injection of the cfos_pm3285 linearised plasmid, except it is less mosaic (Figures 6.1 and 6.2). Expression is observed in the cranial ganglia as well as in primary neurons of the hindbrain and spinal cord, with expression often differing between individuals both in the number of cells expressing and in the intensity of expression. Whilst the background ectodermal expression was visible from early stages and often lost intensity in later stages (Figure 6.1), the earliest neural expression was seen at stage 21 in a low number of cells, becoming more expansive in later stages. 35 embryos had expression in the hindbrain or spinal cord, with 10 of these 35 also expressing GFP in the cranial ganglia. Surprisingly, pm3299 was also found to up-regulate GFP in a similar pattern to pm3285, with background expression in the ectoderm and enhancer-dependent expression in primary neurons (Figure 6.1). However, unlike pm3285, pm3299 was never seen to direct expression to the cranial ganglia, whilst clear expression in the

anterior hindbrain was seen in a number of embryos (Table 6.1, Figure 6.3) – an expression pattern that is consistent with its pattern in the zebrafish assay. Thus, pm3285 and pm3299 direct GFP expression to discreet 'enhancer-specific' domains, namely the cranial ganglia and anterior hindbrain respectively, in a low number of injected embryos. However, both elements also drive very similar expression patterns to each other in primary neurons of the hindbrain and spinal cord in a larger number of embryos. As this expression is reliant upon an enhancer (the cfos promoter alone does not drive this expression) but is not different between the two enhancers, it is termed 'enhancer-dependent' expression.

Expression in the cranial ganglia driven by pm3285 is present in all but the most anterior cranial ganglia (Figure 6.2). This is in agreement with the reporter expression of both pm3285 and dr3285 in zebrafish, in which GFP is up-regulated in clusters of cranial ganglia both anterior and posterior to the otic vesicle. The intensity of expression in lamprey differed somewhat between ganglia within an embryo in a manner that was not consistent across embryos, with some embryos also showing GFP expression in regions of the spinal cord. Two lamprey genes showing homology to jawed vertebrate *Meis* genes have been identified and named *pmMeis1/2a* and *b* (Sauka-Spengler *et al*., 2007). The expression patterns of these two genes at this developmental stage are very similar to each other, with both showing clear cranial ganglia expression (Figure 6.2). However, the expression of these genes is restricted to posterior cranial ganglia, so does not entirely overlap the expression driven by pm3285, which extends to more anterior ganglia. Both pm3285 and dr3285 also drive expression in primary neurons of the hindbrain and spinal cord in zebrafish embryos (Figure 6.2). As mentioned above, pm3285 drives expression in these domains in lamprey embryos but it is not clear to what extent this expression is controlled by the enhancer as pm3299 also drives a very similar expression pattern.

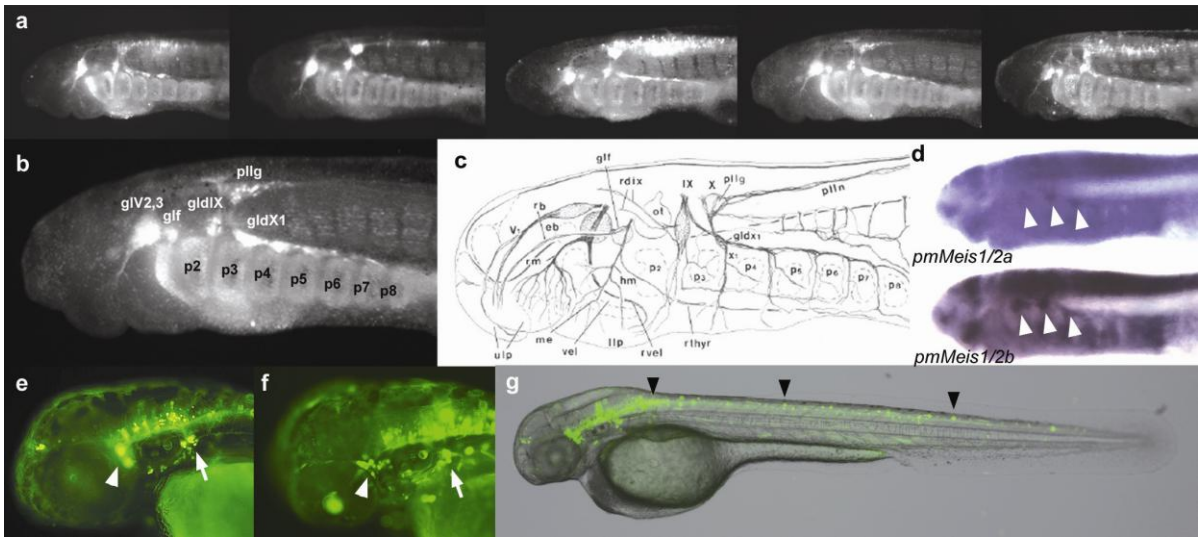


Figure 6.2. Comparison of GFP expression driven by CNE 3285 in lamprey and zebrafish embryos. **a** GFP fluorescence in stage 26 transgenic lamprey embryos, generated by meganuclease-mediated transgenesis with the cfos_IsceI_pm3285 construct. In each embryo shown, expression is seen in the cranial ganglia, whilst the pharyngeal pouches produce autofluorescence. **b,c** Characterisation of the cranial ganglia with GFP expression (**b**) by comparison with an anatomical figure from Kuratani *et al.* (1997) (**c**), in which cranial nerves and their ganglia are described for a stage 27 larval lamprey. p2-8: pharyngeal pouches 2-8; glV2,3: trigeminal ganglion; glf: facial ganglion/anterior lateral line ganglion; gldIX: epibranchial ganglion of the glossopharyngeal nerve; pllg: posterior lateral line ganglion; gldX1: epibranchial ganglion of the first vagal nerve. **d** Expression of *pmMeis1/2a* and *b* genes in stage 25 lamprey embryos, revealed by *in-situ* hybridisation. Arrows highlight expression in cranial ganglia. **e-g** GFP expression of zebrafish (**e**, **g**) and lamprey (**f**) homologs of CNE 3285 in 54hpf zebrafish embryos, using a tol2 assay. Expression in the cranial ganglia is seen for the elements from both zebrafish and lamprey. Ganglia anterior to the otic vesicle are likely to be the trigeminal or facial ganglia (arrowheads), with those posterior to the otic vesicle being the vagal and posterior lateral line ganglia (arrows) (**e,f**). Expression is also seen in primary neurons of the hindbrain and spinal cord (**g** – arrowheads).

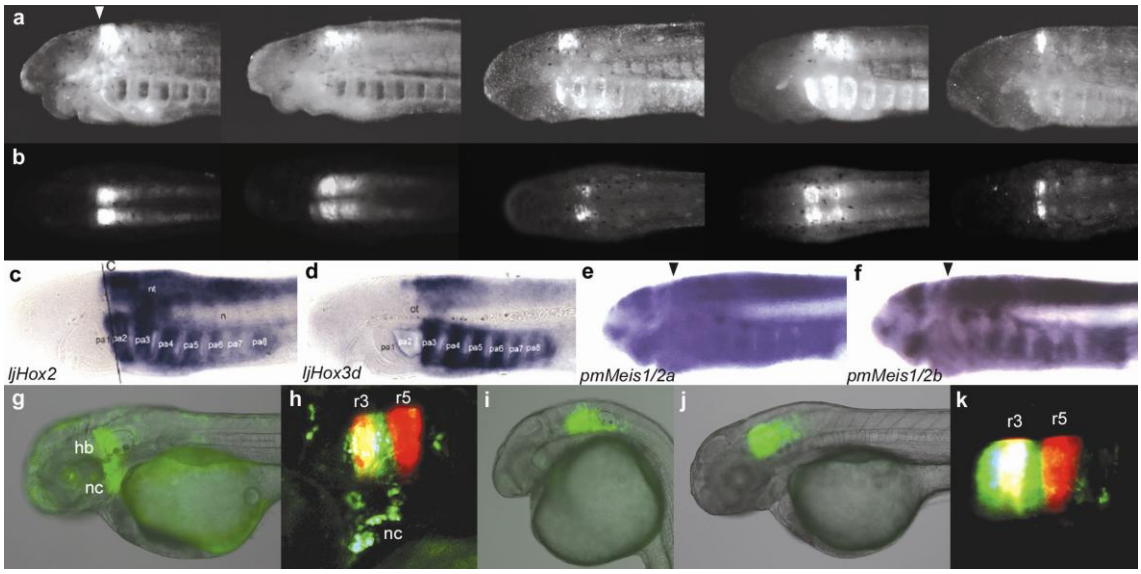


Figure 6.3. Comparison of GFP expression driven by CNE 3299 in lamprey and zebrafish embryos. **a** GFP fluorescence from stage 26 lamprey embryos transgenic for the cfos_IsceI_pm3299 construct, generated by meganuclease-mediated transgenesis. Lateral views show GFP expression in the hindbrain, with a clear anterior limit (arrowhead). **b** Dorsal views of the embryos from **a** show expression on both sides of the neural tube, with low mosaicism and, in the majority of cases, clear anterior limits. **c-d** Expression of *ljHox2* (**c**) and *ljHox3d* (**d**) in stage 26 *Lampetra japonicum* embryos (from Takio *et al.*, 2007). **e-f** Expression of *pmMeis1/2a* (**e**) and *b* (**f**) in stage 25 *Petromyzon marinus* embryos, with anterior limits of hindbrain expression (arrowheads) being consistent with that of *ljHox2*. **g-k** GFP expression driven by the zebrafish (**g,h**) and lamprey (**i,j,k**) homologs of CNE 3299 in 54hpf zebrafish embryos with a tol2 reporter assay (as presented in the previous chapter). GFP expression by the zebrafish element is directed to the hindbrain (hb) and neural crest (nc) (**g**). Hindbrain GFP expression is restricted to rhombomeres (r)3 and 4, in comparison to RFP expression driven by a Krox20 r3r5 enhancer (**h**). GFP expression driven by the lamprey element is seen in the hindbrain at 30hpf (**i**) and 54hpf (**j,k**). Expression from the lamprey element is directed to r2-4 (**k**) but is not seen in the neural crest (**j,k**).

The hindbrain expression driven by pm3299 in lamprey has an anterior limit consistent with that of *ljHox2* in *Lampetra japonicum* (Figure 6.3). This anterior limit of hindbrain expression is in accordance with the expression pattern of the two identified lamprey *meis* genes and also with the pattern of expression driven by pm3299 in zebrafish, which is restricted to r2-4 – the domain of expression of Hoxa2 (Figure 6.3). In zebrafish, the expression of dr3299 differs from that of pm3299, as it is restricted to r3-

4 in the hindbrain and is also present in the neural crest cells migrating into the corresponding pharyngeal arch (Figure 6.3).

Discussion

Development of a reporter assay in lamprey embryos

The I-sceI meganuclease reporter assay described in this chapter is an improvement upon the previously published lamprey assay, due to the higher embryo survival rate and the lower mosaicism of reporter expression. These improvements arise through the use of the I-sce I meganuclease enzyme in the injection protocol. It is likely that this enzyme facilitates early integration of the linearised construct by protecting it from degradation and concatamerisation. This enables a lower concentration of DNA to be injected, which is the probable reason why this assay is less toxic to lamprey embryos than injecting DNA alone. This assay also makes use of the mouse *c-fos* promoter, showing that it has promoter activity in lamprey embryos as well as in gnathostomes.

Enhancer-specific, intense and non-mosaic reporter expression can be observed for a modest, but still significant, number of injected embryos. However, intense reporter expression in primary neurons of the hindbrain and spinal cord is a feature of the expression patterns driven by both enhancers tested, and is seen more frequently than the enhancer-specific expression. This expression is not 'background' expression as the promoter alone only drives weak, mosaic expression in the ectoderm. However, this expression may represent 'promoter bias' for these neurons that is only exerted when the promoter acts in conjunction with a proximal enhancer. Testing other enhancers with this promoter, or other promoters with these enhancers, would shed more light on this issue.

Lamprey represents a useful model organism for investigating the evolution of the vertebrate body plan, due to its phylogenetic position and its lack of certain features that are characteristic of jawed vertebrates, such as a jaw and paired limbs. Whilst regulatory elements from lamprey have been tested in jawed vertebrate reporter assays (Carr *et al*., 1998), to be able to thoroughly answer questions about gene regulatory conservation and divergence it is essential to also test these elements in lamprey. Furthermore, testing the function in lamprey of enhancers involved in the development of gnathostome-

specific traits could give insights into the *cis*-regulatory changes underlying the evolution of these morphological characters in the gnathostome lineage. Thus, the development of the I-sceI meganuclease reporter assay in lamprey embryos promises to have a significant impact on investigations into vertebrate gene-regulatory evolution.

Conservation and divergence of CNE function between lamprey and zebrafish

The enhancer-specific expression patterns driven by pm3285 and pm3299 in lamprey show clear similarity to the expression patterns driven by their orthologs in zebrafish embryos. This suggests that these elements carry out broadly similar gene-regulatory functions in lamprey and zebrafish and that this functional conservation is responsible for their sequence conservation. This may have been predicted, as the gnathostome cranial ganglia and hindbrain have clear homologs in lamprey. Nevertheless, there are some interesting differences between the expression patterns driven by the orthologous CNEs in their respective species.

The only clear difference between pm3285 and dr3285 is their expression in zebrafish embryos, in which dr3285 is restricted to neuronal domains whilst expression driven by pm3285 is often seen in other tissues including muscle and eye. These domains may be attributable to 'background' F0 expression, which is not present in F1 transgenics, yet it is not clear why this would occur only with pm3285 and not dr3285. When pm3285 is tested in lamprey, muscle and eye expression is not observed. This might reflect lineage-specific, compensatory changes in *cis*- and *trans*- which result in tight control of expression when the element is tested in its host species but less well regulated expression when tested heterologously. It would be interesting to test dr3285 in the lamprey assay, to see whether it too drives less tight expression when tested heterologously.

In zebrafish, dr3299 and pm3299 both drive expression in the anterior hindbrain. However, the neural crest expression driven by dr3299 is not seen for pm3299. pm3299 does not drive expression in the neural crest in lamprey embryos either. Therefore, the function of this enhancer in the neural crest has either been gained in the lineage leading to zebrafish or lost in the lamprey lineage. The pattern driven by dr3299 in lamprey would inform us whether this change is purely in *cis*-, in which case dr3299 would drive neural crest expression in lamprey, or whether changes in *trans*- also play a part. The

functional dissection of dr3299 in the previous chapter demonstrated that combined mutation of the first Pbx-Hox site and its proximal Meis site abrogated neural crest expression. Interestingly, this region of the enhancer is highly conserved between lamprey, human and fugu, with the zebrafish enhancer showing the greater divergence. This raises the possibility that the neural crest expression domain was gained by the enhancer in the zebrafish lineage, potentially by the modification of only a modest set of transcription factor binding sites. This could be tested by functional assay of orthologous enhancers from other gnathostomes in zebrafish, coupled with a more detailed functional dissection. However, it must be noted that the functional dissection so far only indicates that these sites may be necessary for neural crest expression, whilst other sites within the enhancer, that may also be divergent between zebrafish and lamprey, might also play key roles.

dr3299 and pm3299 also differ in their expression within the hindbrain in zebrafish, with dr3299 expressing in r3-4 and pm3299 in r2-4. As it was demonstrated in the previous chapter that Pbx-Hox sites within dr3299 are essential for its enhancer activity, these elements are likely to be regulated by hox factors in the hindbrain. The expression patterns of dr3299 and pm3299 correspond to the anterior limits of expression of different PG2 hox genes – *Hoxb2a* and *Hoxa2b* respectively, suggesting that in zebrafish these two elements may be differentially responsive to paralogous hox factors. It is notable that mutation of the first Pbx-Hox and Meis sites in dr3299, as well as removing neural crest expression, also led to a broader hindbrain expression pattern, reminiscent of that driven by the lamprey element, whilst mutation of the second Pbx-Hox and Meis sites abrogated expression completely. This suggests that in dr3299 the first Pbx-Hox and Meis sites may prevent expression in r2. If it is the case that these sites can differentiate between different Hox factors from the same paralogy group, the mechanism underlying this would be interesting to investigate. In lamprey, pm3299 expression has an anterior limit in the hindbrain that is consistent with that of the PG2 hox factor *ljHox2* in *Lampetra japonicum*, suggesting that this element is regulated by homologous hox factors in lamprey and zebrafish. However, it is not clear how many PG2 hox genes are present in the *P. marinus* genome, nor what their precise expression patterns are, so the evolutionary significance of the different expression patterns driven by dr3299 and pm3299 in zebrafish is difficult to address. Again, the expression pattern driven by dr3299 in lamprey embryos would be interesting in this regard, as would those of other gnathostomes in zebrafish.

In gnathostomes the majority of anterior hox factors have anterior expression limits that are in register with rhombomere boundaries; indeed, hox gene expression is often used as a marker for rhombomere boundaries (Alexander *et al*., 2009). In *Lampetra japonicum*, *Krox20* and *EphC* are expressed in the hindbrain in the same rhombomere-specific manner as they are in gnathostomes (Murakami *et al*., 2004). However, in contrast to the registry seen in gnathostomes, some lamprey hox factors have anterior expression limits that do not coincide with rhombomere boundaries, despite them showing collinear spatial expression in the hindbrain (Murakami *et al*, 2004; Takio *et al*., 2004, 2007). For instance, the anterior limit of *LjHox2* expression is in r2 but not at the r1/2 boundary and the anterior limit of *LjHox3d* expression is within r3 - not at the r4/5 boundary as in gnathostomes. Furthermore, application of retinoic acid (RA) to *Lampetra japonicum* embryos results in an anterior shift of hox expression and the positions of branchiomotor neurons, but does not have an effect upon *Krox20* expression or reticulospinal neuron positioning. In light of this, the anterior limit of GFP expression driven by pm3299 in lamprey embryos, which appears to correlate closely with hox PG2 expression, is thus likely to be out of register with the r1/r2 boundary, in contrast to the expression pattern driven by pm3299 in zebrafish. If this is indeed the case, then this would represent a difference in *trans*- between zebrafish and lamprey, which changes the expression pattern of CNE 3299 in each species. Thus, the expression driven by this CNE in lamprey and zebrafish appears to be influenced both by changes in *cis*-, resulting in neural crest expression for the zebrafish element but not for the lamprey element, and by changes in *trans*-, due to the different expression patterns of hox factors in lamprey and zebrafish.

