The role of vertebrate conserved non-coding elements in hindbrain development and evolution

Joseph Martin Grice

UCL

Submitted for the degree of Doctor of Philosophy

Division of Systems Biology, NIMR, Mill Hill, London

I hereby certify that all the work contained in this thesis is my own. Where information has been derived from other sources, this has been clearly indicated in the manuscript.

Joseph Martin Grice

For Johanna and Lilly.

ACKNOWLEDGEMENTS

Thanks to Greg for teaching, supervision and guidance. Thanks to everyone in Systems Biology at the NIMR, past and present, for making it such a great division to work in. Thanks to my family for love and support. Thanks to my teachers and students, and those who were both. I am indebted to the Medical Research Council for funding my training and supporting my work.

Megan Addison, Scott Barolo, James Briscoe, Brook Cooper, Eric Davidson, Laura Doglio, Elena de Domenico, Greg Elgar, Johanna Fischer, Alison Gilchrist, Rosa Gomes-Faria, Ian Grant, Katherine Hill, Peter Holland, Lilian Hunt, Nick Kent, Malcolm Logan, Dave McClay, Jim Murray, Boris Noyvert, Nick Owens, Stefan Pauls, Isabelle Peter, Elsie Place, Simon Scofield, Steve Small, Jim Smith, Hannah Stanforth, Anna Strobl, Glen Sweeney, David Wilkinson, Steve Wilson, Qiling Xu.

ABSTRACT

Vertebrate conserved non-coding elements (CNEs) act as *cis*-regulatory modules of developmental genes. To assess their roles in coordinating gene expression during embryogenesis, CNEs were subjected to motif searches. Using reporter gene assays in zebrafish (*Danio rerio*) embryos, Pbx-Hox (TGATNNAT) motifs are demonstrated to be poor predictors of hindbrain enhancer activity. Hindbrain enhancer CNEs are distinguished from hindbrain negative CNEs accurately by virtue of co-occurring Pbx-Hox (TGATDDATKD) and Meis/Pknox (CTGTCA) motifs. The grammar of these motifs was investigated using a bioinformatic pipeline for the detection of multiple conserved motifs, revealing no patterns in their relative organisation aside from spatial co-occurrence. These motifs were then used to identify additional conserved hindbrain enhancers with high efficacy (89%). Substitutions targeted to either motif abrogate expression by the enhancer or generate ectopic reporter gene expression, suggesting that motif co-occurrence is required for efficient and segment-specific hindbrain activation. Pbx-Hox and Meis/Pknox motifs are enriched in gnathostome CNE sets but are not detected in invertebrate chordate CNEs. Furthermore the presence (or absence) of the hindbrain syntax correlates with the conservation (or lack thereof) of segment-restricted enhancer activity in orthologous CNEs from the sea lamprey. A library of zebrafish hindbrain regulatory elements is made available. The heterogeneity of function and the loose grammar of motifs are consistent with combinatorial factor binding; a model of CNEs as exceptionally well-conserved billboard enhancers is presented (inflexible billboard model). The implications of these data for models of the evolution of the vertebrate hindbrain are discussed. Several components of the hindbrain gene regulatory network are shared-derived characters of gnathostomes, suggesting the establishment and elaboration of the conserved regulatory code controlling hindbrain development on the vertebrate and gnathostome stems, respectively.

CONTENTS

PREFACE 12

CHAPTER 1

Introduction 15

CHAPTER 2

Conserved Pbx-Hox motifs are poor predictors of hindbrain enhancers 39

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

APPENDIX

PREFACE

i Abbreviations and acronyms

i.i Anatomical terms

cg:	cranial ganglia
ey:	eye
fb:	forebrain
hb:	hindbrain
ht:	heart
mb:	midbrain
mhb:	midbrain-hindbrain boundary
msc:	trunk musculature
nc:	neural crest
pa:	pharyngeal arches
r1-7:	(variously) rhombomeres 1-7
sc:	spinal cord

i.ii CNE sets

AcCNEs:	Actinopterygian (ray-finned fish) CNEs
OsCNEs:	Osteicthyan (bony fish) CNEs
GnCNEs:	Gnathostome (jawed vertebrate) CNEs
VeCNEs:	Vertebrate CNEs
OlCNEs:	Olfactores (vertebrates and urochordates) CNEs
ChCNEs:	Chordate CNEs
CiCNEs:	*Ciona* CNEs

i.iii Molecular biology, developmental biology and bioinformatic terms

ChIP:	chromatin immunoprecipitation
CNE:	conserved non-coding element
CONDOR:	(database of) conserved non-coding orthologous regions
CRISPR:	clustered regularly interspaced short palindromic repeats
CRM:	*cis*-regulatory module
dpc:	days post-coitum
EMSA:	electrophoretic mobility shift assay

GA:	gnathostome ancestor (reconstructed sequence)
GRN:	gene regulatory network
hpf:	hours post-fertilisation
ISH:	*in situ* hybridisation
OA:	osteichthyan ancestor (reconstructed sequence)
ORF:	open reading frame
PBM:	protein-binding microarray
PCR:	polymerase chain reaction
PWM:	position weight matrix
SNP:	single nucleotide polymorphism
TF:	transcription factor
TFBS:	transcription factor binding site
VA:	vertebrate ancestor (reconstructed sequence)

i.iv Protein/*gene* nomenclature

Egfp/*egfp*:	enhanced green fluorescent protein (fluorophore/*reporter gene*)
Hox/*hox*:	homeodomain/*homeobox* (homeobox family)
mCherry/*mCherry*:	mCherry red fluorescent protein (fluorophore/*reporter gene*)
Meis/*meis*:	myeloid ecotropic viral integration site (TALE homeobox family)
Pbx/*pbx*:	pre-B-cell leukaemia homeobox (TALE homeobox family)
Pknox/*pknox*:	prep/knotted homeobox (TALE homeobox family)
Pou/*pou*:	pit/oct/unc-specific domain (homeobox family)
TALE:	three amino-acid loop extension (homeobox gene class)
bZIP:	basic leucine zipper (transcription factor family)

i.v Species

Bf:	the lancelet/amphioxus, *Branchiostoma floridae*
Ci:	the sea squirt, *Ciona intestinalis*
Cm:	the elephant shark/chimaera, *Challorhinchus milii*
Dr:	the zebrafish, *Danio rerio*
Gg:	the chicken, *Gallus gallus*
Mm:	the mouse, *Mus musculus*
Pm:	the sea lamprey, *Petromyzon marinus*
Tr:	the pufferfish/fugu, *Takifugu rubripes*

ii Notes on nomenclature

There is no accepted unifying nomenclature for vertebrate conserved non-coding elements (CNEs). In an attempt to give individual elements unique and concise identifiers, the following scheme is followed throughout this thesis.

CNE identifiers are composed of three parts: a species identifier, a gene name and a unique code. The species identifier states the orthologue of the CNE. The gene name is the accepted name for the gene that the element is thought to regulate or, when this is unknown, a nearby gene that is the best candidate (many CNEs have unknown functions; these identifiers are best interpreted as an indication of locus in the first instance). The unique identifier is either the last 4 or 5 digits of the CONDOR accession number (in the format "CRCNE*########*") or a suitable alternative where this is unavailable.

For example, the CONDOR CNE "CRCNE00001102" associating with the sea lamprey (*Petromyzon marinus*) *meis2* gene would be written as:

Pm.*meis2*.1102

The CONDOR CNE "CRCNE00010876" associating with the zebrafish (*Danio rerio*) *znf703* gene would be written as:

Dr.*znf703*.10876

Where CONDOR identifiers do not exist, an appropriate alternative (usually the published name of the element) is substituted for the final identifier.
For example, the mouse hoxb1 r4-specific autoregulatory element from Popperl *et al.* 1995[245] would be written as:

Mm.*hoxb1*.r4

The chicken *krox20* element C from Wassef *et al.* 2008[195] would be written as:

Gg.*egr2*.C

CHAPTER 1

Introduction

1.1 The evolution of the vertebrate body plan

Vertebrate embryos show striking morphological and molecular similarity, most evident at the phylotypic stage during mid-embryogenesis (reviewed by Irie and Kuratani[1]). According to the developmental hourglass model, the phylotypic stage is the most constrained portion of development. This is regarded to most closely resemble the ancestral body plan of vertebrates[2-4]. There has been a long-standing interest in the origin of the ancestral vertebrate body plan and its evolution from its chordate ancestor, highlighted by numerous studies and reviews on the subject[5, 6], many of which focus on the origins of the central nervous system (CNS)[7-11].

The common ancestor of vertebrates evolved from a chordate species approximately, according to molecular estimates, 750 million years ago (MYA)[12, 13]; paleontological evidence suggests a divergence time prior to 525 MYA, which is the earliest time that chordates and putative vertebrates can be found together in the fossil record[14, 15]. Morphologically speaking, the closest extant proxies of this ancestor are protochordates; either urochordates (tunicates), which resemble vertebrates most closely in their larval forms, or cephalochordates (lancelets). Another interesting group are the cyclostomes (lampreys and, possibly, hagfish[16-18]), the most basal vertebrates. These might represent proxies of transitional forms between cephalochordates and vertebrates[18, 19] (discussed below). A phylogeny of chordates is displayed in figure 1.1.

Vertebrates are distinct from invertebrate chordates in many aspects of embryonic development, resulting in derived adult morphology. Gans and Northcutt proposed that these distinctions were ecologically associated with a behavioural switch from passive to active feeding[20, 21]. In this model of vertebrate origins, most aspects of vertebrate embryology are thought to underlie suites of characters for the detection, capture and consumption of prey, largely focused at the anterior of the organism in the head. This is an evolutionary trend referred to as cephalisation. In the intervening years, molecular genetic and developmental studies have supported Gans and Northcutt's assertion that the developmental mechanisms underlying these characters are shared-derived in the vertebrates. These include:

1. **cranial neural crest** (nc), a pluripotent population of migratory cells which populate and contribute to numerous cranial, pharyngeal and maxillofacial structures[22-25];
2. enlargement of the **trunk musculature** enabling undulatory locomotion;

3. **hypopharyngeal muscle**, which enables swallowing[26];
4. **teeth** and a **digestive system** lined with smooth muscle to mechanically break large items of food;
5. **epithelial placodes**, which develop in to paired sense organs and the ganglia of cranial nerves[27, 28];
6. an elaborated and **segmented CNS**[29, 30] for both the integration of sensory stimuli and motor control of newly derived structures.

These traits are all essential aspects of adult morphology required for active feeding behaviour. In this sense lampreys are apparently distinct from other vertebrates: they filter-feed (like cephalochordates) as juveniles and metamorphose in to active feeders (like other vertebrates) only in adulthood[18, 19]. Thus, the lamprey could be thought of as an ecological and morphological intermediate between cephalochordates and vertebrates.

However, the genetic basis for the cranial elaboration that marked early vertebrate evolution is not entirely clear. Tandem gene duplications (TGDs) have been hypothesised to underlie a proportion of genome, and therefore morphological, evolution since the 1960s[31], by generating redundancy amongst duplicates relaxing selection pressure on one (neofunctionalisation) or both (subfunctionalisation) copies of the gene. Gene duplications can generate new functions by altering exons, regulatory sequences or both. It is also demonstrable that vertebrates have undergone two whole-genome duplications (WGDs) [32, 33] with respect to protochordates, although precisely when these occurred is unclear[18, 34, 35]. The prevailing model of vertebrate origins (the 2R hypothesis) posits that two WGDs occurred at the base of vertebrates. With this new raw material (i.e. swathes of redundancy amongst duplicated regulatory proteins) *cis* and *trans*-regulatory mutations would have been permitted, allowing and underlying the re-wiring of gene regulatory networks (GRNs, discussed below)[36, 37]. This re-wiring created novel interactions and, by altering development, morphology upon which selection then acted.