Evolution of the vertebrate hindbrain GRN

The expression pattern driven by pm3299 in the anterior hindbrain of lamprey embryos, which is in good agreement with the expression driven by its zebrafish ortholog in zebrafish embryos, verifies that these elements act within a GRN for hindbrain development that is conserved (in some respects) between all vertebrates. This GRN is likely to also involve many of the other Pbx-Hox +ve CNEs characterised in the previous chapter. This is significant as it enables these CNEs to be placed within a developmental and evolutionary context. By comparing hindbrain development between amphioxus, lamprey and gnathostomes, and correlating this with the conservation signal

of Pbx-Hox +ve CNEs, a model can be formulated that infers the role of many CNEs in the evolution of vertebrate head patterning.

Amphioxus shows a nested pattern of *Hox* gene expression in the neural tube (Wada *et al*., 1999; Schubert *et al*., 2006), suggesting that a rostral region of the neural tube is homologous to the vertebrate hindbrain. The re-iterated and co-incident expression of a number of genes - including *shox*, *AmphiKrox* (Jackman & Kimmel, 2002), *islet* (Jackman *et al*., 2000), *AmphiMnx* (Ferrier *et al*., 2001) and *AmphiERR* (Bardet *et al*., 2005) - in territories within the putative hindbrain region hints towards a division of this region into serially homologous domains, similar to the gnathostome hindbrain. However, the overt morphological segmentation that is characteristic of the vertebrate hindbrain is not seen in amphioxus. The amphioxus hindbrain also lacks the expression of genes such as *krox-20* in the stripes that are crucial for rhombomere development in vertebrates. The amphioxus homologs of other crucial regulators of hindbrain patterning and segmentation, such as *kreissler*, also appear to lack the striped hindbrain expression of their vertebrate counterparts (Jackman & Kimmel, 2002). Analysis of *cis*-regulatory elements associated with amphioxus anterior *Hox* genes by reporter assay in mouse and chick showed that they are regulated by RA signalling in a similar manner to those of vertebrates (Manzanares et al., 2000). Yet the enhancers that, in vertebrates, respond to regulation by *krox-20* and *kreissler* and to auto-regulation by the *hox* genes, appear to be absent in amphioxus (Manzanares, 2001). Additionally, the source of the segmentation signal in amphioxus is from the underlying segmented mesoderm (Mazet & Shimeld, 2002), whilst in gnathostomes, hindbrain segmentation is dependent on an A-P RA gradient within the neurectoderm, which influences segmentation factors such as *krox-20*. Thus, in the amphioxus hindbrain, the segmentation mechanism may be independent of RA signalling and hox-expression domains.

As alluded to earlier, the developing lamprey hindbrain is divided into rhombomeres, with segmentation genes such as *krox-20* and *EphC* showing rhombomere-specific expression as seen in gnathostomes. However, the tight registration between hox expression domains and rhombomere boundaries that is characteristic of the gnathostome hindbrain is only apparent for *Hox1* in lamprey r4, with other Hox factors showing anterior limits of expression that are out of register with rhombomere boundaries (Murakami *et al*., 2004; Takio *et al*., 2007). Two patterning mechanisms appear to be employed in the lamprey hindbrain – one correlating with rhombomeres,

which controls the pattern of reticular neurons and involves *Krox20* and *EphC*, and the other being dependent on the nested pattern of Hox expression, which controls specification of branchiomotor neurons and is influenced by RA signalling. These networks seem to only partially interact in lamprey, in contrast to their extensive integration in gnathostomes.

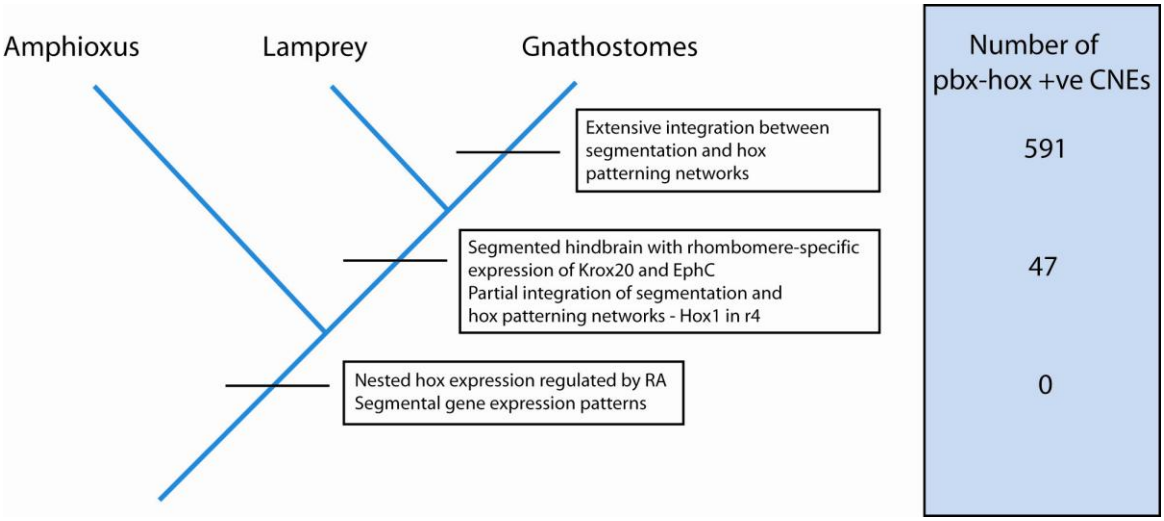


Figure 6.4. Hypothetical scenario for the evolution of the gnathostome hindbrain, with reference to patterns of CNE conservation in the chordate lineage. The inferred hindbrain developmental characteristics of the chordate, vertebrate and gnathostome common ancestors are placed upon a simplified chordate phylogeny, which excludes the derived tunicates. The numbers of Pbx-Hox motif-containing CNEs that have been identified in these common ancestors are shown on the right.

Thus, the gnathostome mechanism for hindbrain development appears to have evolved through sequential steps that can be inferred by comparison with amphioxus and lamprey (Figure 6.4). The key transitions include the evolution of morphological segmentation (rhombomeres) and a multi-step integration between networks governing segmentation and hox-dependent patterning. This integration, combined with the loss of segmentation in the head mesoderm, may have set the stage for further elaboration of downstream, hox-dependent head patterning mechanisms. The integration between networks is likely to have involved the evolution of new regulatory links between hox-factors and components of the GRN for hindbrain segmentation, through the elaboration of *cis*-regulatory elements. Within this context, it is interesting to observe that large numbers of CNEs containing conserved Pbx-Hox motifs become apparent in the vertebrate lineage, before and after the divergence of agnathans and gnathostomes (Figure 6.4). Considering their association with hindbrain and pharyngeal arch

expression (chapter 5), it is likely that the evolution of a more complex, integrated GRN for hindbrain and head patterning involved the evolution and fixation of some of these CNEs. If this is the case, then the conservation pattern for these CNEs across chordates may represent the evolution of new regulatory interactions that are crucial for the development of the complex vertebrate head. This model could be evinced by investigating the expression patterns and regulatory interactions between components of the gnathostome hindbrain GRN in lamprey and amphioxus.

Conclusion

An important issue relating to CNEs is whether their regulatory functions are conserved between the species that contain them. This chapter describes the development of a reporter assay in lamprey embryos that can generate non-mosaic, enhancer-specific expression patterns. A key task remaining in the development of this assay is to clarify whether certain patterns of expression are due to promoter bias. This assay has been used to demonstrate for two lamprey CNEs that the expression patterns they drive in lamprey and zebrafish embryos show clear similarity, mirroring the findings from studies comparing CNE expression patterns between gnathostomes. The implication from this is that the functional roles of CNEs in different species are broadly the same, verifying that CNEs represent conservation of underlying GRNs. Nevertheless, differences between the functional roles of orthologous elements in their respective species were seen, and are likely to have occurred through changes in both *cis*- and *trans*-. This implies that these highly constrained regulatory elements can nevertheless evolve lineage-specific functions through accumulating modest numbers of mutations in their sequences, providing significant scope for evolutionary tinkering. Finally, the lamprey assay provides *cis*-regulatory evidence that aspects of a GRN for hindbrain development are conserved across all vertebrates. This leads to a model of hindbrain evolution in chordates, which hypothesises that many CNEs evolved in the vertebrate lineage to co-ordinate development of the complex vertebrate head. A key question is whether these elements arose afresh in the genome or whether they evolved from previously existing *cis*-regulatory modules.

7 *De-Novo* Motif Discovery in CNEs

Abstract

A major hurdle preventing CNEs from being placed within the context of developmental GRNs is the paucity of information available regarding the TFBSs that they contain. The discovery of enrichment for Pbx-Hox motifs within CNEs, detailed in chapter 5, represented a major step toward de-coding CNEs. The next step is to identify other enriched motifs, enabling a library of motifs to be curated, which could be used to characterise CNEs *in-silico*. The lack of data from *de-novo* motif discovery approaches on CNEs is at odds with our finding of strong Pbx-Hox motif enrichment. In this chapter I use a *de-novo* approach to identify enriched motifs in CNEs. I find a significant number of enriched motifs, some of which correspond to motifs recognised by well characterised TFs. I relate the abundance of these motifs in CNEs to the properties of their predicted binding factors, leading to a general model of CNE action involving the co-operative interaction of facilitator and specifier factors. I also search for a regulatory language within a set of CNEs that are conserved amongst urochordates, finding no overt similarity between the enriched motifs of vertebrate and urochordate CNEs.

Introduction

De-coding CNEs promises to provide a wealth of information pertaining to the GRNs in which they operate and the mechanism by which their *cis*-regulatory language controls gene expression. The identification of enriched motifs within CNEs represents an important step in this characterisation as these motifs would provide a means by which to link CNE sequence with *cis*-regulatory function. Further, by classifying CNEs according to the presence of particular sequence signatures, they could be placed within the context of gene regulatory networks, either known or predicted.

Strategies for identifying motifs within a set of sequences involve either targeted searches, in which sequences are searched for matches to defined motifs, or *de-novo* motif discovery approaches, which characterise all motifs within the set that occur more frequently than would be expected (reviewed by Pavesi *et al*., 2004a). These approaches

are complementary, with findings from one strategy being useful in refining the other. An advantage of *de-novo* strategies is that they require minimal prior knowledge of the motifs that should be searched for, enabling the identification of novel motifs. This is important for CNEs as they may contain motifs that are recognised by complexes of factors binding in combination, which could result in binding-site preferences that are not represented in TF binding profile databases, as they often derive their PWMs from studies in which these factors are considered individually. However, the ultimate aim of *de-novo* motif discovery, when applied to CNEs, is to identify what these motifs are recognised by, requiring either comparison to known motifs or targeted protein-binding assay techniques. Nevertheless, by identifying enriched motifs within CNEs we can address to what extent these elements can be placed within the context of characterised mechanisms of *cis*-regulation.

The finding in chapter 5 that gnathostome CNEs are enriched for Pbx-Hox and Meis motifs confirmed the notion that many of these elements are regulated by transcription factors via recognisable binding-sites, which can be identified through sequence analysis. This is supported by the findings of Bailey *et al*. (2006), who performed a targeted search for Sox, Pou and homeo-domain (HD) motifs within a set of CNEs conserved between human, mouse and Fugu. Their motif scan revealed a general enrichment for these motifs across CNEs, with an association with genes involved in the developing CNS.

The identification of enrichment for these motifs through targeted approaches is hard to reconcile with the paucity of published *de-novo* motif discovery data pertaining to vertebrate CNEs (Pennacchio *et al*., 2006; Li *et al*., 2010). Whilst there are a large number of available tools for motif finding, there is little clear guidance in choosing the best tool for a given set of sequences. The reason for this is that assessment of the performance of tools is not a straightforward task, due to a lack of clear understanding of the regulatory mechanisms of these elements (Tompa *et al*., 2005). Our finding that Pbx-Hox motifs are highly enriched within CNEs provides a standard by which to assess these tools for motif identification. This enables us to select an appropriate motif-discovery tool to address the question of whether the few enriched motifs that have been found in CNEs represent special cases or whether there are others.

Results

CisFinder identifies Pbx-Hox motifs in CNEs

A variety of software tools have been developed to identify enriched motifs within sequence sets (Reviewed by Wei & Yu, 2007). Consensus-based methods (e.g. Weeder (Pavesi *et al.*, 2004b)) enumerate all the possible oligos of a certain length and count the occurrences of each oligo within the test set, allowing for a certain number of substitutions. These counts are then compared to expected counts, for instance within a 'randomly' generated control set, to obtain a statistical measure of the significance of enrichment.

Tompa *et al*. (2005) performed an evaluation of the efficacy of different motif discovery tools using synthetic *cis*-regulatory sequences, finding Weeder to frequently outperform other methods. Recently, a tool has been developed – CisFinder – that estimates PFMs from counts of 8-mer words and clusters them to generate sets of motifs (Sharov & Ko, 2009). This consensus-based method is similar to the Weeder algorithm, but has a significantly faster processing speed, meaning that it can effectively process large sequence sets. It has also been shown to be able to process sequence data containing a low-level enrichment for motifs. These characteristics make CisFinder a suitable candidate tool for searching for motifs within CNEs, as the CNE set is relatively large (the human CNE set comprises 776 Kb, compared to the <32 Kb recommended for efficient use of Weeder (Sandve *et al.*, 2007)) and is likely to contain a functionally heterogeneous mix of elements, thus a relatively low enrichment for motifs compared to other types of data-sets (such as those derived from chip-seq (Valouev *et al*., 2008)). I have utilised CisFinder to search for enriched motifs within a set of 6693 human CNE sequences.

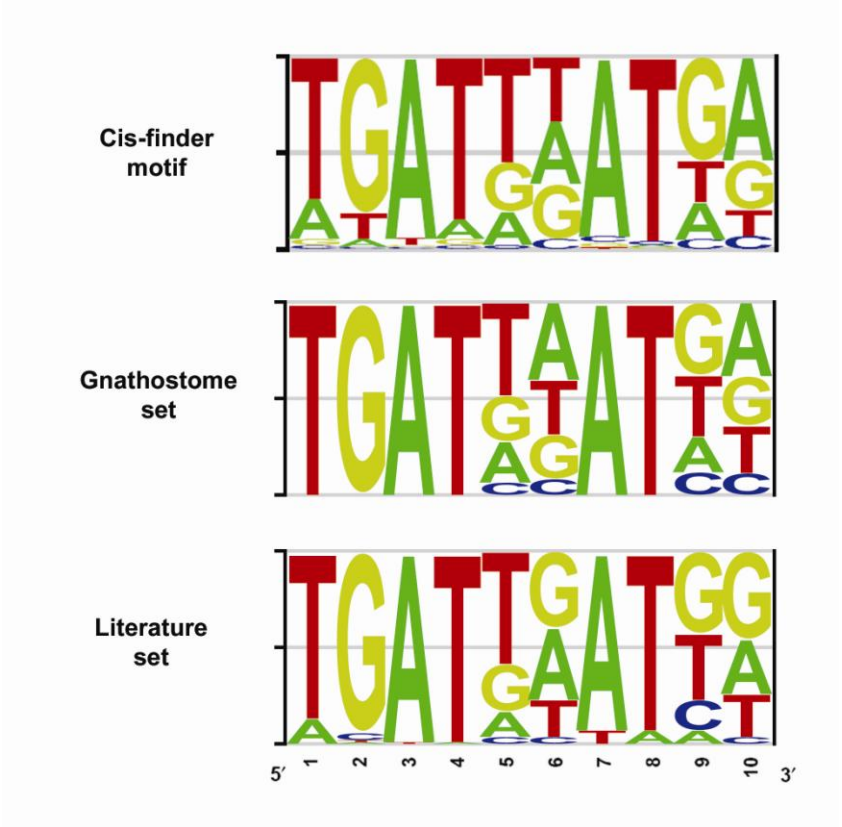


Figure 7.1. CisFinder identifies an enriched motif that closely resembles the Pbx-Hox motif in a set of gnathostome CNEs. 6693 human CNE sequences used for *de-novo* motif identification with *cis*-finder, using clustering similarity of 0.7. Shown are frequency logos for the Pbx-Hox resembling motif found by *cis*-finder, the Pbx-Hox motif hits from the gnathostome CNE set (chapter 5) and from 32 characterised Pbx-Hox motifs (Mann *et al*., 2009; Wassef *et al*., 2008).

Due to our prior knowledge of the enrichment of Pbx-Hox motifs within CNEs, we can judge the efficacy of a *de-novo* motif search tool by asking whether it is able to identify this motif as being enriched in the human CNE set. CisFinder satisfies this criterion, identifying the Pbx-Hox motif as one of its top-ranking (according to z-score and enrichment) motifs within the human CNE set. Comparison of the frequency logos shows that the motif identified by CisFinder is very similar to that derived from the bottom-up search of the gnathostome set, and to previously characterised Pbx-Hox motifs (Literature set) (Figure 7.1). Specifically, these profiles all share a low frequency of C at positions 5 and 6, with a tendency for A/G and T/G at positions 9 and 10. The CisFinder motif has a notable frequency of A at position 1, which is in keeping with the literature set, but was not a part of the Pbx-Hox consensus used for the bottom-up search.

CNEs contain enriched motifs besides Pbx-Hox

When the enriched motifs discovered by CisFinder are ranked according to z-score and enrichment ratio, the Pbx-Hox motif is the third highest ranked motif. Sequence logos for the 27 enriched motifs identified by CisFinder are shown in Figure 7.2. By comparing these motifs to those in the CisView database of vertebrate TFBSs (comprising TFBSs from Jaspar and other literature sources), implementing a comparison tool provided by CisFinder and using a similarity cutoff of 0.8, it can be seen that many of these motifs can be strongly matched with previously characterised motifs.

Motifs 1 and 11 show a strong resemblance to the well characterised 'octamer' motif, recognised by a subset of POU domain proteins (Phillips & Luisi, 2000). Members of this large protein family are broadly expressed during development, particularly in the central nervous system, where they play a role in neural cell differentiation. These proteins bind to the octamer sequence with their POU domain, which consists of two sub-domains that make independent interactions with half-sites on opposite faces of the DNA, with a linker between these sub-domains making contact along the minor groove (Klemm *et al*., 1994). They are able to interact with other factors to carry out their *trans*-regulatory functions, including Sox1-3 in neural progenitor cells (Kondoh & Kamachi, 2010), with co-operative DNA binding between these factors being dependent upon tightly linked TFBSs (Ambrosetti *et al*. 1997). The strong enrichment for this motif in CNEs confirms the findings of Bailey *et al.* (2006).

Motif 4 is a close match to the 'E-box' motif (CANNTG), which represents the consensus binding-site for basic helix-loop-helix (bHLH) factors, which bind as homo- and hetero-dimers to DNA. Members of this family of transcription factors are involved in a variety of developmental pathways such as neural patterning (Olig2) (Guillemot, 2007), neurogenesis (Neurogenin) and myogenesis (MyoD) (Rudnicki et al., 1993).

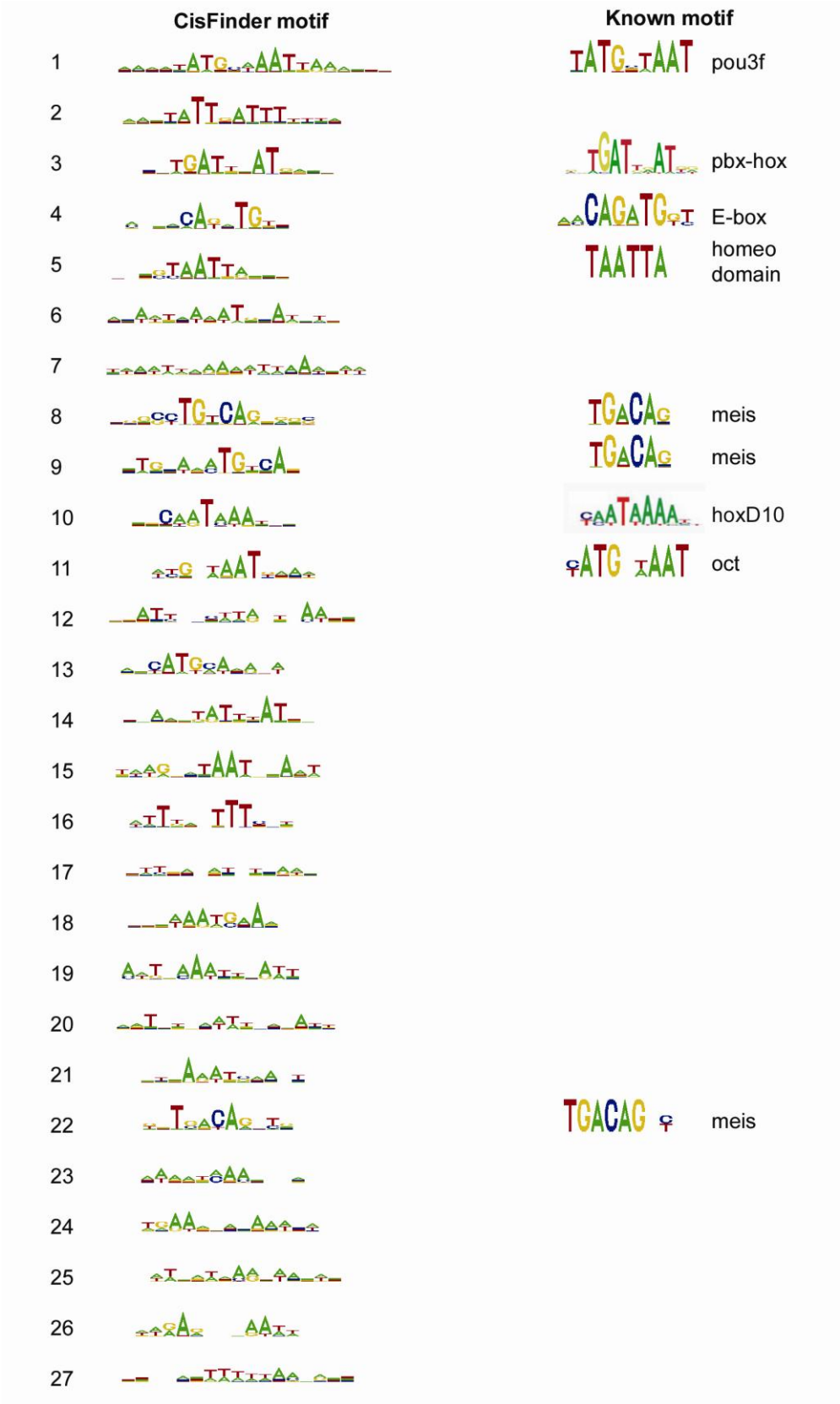


Figure 7.2. CisFinder identifies enriched motifs within human CNEs. Sequence logos representing the 27 motifs enriched within 6993 human CNE sequences as identified by CisFinder are shown, ranked according to enrichment score. Sequence logos of matching TFBS motifs that have been previously characterised are shown for comparison.

Motif 5 resembles a typical homeo-domain binding-site motif that is the potential binding-site of a very large number of proteins including members of the Hox, Phox, Six, Pax, Nkx and Dlx families (Berger *et al*., 2008).

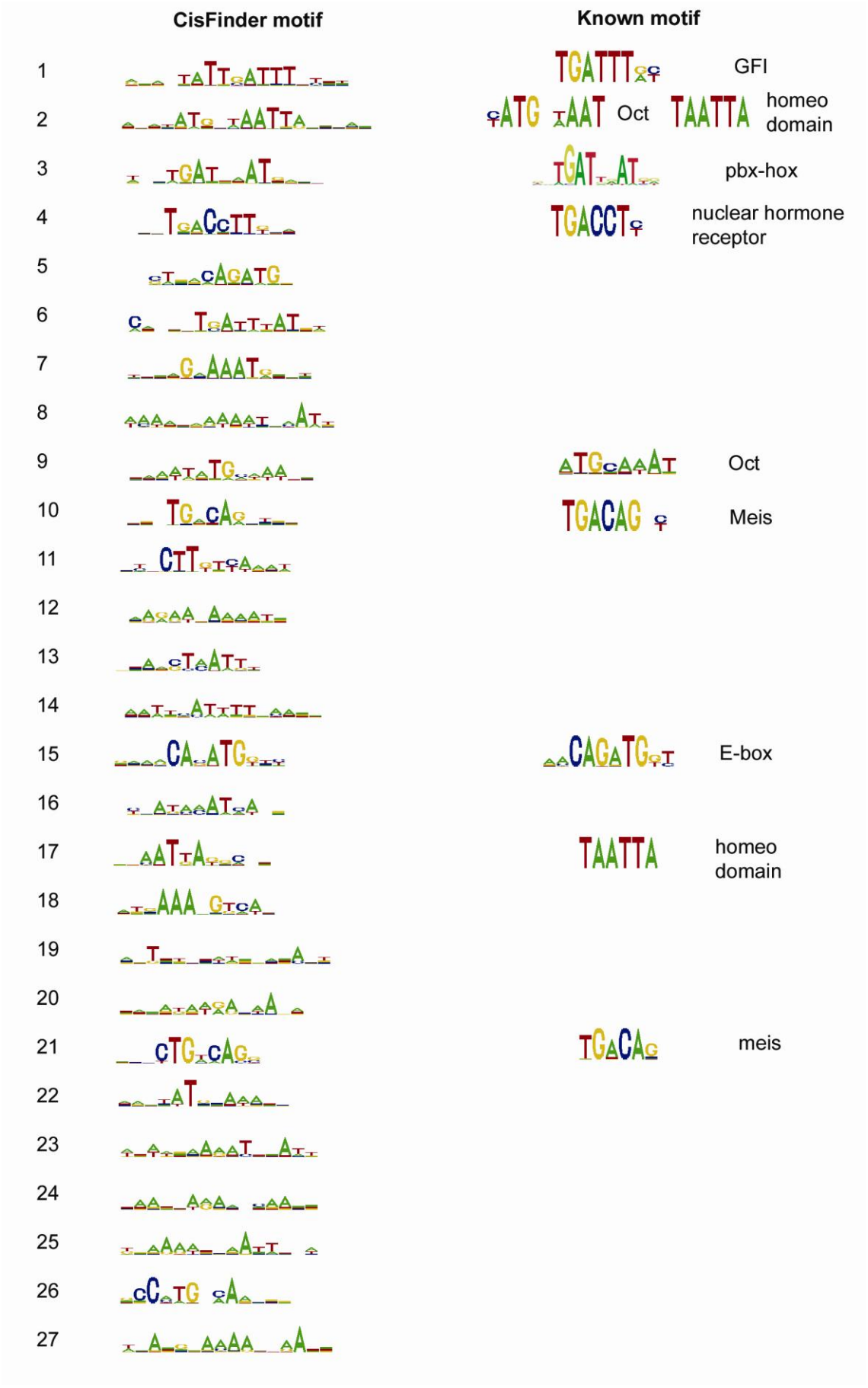
Motifs 8, 9 and 22 closely match those characterised for the Meis/Prep/Tgif family of homeo-domain-containing factors. Meis and Prep proteins are broadly expressed during development, acting as Hox co-factors as well as carrying out Hox-independent roles (Moens & Selleri, 2006; Mercader *et al*., 2000). The enrichment for this motif in CNEs corroborates the enrichment found from the targeted search for canonical Meis motifs (TGACAR) in the previous chapter.

Motif 10 resembles the TFBS motif of a number of posterior Hox factors including HoxD10 (Berger *et al*., 2008), which have roles in patterning the skeleton, the peripheral nervous system and in limb development (de la Cruz *et al*., 1999; Wahba *et al*., 2001).

Interestingly, one of the enriched motifs most frequently occurring in CNEs – motif 2 - does not appear to closely match any characterised motif. It would be of great interest to discover the factors that recognise it, as its high enrichment in CNEs suggests that these factors play important and evolutionarily conserved roles in the regulation of many genes.

I next sought to investigate whether other sets of vertebrate CNEs also contain similar enrichment for these motifs. I used CisFinder to search the shark CNE set (introduced in chapter 5), comprising 4782 elements conserved between human and the elephant shark, for enriched motifs using the same search parameters as used for the Condor human CNE set. This search resulted in the identification of 42 enriched motifs in the shark CNE set (Figure 7.3). Despite the shark set overlapping with only 34% of the Condor set, CisFinder identified many of the same motifs to be enriched in both sets. Indeed, using CisFinder's motif comparison tool, 18 of the motifs from the Condor set had matches to motifs in the shark set (above a match threshold of 0.7). Furthermore, the matching motifs, when ranked according to enrichment, occur in a very similar order – thus, the most highly enriched motifs in the Condor set are also amongst the most highly enriched in the shark set. Other than the different number of enriched motifs

identified, the most notable difference between sets is that a retinoic-acid response element (RARE)/nuclear hormone receptor motif – AAGGTCA – ranks prominently in the shark motif set but is not present in the Condor motif set. The strong agreement between enriched motifs within the different vertebrate CNE sets is likely to reflect shared regulatory properties of the elements they contain, and confirms that these patterns of motif enrichment are significant within a broader context than just the Condor CNE set.


CisFinder motif
Known motif
1
GFI
2
Oct
homeo domain
3
pbx-hox
4
nuclear hormone receptor
5
6
7
8
9
Oct
10
Meis
11
12
13
14
15
E-box
16
17
homeo domain
18
19
20
21
meis
22
23
24
25
26
27

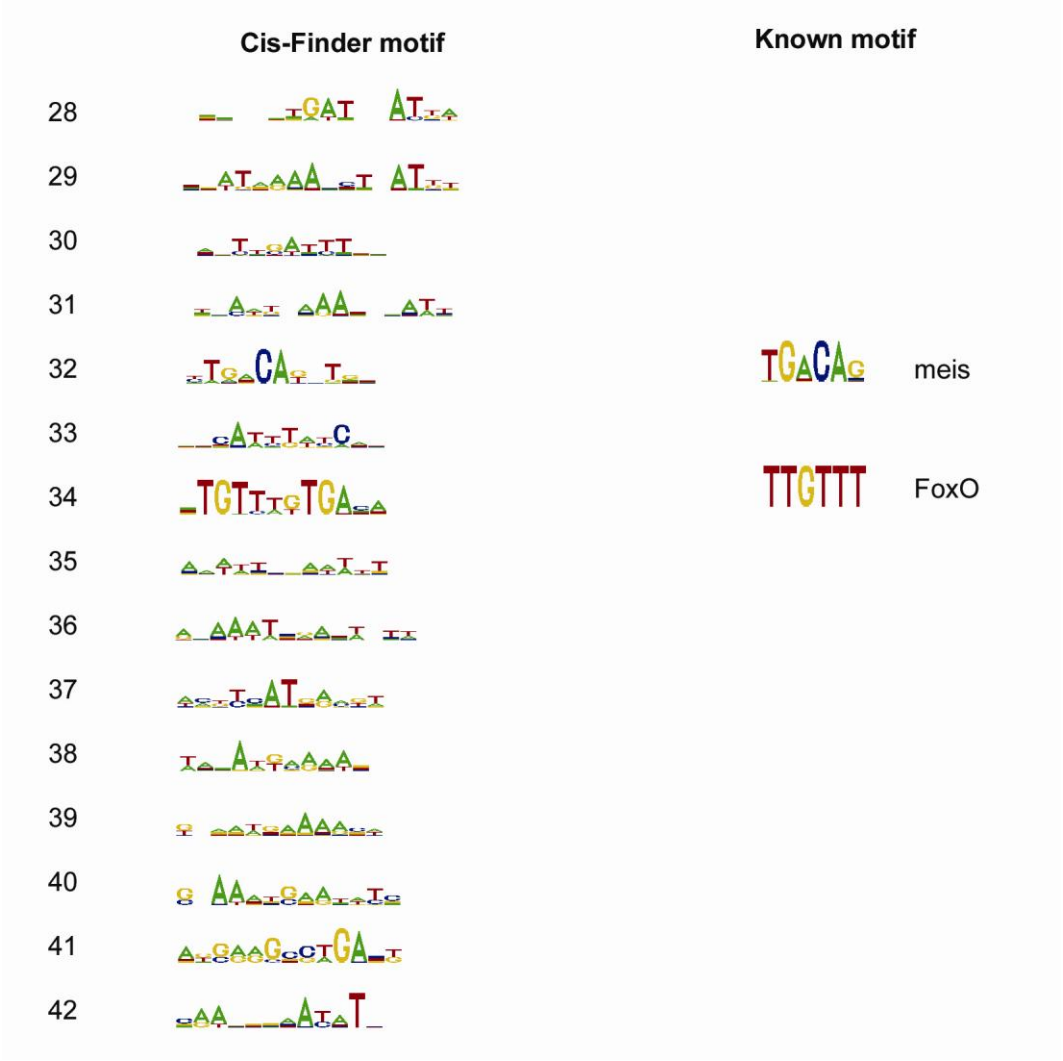


Figure 7.3. Enriched motifs within human-shark CNEs as identified by CisFinder. Sequence logos representing the 42 motifs enriched within 4782 human-shark CNEs are ranked according to enrichment score. Sequence logos of corresponding TFBS motifs that have been previously characterised are shown for comparison.

Pbx-Hox and Oct motifs associate with different gene regions

(This data was obtained by P. Piccinelli)

In chapter 5 I suggested that Pbx-Hox motifs show a non-random distribution across CNEs of different gene regions, being preferentially associated with genes involved in hox-dependent developmental processes. I wished to test whether the same could be true for other enriched motifs in CNEs. To this end, I have tabulated the number of octamer (Oct) motif hits occurring within CNEs of different gene regions of the Condor human CNE set and compared the distribution to that of Pbx-Hox motif hits (Table 7.1).