Furthermore, the evolution of the vertebrate body plan was concomitant with the emergence and fixation of putative gene regulatory elements, detectable by their conservation in extant vertebrates[38-41]. Since many conserved elements are preserved in all vertebrates, these might be required for vertebrate-specific aspects of development. Studying these conserved sequences therefore might provide insight in to the innovations which evolved in early vertebrates.

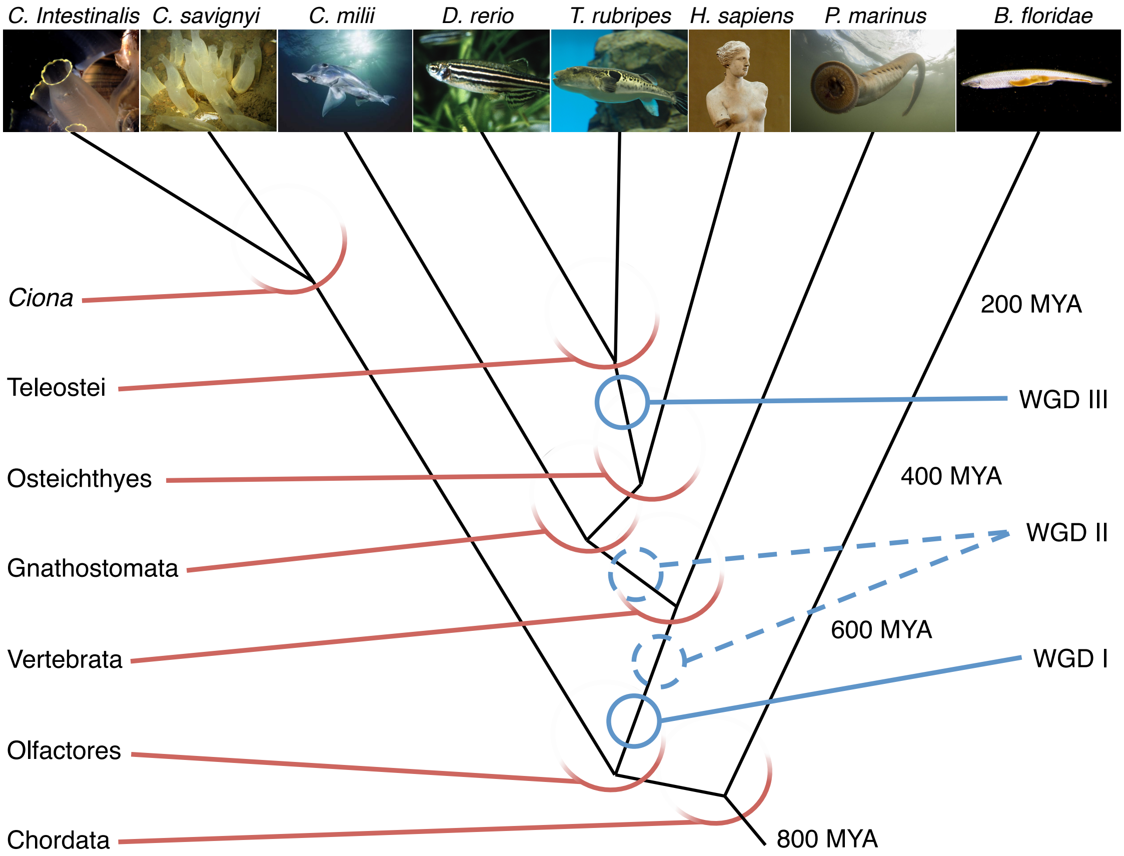


Figure 1.1: Cladogram of chordate genomes. Selected species for which genomic data exist (from left to right: *Ciona intestinalis*; *Ciona savignyi*; *Challorhinchus milii*; *Danio rerio*; *Takifugu rubripes*; *Homo sapiens*; *Petromyzon marinus*; *Branchiostoma floridae*) and their interrelationships are shown. Taxa are indicated with red arcs. Occurrences of whole genome duplications (WGDs) are indicated with blue circles. According to the 2R hypothesis, two WGDs took place in early vertebrates, but whether the second of these occurred before or after the gnathostome/cyclostome divergence is a matter of contention[5, 35] (WGD I, solid line, and possible timings of WGD II, dashed lines). A third WGD took place in stem teleosts (WGD III)[42]. Estimated divergence times, inferred from molecular data, are shown below in millions of years ago (MYA)[12, 43, 44].

The integration of evolutionary and developmental studies (termed evolutionary developmental biology or evo-devo for short[45-47]) has proven useful for understanding the evolution of animal body plans. These often concern detailed molecular studies of single or small groups of genes. Simultaneously, the increasing use of genomic and functional data together using a systems biology approach[48, 49] has allowed the mapping of GRNs in model organisms such as the sea urchin, *Drosophila* and, more recently, model vertebrates. If modern biology seeks to understand how changes in genomes lead to changes in adult morphology over evolutionary time then comparative genomic, molecular and developmental studies must be performed in concert.

1.2 *Cis*-regulatory modules in development

The development of multicellular organisms depends upon the establishment of precise gene expression patterns in both space and time. These are established by regulatory links between genes encoding developmental regulatory proteins (TFs, morphogens, and receptors) or the pathways necessary for the synthesis of hormones. Together, these genes and their links comprise a gene regulatory network (GRN). The nature and topology of the links together with the timing of molecular events (diffusion of morphogens, signal transduction cascades, nuclear localisation, transcription, translation *et c.*) determine the patterns formed, and each link can be represented as part of a spatiotemporal hierarchy. Each level of the hierarchy can be summarised as being composed of two parts:

1. **The regulatory state.** This is the nuclear environment that the genome finds itself in a given cell at a given developmental time. The regulatory state is the summation of all the molecular regulatory events occurring in the cell: the signal transduction networks activated; the TFs present in the nucleus and their activity states; the state of chromatin affecting DNA accessibility *et c.*. As such, the regulatory state records the molecular asymmetries inherited from parent cells and integrates these with external signals. By setting up unique regulatory states in space and time, the embryo is able to activate unique repertoires of genes (via their cis-regulatory modules) in appropriate locations to form morphological patterns.
2. **Target *cis*-regulatory modules (CRMs).** The components of the regulatory state determine which genes the cell will express by activating or repressing them; the information encoding these interactions is encoded in the CRMs of target genes. The regulatory state determines which CRMs to interact with and whether these interactions will be activatory or repressive. These target genes could be other regulatory molecules, altering the regulatory state for the next step in the hierarchy. At the termini of the GRN, however, will be genes that perform some downstream purpose. These effector gene batteries (EGBs) are deployed to determine all the necessary aspects of cell behaviour (cell cycle control, cytoskeletal elements, metabolism *et c.*) underlying differentiation and mature cell function.

Pattern formation events are therefore encoded in the developmental GRNs that unfold during embryogenesis; CRMs respond to the regulatory state in order to activate target genes in patterns informative for the embryo. The final form of the embryo is therefore encoded in regulatory genes, their CRMs, and the topology of the resulting network. By gain, loss or alteration of protein-coding or regulatory sequences, the topology of the network, and therefore the form of the embryo, can change over the course of evolution. The concept of GRNs is reviewed by Howard and Davidson[50] and Peter and Davidson[51].

As the majority of CRMs are thought to operate by binding transcription factors, mutations within CRMs could potentially affect TF binding sites (TFBSs) and thus change the input to the element and the output of its target gene. The arrangement of TFBSs in regulatory elements has been mapped in detail in a few model cases, such as that of the *sparkling* enhancer of *Drosophila*[52, 53], but there is a lack of a general understanding of the rules governing the structure of CRMs (often referred to as *cis*-regulatory syntax or grammar), making the interpretation of mutations difficult. These grammars may differ depending on the function of the elements in question or the factors that mediate such functions. Generating grammars for different functional classes of CRMs is thus vital in understanding and interpreting non-coding variation. Indeed, the very existence of such grammars is a subject of debate (discussed below).

There are two main models of CRM function: the “billboard” model, wherein the presence of binding sites is required to recruit TFs in combinatorial manner but their arrangement is irrelevant; and the “enhanceosome” model, wherein the presence of binding sites in a particular arrangement is necessary to ensure the correct assemblage of a protein complex at the enhancer[54]. Data from studies of many individual CRMs from metazoan genomes studied to date are consistent with either the billboard or enhanceosome model or exist somewhere on the spectrum between them, suggesting that both models apply to different CRMs to varying degrees.

Mutations in CRMs are thought to underlie many adaptive phenotypic changes[55-58], consistent with the notion that coding mutations in developmental genes are very likely to be deleterious and associated with high pleiotropy; a typical developmental regulatory protein is involved in numerous sub-networks during development[59, 60]. Examples where cis-regulatory changes have been implicated in generating novel phenotypes include: Alterations to trichome morphology in *Drosophila* caused by substitutions in a *shaven baby* CRM[61, 62]; diversity of sex-linked pigmentation caused by variation in a conserved *bric-a-brac* CRM in drosophilids[63]; several cases from threespine sticklebacks such as pelvic structure loss caused by deletion of a *pitx1* CRM[64] and tooth gain caused by an allele of a *bmp6* CRM[65]; and red feather pigmentation caused by the deletion of a *sox10* CRM in pigeons[66]. Whilst these cases typically focus on a single variant affecting a single trait, others focus upon detecting non-coding quantitative trait loci (QTLs) and polygenic *cis*-regulatory evolution in, for example, drosophilids[67] sticklebacks[56, 68, 69] and mice[70]. Evidence of this sort suggests that multiple, cumulative *cis*-regulatory changes at many genes could affect suites of traits to generate novel morphology and contribute to the evolution of novel structures over time.

The majority of metazoan developmental regulatory genes have a deep phylogenetic origin, demonstrated by the extent of orthology relationships; however, few of these are retained as 1:1 orthologues in most lineages[71, 72]. This suggests that changes underlying the evolution of morphology are caused by alterations to GRNs. This could be achieved by both the acquisition of new genes (TGDs or WGDs); or by modification of, or addition to, the *cis*-regulatory sequences of existing genes. This enshrines post-WGD regulatory evolution as a plausible route for the emergence of vertebrate phylotypic development.

Deleterious mutations affecting the function of CRMs at key developmental genes also contribute to the etiology of human diseases[73-75]. These can be point mutations, small deletions and rearrangements, collectively termed '*cis*-ruptions'[76] or 'enhanceropathies'[77]. Examples include: *sox9* CRM loss causing Pierre-Robin sequence[78, 79]; enhancer gain by *shh* causing limb defects[80, 81]; preaxial polydactyly caused by a deletion in the *shh* ZRS enhancer[82]; a single nucleotide polymorphism (SNP) in a *tbx5* heart enhancer causing congenital heart disease[83]; and rearrangements at the *pax6* locus causing aniridia[84, 85]. Therefore, elucidating *cis*-regulatory grammars also has implications for understanding the mechanisms of human disease.

1.3 Comparative genomics as a means of CRM identification

The large size and low information content of metazoan genomes typically makes large-scale functional validation of non-coding DNA impractical and uninformative. However, regions of regulatory DNA can be predicted in several ways, reducing the amount of sequence that needs to be cloned and validated. Classically, functional regions of the genome have been detected by evolutionary conservation using sequence alignments. Sequence alignments can detect CRMs because retention of non-coding sequences at high identity in divergent organisms indicates constraint of regulatory function[40, 41, 86-90]; any DNA not under selection will rapidly accumulate mutations and degrade in to meaningless sequence.