GENE	size /kb	Pbx-Hox hits	Oct hits	Pbx-Hox/Oct		Gene	size /kb	Pbx-Hox hits	Oct hits	Pbx-Hox/Oct
ESRRB	4.54	7	0	-		BCL11B	3.26	1	9	0.111
EVI1	4.02	6	0	-		PITX2	5.81	1	4	0.25
POU3F2	3.30	6	0	-		POU3F3	4.02	1	4	0.25
POU4F2	5.81	3	0	-		SHH	5.02	1	3	0.333
PAX5	1.05	2	0	-		LHX1	4.44	2	5	0.4
ARX	2.58	1	0	-		EN1	2.24	2	4	0.5
SOX3	2.05	1	0	-		LMO4	6.74	2	4	0.5
TSHZ1	10.4	16	2	8		LMO1	4.73	1	2	0.5
GLI3	3.43	8	1	8		SOX1	1.80	1	2	0.5
ZNF703	2.99	8	1	8		PAX2	9.19	5	9	0.555
SALL3	11.4	12	2	6		EBF1	4.03	4	7	0.571
MEIS1	9.29	9	2	4.5		ZFHX4	12.7	4	7	0.571
ZNF503	27.8	36	9	4		FOXP2	17.8	10	16	0.625
OTP	8.99	8	2	4		TFAP2A	11.1	5	8	0.625
POU6F2	3.10	4	1	4		EBF3	26.2	10	15	0.666
PRDM16	3.73	4	1	4		FIGN	8.81	4	6	0.666
TCF7L2	11.0	7	2	3.5		BHLHB5	6.94	2	3	0.666
TSHZ3	23.3	29	9	3.22		CST	8.39	2	3	0.666
ZIC1	8.33	6	2	3		DLX1	6.28	2	3	0.666
SOX14	4.29	6	2	3		BNC2	10.6	4	6	0.666
HOXD9	17.8	16	6	2.67		AUTS2	7.04	2	3	0.666
SHOX2	7.61	8	3	2.67		EYA1	5.54	2	3	0.666
NKX6-1	6.85	10	4	2.5		BARHL2	13.0	5	7	0.714
SOX6	10.5	5	2	2.5		MAB21L1	4.84	3	4	0.75
FOXP1	15.9	12	5	2.4		POU3F1	4.13	3	4	0.75

Table 7.1. Comparison of the distribution of 561 Pbx-Hox and 389 Oct motif hits across gene regions in the human CNE set. The results derive from strict searches for Pbx-Hox KR motif hits and Oct (ATGCWAAT) hits in the Condor human CNE set. The hits are organised according to the gene regions in which the CNEs containing them are situated. For the CNEs of each gene region, the ratio of Pbx-Hox to Oct hits is shown. The table on the left hand side describes the 25 gene regions with the highest ratio of Pbx-Hox to oct motifs, whilst the table on the right lists the 25 regions with the lowest ratio.

The distribution of Oct motifs across gene regions is interesting when compared to that of Pbx-Hox motifs. Whilst they overlap for many gene regions, some genes have many more Pbx-Hox motifs in their CNEs, such as *esrrb*, *tshz1* and *znf703*, and others many more Oct motifs, such as *bcl11b*, *pitx2*, *ebf1*. These contrasting distributions of Pbx-Hox and Oct motifs may correlate with roles of Hox and Oct factors in different developmental processes, such as in different pathways of patterning and differentiation.

Motifs enriched in *ciona* CNEs show no overlap with those of vertebrate CNEs

Despite the urochordates being the closest invertebrate relatives of vertebrates (Delsuc *et al*, 2006), their genomes contain no detectable traces of the CNEs that are found in vertebrates. However, in keeping with the existence of clade-specific CNEs in worms and flies (Vavouri *et al*., 2007), 2,336 urochordate-specific CNEs have been identified by genome comparison of *C. intestinalis* and *C. savignyi* (Vavouri, unpublished data). The sequence divergence between these two urochordate species is between that of human-chicken and human-frog (Johnson *et al*., 2004). The *ciona* CNEs are on average 182bp in length and share an average of 82% identity between *C. intestinalis* and *C. savignyi*, which is comparable to the first set of human-Fugu CNEs identified by Woolfe *et al*. (2005) (mean length 200bp, average 84% identity). The *ciona* CNE-associated gene set was found to be enriched for transcription factors and signalling genes, as seen for vertebrate CNEs, with a significant proportion of these genes being orthologous to human CNE-associated genes. This hints that *ciona* and vertebrate CNEs may function in parallel, lineage-specific GRNs involving some of the same factors.

I sought to test whether the same sequence motifs that we have found to be enriched in vertebrate CNEs are also strongly enriched in *ciona* CNEs. A search for enriched motifs within the *ciona* CNE set using CisFinder, with the same parameters as were used for the human CNE set, identified 19 motif clusters (Figure 7.4). CisFinder's motif comparison tool is unable to detect any overt similarity between the *ciona* motifs and those of the Condor human CNE set (using a lenient match threshold of 0.7). Nevertheless, some of the *Ciona* motifs do match closely to known TFBS motifs, with particularly strong matches to Atf1/CREB and Myc motifs. Interestingly, the proteins characterised as binding to CREB and Myc motifs are components of signalling pathways. This high enrichment for putative signal-transduction motifs in *ciona* CNEs is in contrast to the enriched motifs in vertebrate CNEs. Thus, whilst enriched motifs are identifiable within ciona CNEs, some of which correspond closely to previously characterised binding preferences of vertebrate transcription factors, these motifs provide no evidence that ciona and vertebrate CNEs are regulated by the same factors.

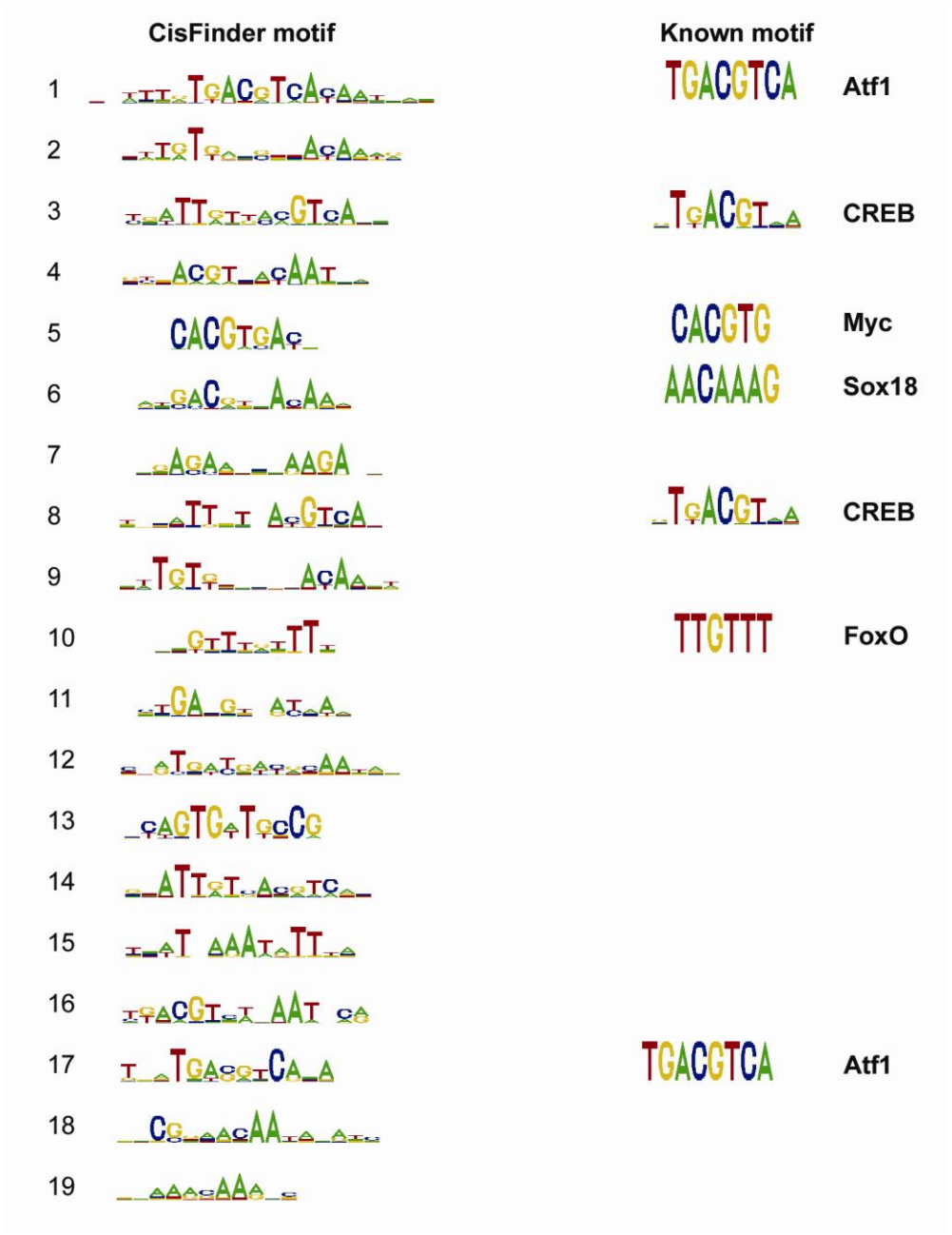


Figure 7.4. Motifs enriched within *ciona* CNEs, as identified by CisFinder. Motifs are represented as PWMs. Motifs are ranked according to enrichment and z-score. Matches to characterised PWMs from the literature are shown.

Discussion

de-novo tools can identify enriched motifs in CNEs

I set out to address whether CNEs are enriched for any other motifs besides Pbx-Hox and Meis motifs. I approached this by utilising a top-down, *de-novo* motif discovery approach. Interestingly, these approaches have not yielded any strong enrichment for motifs within CNEs up until now. I reasoned that the enrichment for Pbx-Hox motifs in CNEs suggested that *de-novo* motif discovery using the human CNE set should be possible, and that the ability of a *de-novo* motif discovery tool to identify these motifs could be used as a criterion by which to assess its suitability to the task. CisFinder was selected as a suitable candidate *de-novo* motif discovery tool for three reasons, it is fast (so can be used on large sets in an effective way), it can identify relatively weak enrichment, and it is based on a similar approach used by Weeder, which has been shown to frequently outperform other algorithms in a previous evaluation (Tompa *et al*., 2005).

CisFinder identified 27 enriched motifs within the Condor CNE set, many of which match to PWMs for factors that have been previously characterised, either *in-vivo* or *in-vitro*. Indeed, there is good agreement between *in-vitro* characterised motifs (such as from protein-binding micro-array data) and enriched motifs within putative regulatory elements (CNEs). This is significant for two reasons: firstly, it indicates that *in-vitro*-characterised binding profiles can faithfully represent many *in-vivo* interactions; secondly, it strongly suggests that CNEs are regulated by a range of known - and, in some cases, well characterised - TFs.

The existence of a highly enriched, novel motif (motif 2) suggests either that an important family of TFs have not had their binding characterised yet, or that the characterised binding profiles of the factors involved do not represent their preferences *in-vitro* – possibly due to co-operative binding. The Oct and Pbx-Hox motifs have been a source of much interest and have generated many principles regarding TF binding. Motif 2 appears to have a similar prevalence within CNEs, so the discovery of the factor/s binding to it is an important problem.

It is noteworthy that of the 6 characterised PWMs that can be matched to enriched motifs in CNEs, 5 of them are for homeodomain factors. The interpretation of the biological significance of this finding is clouded by a number of factors that may introduce ascertainment bias into the motif discovery and characterisation process. For example, it is possible that an ascertainment bias arises from the limited breadth of characterised PWMs available for comparison with the enriched CNE motifs. However there are numerous PWMs within the Jaspar database that describe the binding preferences of non-homeodomain factors, such as those from the forkhead (Fox), Sox, bHLH, nuclear hormone receptor and zinc finger families. Of these, the consensus bHLH motif matches an enriched motif in the Condor CNE set, whilst a nuclear hormone receptor motif matches to an enriched motif in the shark CNE set, yet the PWMs of members of the other TF families do not match to any enriched CNE motifs. This could be attributable to the *de-novo* motif discovery approach having certain limitations; for instance, its exhaustive enumeration of all possible words within the CNE set is based on 8mers, so some shorter motifs are likely to be missed. Further, it is probable that many TFs do not have such consistent binding preferences as those recognising the enriched motifs found here, or their binding profiles may not be so amenable to representation as PWMs, which assume positional independence across the motif. Additionally, if factors show highly promiscuous DNA binding preferences depending on the context of their binding (e.g. through protein-protein interactions), they will bind to a number of different motifs that will not then be identifiable by significant enrichment within CNEs. These limitations mean that caution must be taken in inferring the relative prominence of a factors role in CNE-regulated developmental processes from the frequency of occurrence of its binding-motif in CNEs. Thus, whilst the identification of these motifs is a major step toward de-coding CNEs, this set of motifs is likely to represent only a sub-set of TFBSs within CNEs.

Enriched motifs in CNEs point towards combinatorial transcriptional regulation

This *de-novo* motif discovery approach has found CNEs to be enriched in TFBSs corresponding to a variety of TFs. Many of these, such as Oct, Pbx and Meis, show broad expression patterns during development and have been shown to interact with many different TF partners (Ravasi *et al.*, 2010). These characteristics have led to a model of TF interaction networks, in which such broadly expressed factors act as ‘facilitators’ of transcription, interacting with other tissue/domain-specific ‘specifiers’

to create tightly defined gene expression patterns. Under this model, specific expression domains are defined by the domain-specific interactions between these factors. The enrichment for the TFBS motifs of these facilitators in CNEs may therefore represent their broad roles in *cis*-regulation, collaborating with many different specifiers to regulate expression in different domains. In some cases, such as with many of the homeodomain TFBSs, the homeodomain-containing factor may play the role of a specifier. Thus, the enrichment for the TAATTA homeo-domain TFBS motif may also reflect the large number of factors that can bind to it, rather than being due to these factors being facilitators. If it is the case that homeodomain proteins acting as specifiers bind onto this motif then this raises the question of how the enhancer attracts only the correct homeodomain protein. Combinatorial interactions between proteins that assemble upon the CNE are likely to play an important role in this specificity. Addressing the nature of these interactions is an important task for the future, which goes beyond investigating just the protein-protein interactions, their binding to DNA adding a further degree of complexity.

The Pbx-Hox hetero-dimer is an interesting case to classify according to the facilitator/specifier model. Pbx factors have broad expression patterns and many interactions with other factors, so they can be considered as facilitators. The interaction of Pbx and Hox factors creates a heterodimer with potential to regulate transcription within only a limited A-P region. Thus the Pbx-Hox interaction could be viewed as a facilitator-specifier interaction, with a characteristic binding motif shared amongst many different Pbx-Hox heterodimers. From the expression patterns of Pbx-Hox +ve CNEs, it is apparent that further specification is being carried out, presumably by other factors binding to neighbouring TFBSs. Thus, CNEs could be viewed to define tightly controlled expression patterns by using facilitators in conjunction with multiple specifiers.

The identification of these motifs presents a number of opportunities for further characterising their distribution and conservation across CNEs. Interesting questions are whether any of these motifs show patterns of co-occurrence with each other, or significant associations with particular gene regions, as may be the case for the Pbx-Hox and Oct motifs. By grouping CNEs with common motifs, it may be possible to enrich these new sets for other motifs that were not picked up by CisFinder over the whole CNE set. As the efficacy of CisFinder decreases for smaller sequence sets (Sharov &

Ko, 2009), it may be necessary to utilise other tools for this task. Finally, it would be interesting to identify what proportion of CNE sequence can be 'explained' by these motifs, and conversely, what proportion is unexplained. These questions highlight how little is known about the regulatory language contained within CNEs; however, the identification of common words constituting this language is a critical step toward de-coding these crucial *cis*-regulatory elements.

ciona CNE motifs do not overlap with those of vertebrate CNEs

Despite the lack of conservation of vertebrate CNEs in invertebrates, invertebrate phyla have their own, lineage-specific CNEs (Bejerano *et al*., 2005; Vavouri *et al*., 2007). As many of these elements are associated with the same genes that have CNEs in vertebrates, they have been hypothesised to function in GRNs that evolved in parallel between the different phyla during the diversification of metazoans, which contribute to the different morphological characteristics of each phylum (Vavouri & Lehner, 2009). Comparison of urochordate genomes has uncovered a set of '*ciona* CNEs', which are associated with many of the same genes as vertebrate CNEs, including those orthologous to Pax, Tshz, Sall, and Zfhx factors (Vavouri, unpublished). If vertebrate and *ciona* CNEs, and their associated transcription-factors, belong to equivalent but distinct GRNs that produce alternative body plans, they may be expected to share a common *cis*-regulatory language.

Despite the apparent similarities between vertebrate and *ciona* CNEs – the size and conservation of the elements, and the overlapping sets of genes with which they are associated – there is no overt similarity between the enriched motifs that can be identified within them. As CisFinder was used with the same settings to uncover enriched motifs in both the vertebrate and *ciona* CNE sets, the sets of motifs identified are directly comparable and differences between them should be attributable to biological differences between the urochordates and vertebrates. Potential differences between urochordate and vertebrate transcription factors that could explain these different motifs include clade-specific TF repertoires and DNA binding-specificities.

It is unlikely that clade-specific transcriptional repertoires account for the non-equivalence of enriched motifs in urochordate and vertebrate CNEs as the factors that bind to the Pbx-Hox, Oct and HD motifs - highly enriched in vertebrate CNEs – are also

utilised by enhancers in invertebrates (e.g. Ryoo *et al*., 1999; Kitamoto & Salvaterra, 1995; Mann *et al*., 2009). Indeed, a survey of developmentally relevant genes in *Ciona intestinalis* identified homeobox genes related to a variety of classes of vertebrate factors that would recognise these motifs (including Hox, TALE, pou and NK) (Wada *et al*., 2003). It is possible that the binding preferences of these factors have diverged in urochordates, but the conservation of their characteristic DNA-binding domains suggests otherwise (Wada *et al*., 2003).

The different motifs in these two CNE sets may also reflect differences between the developmental programs of urochordates and vertebrates. Importantly, vertebrate CNEs represent only a sub-set of *cis*-regulatory elements, as not all elements are highly conserved. This sub-set may be involved in particular aspects of vertebrate development – those aspects that are highly conserved between vertebrates, and that rely on complex, highly structured, regulatory elements. Different aspects of vertebrate development might involve the use of different sets/classes of transcription factors, such that enhancers involved in the development of different tissues may be enriched for different motifs. Thus, it is possible that *ciona* CNEs represent a sub-set of regulatory elements that are involved in aspects of urochordate development that are highly conserved within that clade, but are only partially equivalent to those aspects of vertebrate development that are represented by vertebrate CNEs.