Sequence conservation can be detected most easily using BLAST[91] or BLAT[92], which use (typically short) query sequences from one organism to search for regions sharing high identity in another genome. The algorithm first finds short regions of complete identity ("words") which act as seeds to compute local alignments. These algorithms must set high identity thresholds when searching whole genomes for short sequences in order to obtain significant alignments. Variations of BLAST such as MegaBLAST[93] and Discontiguous MegaBLAST[94, 95] are more efficient at searching two whole genomes against one another because the queries are concatenated prior to the search. MegaBLAST has been used to identify large sets of sequences shared between two species[96, 97]. Again, these alignments will only detect conservation at high identity and comparing only two genomes might detect sequences sharing identity by chance. Smaller, orthologous regions can be aligned with algorithms such as LAGAN[98]. In these methods, regions known to be orthologous (such as stretches of collinear genes) are aligned. This more targeted approach to alignment can loosen the identity criteria whilst maintaining significance by shrinking the search space. Additionally, LAGAN can optionally anchor alignments to exons, making it ideal for comparing loci containing long stretches of collinear genes. Multiple species can also be aligned simultaneously with multi-LAGAN[98] and shuffle-LAGAN[99], rendering the detection of conservation by chance very unlikely. Depending on the species used, these alignments can be highly sensitive (using species with more recent divergence times) or highly specific (using species with less recent divergence times)[89].

Despite their utility, traditional sequence alignments are likely to fail to identify a proportion of CRMs due to the nature of rigid bioinformatic definitions and evolutionary re-wiring of enhancers[52, 53]; selection operates on function and sequence alignments detect only one aspect of structure. This precludes the identification of all functional enhancers by sequence alignment alone. Methods such as EDGI[100] and the more recent DREiVe[101] aim to detect modules at orthologous locations with conserved sites which have undergone local rearrangement; these algorithms use similar approaches and are optimised for use with *Drosophila* and vertebrate genomes, respectively. These algorithms identify the conservation of clustered motifs rather than just sequence identity, ideal for detecting elements where the defining grammar is thought to be quite loose, consistent with the billboard model. An algorithm called CORECLUST[102] uses the converse approach; beginning with a set of regulatory regions or co-regulated genes, it detects preferred site arrangements (grammar elements) that can be subsequently applied to predict enhancers with similar functions. However, the ability of such methods to detect conserved, or indeed derived, function requires detailed experimental validation in cross-species analyses. Thus, a deeper understanding of the grammars of different functional classes of enhancers will aid the identification of functionally and structurally constrained elements outside of those defined by sequence identity alone.

Other approaches typically detect biochemical events rather than evolutionary signatures. As is discussed below, biochemical assays tend to overestimate the amount of functional (i.e. adaptive) regulatory DNA, raising concerns over the interpretation of these data. Sources of such error of interpretation are discussed in detail in two articles, one by Graur and colleagues[103] and another by Doolittle[104]. Further overestimation arises from defaulting to an adaptationist standpoint and failure to recognise neutral or mechanical explanations for these biochemical phenomena. These ideas are noted in the aforementioned reviews, but are discussed conceptually by Gould and Lewontin[105].

Firstly, experiments that detect the proteins bound to the genome have been used to infer CRMs. One class of methods are referred to as genomic footprinting assays. In these types of assays native chromatin is isolated, DNA and protein are cross-linked and digested by DNase I[106, 107] or MNase[108]. Following this the cross-links are removed and the resulting undigested DNA is sequenced. This creates sequencing reads that stack over regions of the genome bound by all native proteins (footprints). These methods require large numbers of cells because the majority of the DNA is digested away, making them useful for studies of bacterial, yeast or cell line genomes, but less useful for studying development. Conversely, open chromatin can be selectively sequenced by methods such as ATAC-seq[109, 110], which exploits steric hindrance caused by bound proteins to prevent the transposon-mediated insertion of sequencing tags. This process simultaneously fragments the DNA ready for sequencing. ATAC-seq requires less starting material than genomic footprinting so is perhaps more applicable to developmental biology, where raw material is typically more scarce. Neither genomic footprinting nor ATAC-seq can provide information as to which protein is bound; they are all versus all methods and generate no information alluding to the bound factors aside from the size of the footprint. A more targeted approach is ChIP-seq. As with genomic footprinting, this method requires protein and DNA be cross-linked, but then utilises factor-specific antibodies to pull down regions of DNA bound by the target, thereby detecting individual binding events[111-113]. This method detects interactions of a particular factor of interest against the whole genome (one versus all) and thus can be useful for the detection of cell-type or tissue specific CRMs. The difficulty with all of these methods is that, whilst they detect genuine and reproducible biochemical events, they cannot distinguish between those that are biologically meaningful and those that are inconsequential. ChIP-seq for a single factor usually generates a high proportion of false positives when peaks are functionally assayed in embryos but this can be improved by introducing additional peak size and motif criteria[112] or by finding overlap between peaks for many functionally related factors[114].

CRMs may also be predicted by performing ChIP-seq using an antibody for more general factors: common cofactors such as p300[113] or modified histones[4, 115, 116]. These are qualitatively different from ChIP-seq for particular TFs, requiring an additional assumption that some effector TFs are co-binding at these peaks. The histone post-translational modifications h3k4me1, h3k4me3 and h3k27ac[115, 117-119] have been found to associate with a proportion of enhancers and are frequently referred to as "enhancer marks"[120, 121]. These marks tend to associate with open chromatin, and as such regulatory function is ascribed to DNA exhibiting such marks. However, these modifications do not associate with all enhancers at all times. Even if this assertion is true, the argument is non-sequitur in the following form: all enhancers are marked by h3k4me3; sequence A is marked by h3k4me3; therefore, sequence A is an enhancer[103]. Whether histone modifications are a cause or effect of enhancer activity is not clear; they may simply be a mechanical requirement for DNA accessibility. For example, at the mouse beta-globin cluster, deletion of the beta-globin 2 promoter or a characterised enhancer fails to affect the formation of hyperacetylated domains over the beta-globin genes, even though transcription is abolished[122]. This raises doubts as to the nature of such marks and demonstrates the problematic nature of using epigenetic marks to predict CRMs.

Finally, STARR-seq[67, 123] can perform prediction and validation of enhancers simultaneously. In this method, DNA fragments are cloned in to the first intron of a reporter gene within an expression vectors such that they control their own transcription. This library is then transfected in to cultured cells, mRNA is isolated, and finally cDNA is synthesised and sequenced. The resulting reads then stack over the regions that generated them, providing quantitative information on how many transcripts the sequence has generated in the cell type used. This is the method of functional validation with the highest throughput, but is limited to studying particular cell types and thus fails to provide information on the spatiotemporal patterns driven by developmental enhancers. Furthermore the existence of enhancer activity does not mean that this activity is biologically meaningful, even if this has a statistically significant effect on gene expression.

Recent results from the Encyclopedia of DNA elements (ENCODE) project suggests that there is significant overlap of the regulatory regions predicted by conservation, DNAse I hypersensitivity, histone marks or ChIP-seq peaks (up to 70%)[124, 125]. However, the recent criticisms of ENCODE have suggested that their definitions and methods systematically overestimate the amount of functional regulatory DNA, so these results should be interpreted with caution[103, 104]. Evolutionary conservation remains the most reliable way to identify functional non-coding DNA, with the caveat that what that function is requires experimental validation.

Whilst the detection of CRMs can be challenging, assigning a target gene to a CRM can be even more difficult[126]. CRMs can be 5', intronic or 3' of the target gene, and have been found to be located up to 1Mb distal from the promoter[127, 128]. Furthermore, CRMs are often found within introns of neighbouring genes[129, 130], and occasionally exonic sequences have been found to act as CRMs[121, 131, 132]. However, distal or cryptic CRMs appear to regulate developmental genes (TFs, receptors and morphogens) in most instances, particularly where these are highly conserved. This suggests that complex landscapes of CRMs have become fixed around a number of key developmental regulators because they coordinate shared aspects of vertebrate embryogenesis; indeed, this model might account for the maintenance of these conserved elements by purifying selection[133]. Understanding the functions of conserved CRMs is pivotal to enriching our knowledge of development and disease. Furthermore, the elucidation of key lineage-specific innovations in developmental CRMs is needed if we are to understand the effects of these in generating pathologies or adaptive phenotypes.

1.4 Gnathostomes share thousands of ancestral non-coding elements

Many lineage-specific sets of conserved non-coding elements (CNEs) have been identified through comparative genomic studies. These associate with genes involved in the transcriptional regulation of development (trans-dev genes) and are therefore thought to represent indispensible CRMs that play a part in defining the ancestral body plan of the lineage[134-136]. Vertebrate CNEs, and later those shared between vertebrates and invertebrate chordates, have been identified in a series of studies[35, 38-41, 137, 138]. A summary of non-coding sequence conservation amongst chordates is shown in table 1.1.

Although bony fish have been used as a reference for the identification of conserved regulatory elements since the mid 1990s[86, 139, 140], the first genome-wide sets of vertebrate CNEs were not identified until almost a decade later[38, 39]. Woolfe *et al.* performed MegaBLAST of orthologous genomic regions from human and fugu to identify thousands of vertebrate CNEs[40]. These sequences have therefore been conserved since the divergence of ray-finned fish (actinopterygii) and lobe-finned fish (sarcopterygii), estimated at 450 million years ago (MYA). In these studies, CNEs were defined as any nonexonic element longer than 40 bp with greater than 60% identity between mammals and fugu[40, 41]. CNEs appear at loci containing developmental genes almost without exception, and show little or no evidence of transcription[141], suggesting that they act as CRMs of developmentally important genes. 6670 sequences derived from subsequent mammal-fish mLAGAN[98] and sLAGAN[99] alignments of orthologous human, mouse, rat and fugu regions are now stored in the CONDOR database (available at http://condor.nimr.mrc.ac.uk/)[142]. These sequences map to ~0.8Mb of the human genome.

Table 1.1: Non-coding sequence conservation amongst chordates. Table showing a summary of studies searching for conserved non-coding elements (CNEs) in chordate genomes. Thousands of CNEs are detected in alignments of gnathostome (jawed vertebrate) genomes, with fewer detectable in cyclostomes and protochordates. A parallel (but distinct) set of CNEs are detectable in *Ciona*. The total length refers to the amount of sequence mapping to the first species listed in the 'Genomes' column (human or *C. intestinalis*). The associated references are, from top to bottom, Woolfe et al. 2007[142]; Venkatesh et al. 2006[143]; Smith et al. 2013[35]; Sanges et al. 2013[138]; Putnam et al. 2008[137] and Doglio et al. 2013[97].

Genomes	Clade	Algorithm	Criteria	No. CNEs	Total length
Human, Mouse, Rat, Fugu	Osteichthyans (bony vertebrates)	Multi-LAGAN	>60% ID, >40bp	6670	0.8Mb
Human, Chimaera	Gnathostomes (jawed vertebrates)	Discontiguous MegaBLAST	E = <1e-4, word size 16, mismatch penalty -2, >100bp	4782	1Mb
Human, Mouse, Rat, Fugu, Chimaera, Lamprey	Vertebrates	BLASTn	E = <5e-3, word size 5, gap penalty -1	476	38kb
Human, Mouse, Dog, Fugu, Stickleback, Medaka, *C. intestinalis, C. savignyi*	Olfactores (vertebrates and urochordates)	Various	>50% ID, >35bp	183	8kb
Human, Amphioxus	Chordates	Shuffle-LAGAN	>60% ID, >50bp	77	3kb
C. intestinalis, C. savignyi	*Ciona* sp.	MegaBLAST	E = <1e-3, word size 20, mismatch penalty -2, >100bp	2336	0.4Mb

Subsequent searches for conservation of human sequences in the genome of the elephant shark *Callorhinchus milii*, a cartilaginous chimaera, have also revealed thousands of CNEs. In this case the chimaera genome was aligned to the human genome using Discontiguous MegaBLAST. Using this method, the elephant shark has more extensive sequence conservation with human than does zebrafish or fugu[143]. Due to a further WGD in the ray-finned fish lineage[42], CNEs have been evolving rapidly in this lineage, and many may have diverged beyond the limits of what can be recognised by sequence comparison[96]. This could be considered disadvantageous as some ancestral elements may have been lost, but the increased rate of evolution in teleosts may also highlight the most developmentally important sequences that have been retained in all vertebrate lineages. In either case, these comparisons with the elephant shark indicate that many elements predate the divergence of cartilaginous fish (chondrichthyes) and bony fish (osteichthyes), estimated at 550 MYA. This CNE set maps to around 1Mb of the human genome.

Hundreds of these elements, from the mammal-fugu set and the human-chimaera set, are also detectable by BLAST in the genome of the sea lamprey *P. marinus*, a member of the jawless fish (cyclostomata) lineage, the most basal extant vertebrates[35]. Cyclostomata last shared an ancestor with jawed vertebrates (gnathostomata) 600 MYA. Therefore, this subset of vertebrate CNEs arose prior to the divergence of crown-group vertebrates. It remains possible that even more of these elements have orthologues in the lamprey, but they cannot be detected because the lamprey genome is poorly assembled and currently phylogenetically isolated[144], rendering detection of CRMs by conservation difficult; sequence data from other lampreys and hagfish are currently limited. The nonredundant subset of osteichthyan and chondrichthyan CNEs with known orthologues in the lamprey genome map to approximately 38kb of the human genome.

Few of these elements are detectable in protochordate genomes. Only 183 putative orthologous elements are detectable in urochordate genomes (*Ciona intestinalis* and *Ciona savignyi*), despite the fact that this lineage is the sister group to vertebrates. These sequences are on average very short and share low sequence identity with vertebrate CNEs (typically 40-50%), totalling only ~8kb. These were also detected at non-syntenic loci[138], and as such their relationship to vertebrate CNEs is dubious. Perhaps this is due to an increased substitution rate in tunicates[145]and a number of tunicate-specific rearrangement events[146, 147]. Another possibility is that these elements could represent genes with RNA products (as many of these elements show evidence of transcription), and thus could tolerate rearrangement. This suggests that these are qualitatively different from vertebrate CNEs, which rarely tolerate rearrangement. Contrastingly, another study argues that urochordate and vertebrate CNEs arose largely independently, with parallel CNEs associated with orthologous genes sharing no identity and thus distinct origins[97].

There are some CNEs with orthologues in both human and amphioxus, tracing the origin of these elements back to the ancestral chordate. Alignment of the *B. floridae* and human genomes reveals 77 CNEs, though these are typically at much lower sequence identity than amongst vertebrates, indicating that there are very few conserved regulatory elements shared amongst all chordates[137]. These elements only total ~3kb, ~500bp of which is a single element, a *sox21* CNE; this is the most deeply conserved element amongst all animals, detectable even in echinoderms and cnidaria[148]. One other study claims to have found more extensive conservation between vertebrates and amphioxus[149], but these sequences are very short and share little identity to vertebrate sequences. Therefore, their homology to vertebrate CNEs appears doubtful.

In summary, the large repertoire of CNEs shared amongst gnathostomes is largely undetectable in the lamprey and other chordates. Thus, the fixation of most of these CNEs appears to have taken place in the ancestral gnathostome, after this ancestor had undergone 2 WGDs. Elucidation of the functions of these CNEs, and the mechanisms by which they operate could generate insight in to the ancestral gnathostome. This will also contextualise the roles of CNEs in development and evolution. Furthermore, detailed study of molecular events at these sequences will allow the interpretation of the consequences of sequence variation in CNEs. Finally, generating a catalogue of cryptic grammars of binding sites within subsets of functionally similar elements makes the prediction of tissue-specific gene expression profiles a tangible possibility, and will furnish our understanding of the gene-regulatory complexity of the ancestral vertebrate.

1.5 Functional analysis of vertebrate conserved non-coding elements

Much understanding of the functions of CNEs comes from data from *in-vivo* functional assays. For the most part, these test for enhancer function using a reporter gene, as gain of function assays are simple to perform and the results are easy to interpret. One commonly used method is *Tol2* mediated transgenesis, where a putative regulatory sequence is placed upstream of *egfp* and a minimal promoter flanked by recognition sequences for the *Tol2* transposon. This construct is then co-injected with *Tol2* transposase mRNA in to zebrafish embryos where the transposase catalyses the insertion of the constructs in to the genome[150, 151]. This method has also been used, to a lesser extent, to generate transgenic mice, chicks and frogs[151]. Other methods include comparable assays in mice using a lacZ reporter[87, 88] (the results from such assays form the data stored in the VISTA enhancer browser[152]) or assays in various cell lines[67, 123, 153, 154].

Assays in zebrafish have proven tractable and informative, as these are much more rapid than comparable assays in mice[155], whilst having advantages over cell-based assays as one can screen multiple time-points, tissues and cell types. Furthermore, efficient *Tol2*-mediated transposition means that the generation and maintenance of stable transgenic lines is generally unnecessary, saving time[151]. Enhancers isolated from a variety of vertebrates have been tested and found to generate identical expression patterns in zebrafish and their host organisms, including mouse and lamprey sequences[156, 157], even when these sequences are not highly conserved[158, 159]. This indicates that zebrafish might provide an informative readout of expression from enhancers belonging to any vertebrate. Since the *Tol2* transposon incorporates randomly, there is the potential for reporter constructs to receive input from other regulatory elements or the chromatin state at the insertion site, collectively termed ‘positional effects’. This creates a certain amount of variability between embryos. By injecting many embryos, shared aspects of the expression pattern can be revealed, highlighting the regulatory ability of the element under scrutiny regardless of its genomic location. This is preferable to the consistent bias introduced by targeted approaches.

Nevertheless, there are some problems with these sorts of assays. First, this method of validation is still fairly low-throughput. Second, EGFP must accumulate to a certain concentration in a cell before it becomes visible, which might lead to false negatives for very weak enhancers. Finally, EGFP has a half-life in excess of 24 hours, making the inference of precise timings of expression difficult; however, this might allow some weak enhancers to generate visible expression over time. Greater accuracy can be achieved using *in situ* hybridisation (ISH) for *egfp* mRNA on either transient or stable transgenic embryos.

Recent studies also suggest that there may be some disparity in reporter gene expression when testing orthologous CNEs in cross-species comparisons, such as mouse and zebrafish[160, 161]. Firstly, these cases appear to be far more infrequent than cases where the expression patterns driven by orthologous elements agree. Secondly, disparity between orthologous enhancers could highlight rather than mask important functions by tracking lineage-specific changes to activity[162]. This has been observed in the case of duplicated CNEs in teleost genomes, which often display divergent functions[163, 164], highlighting their potential roles in the regulatory subfunctionalisation of duplicated developmental genes[165]. Therefore, these sorts of comparative studies might provide valuable information about both *cis* and *trans* regulatory mutations underlying the evolution of new gene regulatory networks underlying development.

Around half of all CNEs tested in such assays, either in mice or zebrafish, act as developmental enhancers[40, 41, 88]. Whilst this is a key technique for the visualisation of expression patterns, the mechanisms of action of most characterised enhancers is unknown. Where this has been tested, these elements have been shown to be required for normal development[166-168]. Although these data are encouraging, only a tiny fraction of CNEs have had their roles during development properly elucidated, preventing a thorough meta-analysis.

The lines of evidence in support of the model that vertebrate CNEs control phylotypic development are largely indirect. Since the most common form of data available are expression data, this is often used as a proxy for developmental function. Whilst extensive, expression data stored in online databases are far from comprehensive. It has however been demonstrated that CNEs are statistically enriched for brain enhancers over enhancers of other tissues[169]. Amphioxus, for the most part, lacks detectable orthologues of these sequences. During early vertebrate evolution, the acquisition of such enhancers might have co-opted duplicated genes to novel GRNs underlying morphogenesis of the head and brain[170]. This conjecture, whilst supported by the available evidence, requires further scrutiny.

CNEs are thought to be arrays of TF binding sites (TFBSs)[141]; however, the underlying sequence level grammar of such sites is in most cases not known; in fact, the very existence of such grammars within CRMs is a matter of contention[53, 102, 171, 172]. Despite this, several studies have attempted to search for grammatical signatures within regulatory elements. This is one plausible approach to elucidate generic mechanisms of action for different functional sets of tissue-specific enhancers.

1.6 Prediction of tissue-specific enhancers in metazoan genomes

The prediction of tissue-specific CRMs from sequence and biochemical data has become an emerging area of interest over the past decade or so[155, 173]. Many studies have attempted to de-code the information content of sets of tissue-specific CRMs and/or predict their tissue specificity using combinatorial approaches. Typically, these studies take an initial set of known tissue-specific enhancers, attempt to discover and model grammatical signatures within them, and then use these signatures to predict functionally similar enhancers before subjecting them to validation. These models of grammatical signatures typically arise from two or more of the following: i. evolutionary conservation data; ii. presence of TFBS motifs; iii. chromatin state/accessibility data; iv. ChIP-seq data; v. nucleotide composition and/or k-mer frequencies.

Two early studies identified certain sequence features associated with tissue specificity, but after initial successes did not validate their models by testing additional predictions. Davies *et al.* predicted four cartilage regulatory elements by searching for enriched motifs within loci containing cartilage-expressed genes, and searched for clusters of these motifs to predict likely regulatory elements. Thereafter, their predicted motifs were mutated and shown to be required for activity during luciferase assays in cartilage cells[153]. Rastegar *et al.* identified a pair of motifs within 5 conserved notochord enhancers that were both necessary and sufficient for notochord expression in transgenic zebrafish. However these two motifs were variably arranged, suggesting a lack of any definite grammar[174].

In a later study, Papatsenko *et al.* studied a heterogeneous set of nearly 100 previously characterised *Drosophila* enhancers in great depth, and contrastingly found numerous patterns in their organisation. These included helical or anti-helical phasing of paired binding sites and separation of binding sites by nucleosome positioning sequences. This study demonstrated that even enhancers that drive very different tissue-specific expression patterns might possess common sequence-level grammar elements[171], shedding doubt on the association between common grammar elements and tissue specificity. Despite these interesting results, none of the aforementioned studies attempted to identify additional enhancers using the generated models.

Later, studies began to use their models to predict tissue-specific enhancers. Kantorovitz *et al.* began with small training sets of validated CRMs, and utilised a motif-blind statistical method to search for elements with similar nucleotide compositions and k-mer content. Only 7 enhancers were functionally tested (5 from *Drosophila* and 2 from mouse), but all were predicted correctly, demonstrating this as an effective approach across bilateria[175]. Narlikar *et al.* generated a sequence-level model of heart enhancer activity based on the composition of 50 human heart enhancers, and successfully validated 16/26 (62%) in transgenic assays using mouse embryos[176].

Motif-based approaches, whilst simple, have been successful in a number of cases. Haeussler *et al.* detected enrichment for Otx binding site motifs in anterior neurectodermal enhancers in *C. intestinalis*, and used the presence of this motif to predict similar enhancers, successfully validating 10/23 (43%) in *C. intestinalis* embryos[177]. Mongin et al. used both mammal-fish evolutionary conservation and the presence of motifs resembling TFBSs, described by position-weight matrices (PWMs) from TRANSFAC[178] and JASPAR[179], to predict CRMs for subsequent functional assays. These elements were shown to act as enhancers in the nervous system in 95% of cases[180]. It should be noted that in this case no attempt was made to predict neuronal enhancers over other tissue-specific elements, suggesting that this work has simply identified a general enrichment for neuronal enhancers within mammal-fish CNEs. Kwon *et al.* attempted to predict muscle-specific elements in the human genome using PWMs for TFs with known roles during myogenesis. Only 6% of their predictions proved positive in functional assays, leading this group to suggest a combinatorial approach for future studies, incorporating sequence conservation, nucleotide composition and chromatin state with functional sets of PWMs[154]. Parker *et al.* detected conserved Pbx-Hox motifs in 4 hindbrain enhancers associating with the vertebrate *meis2* gene, and used the presence of conserved Pbx-Hox motifs to predict further hindbrain enhancers of other genes, validating 12/21 (57%) in functional assays in zebrafish[157].