Two of the most enriched motifs in *ciona* CNEs correspond to characterised TFBSs for factors involved in signal transduction in vertebrates, supporting the notion that the *ciona* CNEs may be regulated by different types of factors to those regulating vertebrate CNEs. Indeed, despite the sets of genes containing vertebrate and *ciona* CNEs showing some overlap, they are not entirely equivalent, with *ciona* CNEs being associated with a relatively high proportion of signalling genes compared to vertebrate CNEs. Nevertheless, genes encoding proteins with POU and HD domains are enriched within the *ciona* CNE-associated gene set, yet the oct and HD motifs are not identified as being enriched in *ciona* CNEs, in contrast to their strong enrichment in vertebrate CNEs. It is possible that some *ciona* CNEs do indeed contain the same motifs as are enriched in vertebrate CNEs and contribute to equivalent developmental processes as some vertebrate CNEs but that these are diluted by other *ciona* CNEs with different motifs that contribute to other aspects of development.

A direct comparison between the contribution of particular factors to vertebrate and urochordate embryonic development can be made for the Hox family. As described in chapter 5, Hox factors play crucial and conserved roles in AP patterning during vertebrate development and their binding motifs (Pbx-Hox) are enriched in vertebrate CNEs. In contrast, it has recently been shown that Hox genes have limited functions during the larval development of *ciona intestinalis*, based upon evidence from knock-down experiments (Ikuta *et al*. 2010). This finding may explain the lack of enrichment for the Pbx-Hox motif in *ciona* CNEs and supports the notion that *ciona* and vertebrate CNEs are involved in regulatory networks that are only partially equivalent. Thus the different enriched motifs in the two CNE sets could be partially due to differential enrichment of motifs in each set, due to inter-clade differences in the contributions of particular factors to development. With this in mind, it would be interesting to investigate whether the enriched motifs of the vertebrate CNE set can be identified as enriched in *ciona* CNEs through targeted motif searches, which may be more sensitive than *de-novo* strategies. Further it would be of interest to investigate the types of motifs that are enriched within CNEs of other invertebrate phyla, and within different vertebrate CNE sets (e.g. those defined by human-mouse or human-frog genome comparisons).

Conclusion

I set out to identify enriched motifs within CNEs using a *de-novo* motif discovery approach. I predicted that such an approach should find at least some significantly enriched motifs, as we have already characterised the enrichment of the Pbx-Hox motif in the human CNE set. CisFinder is a motif discovery tool that appears well suited to the task, identifying 27 enriched motifs. A substantial proportion of these motifs match closely to the binding preferences of well characterised transcription factors, suggesting that many CNEs are regulated by known factors. This finding also implies a generally good agreement between *in-vitro* characterised binding profiles and *in-vivo* TFBSs. However, there are some motifs that, despite occurring frequently in CNEs, do not match the characterised binding profiles of any factors. These motifs may indicate hitherto uncharacterised transcription factors or structural elements, or may point toward novel combinatorial interactions between known factors. Identifying what binds to these motifs would be of great interest. Whilst there are a number of possible reasons for the enrichment for these specific motifs in CNEs, many of the TFs that are predicted

to recognise these motifs share common roles as transcriptional ‘facilitators’, hinting at a model of combinatorial protein interaction through which CNEs carry out their regulatory functions. An intriguing finding is the lack of overt similarity between motifs enriched in vertebrate and *ciona* CNEs. Whilst the biological significance of this is hard to interpret, this finding hints at differences between the aspects of development that are highly conserved within the vertebrate and urochordate lineages. The identification of these enriched motifs is a crucial step forward in de-coding CNEs and helps to further place these elements within the context of developmental gene regulatory networks. Key issues to be resolved are whether these motifs show any significant patterns of co-occurrence, or associations with particular genes, whether further motifs can be identified through other, complementary motif discovery approaches, and the identity of the factors that recognise the novel motifs.

8 Discussion

This thesis documents the use of the sea lamprey as a model organism for investigating the role of deeply conserved non-coding genomic elements in vertebrate development and evolution. Through comparative genomics, I have described the pattern of sequence conservation of these elements across the vertebrate phylogeny. I have addressed the biological meaning of this sequence conservation pattern using developmental biology approaches in zebrafish and lamprey embryos. The major findings presented herein are fourfold: firstly, a significant number of ancient CNEs are shared across all vertebrates; secondly, a large proportion of vertebrate CNEs contain Pbx-Hox TFBS motifs, which correlate with regulatory functions in hindbrain and head patterning; thirdly, distantly related orthologous CNEs perform similar, but evolvable, roles in the development of their respective species; fourthly, CNEs contain components of a regulatory language that is also found in less constrained regulatory elements. In this chapter, these findings will be discussed in the context of the broader aspects of evo-devo and gene regulation.

CNEs and gene regulation

Broad-scale surveys of the regulatory activity of CNEs have now provided overwhelming evidence that the majority of them can function as *cis*-regulatory elements during vertebrate development (Pennacchio *et al*., 2006; Woolfe *et al*., 2007; Li *et al*., 2010). These databases of reporter expression represent incredibly useful resources for systematically investigating the relationship between CNE sequences and their regulatory functions. The development of new reporter assay methodologies, particularly the tol2 system in zebrafish embryos (Fisher *et al*., 2006), has facilitated faster and more detailed characterisation of the reporter expression patterns that CNEs can drive. This was evident in chapter 5, where the use of the less mosaic tol2 system significantly aided the characterisation of rhombomere-specific reporter expression patterns, and the dissection of the TFBSs responsible for these patterns.

The ability for CNEs to act as enhancers in reporter assays does not necessarily indicate that this is their only biological function. The hypothesis that CNEs may carry out multiple overlapping functions was posited in order to account for their unusually high sequence constraint. Recently, a large number (~3,000) of neuronal enhancer sequences have been shown to give rise to short ‘enhancer RNAs’ (eRNAs) in mouse cell cultures,

whilst also up-regulating expression of their target gene (Kim *et al.*, 2010). Roughly half of these elements were found to be under evolutionary constraint (although not to the level of human-Fugu CNEs). Although the functional significance of these eRNAs was not established, they may play a role in regulating transcription. This hints that some CNEs may also produce RNA molecules that have yet to be characterised. It is possible that such RNAs may be transcribed locally, making them difficult to detect without more sensitive methods. However, we identified lamprey CNEs showing patterns of sequence conservation that are clearly consistent with TFBSs, where the majority of the conserved regions can be explained by Pbx-Hox and Meis motifs. Thus, it is possible that an alternative explanation to constraint upon TFBSs may not need to be evoked in order to explain the patterns of deep sequence conservation of these elements.

The identification of a high enrichment for Pbx-Hox motifs in vertebrate CNEs represented a major breakthrough in this thesis. Prior to the identification of these elements, it was difficult to place CNEs into GRNs as, without any TFBS motif information, there was no means by which to predict the factors regulating CNEs or what the functional output (reporter expression) of a CNE could be. The identification of many other enriched motifs within CNEs confirmed that these elements contain TFBS motifs that correspond to those of well characterised TFs, further placing CNEs within a developmental GRN context. The crucial task now is to find a way to utilise these motifs, and our knowledge of the factors that are predicted to recognise them, to de-code the *cis*-regulatory functions of these elements. This will require the identification/confirmation of factors that bind to CNEs, as well as characterisation of their interactions.

A recent study used a systematic approach to identify many interactions between developmental TFs in cell culture, characterising ~800 interactions from a set of ~1200 human TFs (Ravasi *et al.*, 2010). This led to insights regarding the TF interaction networks that occur in different tissue types. A major theme arising from their networks was that of specificity through interaction, which is in agreement with the patterns of motifs that we find in CNEs – individual motifs within CNEs may have only modest predictive power regarding reporter expression patterns, whilst specific combinations of motifs are likely to enable more accurate predictions. The TF-TF interactions characterised by that study are just a first step – further progress must be made in

discovering TF-DNA-TF interactions, which are likely to be a critical aspect of the mechanism of complex *cis*-regulatory elements. CNEs could be used as a starting point to identify TF combinations and their TFBSs, through *in-silico* and proteomics methods.

CNEs were first searched for as a means to identify functional non-coding elements, particularly *cis*-regulatory elements. Recent advances in sequencing technology mean that chip-seq can now be used to identify *cis*-regulatory elements, with this approach being complemented in some instances by the use of sequence conservation to further refine predictions of functional elements. Two recent studies using chip-seq to identify enhancers active during development of the heart in mouse (McCulley *et al*., 2010) and in human and mouse ES cells (Kunarso *et al*., 2010) have shed more light on the degrees of sequence constraint that developmental enhancers are under. In the first case, predicted mouse forebrain enhancers were found to be on average three times more deeply conserved (most showing conservation between humans and birds) than predicted heart enhancers. In the second study, the binding profiles of three TFs were obtained in human and mouse ES cells. It was found that CTCF binding was highly conserved between species, whilst the binding profiles of OCT4 and NANOG were different between human and mouse cells, suggesting turnover of TFBSs and re-wiring of regulatory circuits. These studies are a reminder that CNEs represent only a subset of *cis*-regulatory elements, and this subset is likely to be biased towards elements that are involved in the development of certain tissue types and that are regulated by particular combinations of TFs. Chip-seq will be an incredibly useful tool for systematically predicting *cis*-regulatory elements, and can be used to identify elements in a general or a TF-specific manner. Nevertheless, CNEs provide a complementary source of information representing many elements that may not be immediately identifiable by chip-seq due to the unavailability of particular antibodies. Furthermore, patterns of motifs that may be identifiable through *in-silico* investigation of CNE sequences could shape investigations into TF-binding using chip-seq.

Recently, a targeted motif-based strategy has been used to predict tissue-specific vertebrate enhancers (Narlikar *et al*., 2010). This strategy leveraged the functional data from the EB database to create a set of heart enhancers, which was used to train an algorithm to identify similar motif patterns in constrained genomic sequence (conserved between mammals). The success of this approach in identifying heart enhancers highlights the utility of CNE functional datasets (such as EB and Condor) and implies

that CNEs can be useful subjects for discerning regulatory codes predictive of tissue-specific enhancer function in less deeply conserved sequences. This same approach could be applied to the Pbx-Hox +ve CNEs that show hindbrain enhancer activity, in order to identify further Hox-responsive vertebrate hindbrain enhancers. Another recent study made use of homotypic clusters of TFBSs to predict vertebrate tissue-specific enhancer elements (Gotea *et al*., 2010), further showing that the knowledge gleaned from characterising CNEs can be used to predict other *cis*-regulatory elements in the genome.

In each of these cases, enhancer verification required the use of a reporter assay, which, despite the relatively high throughput tol2 assay in zebrafish, is still a limiting factor for systematic characterisation of enhancers. This is particularly true for investigating the language of enhancers, which could require extensive fine-scale perturbation analyses. Automated screening of zebrafish embryos promises to make the process faster (Gehrig *et al*., 2010). It is also likely that alternative high-throughput reporter assay methodologies involving DNA barcoding can be developed for vertebrate models, based on an approach introduced in sea urchin embryos (Nam *et al*., 2010).

The over-riding message from this section is simple: whilst CNEs do not represent the be-all and end-all of vertebrate *cis*-regulation, their mechanisms are still likely to apply to a much larger group of *cis*-regulatory elements, making them an incredibly useful stepping stone toward the de-coding of vertebrate gene regulation. In order to use CNEs for this purpose, approaches combining multiple sources of information – protein-protein interactions, protein-DNA binding, *cis*-regulatory activity – must be utilised.

CNEs and evo-devo

A point frequently emphasised in this thesis is that in order to infer the roles of CNEs in evolution we must first know their roles in development. The approach to address the significance of the lack of many gnathostome CNEs in the lamprey genome in chapter 4 was limited by a lack of knowledge of how the missing CNEs worked and what their developmental significance was. This made it hard to interpret their absence in lamprey in terms of developmental GRNs. The identification of conserved Pbx-Hox motifs in many CNEs, coupled with the correlation with hindbrain and pharyngeal arch enhancer function, makes it possible to predict (roughly) the function of a large cohort of CNEs

and to place them within a GRN for hindbrain/head patterning. With these CNEs placed into a developmental context it is possible to address the functional significance of their patterns of conservation in chordates by referring to other sources of data (e.g. gene expression patterns) regarding the GRN in which they are predicted to act, leading to the formulation of a testable model of their role in vertebrate evolution. Importantly, with increased knowledge of the motifs underlying CNE functions, it will be possible to search in the lamprey and amphioxus genomes for homologous elements that are not highly conserved, thereby further addressing how these elements, and the GRNs in which they act, evolved.

The finding in chapter 7 that vertebrate and *ciona* CNEs do not share the same enriched motifs, as characterised by a *de-novo* motif search, is intriguing. The discovery that CNEs from different metazoan lineages were associated with overlapping sets of developmental regulatory genes originally led to the hypothesis that these CNEs are associated with equivalent GRNs, composed of similar genes, that have diverged in each lineage. If this is the case, CNEs from different lineages might be expected to be regulated by homologous sets of factors and to utilise equivalent sequence vocabulary. However, the different enriched motifs in vertebrate and *ciona* CNEs do not provide support to this notion. In chapter 7, I proposed two reasons for these differences. Firstly, the two CNE sets are associated with only partially overlapping sets of genes so the factors regulating the CNEs of each set could differ significantly between sets. Secondly, even for factors that do have CNEs associated with them in both clades, there may be significant differences between their contributions to ascidian and gnathostome development, leading to differential enrichment of motifs within the two CNE sets. The combination of both of these factors could explain the differences between the enriched motifs in each CNE set. Thus, the clade-specific GRNs in which these CNEs act may be partially equivalent - comprising interactions between some common and some different factors - and the involvement of each CNE set with some different genes may reflect the different developmental trajectories of each clade. CNEs represent only a sub-set of regulatory elements in each clade, and the motifs identified to be enriched within them are likely to represent only a sub-set of the TFBSs that they contain, so care must be taken in inferring the degree of equivalence of GRNs from different metazoan lineages based on the characteristics of their CNEs.

The lamprey reporter assay, developed in chapter 6, confirmed the functional conservation of CNEs between zebrafish and lamprey. As well as showing functional conservation, this approach also suggested that significant functional divergence is possible between orthologous CNEs, and that it can involve only a relatively modest change in enhancer sequence. With greater knowledge of the TFBSs in CNEs, these elements will provide a rich source of information regarding this type of lineage-specific evolutionary tinkering, which is likely to play a crucial role in generating variations of the vertebrate body plan (Carroll, 2008; Prabhakar *et al*., 2008).

An important question in evo-devo is the relative contributions to morphological evolution of *cis*-regulatory versus genetic changes. The current model invokes the negative pleiotropic effects of divergence in developmental proteins as prohibiting their evolution, with changes in developmental GRNs predicted to arise mainly through *cis*-regulatory divergence (Carroll, 2008). However, TFs are likely to function in a modular manner, with the evolution of new protein-protein interactions being possible without altering the other functions of the protein (Lynch & Wagner, 2008). In order to obtain empirical evidence regarding a role for the evolution of new protein-protein interactions in morphological evolution, the complexes of proteins that form on developmental enhancers need to be further characterised. As mentioned above, CNEs could be useful subjects for this characterisation.

A final evolutionary question regards how CNEs, and *cis*-regulatory elements in general, arise in the genome. Chapter 6 hinted at the functional co-option of an existing hindbrain enhancer to drive additional expression in the neural crest in zebrafish, in keeping with a characterised mechanism whereby additional TFBSs can evolve within previously existing *cis*-regulatory elements (e.g. Gompel *et al*., 2005). Other *cis*-regulatory elements have been identified to have evolved through the action of transposable elements (Bejerano *et al*., 2006), with the conserved nature of CNEs being crucial for this identification. Whilst insightful, this exaptation of regulatory elements does not answer how functional elements first arise from non-functional genomic sequence. A recent characterisation of functionally constrained genomic regions of vertebrates and invertebrates has found 200-300MB of the human genome to be under functional constraint, with constrained non-coding bases representing 5-8 times the number of constrained protein-coding bases. In contrast, only ~60MB of the *D melanogaster* genome was similarly constrained, with a ratio of non-coding to protein-

coding constrained bases of roughly 2. This confirms the notion that many new *cis*-regulatory elements must have emerged during metazoan diversification, rather than *cis*-regulatory evolution acting through divergence of pre-existing regulatory elements. Whilst potential mechanisms of *de-novo cis*-regulatory element evolution have been imagined (e.g. Cameron & Davidson, 2009), there is scant evidence for their veracity. The detailed characterisation of the mechanism of action of CNEs, identifying the crucial functional motifs within them, may prove enlightening in this regard.

Thus, CNEs are likely to prove useful subjects for investigating many aspects of evo-devo research. They can be used to identify certain ancient conserved GRN circuits and to trace how these circuits have evolved. Because CNEs, by definition, represent sets of orthologous (and sometimes paralagous) regulatory elements, they are a useful resource for investigating lineage-specific *cis*-regulatory changes, as well as the *cis*-regulatory changes that can occur after genome duplication. These investigations are facilitated by the development of reporter assay techniques in multiple model organisms. As subjects of detailed *cis*-regulatory characterisation to identify the mechanism of action of enhancer elements, CNEs may also represent key sources of information regarding the role of protein-protein interactions in evolution, and the mechanisms by which regulatory elements first arise in the genome.

References

Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA and Rubin EM. (2007). Deletion of ultraconserved elements yields viable mice. PLoS Biol. 5: 234.