In a more recent study, Burzynski *et al.* used a machine-learning approach to generate a sequence-level model of hindbrain enhancer activity from a training set of 211 validated enhancers, mostly from the VISTA Enhancer Browser[152]. This model was then applied to predict potential hindbrain enhancers in the human genome. They predicted over 40,000 hindbrain enhancers. A sample of these were tested first transiently, and then in stable transgenic zebrafish. 30/34 (88%) elements in stable transgenics enhanced expression in patterns including the hindbrain. However, those tested were the most high-scoring predictions, rather than a sample from across the full breadth of scores. The classifier was also incapable of distinguishing between more specific patterns of expression i.e. anterior versus posterior hindbrain, perhaps reflecting the heterogeneity of the training set[120].

An alternative to using sequence-level models is to use ChIP-seq, advantageous because this detects actual biochemical events rather than relying on the occurrence of short motifs resembling TF binding preferences. However, whether a reproducible binding event is biologically meaningful is difficult to assess, often causing inaccuracy of predictions. Zinzen *et al.* used ChIP-seq for five *Drosophila* mesodermal TFs to predict mesoderm enhancers, and found that elements where these factors were co-bound drove predictable expression in over 70% of cases[114]. Wilczynski *et al.* combined genome-wide data on TF occupancy and chromatin state to attempt to predict the tissue-specificity of *Drosophila* enhancers, with a 50% success rate[118]. More recently, Visel *et al.* performed chip-seq for p300 on mouse forebrain (fb) tissue to predict ~4,500 fb enhancers, and functionally validated 105/329 (32%) of these. They then developed motif models of fb subregions; three distinct modes of activity were shown to associate with their own sets of motifs, but these models were not used to predict additional tissue-specific elements[113].

Despite their successes, the aforementioned studies frequently validate only a small fraction of the predicted elements, validate only the highest scoring predictions of their models and/or lack specificity (they produce a high proportion of false positives). The ability to generate more effective models with greater predictive power remains a difficult challenge in the post-genomic era. This must be overcome if we are to decrypt the information content of the regulatory portion of animal genomes and construct artificial tissue-specific enhancers. This is a vital step in understanding vertebrate development, evolution, and human disease.

Databases such as TRANSFAC[178], Uniprobe[181], JASPAR[179] and the ENCODE project's factorbook[182] contain thousands of experimentally validated TF binding preferences. These preferences are key to understanding the information content of CRMs, and bring us closer to understanding the eukaryotic regulatory code[183]. Given these expansive databases, prediction of tissue-specific enhancers appears more plausible than ever. Utilising these in combination with results from functional assays will allow us to build motif models describing the tissue specificity of enhancers. Subsequently, these can be used to predict enhancers with similar functions, including those that cannot be detected by sequence identity alone. However, it should be noted that there are a large proportion of DNA binding proteins in vertebrate genomes that remain to have their binding preferences and regulatory functions elucidated[107].

1.7 Vertebrate CNEs in the hindbrain gene regulatory network

The hindbrain is a vertebrate shared-derived structure composed of 7 to 8 segments known as rhombomeres (r1-8). Many studies have elucidated a number of genes and regulatory elements that form a GRN underlying hindbrain segmentation which is regarded as, for the most part, conserved amongst all vertebrates[184] (reviewed by Tumpel *et al.*[185] and Phillipidou and Dasen[186]). Normal hindbrain development is dependent upon Hox proteins of paralogous groups 1-4, which are expressed in nested patterns in the hindbrain posterior of the r1/r2 boundary. The Hox proteins then determine the expression of downstream, segment-specific regulatory genes. The Hox proteins lie upstream of both morphological segmentation and the identities of branchiomotor and reticulospinal neurons arising from each segment. The proper interpretation of Hox expression patterns requires the Hox cofactors Pbx and Meis/Pknox, several of which are expressed throughout the hindbrain. Loss of function of various Hox, Pbx, or Meis genes leads to failure to form boundaries between, and homeotic transformations of, rhombomeres[187-190], demonstrating the essential roles of these factors in both hindbrain segmentation and patterning.

Pbx, Hox and Meis/Pknox factors form trimers at enhancers in order to activate transcription of their target genes. Pbx-Hox heterodimers bind to inseparable half-sites with the consensus TGATNNAT[191], whereas Meis/Pknox proteins bind the consensus CTGTCA at proximal, but potentially gapped and/or inverted, sites[192]. Several segment-specific enhancers containing these sites have been identified, and using mutagenesis it has been shown that loss of either site abrogates expression by the enhancer[188, 191-196]. The inability for Hox/Pbx/Meis/Pknox null mutants to form appropriate morphology therefore appears to be caused by a failure to activate appropriate enhancers and, thereafter, the regulatory network downstream of Hox proteins.

The ability for Hox proteins to dimerise with Pbx is probably a bilaterian novelty[197], suggestive of a scenario where different Pbx-Hox target enhancers specify unique structures in different bilaterian lineages and, by activating unique downstream GRNs, lead to the formation of diverse morphology in the Hox-dependent segments. With regard to gnathostomes, loss of function phenotypes demonstrate that Hox proteins and their cofactors activate unique repertoires of downstream genes in each segment, subsequently giving rise to the individual identities of rhombomeres and their derivatives. This necessitates the existence of segment specific genes and regulatory elements capable of responding to particular Hox proteins in order to coordinate their expression.

The presumptive hindbrain is first determined by the expression of the 4 most anterior *hox* paralogous groups, which are positioned by opposing gradients of retinoic acid (RA) from the posterior and FGF from the anterior. These signals are integrated through the action of *cdx* genes that repress hindbrain fates[198, 199]. The anterior *hox* code subsequently activates a downstream network of regulatory proteins, many of which are segmentally restricted. For example, the transcription factors *egr2* (*krox20*) in r3/r5[196, 200, 201], *vhnf1* followed by *mafb* (*kriesler*/*valentino*) in r5r6, and *znf703/znf503* (*nlz1/nlz2*) anterior of the r4/r5 boundary[202, 203] all respond to Hox proteins and accordingly fail to activate under the loss of Hox function. The members of this downstream network subsequently interact with one another and provide feedback to the *hox* code to define a series of sharp expression boundaries[185]. This furnishes the hindbrain with a series of genetic segments, no two of which express the same set of transcription factors (i.e. they possess different regulatory states). The network also activates segment-specific patterns of *eph* and *ephrin* genes, which go on to mediate morphological segmentation[204]. As such, the *hox* code ultimately determines the regulatory states of each rhombomere and the subsequent activation of two EGBs in each segment: one mediating like-with-like cell sorting in response to *eph/ephrin* signalling (the segmentation EGBs); and another determining the function and morphology of cells within and neurons arising from each segment (the identity EGBs). This network is summarised in figure 1.2.

Effectors of HoxB1 have been identified in mouse ES cell-derived neurons using microarrays[205], but cell lines lack their usual developmental context. ChIP-seq for HoxA2 in the developing second branchial arch of the mouse identified numerous putative targets[206], but the lack of comparative data from other segments prevents the identification of genes specifying segment identity. Rhombomere-specific mRNAs have also been determined in the mouse using microarrays[207], but this does not distinguish between direct and indirect Hox targets. Relatively few direct and segment-specific Hox targets have been unequivocally identified. For others, putative network interactions have yet to be validated in perturbation experiments and the CRMs controlling their segment-specific expression remain to be identified. This hinders the construction of a full gene regulatory network downstream of Hox proteins for different model organisms.

The vertebrate hindbrain plays a role in the integration of sensory information[208] as well as the motor control of respiration[209], locomotion[210] and swallowing[211]. These abilities are essential for the switch from passive to active feeding that characterized early vertebrate evolution, enabling the detection, capture and consumption of prey[21]. Whilst it is commonly regarded that the gnathostome hindbrain GRN is largely conserved (figure 1.2, figure 1.3 A), the conservation of this network in jawless vertebrates is less clear; there are both similarities and differences and not all of the orthologous genes have been found or studied (figure 1.3 B). Furthermore, the region which best corresponds to the vertebrate hindbrain in invertebrate chordates is not segmented; *hox* 1-4 patterns in this region do not have sharp boundaries and do not overlap expression of genes such as *egr2* (*krox20*) and *mafb* (*kreisler*), which act as key segmentation genes in vertebrates (figure 1.3 C). This raises questions about how the GRNs controlling hindbrain segmentation and, subsequently, rhombomere specialisation arose in early vertebrates[6, 212, 213]. Therefore, the identification of conserved Hox targets involved in hindbrain development has implications for understanding the evolution of the vertebrate brain.

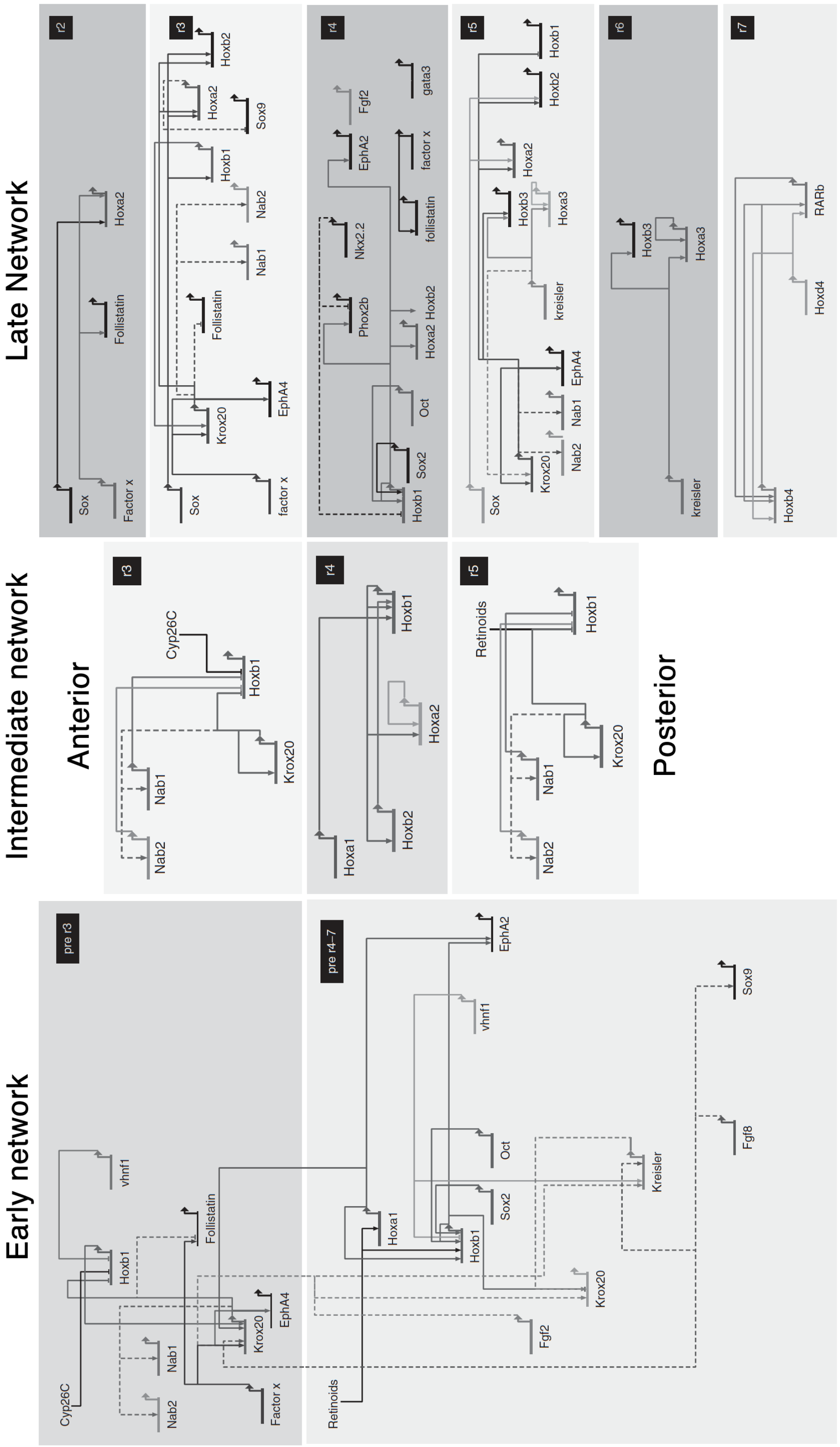


Figure 1.2: The gnathostome hindbrain gene regulatory network. The figure shows direct (solid lines) and indirect/unknown (dotted lines) interactions between genes in the developing hindbrain. The diagram was created using biotapestry[214]. Patterns established at three time-points are shown: early, intermediate and late (equivalent to 8.0, 8.75, and 9.5 dpc in the mouse and 90% epiboly, 3s and 10s in the zebrafish). (Presumptive) rhombomeres are shown as distinct boxes. This figure was adapted from Tümpel *et al.* 2009[185]).

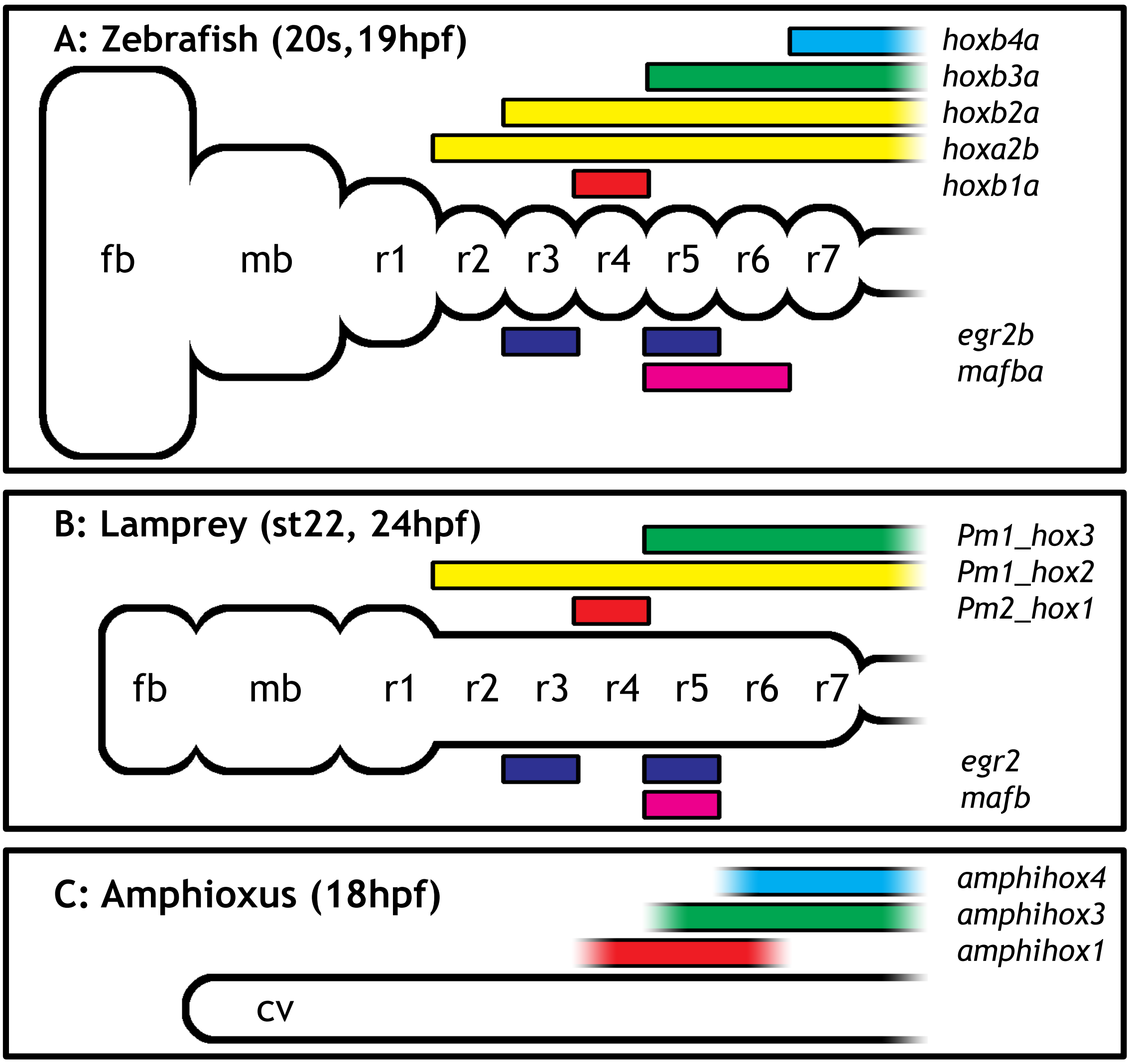


Figure 1.3: Expression patterns of gnathostome hindbrain segmentation genes and their orthologues in cyclostomes and cephalochordates. The figure shows schematic representations of adult brains of a gnathostome (A: zebrafish) and a cyclostome (B: lamprey), and the anterior nerve cord of a cephalochordate (C: amphioxus) with the anterior to the left. These are annotated with expression patterns of genes known to act upstream of hindbrain segmentation and patterning in gnathostomes, at the time points noted (19hpf for zebrafish, 24hpf for lamprey and 18hpf for amphioxus). The segments r2-r7 are determined by *hox* genes of paralogous groups 1 (red), 2 (yellow), 3 (green) and 4 (blue). Gnathostomes exhibit sharp *hox* expression boundaries coincident with segment interfaces, as well as expression of segment-specific transcription factors: *egr2* in r3r5 and *mafb* in r5/6[215] (A). Cyclostomes possess segmental expression of *hox* genes, *egr2* and *mafb*, but their hindbrain is not overtly segmented. Furthermore, their *mafb* gene is restricted to r5 only, unlike r5r6 in gnathostomes[184] (B). Amphioxus lacks both sharp boundaries of *hox* expression and morphological segmentation in this region, and there are no *egr* or *maf* genes expressed in this region[216] (C). fb: forebrain; mb: midbrain; r1-r7: rhombomeres 1-7; sc: spinal cord; cv: cerebral vesicle.

Parker *et al.* (2011) demonstrated that CNEs are enriched for Pbx-Hox binding site motifs (TGATNNATKR) and that CNEs containing this motif preferentially upregulate expression in the hindbrain. A total of 23 CNEs were subjected to an enhancer assay, and 15 (65%) enhance reproducible expression in the hindbrain, pharyngeal arches (PAs) or both. These findings strongly implicate several CNEs in patterning the hindbrain and pharyngeal region, and suggest a direct mechanism linking CNEs to the elaboration of the vertebrate hindbrain during evolution. A model was proposed whereby many CNEs specify Hox-dependant regulatory interactions underlying the ancestral morphogenesis of the hindbrain, encoded by their Pbx-Hox motifs. Finally, it was postulated that a large number of hindbrain enhancers might be predicted by virtue of their containing the Pbx-Hox motif and this might allow the identification of immediate downstream targets of Hox proteins[157]. Subsequently, Burzynski et al. used Meis/Pknox motifs as part of a classifier predicting hindbrain regulatory elements[120], though it was unclear to what extent these motifs contributed to the accuracy of predictions. Taken together, these results suggest that binding motifs for Hox proteins and their cofactors can be used to predict hindbrain enhancers.

Based on the evidence collected thus far, there is a correlation between the number of conserved enhancers and the similarity of the genetic network underlying hindbrain development amongst vertebrates (and lack thereof in chordates). There is conserved hindbrain morphology amongst gnathastomes, alongside thousands of CNEs; the lamprey has more divergent hindbrain morphology, lacking the overt segmentation easily visible in gnathostome hindbrains, with fewer CNEs identifiable; and amphioxus lacks any segment boundaries and possesses very few CNEs. Validating the functions of CNEs from each lineage could indirectly investigate the hypothesis that CNEs play a role in determining downstream morphology.

How common this mechanism of action is in vertebrate CNEs has not been thoroughly investigated, with only 15 such CNE hindbrain enhancers characterised in published literature to date. Furthermore, the role of CNEs during the evolution of vertebrate characteristics has been suggested[157] but many elements remain to have their evolutionary significance evaluated. Nevertheless, that vertebrate CNEs are enriched for Pbx-Hox binding motifs suggests that many may specify key regulatory interactions in the developing vertebrate hindbrain, and direct an ancient GRN for hindbrain and pharyngeal development. Searching for incidences of these binding motifs in putative enhancers might identify additional transcriptional targets of Hox proteins and further embellish our knowledge of the downstream GRN involved in patterning the gnathostome hindbrain.

1.8 Overall aims and hypotheses

Convergence of findings from several disciplines has begun to build a picture of early vertebrate cephalisation[14]. Research is being carried out on many chordate models, several of which now have whole-genome assemblies[35, 137, 217, 218]. Refining our understanding of these pivotal evolutionary events depends upon understanding the genetic basis for the emergence of vertebrates. This requires the elucidation of the developmental mechanisms that contributed to the emergence of the head as a novel vertebrate unit.

Previous studies have used binding sites for Hox proteins and/or their cofactors to predict hindbrain enhancer activity with some success[120, 157]. Concomitantly, the connection between vertebrate CNEs and hindbrain development had been suggested[157]. In light of these data and hypotheses, this study aims to:

- Identify additional hindbrain enhancers from a set of CNEs stored in CONDOR[142] using a functional assay in zebrafish embryos[150];
- Develop a bioinformatic pipeline that can distinguish hindbrain enhancers from hindbrain negative sequences with high accuracy;
- Further dissect the function of CNE hindbrain enhancers experimentally using site-directed mutagenesis;
- Place a larger number of CNEs in to their developmental and evolutionary context.

Using these approaches, this study will test the following hypotheses:

- That lineage-specific CNEs control phylotypic aspects of the group concerned[135, 219] (generally) and that vertebrate CNEs control phylotypic aspects of vertebrate development (specifically), particularly hindbrain development[157];
- That hindbrain enhancers share sequence features (grammar) that can be identified using bioinformatic methods;
- That these shared sequence features can be applied to identify additional hindbrain enhancers.
- That Pbx-Hox motifs contributed to the evolution of novel hindbrain enhancers on the vertebrate and gnathostome stems.

CHAPTER 2

Conserved Pbx-Hox motifs are poor predictors of hindbrain enhancers

2.1 BACKGROUND

Previous work used Pbx-Hox motifs to predict sequences capable of acting as hindbrain or pharyngeal arch enhancers in a reporter assay in zebrafish embryos. This led the authors to posit that the presence of Pbx-Hox motifs within CNEs could indicate an encoded ability to generate segment-specific expression patterns in the hindbrain. The authors hypothesised that conserved Pbx-Hox motifs might identify novel Hox target genes[157]. In order to create a large set of experimentally validated hindbrain enhancers, upon which a sequence model of hindbrain enhancer activity could be based, it was decided that a variety of CNEs containing conserved Pbx-Hox motifs (henceforth Pbx-Hox CNEs) should be assayed.

An identical enhancer assay[150] to the previous study was used. Similarly, only CNEs containing Pbx-Hox motifs conserved between human and zebrafish were selected. However, lamprey sequences were not used as a reference in this case because of the relatively few CNEs detected in the lamprey genome[35], which limits the number of available elements. CNEs were selected to represent a range of gene loci. Only CNEs containing a single Pbx-Hox motif were selected to facilitate future mutagenesis experiments.

2.2 AIMS AND HYPOTHESES

- **Aim:** to identify novel hindbrain enhancers by virtue of their containing Pbx-Hox motifs, in order to expand the set of experimentally validated hindbrain enhancers.
- **Hypothesis:** Pbx-Hox CNEs are sufficient to activate reporter gene expression in the hindbrain in the context of the functional assay.
- **Hypothesis:** Pbx-Hox CNEs correspond to previously uncharacterised hindbrain enhancers and, by spatial association, can identify putative Hox target genes.