Alexander T, Nolte C and Krumlauf R. (2009). Hox genes and segmentation of the hindbrain and axial skeleton. Ann Rev Cell Dev Biol. 25: 431-456.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389-3402.

Ambrosetti DC, Basilico C and Dailey L. (1997). Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. Mol Cell Biol. 17: 6321-9.

Andermann P and Weinberg ES. (2001). Expression of zTlxA, a Hox11-like gene, in early differentiating embryonic neurons and cranial sensory ganglia of the zebrafish embryo. Dev Dyn. 222: 595-610.

Arnone M.I. and Davidson E.H. (1997). The hardwiring of development: organisation and function of genomic regulatory systems. Development. 124:1851-64.

Bailey PJ, Klos JM, Andersson E, Karlén M, Källström M, Ponjavic J, Muhr J, Lenhard B, Sandelin A and Ericson J. (2006). A global genomic transcriptional code associated with CNS-expressed genes. Exp Cell Res. 312: 3108-19.

Bardet PL, Schubert M, Horard B, Holland LZ, Laudet V, Holland ND and Vanacker JM. (2005). Expression of estrogen-receptor related receptors in amphioxus and zebrafish: implications for the evolution of posterior brain segmentation at the invertebrate-to-vertebrate transition. Evol Dev. 7: 223-33.

Barik S. (2002). Megaprimer PCR. Methods Mol Biol. 192: 189-96.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS and Haussler D. (2004). Ultraconserved elements in the human genome. Science. 304: 1321–1325.

Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Peña-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET, Khalid F, Zhang W, Newburger D, Jaeger SA, Morris QD, Bulyk ML and Hughes TR. (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell. 133: 1266-76.

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM and Eisen MB. (2002). Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci USA. 99: 757-62.

Berthelsen J, Kilstrup-Nielsen C, Blasi F, Mavilio F and Zappavigna V. (1999). The subcellular localization of PBX1 and EXD proteins depends on nuclear import and export signals and is modulated by association with PREP1 and HTH. Genes Dev. 13: 946–953.

Bessa J, Tavares MJ, Santos J, Kikuta H, Laplante M, Becker TS, Gómez-Skarmeta JL and Casares F. (2008). meis1 regulates cyclin D1 and c-myc expression, and controls the

proliferation of the multipotent cells in the early developing zebrafish eye. Development 135: 799 -803.

Biemar F, Devos N, Martial JA, Driever W and Peers B. (2001). Cloning and expression of the TALE superclass homeobox Meis2 gene during zebrafish embryonic development. Mech Dev. 109: 427 -431.

Bilioni A, Craig G, Hill C and McNeill H. (2005). Iroquois transcription factors recognize a unique motif to mediate transcriptional repression in vivo. Proc Natl Acad Sci USA. 102: 14671-6.

Blair JE and Hedges SB. (2005). Molecular phylogeny and divergence times of deuterostome animals. Mol Biol Evol. 22: 2275-84.

Bosse A, Zülch A, Becker MB, Torres M, Gómez-Skarmeta JL, Modolell J and Gruss P. (1997). Identification of the vertebrate Iroquois homeobox gene family with overlapping expression during early development of the nervous system. Mech Dev. 69: 169-81.

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E; NISC Comparative Sequencing Program, Green ED, Sidow A and Batzoglou S. (2003). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. 13: 721-31.

Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B and Sandelin A. (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res. 36: D102-6.

Cameron RA and Davidson EH. (2009). Flexibility of transcription factor target site position in conserved *cis*-regulatory modules. Dev Biol. 336: 122-35.

Campbell NA, Reece JB and Mitchell LG. (1999). *Biology: Fifth Edition*. Addison Wesley Longman U.S.A.

Capdevila J, Tsukui T, Rodriguez-Esteban C, Zappavigna V and Izpisua-Belmonte JC. (1999). Control of vertebrate limb outgrowth by the proximal factor Meis2 and distal antagonism of BMPs by Gremlin. Molecular Cell 4: 839-849.

Carr JL, Shashikant CS, Bailey WJ and Ruddle FH. (1998). Molecular evolution of Hox gene regulation: cloning and transgenic analysis of the lamprey HoxQ8 gene. J Exp Zool. 280:73-85.

Carroll, S.B. (2000). Endless forms: The evolution of gene regulation and morphological diversity. Cell 101, 577–580.

Carroll SB. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell. 134: 25-36.

Cecconi F, Proetzel G, Alvarez-Bolado G, Jay D, Gruss P (1997). Expression of Meis2 , a knotted -related murine homeobox gene, indicates a role in the differentiation of the forebrain and the somitic mesoderm. Dev Dyn 210: 184-190.

Chan SK and Mann RS. 1996. A structural model for a homeotic protein-extradenticle-DNA complex accounts for the choice of HOX protein in the heterodimer. Proc Natl Acad Sci USA. 93: 5223-8.

Chan SK, Ryoo HD, Gould A, Krumlauf R and Mann RS. (1997). Switching the in vivo specificity of a minimal Hox-responsive element. Development. 124: 2007-14.

Chang CP, Brocchieri L, Shen WF, Largman C and Cleary ML. (1996). Pbx modulation of Hox homeodomain amino-terminal arms establishes different DNA-binding specificities across the Hox locus. Mol Cell Biol. 16: 1734-45.

Chang CP, Jacobs Y, Nakamura T, Jenkins NA, Copeland NG and Cleary ML. (1997). Meis proteins are major in vivo DNA binding partners for wild-type but not chimeric Pbx proteins. Mol Cell Biol. 17: 5679–5687.

Choe SK, Vlachakis N and Sagerström CG. (2002). Meis family proteins are required for hindbrain development in the zebrafish. Development 129: 585-595.

Chung AC and Cooney AJ. (2003). The varied roles of nuclear receptors during vertebrate embryonic development. Nucl Recept Signal. 1:e007.

Coré N, Caubit X, Metchat A, Boned A, Djabali M and Fasano L. Tshz1 is required for axial skeleton, soft palate and middle ear development in mice. Dev Biol. 308: 407-20.

Crozatier M, Valle D, Dubois L, Ibnsouda S and Vincent A. (1996). Collier, a novel regulator of Drosophila head development, is expressed in a single mitotic domain. Curr Biol. 6: 707-18.

Davidson EH. (2006). *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution.* Academic Press, Oxford.

Davidson EH and Erwin DH (2006). Gene regulatory networks and the evolution of animal body plans. Science. 311: 796-800.

Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C, Otim O, Brown CT, Livi CB, Lee PY, Revilla R, Rust AG, Pan Z, Schilstra MJ, Clarke PJ, Arnone MI, Rowen L, Cameron RA, McClay DR, Hood L and Bolouri H. (2002). A genomic regulatory network for development. Science. 295: 1669–1678.

de la Calle-Mustienes E, Feijoo CG, Manzanares M, Tena JJ, Rodriguez-Seguel E, Letizia A, Allende ML and Gomez-Skarmeta JL. (2005). A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. Genome Res. 15: 1061–1072.

de la Cruz CC, Der-Avakian A, Spyropoulos DD, Tieu DD and Carpenter EM. (1999). Targeted disruption of Hoxd9 and Hoxd10 alters locomotor behavior, vertebral identity, and peripheral nervous system development. Dev Biol. 216: 595-610.

Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C and Antonarakis SE. (2003). Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). Science. 302: 1033–1035.

Distel M, Wullimann MF and Köster RW. (2009). Optimized Gal4 genetics for permanent gene expression mapping in zebrafish. Proc Natl Acad Sci USA.

Dong X, Navratilova P, Fredman D, Drivenes Ø, Becker TS and Lenhard B. (2010). Exonic remnants of whole-genome duplication reveal *cis*-regulatory function of coding exons. Nucleic Acids Res. 38: 1071-85.

Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET and Hirschhorn JN. (2006) Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet. 38, 223–227.

Ebner A, Cabernard C, Affolter M and Merabet S. (2005). Recognition of distinct target sites by a unique Labial/Extradenticle/Homothorax complex. Development. 132: 1591-600.

Elgar G and Vavouri T. (2008). Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends Genet. 24: 344-52.

Erwin DH and Davidson EH. (2009). The evolution of hierarchical gene regulatory networks. Nat Rev Genet. 10: 141-8.

Ferretti E, Marshall H, Pöpperl H, Maconochie M, Krumlauf R, Blasi F. (2000). Segmental expression of Hoxb2 in r4 requires two separate sites that integrate cooperative interactions between Prep1, Pbx and Hox proteins. Development. 127: 155-66.

Ferrier DE, Brooke NM, Panopoulou G and Holland PW. (2001). The Mnx homeobox gene class defined by HB9, MNR2 and amphioxus AmphiMnx. Dev Genes Evol. 211: 103-7.

Fisher S, Grice EA, Vinton RM, Bessling SL and McCallion AS. (2006) Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science 312:276–279.

Freitas R, Zhang G and Cohn MJ. (2006). Evidence that mechanisms of fin development evolved in the midline of early vertebrates. Nature. 442: 1033-7.

Gehrig J, Reischl M, Kalmár E, Ferg M, Hadzhiev Y, Zaucker A, Song C, Schindler S, Liebel U and Müller F.(2009). Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos. Nat Methods. 6: 911-6.

Glazov EA, Pheasant M, McGraw EA, Bejerano G and Mattick JS. (2005). Ultraconserved elements in insect genomes: A highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. Genome Res. 15: 800–808.

Gómez-Skarmeta JL and Modolell J. (2002). Iroquois genes: genomic organisation and function in vertebrate neural development. Curr Opin Genet Dev. 12: 403-8.

Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA and Ovcharenko I. (2010). Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res. 20: 565-77.

Guillemot F. (2007). Spatial and temporal specification of neural fates by transcription factor codes. Development. 134: 3771-80.

Hahn MW and Wray GA. (2002). The g-value paradox. Evol Dev. 4: 73-75.

Halder G, Callaerts P and Gehring WJ. (1995). Induction of ectopic eyes by targeted expressin of the eyeless gene in Drosophila. Science. 267: 1788-92.

Hare EE, Peterson BK, Iyer VN, Meier R and Eisen MB. (2008). Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet. 4: e1000106.

Heine P, Dohle E, Bumsted-O'Brien K, Engelkamp D and Schulte D. (2008). Evidence for an evolutionary conserved role of homothorax/Meis during vertebrate retina development. Development 135: 805 -811.

Hertz GZ and Stormo GD. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics. 15: 563-77.

Hinman VF and Davidson EH (2004). Evolutionary plasticity of developmental gene regulatory network architecture. Proc Natl Acad Sci USA. 104: 19404-9.

Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, Kennedy BP and Wasserman WW. (2005). Nucleic Acids Res. 33: 3154-64.

Holland LZ, Albalat R, Azumi K, Benito-Gutiérrez E, Blow MJ, Bronner-Fraser M, Brunet F, Butts T, Candiani S, Dishaw LJ, Ferrier DE, Garcia-Fernàndez J, Gibson-Brown JJ, 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 GW, McHardy AC, Meulemans D, Nonaka M, Olinski RP, Pancer Z, Pennacchio LA, Pestarino M, Rast JP, 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 JK, Zhang Q, Zmasek CM, de Jong PJ, Osoegawa K, Putnam NH, Rokhsar DS, Satoh N and Holland PW. (2008). The amphioxus genome illuminates vertebrate origins and cephalochordate biology. Genome Res. 18: 1100-11.

Hossain AM, Rizk B, Behzadian A and Thorneycroft IH. (1997). Modified guanidinium thiocyanate method for human sperm DNA isolation. Molecular Human Reproduction. 3(11): 953-6.

Howard M. L. and Davidson E. H. (2004). *cis*-Regulatory control circuits in development. Dev. Biol. 271:109 -118.

Hoyle J, Tang YP, Wiellette EL, Wardle FC and Sive H. (2004). nlz gene family is required for hindbrain patterning in the zebrafish. Dev Dyn. 229: 835-46.

Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, Kim J, Gordon L, Branscomb E and Stubbs L. (2006). A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. Genome Res. 16: 669-77.

Ikuta T, Satoh N and Saiga H. (2010). Limited functions of Hox genes in the larval development of the ascidian Ciona intestinalis. Development. 137: 1505-13.

Jacobs Y, Schnabel CA and Cleary ML. (1999). Trimeric association of Hox and TALE homeodomain proteins mediates Hoxb2 hindbrain enhancer activity. Mol Cell Biol. 19: 5134–5142.

Jackman WR, Langeland JA and Kimmel CB. (2000). islet reveals segmentation in the Amphioxus hindbrain homolog. Dev Biol. 220: 16-26.

Jackman WR and Kimmel CB. (2002). Coincident iterated gene expression in the amphioxus neural tube. Evol Dev. 4: 366-74.

Jaillon O, Aury JM, Brunet F, Petit JL, 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 V, Jubin C, Castelli V, Katinka M, Vacherie B, Biémont C, Skalli Z, Cattolico L, Poulain J, De Berardinis V, Cruaud C, Duprat S, Brottier P, Coutanceau JP, Gouzy J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M, Volff JN, Guigó R, Zody MC, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quétier F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, Roest Crollius H. (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature. 431: 946-57.

Johnson DS, Davidson B, Brown CD, Smith WC and Sidow A. (2004). Noncoding regulatory sequences of Ciona exhibit strong correspondence between evolutionary constraint and functional importance. Genome Res. 14: 2448-56.

Joshi R, Passner JM, Rohs R, Jain R, Sosinsky A, Crickmore MA, Jacob V, Aggarwal AK, Honig B and Mann RS. (2007). Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell. 131: 530-43.

Kadonaga JT. (2004). Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. Cell. 116(2):247-57.

Karpinski BA, Morle GD, Huggenvik J, Uhler MD and Leiden JM. (1992). Molecular cloning of human CREB-2: an ATF/CREB transcription factor that can negatively regulate transcription from the cAMP response element. Proc Natl Acad Sci USA. 89: 4820-4.

Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engström PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K, Ghislain J, Pezeron G, Mourrain P, Ellingsen S, Oates AC, Thisse C, Thisse B, Foucher I, Adolf B, Geling A, Lenhard B and Becker TS. (2007). Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res. 17: 545-55.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G, Greenberg ME. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature. 465: 182-7.

Kimura A, Nishiyori A, Murakami T, Tsukamoto T, Hata S, Osumi T, Okamura R, Mori M and Takiguchi M. (1993). Chicken ovalbumin upstream promoter-transcription factor (COUP-TF) represses transcription from the promoter of the gene for ornithine transcarbamylase in a manner antagonistic to hepatocyte nuclear factor-4 (HNF-4). J Biol Chem. 268: 11125-33.

Kitamoto T and Salvaterra PM. (1995). A POU homeo domain protein related to dPOU-19/pdm-1 binds to the regulatory DNA necessary for vital expression of the Drosophila choline acetyltransferase gene. J Neurosci. 15: 3509-18.

Klemm JD, Rould MA, Aurora R, Herr W and Pabo CO. (1994). Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. Cell. 77: 21-32.

Koebernick K, Kashef J, Pieler T and Wedlich D. (2006). Xenopus Teashirt1 regulates posterior identity in brain and cranial neural crest. Deb Biol. 298: 312-26.

Kondoh H and Kamachi Y. (2010). SOX-partner code for cell specification: Regulatory target selection and underlying molecular mechanisms. Int J Biochem Cell Biol. 42: 391-9.

Kumar S and Hedges SB. (1998) A molecular timescale for vertebrate evolution. Nature. 392: 917–920.

Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH and Bourque G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 42: 631-4.

Kuraku S and Kuratani S. (2006). Time scale for cyclostome evolution inferred with a phylogenetic diagnosis of hagfish and lamprey cDNA sequences. Zoolog Sci. 23: 1053-64.

Kuraku S, Meyer A and Kuratani S. (2009). Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? Mol Biol Evol. 26: 47-59.

Kuratani S, Ueki T, Aizawa S and Hirano S. (1997). Peripheral development of cranial nerves in a cyclostome, Lampetra japonica: morphological distribution of nerve branches and the vertebrate body plan. J Comp Neurol. 384: 483-500.

Kusakabe R, Tochinai S and Kuratani S. (2003). Expression of foreign genes in lamprey embryos: an approach to study evolutionary changes in gene regulation. J Exp Zool B Mol Dev Evol. 15: 87-97.

Lecaudey V, Anselme I, Dildrop R, Rüther U and Schneider-Maunoury S. (2005). Expression of the zebrafish Iroquois genes during early nervous system formation and patterning. J Comp Neurol. 492: 289-302.

Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE and de Graaff E. (2003). A long range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet. 12: 1725–1735.

Levine M and Davidson EH. (2005). Gene regulatory networks for development. Proc. Natl. Acad. Sci. 102: 4936–4942.

Levine M. and Tjian R. (2003). Transcription regulation and animal diversity. Nature. 424: 147-51

Li X, Veraksa A and McGinnis W. (1999). A sequence motif distinct from Hox binding sites controls the specificity of a Hox response element. 126: 5581-9.

Li Q, Ritter D, Yang N, Dong Z, Li H, Chuang JH and Guo S. (2010). A systematic approach to identify functional motifs within vertebrate developmental enhancers. Dev Biol. 337: 484-95.

Loots, G.G. and Ovcharenko, I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. Nucleic Acids Res., 32, W217–W221

Lunter G, Ponting CP and Hein J. (2006). Genome-wide identification of human functional DNA using a neutral indel model. PLoS Comput Biol. 2: e5.

Lynch VJ and Wagner GP. (2008). Resurrecting the role of transcription factor change in developmental evolution. 62: 2131-54.

Maeda R, Ishimura A, Mood K, Park EK, Buchberg AM and Daar IO. (2002). Xpbx1b and Xmeis1b play a collaborative role in hindbrain and neural crest gene expression in Xenopus embryos. Proc Natl Acad Sci. USA 99: 5448 -5453.