2.3 METHODS

2.3.1 Identification of candidate CNEs

Previously, a list of Pbx-Hox motifs (TGATNNAT) within human CONDOR CNEs[142] and a table describing the presence or absence of the motif in orthologous CNEs was generated[157]. In order to ensure that zebrafish orthologous of the CNE also contained the relevant motif, the list was filtered to include only CNEs containing Pbx-Hox motifs conserved between human and zebrafish. There are 465 instances of the motif distributed amongst 394 CNEs. 29 CNEs were selected for functional characterisation. The candidates were chosen from a variety of loci and contained a single conserved instance of the Pbx-Hox motif. In each case, 16-24 orthologues of the CNE from different vertebrates were downloaded from CONDOR (http://condor.nimr.mrc.ac.uk), trimmed to prevent unaligned sequence ends and aligned using ClustalW2[220] (http://www.ebi.ac.uk/Tools/msa/clustalw2/) using default settings to confirm the conservation of the Pbx-Hox motif. CONDOR uses some genome assemblies that are now out of date; therefore, in some instances alignments were augmented with BLAST hits against the up-to-date genome assemblies from Ensembl[221] using CNEs from the most closely related species as queries (e.g. the fugu sequence was used as a query against zebrafish).

2.3.2 Cloning of candidate CNEs

Zebrafish genomic DNA was prepared from a single adult female zebrafish from the NIMR aquatics facility using the ISOLATE genomic DNA kit (Bioline) according to the manufacturer's guidelines. Oligonucleotide primers targeting the zebrafish orthologue of the candidate CNEs were designed using primer3[222] and synthesised by Sigma-Aldrich. Polymerase chain reactions (PCRs) were set up using element specific primers, zebrafish genomic DNA as a template and BIOTAQ™ *taq* polymerase and buffers (Bioline) according to the manufacturer's guidelines. PCR products were visualised by agarose gel electrophoresis to confirm product size and purity. PCR products were column purified using the GFX PCR DNA and Gel Band Purification kit (GE Healthcare) according to the manufacturer's guidelines to remove excess primers and/or undesired products. PCR products were eluted from the column in double-distilled water.

Purified products were cloned in to the pCR™8/GW/TOPO™ vector (Invitrogen) using the manufacturer's guidelines and were transformed in to Oneshot TOP10 chemically competent cells (Invitrogen) according to the manufacturers guidelines. Outgrown cultures were spread on spectinomycin agar plates and grown at 37°C overnight. Tubes of 3ml lysogeny broth (LB) plus spectinomycin were inoculated with freshly picked colonies and incubated at 37°C overnight with agitation. Plasmids were obtained from 2ml of culture using the QIAprep Spin Miniprep Kit (Qiagen) according to the manufacturer's guidelines. DNA was eluted from the column in double-distilled water.

Entry clone preparations then underwent Gateway® LR recombination (Invitrogen) with the pGW_*tol2:cfos*:*egfp* vector according to the manufacturer's guidelines and were transformed in to Oneshot TOP10 chemically competent cells (Invitrogen) according to the manufacturers guidelines. Outgrown cultures were spread on ampicillin agar plates and grown at 37°C overnight. Tubes of 3ml lysogeny broth (LB) plus ampicillin were inoculated with freshly picked colonies and incubated at 37°C overnight with agitation. Plasmids were obtained from 2ml of culture using the QIAprep Spin Miniprep Kit (qiagen) according to the manufacturer's guidelines and eluted from the column with double-distilled water. Inserts were confirmed by sanger sequencing (Source Bioscience).

2.3.3 Enhancer assay in zebrafish embryos

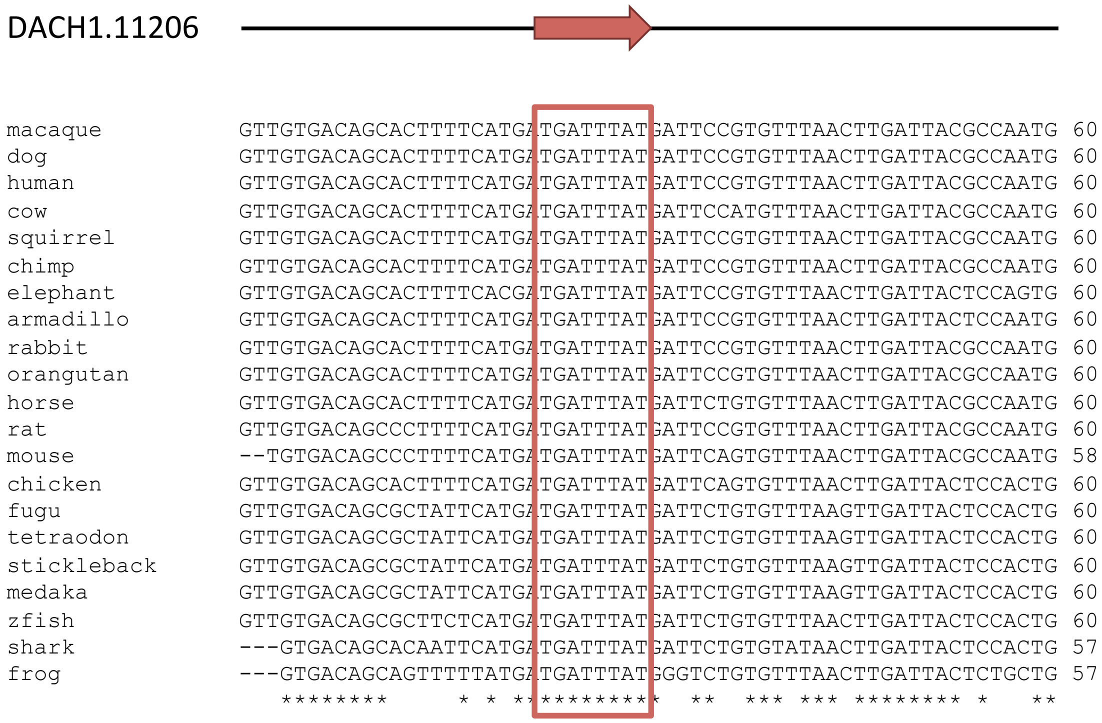
The principal of the zebrafish enhancer assay has been described previously[150]. After cloning, CNEs are placed upstream of the mouse *cfos* promoter and the enhanced green fluorescent protein (*egfp*) open reading frame (ORF), flanked by *Tol2* transposase recognition sequences. As such the resulting insertion can be written as Tg(*cne-cfos*:*egfp*). This cassette is henceforth referred to as the 'expression construct'. *Tol2* transposase mRNA was transcribed *in vitro* from a linearised pCS-Tp vector containing the *Tol2* transposase ORF using the mMESSAGE mMACHINE SP6 kit (Invitrogen) according to the manufacturers guidelines. 5µl microinjection mix was prepared using 175ng *Tol2* mRNA, 150ng plasmid, and 0.1% phenol red plus salts (tracer). The mix was injected in to wild-type or Tg(*egr2b*:*kalta4*;*uas*:*mCherry*) zebrafish[223, 224], which express mCherry in rhombomeres 3 and 5 (r3r5), using an Eppendorf Picospritzer microinjector. Embryos were stored in zebrafish embryo medium and incubated at 28°C. Embryos were screened for EGFP expression at three time points: 24-30h, 48-54h and 72-78h using fluorescence microscopy. Embryos were considered to be hindbrain positive if there were any GFP positive cells in r2-7. Elements were considered to act as hindbrain enhancers if at least 20% of embryos were hindbrain positive at one or more of the time points. At least 30 embryos were screened in each case. Transient transgenics generated from wild-type embryos were photographed in greyscale twice, once in bright-field with no filter and once in dark-field with a GFP filter. The dark-field EGFP expression pattern was artificially coloured to the green channel. The bright-field and dark-field images were then overlaid to generate the final image. Transient transgenics generated from *krox20*:RFP embryos were imaged in darkfield in greyscale twice, once using a GFP filter and once using an RFP filter. The GFP pattern was artificially coloured to green and the RFP image was artificially coloured to red before overlaying to produce the final image.

2.4 RESULTS

2.4.1 CNEs containing Pbx-Hox motifs are not sufficient for hindbrain activity

Elements containing Pbx-Hox motifs were cloned and assayed with the aim of identifying novel hindbrain enhancers. 29 zebrafish sequences were selected and the presence of conserved Pbx-Hox motifs was confirmed using ClustalW2 alignments (an example is shown in figure 2.1). These were cloned and subsequently assayed in wild-type or KROX20:RFP zebrafish embryos (see methods). 15/29 (52%) of these acted as enhancers of various tissues according to our criteria. However, just 7/29 (24%) of these were considered to act as hindbrain enhancers (figure 2.2). Other enhancers were active in trunk muscle, heart or spinal cord. A proportion of embryos injected with *tol2:cne:cfos:egfp* reporter constructs express in muscle, seemingly independent of the enhancer (Elgar lab, unpublished observations). This is presumably caused by promoter bias. Two enhancers (*foxd3*.365 and *cst*.9931) drive expression frequently in cranial ganglia and rohon-beard cells (mechanosensory neurons in the spinal cord); this pattern was also driven by a 24mer containing a Pbx-Hox site when using the same promoter[97]. It therefore seems that many elements can drive this expression pattern; it could also be caused by promoter bias. Contrary to this, there has been no noted predisposition for CNEs to generate hindbrain expression, except for when these come from loci containing hindbrain genes. Therefore, there is no reason to doubt that this hindbrain activity is reflective of the *in vivo* function of these CNEs. The tissue specificities of all 29 elements can be found in table 2.1. Raw expression data in the form of embryo counts can be found in the appendix (8.2.2).

Four CNEs are broadly active in r2-7 (*bnc2*.8642, *dachd*.11206, *foxd3*.365, *hmx2*.9713), suggesting that they are activated by multiple Hox proteins or by Hox proteins which are active in broad domains the hindbrain such as Hoxa2. In a continuation of the trend seen in the previous study[157], the remaining three hindbrain enhancers activated patterns of reporter expression restricted to certain regions of the hindbrain (hoxd.10479 in ventral r5-6, hoxd.10482 in ventrolateral r4 and r6 and foxd3.327 in ventral r5-6), suggesting that these elements are activated by particular Hox proteins and/or have their boundaries delimited due to regulation by other factors. Those associating with the *hoxd* cluster may be initiator (or autoregulatory) elements responsible for the establishment (or maintenance) of *hoxd* expression patterns in the hindbrain. Indeed, hoxd.10479 has a pattern of expression in ventrolateral r5 and r6, evocative of part of the characteristic pattern of zebrafish hoxd3 in ventrolateral cells in r5[215]. Elements from other loci could act as Hox-dependant regulatory elements of their nearby genes; this highlights putative *hox* targets.



```
macaque        GTTGTGACAGCACTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACGCCAATG 60
dog            GTTGTGACAGCACTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACGCCAATG 60
human          GTTGTGACAGCACTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACGCCAATG 60
cow            GTTGTGACAGCACTTTTCATGATGATTTATGATTCCATGTTTAACTTGATTACGCCAATG 60
squirrel       GTTGTGACAGCACTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACGCCAATG 60
chimp          GTTGTGACAGCACTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACGCCAATG 60
elephant       GTTGTGACAGCACTTTTCACGATGATTTATGATTCCGTGTTTAACTTGATTACTCCAGTG 60
armadillo      GTTGTGACAGCACTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACTCCAATG 60
rabbit         GTTGTGACAGCACTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACGCCAATG 60
orangutan      GTTGTGACAGCACTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACGCCAATG 60
horse          GTTGTGACAGCACTTTTCATGATGATTTATGATTCTGTGTTTAACTTGATTACGCCAATG 60
rat            GTTGTGACAGCCCTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACGCCAATG 60
mouse          --TGTGACAGCCCTTTTCATGATGATTTATGATTCAGTGTTTAACTTGATTACGCCAATG 58
chicken        GTTGTGACAGCACTTTTCATGATGATTTATGATTCAGTGTTTAACTTGATTACTCCACTG 60
fugu           GTTGTGACAGCGCTATTCATGATGATTTATGATTCTGTGTTTAAGTTGATTACTCCACTG 60
tetraodon      GTTGTGACAGCGCTATTCATGATGATTTATGATTCTGTGTTTAAGTTGATTACTCCACTG 60
stickleback    GTTGTGACAGCGCTATTCATGATGATTTATGATTCTGTGTTTAAGTTGATTACTCCACTG 60
medaka         GTTGTGACAGCGCTATTCATGATGATTTATGATTCTGTGTTTAAGTTGATTACTCCACTG 60
zfish          GTTGTGACAGCGCTTCTCATGATGATTTATGATTCTGTGTTTAACTTGATTACTCCACTG 60
shark          ---GTGACAGCACAATTCATGATGATTTATGATTCTGTGTATAACTTGATTACTCCACTG 57
frog           ---GTGACAGCAGTTTTTATGATGATTTATGGGTCTGTGTTTAACTTGATTACTCTGCTG 57
                  ********     * * ***********  **  *** *** ******** *   **
```

Figure 2.1: Conservation of Pbx-Hox motifs in a vertebrate CNE. A portion of a ClustalW2 alignment of 21 orthologous CNEs (*dach1*.11206) from different vertebrates is shown. The species from which the sequences are derived are indicated to the left. Bases conserved in all aligned species are indicated with an asterisk below the alignment. The conserved Pbx-Hox motif (matching the consensus TGATNNAT) is indicated (red box). A schematic representation of the element is shown at the top, with the Pbx-Hox motif indicated by a red arrow.