Malicki J, Schughart K and McGinnis W. (1990). Mouse Hox-2.2 specifies thoracic segmental identity in Drosophila embryos and larvae. Cell. 63: 961-7.

Mangelsdorf DJ, Thummel C, Beato M, Herrlich P, Schütz G, Umesono K, Blumberg B, Kastner P, Mark M, Chambon P and Evans RM. (1995). The nuclear receptor superfamily: the second decade. Cell. 83: 835-9.

Mann RS and Chan SK. (1996). Extra specificity from extradenticle: the partnership between HOX and PBX/EXD homeodomain proteins. Trends Genet. 12: 258-62.

Mann RS, Lelli KM and Joshi R. (2009). Hox specificity: unique roles for cofactors and collaborators. Curr Top Dev Biol. 88:63-101.

Manzanares M, Wada H, Itasaki N, Trainor PA, Krumlauf R and Holland PW. (2000). Conservation and elaboration of Hox gene regulation during evolution of the vertebrate head. Nature. 408: 854-7.

Mazet F and Shimeld SM. (2002). The evolution of chordate neural segmentation. Dev Biol. 251: 258-70.

McCauley DW and Bronner-Fraser M. (2002). Conservation of Pax gene expression in ectodermal placodes of the lamprey. Gene. 287: 129-39.

McCauley DW and Bronner-Fraser M. (2006). Importance of SoxE in neural crest development and the evolution of the pharynx. Nature. 441: 750-2.

McEwen GK, Woolfe A, Goode D, Vavouri T, Callaway H and Elgar G. (2006) Ancient duplicated conserved noncoding elements in vertebrates: A genomic and functional analysis. Genome Res. 16:451–465.

McEwen GK, Goode DK, Parker HJ, Woolfe A, Callaway H and Elgar G. (2009). Early evolution of conserved regulatory sequences associated with development in vertebrates. PLoS Genet. 5: e1000762.

McGaughey DM, Vinton RM, Huynh J, Al-Saif A, Beer MA and McCallion AS. 2008. Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. Genome Res. 18:252-260.

McGinnis N, Kuziora MA, McGinnis W. Human Hox-4.2 and Drosophila deformed encode similar regulatory specificities in Drosophila embryos and larvae. Cell. 63: 969-76.

Meader S, Ponting CP and Lunter G. (2010). Massive turnover of functional sequence in human and other mammalian genomes. Genome Res. Sep 1 [Epub ahead of print]

Mercader N, Leonardo E, Piedra ME, Martinez AC, Ros MA and Torres M. (2000). Opposing RA and FGF signals control proximodistal vertebrate limb development through regulation of Meis genes. Development 127: 3961-3970.

Merika M and Thanos D. (2001). Enhanceosomes. Curr Opin Genet Dev. 11: 205-8.

Moens CB and Selleri L. (2006). Hox cofactors in vertebrate development. Dev Biol. 291: 193-206.

Morley RH, Lachani K, Keefe D, Gilchrist MJ, Flicek P, Smith JC and Wardle FC. (2009). A gene regulatory network directed by zebrafish No tail accounts for its roles in mesoderm formation. Proc Natl Acad Sci USA. 106: 3289-34.

Müller F, Williams DW, Kobolák J, Gauvry L, Goldspink G, Orbán L and Maclean N. (1997). Activator effect of coinjected enhancers on the muscle-specific expression of promoters in zebrafish embryos. Mol Reprod Dev. 47: 404-412.

Müller F, Blader P, and Strahle U. (2002). Search for enhancers: Teleost models in comparative genomic and transgenic analysis of *cis*-regulatory elements. BioEssays 24: 564-572.

Murakami Y, Ogasawara M, Sugahara F, Hirano S, Satoh N and Kuratani S. (2001). Identification and expression of the lamprey Pax6 gene: evolutionary origin of the segmented brain of vertebrates. Development. 128: 3521-31.

Murakami Y, Pasqualetti M, Takio Y, Hirano S, Rijli FM and Kuratani S. (2004). Segmental development of reticulospinal and branchiomotor neurons in lamprey: insights into the evolution of the vertebrate hindbrain. Development. 131: 983-95.

Nam J, Dong P, Tarpine R, Istrail S and Davidson EH. (2010). Functional *cis*-regulatory genomics for systems biology. Proc Natl Acad Sci USA. 107: 3930-5.

Narlikar L, Sakabe NJ, Blanski AA, Arimura FE, Westlund JM, Nobrega MA and Ovcharenko I. Genome-wide discovery of human heart enhancers. Genome Res. 20: 381-92.

Navratilova P, Fredman D, Hawkins TA, Turner K, Lenhard B and Becker TS. (2009). Systematic human/zebrafish comparative identification of *cis*-regulatory activity around vertebrate developmental transcription factor genes. Dev Biol. 327: 526-40.

Nikitina N, Bronner-Fraser M and Sauka-Spengler T. (2009). The sea lamprey Petromyzon marinus: a model for evolutionary and developmental biology. Cold Spring Harb Protoc. pdb.emo113.

Nobrega MA, Ovcharenko I, Afzal V and Rubin EM. (2003). Scanning human gene deserts for long-range enhancers. Science. 302:413.

Nobrega MA, Zhu Y, Plajzer-Frick I, Afzal V and Rubin EM. (2004). Megabase deletions of gene deserts result in viable mice. Nature 431: 988–993.

Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH and Wolfe SA. (2008). Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell. 133: 1277-89.

Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK and Fraenkel E. (2007). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat Genet. 39: 730-2.

Ogino H, McConnell WB and Grainger RM. (2006). High-throughput transgenesis in Xenopus using I-SceI meganuclease. Nat Protoc. 1: 1703-10.

Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W and Stubbs L. (2005). Evolution and functional classification of vertebrate gene deserts. Genome Res. 15: 137–145.

Pavesi G, Mauri G and Pesole G. (2004a). In silico representation and discovery of transcription factor binding sites. Brief Bioinform. 5: 217-36.

Pavesi G, Mereghetti P, Mauri G, Pesole G. (2004b). Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res. 32: W199-203

Pennacchio LA and Rubin EM. (2001). Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet. 2: 100–109.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A and Rubin EM. (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature. 444: 499-502.

Pereira FA, Tsai MJ and Tsai SY. (2000). COUP-TF orphan nuclear receptors in development and differentiation. Cell Mol Life Sci. 57: 1388-98.

Pfeffer PL, Gerster T, Lun K, Brand M and Busslinger M. (1998). Characterisation of three novel members of the zebrafish Pax2/5/8 family: dependency of Pax5 and Pax8 expression on the Pax2.1 (noi) function. Development. 125: 3063-74.

Phelan ML and Featherstone MS. (1997). Distinct HOX N-terminal arm residues are responsible for specificity of DNA recognition by HOX monomers and HOX.PBX heterodimers. J Biol Chem. 272: 8635-43.

Phillips K and Luisi B. (2000). The virtuoso of versatility: POU proteins that flex to fit. J Mol Biol. 302: 1023-39.

Pöpperl H, Bienz M, Studer M, Chan SK, Aparicio S, Brenner S, Mann RS and Krumlauf R. (1995). Segmental expression of Hoxb-1 is controlled by a highly conserved autoregulatory loop dependent upon exd/pbx. Cell. 81: 1031-42.

Portales-Casamar E, Kirov S, Lim J, Lithwick S, Swanson MI, Ticoll A, Snoddy J and Wasserman WW. (2007). PAZAR: a framework for collection and dissemination of *cis*-regulatory sequence annotation. Genome Biol. 8: R207.

Prabhakar S, Visel A, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Morrison H, Fitzpatrick DR, Afzal V, Pennacchio LA, Rubin EM and Noonan JP. (2008). Human-specific gain of function in a developmental enhancer. Science. 321: 1346-50

Prasad BC, Ye B, Zackhary R, Schrader K, Seydoux G and Reed RR. (1998). unc-3, a gene required for axonal guidance in Caenorhabditis elegans, encodes a member of the O/E family of transcription factors. Development. 125: 1561-8.

Protas ME, Hersey C, Kochanek D, Zhou Y, Wilkens H, Jeffery WR, Zon LI, Borowsky R and Tabin CJ. (2006). Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. Nat Genet. 38: 107-11.

Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK, Benito-Gutiérrez EL, Dubchak I, Garcia-Fernàndez J, Gibson-Brown JJ, Grigoriev IV, Horton AC, de Jong PJ, Jurka J, Kapitonov VV, Kohara Y, Kuroki Y, Lindquist E, Lucas S, Osoegawa K, Pennacchio LA, Salamov AA, Satou Y, Sauka-Spengler T, Schmutz J, Shin-I T, Toyoda A, Bronner-Fraser M, Fujiyama A, Holland LZ, Holland PW,

Satoh N and Rokhsar DS. The amphioxus genome and the evolution of the chordate karyotype. Nature. 453: 1064-71.

Rada-Iglesias A, Ameur A, Kapranov P, Enroth S, Komorowski J, Gingeras TR and Wadelius C. (2008). Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders. Genome Res. 18: 380-92.

Ragvin A, Moro E, Fredman D, Navratilova P, Drivenes Ø, Engström PG, Alonso ME, de la Calle Mustienes E, Gómez Skarmeta JL, Tavares MJ, Casares F, Manzanares M, van Heyningen V, Molven A, Njølstad PR, Argenton F, Lenhard B and Becker TS. (2010). Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3. Proc Natl Acad Sci USA. 107: 775-80.

Rieckhof GE, Casares F, Ryoo HD, Abu-Shaar M and Mann RS. (1997). Nuclear translocation of extradenticle requires homothorax, which encodes an extradenticle-related homeodomain protein. Cell 9: 171–183.

Rohrschneider MR, Elsen GE and Prince VE. (2007). Zebrafish Hoxb1a regulates mlultiple downstream genes including prickle1b. Dev Biol. 309: 358-72.

Robertson LK, Bowling DB, Mahaffey JP, Imiolczyk B and Mahaffey JW. (2004). An interactive network of zinc-finger proteins contributes to regionalization of the Drosophila embryo and establishes the domains of HOM-C protein function. Development. 131: 2781-9.

Rudnicki MA, Schnegelsberg PN, Stead RH, Braun T, Arnold HH and Jaenisch R. (1993). MyoD or Myf-5 is required for the formation of skeletal muscle. Cell. 75:1351-9.

Runko AP and Sagerström CG. (2003). Nlz belongs to a family of zinc-finger-containing repressors and controls segmental gene expression in the zebrafish hindbrain. Dev Biol. 262:254-67.

Ryoo HD and Mann RS. (1999). The control of trunk Hox specificity and activity by Extradenticle. Genes Dev. 13: 1704-16.

Samad OA, Geisen MJ, Caronia G, Varlet I, Zappavigna V, Ericson J, Goridis C and Rijli FM. (2004). Integration of anteroposterior and dorsoventral regulation of Phox2b transcription in cranial motoneuron progenitors by homeodomain proteins. Development. 131: 4071-83.

Sandelin A. Bailey P, Bruce S, Engstrom PG, Klos JM, Wasserman WW, Ericson J and Lenhard B. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics 5: 99.

Sandve GK, Abul O, Walseng V and Drabløs F. Improved benchmarks for computational motif discovery. BMC Bioinformatics. 8: 193.

Santos JS, Fonseca NA, Vieira CP, Vieira J and Casares F. (2010). Phylogeny of the teashirt-related zinc finger (tshz) gene family and analysis of the developmental expression of tshz2 and tshz3b in the zebrafish. Dev Dyn. 239: 1010-8.

Sauka-Spengler T, Meulemans D, Jones M and Bronner-Fraser M. (2007). Ancient evolutionary origin of the neural crest gene regulatory network. Dev Cell. 13: 405-20.

Sauka-Spengler T and Bronner-Fraser M. (2008). A gene regulatory network orchestrates neural crest formation. Nat Rev Mol Cell Biol. 9: 557-68.

Schubert M, Holland ND, Laudet V and Holland LZ. (2006). A retinoic acid-Hox hierarchy controls both anterior/posterior patterning and neuronal specification in the developing central nervous system of the cephalochordate amphioxus. Dev Biol. 296: 190-202.

Serpente P, Tümpel S, Ghyselinck NB, Niederreither K, Wiedemann LM, Dollé P, Chambon P, Krumlauf R and Gould AP. (2005). Direct cross-regulation between retinoic acid receptor β and Hox genes during hindbrain segmentation. Development. 503-13.

Shanmugan K, Green NC, Rambaldi I, Saragovi HU and Featherstone MS. (1999). PBX and MEIS as non-DNA-binding partners in trimeric complexes with HOX proteins. Mol Cell Biol. 19: 7577–7588.

Sharov AA and Ko MS. (2009). Exhaustive search for over-represented DNA sequence motifs with CisFinder. DNA Res. 16: 261-73.

Shen WF, Rozenfeld S, Kwong A, ves Kom LG, Lawrence HJ and Largman C. (1999). HOXA9 forms triple complexes with PBX2 and MEIS1 in myeloid cells. Mol Cell Biol. 19: 3051–3061.

Shi W, Levine M and Davidson B. (2005). Unravelling genomic regulatory networks in the simple chordate, *Ciona intestinalis*. Genome Res. 15: 1668–1674.

Shigetani Y, Sugahara F, Kawakami Y, Murakami Y, Hirano S and Kuratani S. (2002). Heterotypic shift of epithelial-mesenchymal interactions in vertebrate jaw evolution. Science. 296: 1316-9.

Shimeld SM and Holland PW. (2000). Vertebrate innovations. Proc Natl Acad Sci USA. 97: 4449-52.

Sironi M, Menozzi G, Comi GP, Cagliani R, Bresolin N and Pozzoli U. (2005). Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences. Hum Mol Genet. 14: 2533–2546.

Smith JJ, Antonacci F, Eichler EE and Amemiya CT. (2009). Programmed loss of millions of base pairs from a vertebrate genome. Proc Natl Acad Sci USA. 106: 11212-7.

Stedman A, Lecaudey V, Havis E, Anselme I, Wassef M, Gilardi-Hebenstreit P and Schneider-Maunoury S. (2009). A functional interaction between Irx and Meis patterns the anterior hindbrain and activates krox20 expression in rhombomere 3. Dev Biol. 327: 566-77.

Steelman S, Moskow JJ, Muzynski K, North C, Druck T, Montgomery JC, Huebner K, Daar IO and Buchberg AM. (1997). Identification of a conserved family of Meis1-related homeobox genes. Genome Res. 7: 142–156.

Stormo GD and Fields DS. (1998). Specificity, free energy and information content in protein-DNA interactions. Trends Biochem Sci. 23: 109-13.

Taghli-Lamallem O, Gallet A, Leroy F, Malapert P, Vola C, Kerridge S and Fasano L. (2007). Direct interaction between Teashirt and Sex combs reduced proteins, via Tsh's acidic domain, is essential for specifying the identity of the prothorax in Drosophila. Dev Biol. 307: 142-51.

Takio Y, Pasqualetti M, Kuraku S, Hirano S, Rijli FM and Kuratani S. (2004). Evolutionary biology: lamprey Hox genes and the evolution of jaws. 429: 1.

Takio Y, Kuraku S, Murakami Y, Pasqualetti M, Rijli FM, Narita Y, Kuratani S and Kusakabe R. (2007). Hox gene expression patterns in Lethenteron japonicum embryos--insights into the evolution of the vertebrate Hox code. Dev Biol. 308: 606-20.

Thermes V, Grabher C, Ristoratore F, Bourrat F, Choulika A, Wittbrodt J and Joly JS. (2002). I-SceI meganuclease mediates highly efficient transgenesis in fish. Mech Dev. 118: 91-8.

Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Régnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C and Zhu Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol. 23: 137-44.

Toresson H, Parmar M and Campbell K. (2000). Expression of Meis and Pbx genes and their protein products in the developing telencephalon: implications for regional differentiation. Mech Dev. 94: 183-187.

Tümpel S, Cambronero F, Wiedemann LM and Krumlauf R. (2006). Evolution of cis elements in the differential expression of two Hoxa2 coparalogous genes in pufferfish (Takifugu rubripes). Proc Natl Acad Sci USA. 103: 5419-24.

Tümpel S, Wiedemann LM and Krumlauf R. (2009). Hox genes and segmentation of the vertebrate hindbrain. Curr Top Dev Biol. 88: 103-137.

Tvrdik P and Capecchi MR. (2006). Reversal of Hox1 gene subfunctionalization in the mouse. Dev Cell. 11: 239-50.

Uhl JD, Cook TA and Gebelein B. (2010). Comparing anterior and posterior Hox complex formation reveals guidelines for predicting *cis*-regulatory elements. Dev Biol. 343: 154-66.

Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM and Sidow A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods. 5: 829-34.

Vavouri T, McEwen GK, Woolfe A, Gilks WR and Elgar G. (2006). Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key. Trends Genet. 22:5–10.

Vavouri T, Walter K, Gilks WR, Lehner B and Elgar G. (2007). Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. Genome Biol. 8: R15.

Vavouri T and Lehner B. (2009). Conserved noncoding elements and the evolution of animal body plans. Bioessays. 31: 727-35.

Venkatesh B, Kirkness EF, Loh YH, Halpern AL, Lee AP, Johnson J, Dandona N, Viswanathan LD, Tay A, Venter JC, Strausberg RL and Brenner S. (2006). Ancient noncoding elements conserved in the human genome. Science. 314: 1892.

Visel A, Rubin EM and Pennacchio LA. (2009). Genomic views of distant-acting enhancers. Nature. 461: 199-205.

Wada H, Garcia-Fernàndez J and Holland PW. (1999). Colinear and segmental expression of amphioxus Hox genes. Dev Biol. 213: 131-41.

Wada S, Tokuoka M, Shoguchi E, Kobayashi K, Di Gregorio A, Spagnuolo A, Branno M, Kohara Y, Rokhsar D, Levine M, Saiga H, Satoh N and Satou Y. (2003). A genomewide survey of developmentally relevant genes in Ciona intestinalis. II. Genes for homeobox transcription factors. Dev Genes Evol. 213: 222-34.

Wahba GM, Hostikka SL and Carpenter EM. (2001). The paralogous Hox genes Hoxa10 and Hoxd10 interact to pattern the mouse hindlimb peripheral nervous system and skeleton. Dev Biol. 231: 87-102.

Waskiewicz AJ, Rikhof HA, Hernandez RE and Moens CB. (2001). Zebrafish Meis functions to stabilize Pbx proteins and regulate hindbrain patterning. Development 128: 4139-4151.

Wassef MA, Chomette D, Pouilhe M, Stedman A, Havis E, Desmarquet-Trin Dinh C, Schneider-Maunoury S, Gilardi-Hebenstreit P, Charnay P and Ghislain J. (2008). Rostral hindbrain patterning involves the direct activation of a Krox20 transcriptional enhancer by Hox/Pbx and Meis factors. Development. 135: 3369-78.

Wasserman WW, Palumbo M, Thompson W, Fickett JW and Lawrence C E. (2000). Human-mouse genome comparisons to locate regulatory sites. Nature Genet. 26: 225−228

Wasserman WW and Sandelin A. (2004). Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet. 5: 276–287.

Wei W and Yu XD. (2007). Comparative analysis of regulatory motif discovery tools for transcription factor binding sites. Genomics Proteomics Bioinformatics. 5: 131-42.

Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VL, Fisher EM, Tavaré S and Odom DT. (2008). Species-specific transcription in mice carrying human chromosome 21.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJ, Cooke JE and Elgar G. (2005). Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 3: e7.

Woolfe A and Elgar G. (2007). Comparative genomics using Fugu reveals insights into regulatory subfunctionalization. Genome Biol. 8: R53.

Woolfe A, Goode DK, Cooke J, Callaway H, Smith S, Snell P, McEwen GK and Elgar G. (2007). CONDOR: a database resource of developmentally associated conserved non-coding elements. BMC Dev Biol. 7: 100.

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES and Kellis M. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. (2005). Nature. 434: 338-45.

Zerucha T and Prince VE. (2001). Cloning and developmental expression of a zebrafish meis 2 homeobox gene. Mech Dev 102: 247–250.

Zhang X, Friedman A, Heaney S, Purcell P and Maas RL. (2002). Meis homeoproteins directly regulate Pax6 during vertebrate lens morphogenesis. Genes Dev. 16: 2097–2107.

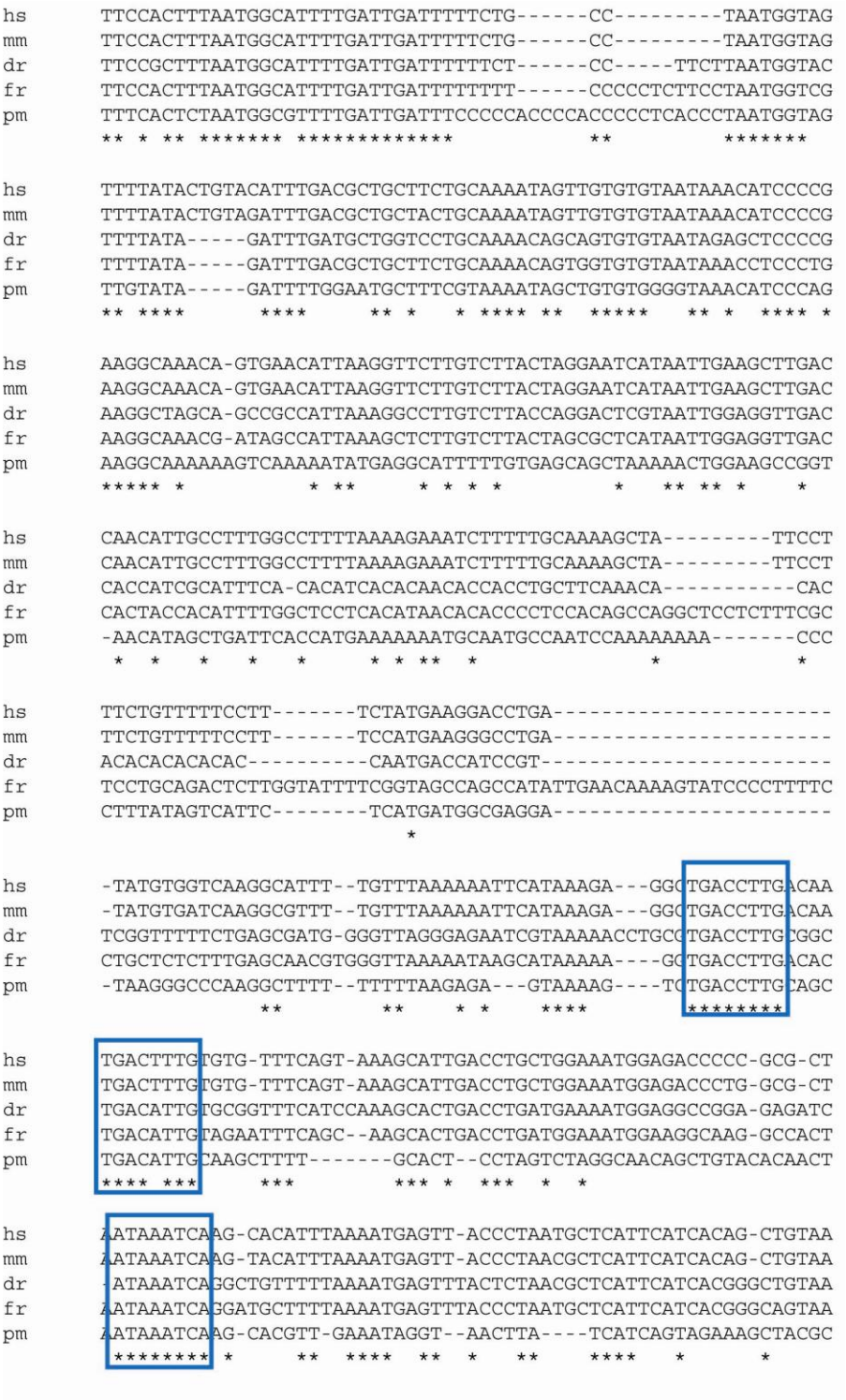
Appendix
Chapter 3.

Table A1. A list of gene regions with lamprey CNEs.

gene region	#lamprey CNEs		gene region	#lamprey CNEs
TSHZ3	16		ZNF703	2
TSHZ1	12		ZIC2	2
IRX5	12		TFAP2A	2
EBF3	12		SOX5	2
PAX2	9		SOX21	2
NR2F2	9		SOX14	2
MEIS2	9		SALL3	2
ESRRG	9		POU6F2	2
ZNF503	8		POU4F2	2
FOXP2	8		POU3F2	2
BNC2	8		MAB21L2	2
NR2F1	6		LMO4	2
ZFHX4	5		FOXB1	2
IRX2	5		FOG2	2
HMX2	5		EYA1	2
ATBF1	5		ESRRB	2
ZFHX1B	4		EMX2	2
TSHZ2	4		EBF1	2
PAX1	4		SP8	1
NR4A2	4		SOX2	1
MEIS1	4		SHOX2	1
BCL11A	4		SHOX	1
TCF7L2	3		PRDM16	1
SOX6	3		PBX3	1
OTP	3		PAX9	1
NKX6-1	3		PAX5	1
HOXD9	3		PAX3	1
GBX2	3		MAF	1
FOXP1	3		MAB21L1	1
EVI1	3		LMO1	1
DACH1	3		FOXD3	1
BCL11B	3		CST	1
ARX	3		AUTS2	1

Chapter 4.

Figure A1. Multiple sequence alignment of CNE 3285-6. Chunks of lamprey sequence conservation are highlighted in the region of CNE 3286. Generated using ClustalW2.



Chapter 5.

List of injected Pbx-Hox +ve CNEs:

Lamprey set -

>drEvi1_40224

>drNR2F2_27254

>drPax2_217

>drZNF503_32799

>drTshz3_24797

>drTshz3_24798

>drTshz3_24800

>drTshz3_24804

>drTshz3_24805-6

>drTshz3_24807

>drTshz3_43509

Gnathostome set -

>drBCL11A_2554

>drFoxp1_886

>drFoxp2_3502

>drGli3_2152

>drNKX6.1_4281

>drPax9_2099

>drPou3f2_9802

>drSP8_1540

>drTCF7L2_5416

>drTshz3_7655

Table A2. The number of Pbx-Hox motif hits in the CNEs of each gene region. Control sets generated by 0-order markov model.

GENE	# hits in test set	kb size	#hits per kb	# hits in control set (average)	standard deviation	z-score	p-value
ZNF503	36	27.781	1.295849681	3.184	1.762425601	18.61979307	0.00E+00
TSHZ3	30	23.323	1.286283926	3.085	1.767420437	15.22840827	0.00E+00
IRX5	27	37.059	0.728567959	5.388	2.32539373	9.293909982	0.00E+00
IRX2	21	23.981	0.875693257	3.1	1.799444359	9.947515139	0.00E+00
HOXD9	16	17.77	0.900393922	2.186	1.439931943	9.593508963	0.00E+00
MEIS2	16	24.553	0.651651529	3.42	1.909869105	6.586838838	4.49E-11
NR2F1	16	25.655	0.623660105	3.724	1.840060869	6.671518432	2.53E-11
NR2F2	16	18.99	0.84254871	2.524	1.587269353	8.490052411	0.00E+00
PBX3	16	17.886	0.8945544	1.89	1.351998521	10.43640195	0.00E+00
TSHZ1	16	10.351	1.545744373	1.627	1.315245604	10.92799699	0.00E+00
ZFHX1B	13	23.275	0.558539205	3.132	1.721213525	5.733164337	9.86E-09
FOXP1	12	15.857	0.756763574	1.731	1.244443249	8.251882927	2.22E-16
SALL3	12	11.405	1.052170101	1.425	1.206803629	8.762817531	0.00E+00
MAF	11	7.334	1.499863649	1.151	1.100090451	8.952900184	0.00E+00
BCL11A	10	13.643	0.732976618	2.143	1.432672677	5.48415568	4.15E-08
EBF3	10	26.18	0.38197097	3.219	1.792495188	3.782994813	1.55E-04
FOXP2	10	17.844	0.560412464	2.296	1.512079363	5.094970667	3.49E-07
NKX6-1	10	6.853	1.459214942	0.82	0.923904757	9.936089116	0.00E+00
MEIS1	9	9.298	0.967950097	1.457	1.214146202	6.212596133	5.21E-10
NR4A2	9	14.765	0.609549611	1.513	1.240899271	6.033527601	1.60E-09
DACH1	8	12.338	0.648403307	1.428	1.22589396	5.360985711	8.28E-08
ESRRG	8	8.743	0.915017728	1.731	1.36258541	4.600812509	4.21E-06
GLI3	8	3.432	2.331002331	0.503	0.702844933	10.66664871	0.00E+00
OTP	8	8.986	0.890273759	1.28	1.133843023	5.926746354	3.09E-09
SHOX2	8	7.615	1.050558109	0.966	0.998420753	7.045125994	1.85E-12
ZNF703	8	2.991	2.674690739	0.372	0.601345159	12.68489467	0.00E+00
EMX2	7	8.736	0.801282051	0.808	0.897293709	6.900750489	5.17E-12
ESRRB	7	4.537	1.542869738	0.664	0.820429156	7.722787463	1.13E-14
TCF7L2	7	11.045	0.633770937	1.514	1.214826737	4.515870317	6.31E-06
TSHZ2	7	7.221	0.969394821	0.861	0.93363751	6.575357068	4.85E-11
ATBF1	6	7.514	0.798509449	1.178	1.119962499	4.305501303	1.67E-05
EVI1	6	4.015	1.494396015	0.648	0.817371397	6.547819048	5.84E-11
FOG2	6	11.306	0.530691668	1.067	1.070752539	4.607040208	4.08E-06
MAB21L2	6	5.733	1.046572475	0.892	0.932918003	5.475293632	4.37E-08
POU3F2	6	3.297	1.819836215	0.47	0.689274981	8.022922852	1.11E-15
SOX14	6	4.286	1.399906673	0.417	0.638052506	8.750063592	0.00E+00
ZIC1	6	8.332	0.720115218	1.137	1.070621782	4.542220308	5.57E-06
BARHL2	5	12.98	0.385208012	1.405	1.167465203	3.079320901	2.07E-03
HMX2	5	7.662	0.65257113	0.977	0.959411799	4.19319421	2.75E-05
PAX2	5	9.193	0.543892092	1.083	1.045041148	3.748177772	1.78E-04
SATB1	5	5.558	0.899604174	0.54	0.755248304	5.905342626	3.52E-09
SOX6	5	10.525	0.475059382	1.377	1.175955356	3.080899271	2.06E-03
TFAP2A	5	11.087	0.450978624	1.251	1.107248391	3.385870803	7.10E-04
BNC2	4	10.59	0.377714825	1.65	1.32721513	1.770624782	7.66E-02
cont.							

GENE	# hits in test set	kb size	#hits per kb	# hits in control set (average)	standard deviation	z-score	p-value
EBF1	4	4.032	0.992063492	0.925	0.955706545	3.217514849	1.29E-03
FIGN	4	8.809	0.454081053	1.02	1.073126274	2.77693322	5.49E-03
FOXB1	4	8.603	0.464954086	1.181	1.106453343	2.547780273	1.08E-02
PAX1	4	9.457	0.422967114	1.048	0.998847336	2.955406592	3.12E-03
PAX6	4	4.913	0.814166497	0.731	0.886926716	3.685761112	2.28E-04
PAX9	4	3.217	1.243394467	0.51	0.742899724	4.697807641	2.63E-06
POU6F2	4	3.103	1.289075089	0.452	0.696918934	5.090979494	3.56E-07
PRDM16	4	3.736	1.070663812	0.737	0.834164852	3.911696822	9.16E-05
UNC4	4	6.23	0.642054575	0.65	0.84231823	3.977119196	6.98E-05
ZFHX4	4	12.676	0.315556958	1.954	1.379813031	1.482809594	1.38E-01
FOXD3	3	9.839	0.304909035	1.286	1.130576844	1.516040249	1.30E-01
MAB21L1	3	4.838	0.620090947	0.738	0.869112191	2.602655934	9.25E-03
POU3F1	3	4.128	0.726744186	0.722	0.847771196	2.687045764	7.21E-03
POU4F2	3	5.809	0.516440007	0.724	0.865923784	2.628406843	8.58E-03
ZIC2	3	7.68	0.390625	0.917	0.974736375	2.136988066	3.26E-02
AUTS2	2	7.036	0.284252416	1.01	1.018773773	0.971756465	3.31E-01
BHLHB5	2	6.937	0.288309067	0.959	0.970215955	1.072956999	2.83E-01
CST	2	8.39	0.238379023	0.959	0.979448314	1.06284322	2.88E-01
DLX1	2	6.28	0.318471338	0.703	0.810426431	1.600392029	1.10E-01
EN1	2	2.244	0.891265597	0.4	0.6244998	2.562050461	1.04E-02
EYA1	2	5.536	0.361271676	0.806	0.889024184	1.343045579	1.79E-01
GBX2	2	2.623	0.762485703	0.463	0.663800422	2.315454991	2.06E-02
LHX1	2	4.442	0.450247636	0.749	0.854399789	1.464185754	1.43E-01
LMO4	2	6.74	0.296735905	0.663	0.832725045	1.605571982	1.08E-01
PAX5	2	1.05	1.904761905	0.271	0.505528436	3.420183467	6.26E-04
PHOX2B	2	4.45	0.449438202	0.612	0.785783685	1.766389435	7.73E-02
SHOX	2	4.128	0.484496124	0.489	0.707021216	2.137135302	3.26E-02
SOX21	2	5.466	0.36589828	0.491	0.682582596	2.210721471	2.71E-02
SOX5	2	3.691	0.541858575	0.729	0.835199976	1.521791231	1.28E-01
SP8	2	4.61	0.433839479	0.507	0.705656432	2.115760492	3.44E-02
ARX	1	2.572	0.388802488	0.368	0.608749538	1.038193806	2.99E-01
BCL11B	1	3.257	0.30703101	0.576	0.747143895	0.567494431	5.70E-01
LMO1	1	4.728	0.211505922	0.444	0.665480278	0.835486819	4.03E-01
PITX2	1	5.806	0.172235618	0.798	0.902882052	0.223728005	8.23E-01
POU3F3	1	4.017	0.248941997	0.677	0.829862037	0.389221323	6.97E-01
SHH	1	5.019	0.199242877	0.726	0.88933908	0.308093961	7.58E-01
SOX1	1	1.802	0.554938957	0.303	0.522676764	1.333520155	1.82E-01
SOX11	1	1.678	0.595947557	0.185	0.422817928	1.927543622	5.39E-02
SOX3	1	2.049	0.488042948	0.257	0.494925247	1.501236812	1.33E-01
HLX1	0	2.925	0	0.369	0.605672354	-0.609240289	1.46E+00
PAX3	0	1.673	0	0.271	0.509469332	-0.531926032	1.41E+00
PAX7	0	2.807	0	0.391	0.60672811	-0.644440225	1.48E+00
PAX8	0	0.19	0	0.03	0.170587221	-0.175863115	1.14E+00
SOX2	0	4.237	0	0.601	0.799874365	-0.751367998	1.55E+00
SOX4	0	1.156	0	0.186	0.430585648	-0.431969809	1.33E+00