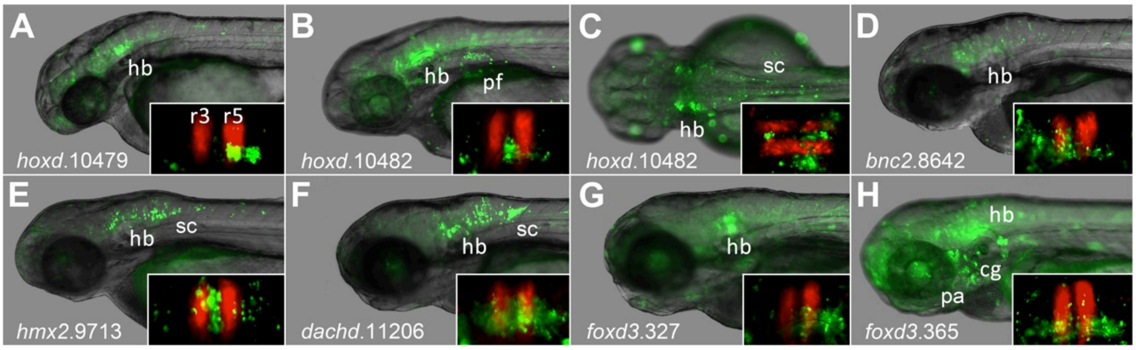


Figure 2.2: Pbx-Hox CNEs drive expression in the hindbrain during the first three days of zebrafish development. Images show F^0 transgenic embryos between 2 and 3 dpf expressing Egfp under the control of CNEs. Insets show comparison with mCherry in rhombomeres 3 and 5. *hoxd*.10479 (A) at ~42 hpf in ventral r5r6; hoxd.10482 lateral (B) and dorsal (C) views at ~56 hpf in lateral r2, r4, r6 and pectoral fin; *bnc2*.8642 (D) at ~60 hpf in hindbrain; *hmx2*.9713 (E) at ~60 hpf in hindbrain and spinal cord; *dachd*.11206 (F) at ~72hpf in hindbrain and spinal cord; *foxd3*.327 (G) at ~72hpf in ventral r5 and r6; *foxd3*.365 in hindbrain, pharyngeal arches/neural crest and cranial ganglia. Embryo counts can be found in the appendix (8.2.2).

cg: cranial ganglia; hb: hindbrain; pa: pharyngeal arches/neural crest; pf: pectoral fin; r3 r5: rhombomeres 3 and 5; sc: spinal cord.

Table 2.1: Pbx-Hox CNEs can act as hindbrain enhancers or enhancers of other tissues during the first three days of zebrafish development. The table shows the name of CNEs containing the hb+ grammar (CNE) and the percentage of injected embryos as a proportion of the total number injected (hb+/total) and as a proportion of GFP positive embryos (hb+/total). The most common regions observed for each element are also displayed (common regions). Hindbrain enhancers are indicated in green. Other enhancers are indicated in red. Embryo counts can be found in the appendix (8.2.2)

	2 dpf			**3 dpf**		
CNE	**hb+ /total**	**hb+ /GFP+**	**Common regions**	**hb+ /total**	**hb+ /GFP+**	**Common regions**
***atbf1*.5817**	0.00%	0.00%	none	0.00%	0.00%	none
***auts2*.8971**	0.00%	0.00%	none	0.00%	0.00%	none
***barhl2*.3932**	0.00%	0.00%	none	0.00%	0.00%	none
***bcl11a*.2446**	0.00%	0.00%	none	0.00%	0.00%	none
***bcl11b*.4483**	0.00%	0.00%	none	0.00%	0.00%	none
***bhlhb5*.9206**	0.00%	0.00%	none	0.00%	0.00%	none
***bnc2*.8642**	52.86%	92.50%	hb	45.45%	83.33%	hb
***cst*.9931**	0.00%	0.00%	sc	0.00%	0.00%	sc
***dachd*.11206**	61.54%	96.00%	hb	26.00%	100.00%	hb
***dachd*.227**	0.00%	0.00%	none	0.00%	0.00%	none
***dlx1*.6882**	0.00%	0.00%	none	0.00%	0.00%	none
***ebf3*.3763**	0.00%	0.00%	none	0.00%	0.00%	none
***emx2*.4543**	0.00%	0.00%	msc	0.00%	0.00%	msc
***esrrg*.9376**	0.00%	0.00%	none	0.00%	0.00%	none
***evi1*.10731**	0.00%	0.00%	msc	0.00%	0.00%	none
***fign*.5108**	0.00%	0.00%	msc	0.00%	0.00%	none
***fog2*.3620**	0.00%	0.00%	ht	0.00%	0.00%	ht
***foxb1*.5486**	0.00%	0.00%	none	0.00%	0.00%	none
***foxd3*.308**	0.00%	0.00%	msc	0.00%	0.00%	msc
***foxd3*.327**	32.89%	92.59%	hb	81.82%	100.00%	hb
***foxd3*.365**	21.43%	83.33%	hb	18.02%	100.00%	none
***foxp1*.885**	0.00%	0.00%	msc	0.00%	0.00%	msc
***foxp1*.887**	0.00%	0.00%	none	0.00%	0.00%	none
***foxp2*.3468**	0.00%	0.00%	none	0.00%	0.00%	none
***gli3*.2148**	0.00%	0.00%	ht	0.00%	0.00%	ht
***hmx2*.9713**	36.94%	91.11%	hb	7.69%	87.50%	none
***hoxd*.10479**	41.18%	89.09%	hb	55.56%	91.67%	hb
***hoxd*.10482**	21.05%	90.91%	hb	16.42%	100.00%	none
***lmo4*.2659**	0.00%	0.00%	none	0.00%	0.00%	none

hb: hindbrain; sc: spinal cord; msc: muscle; ht: heart

2.5 DISCUSSION

The data from this chapter demonstrate that CNEs containing Pbx-Hox motifs can drive reporter gene expression in the hindbrain of developing zebrafish embryos. However, the hindbrain patterns driven by these elements are heterogeneous, exhibiting different segment specificities or lack thereof (figure 2.2). All CNEs from CONDOR[142] are conserved at at least 60% identity and are >38bp long. Aside from Pbx-Hox motifs (figure 2.1), alignments of CNEs contain conserved sequence blocks that presumably correspond to additional TFBSs. This suggests a model where combinatorial binding by numerous factors provides further activatory and repressive inputs to determine the tissue-specific output of these elements.

The proportion of enhancers discovered during these experiments is consistent with previous work, where roughly half of all tested CNEs activate expression in this functional assay[40, 41, 157, 164]. However, the enrichment for hindbrain enhancer activity detected in this test set was considerably lower than that previously reported for a set of Pbx-Hox CNEs. 11/20 (55%) candidate elements acted as hindbrain enhancers in the previous study[157], compared with only 7/29 (24%) identified herein. There are several possible reasons for this discrepancy, mostly relating to differences in the selection criteria for candidate CNEs.

The earlier study appears to have selected CNEs associated with known hindbrain genes; all of the loci from which these candidates were selected contain genes exhibiting strong hindbrain expression (ZFIN gene expression database), for example *meis2a*, *tshz3.1*, and *znf503.1*. Contrastingly, an unbiased approach to candidate selection was taken herein; CNEs from any locus were considered as candidates as long as they contained a conserved Pbx-Hox motif. This was to ensure fair testing of the previous authors' hypotheses. As evidenced by the above data, conserved Pbx-Hox motifs perform inadequately; 3/4 assayed elements are hindbrain negative when candidate CNEs are selected without considering the expression patterns of nearby genes. Indeed, all of the hindbrain enhancers identified above are located near genes known to be strongly expressed in the zebrafish hindbrain. However, when considering CNEs associating with known hindbrain genes, Pbx-Hox motifs are still poor at identifying hindbrain enhancers. For example, in these experiments some CNEs associated with the hindbrain genes *dachd*, *foxd3*, *foxp1*, *foxp2* and *gli3* fail to act as hindbrain enhancers despite containing Pbx-Hox motifs (table 2.1). This is also seen in the previous study, where 9/20 Pbx-Hox CNEs fail to act as hindbrain enhancers. Nevertheless, considering gene expression data during the candidate selection process might increase the specificity of predictions.

The first study selected CNEs containing at least one Pbx-Hox motif, whereas the current study selected CNEs containing just a single Pbx-Hox motif. Including elements with more than one Pbx-Hox motif increases the likelihood that at least one site per element will be functional. Generally speaking, it is commonly observed that strong enhancers contain repeated binding motifs for one or more activating factors, referred to as homotypic clusters[225]. Specifically, artificial elements containing multiple tandem Pbx-Hox motifs can act as hindbrain enhancers but there is little evidence that Pbx-Hox sites alone can direct hindbrain expression in a non-artificial setting. Therefore, CNEs with multiple Pbx-Hox sites might be more likely to act as enhancers in the hindbrain and also more likely to generate visible levels of GFP expression.

The use of lamprey CNEs as a reference by the earlier study could have increased the proportion of hindbrain enhancers found. This could reflect the functions of vertebrate versus gnathostome CNEs. Nevertheless several hindbrain enhancers discovered in the earlier study are not conserved in lamprey, indicating that restricted to the gnathostome lineage also function to control hindbrain development. Furthermore, using a more distant out-group detects even more ancient conservation of Pbx-Hox sites; thus, the use of lamprey CNEs as a reference might be more likely to identify functional motifs.

During these experiments, primers were designed to be as close to the bioinformatically defined conservation peak (from CONDOR) as possible. Whilst many of the sequences tested above fail to act as hindbrain enhancers in the context of this assay, these might act as hindbrain enhancers *in vivo* where local sequence context could contribute to their function. Sequence flanking these CNEs might represent poorly conserved or lineage-specific, but nonetheless indispensible sites. These sorts of flanking sequences might have been missed due to primer placement. This might suggest some discrepancy between the bioinformatic definition of CNEs and the boundaries of the corresponding enhancers in the genome. Repeating the experiments using primers that target a greater amount of flanking sequence either side of the conservation peak could test this hypothesis. Indeed, including larger flanking sequences has been known to alter the function of some CNEs in functional assays[163].

All the hindbrain enhancers come from loci containing known hindbrain genes and the pattern of reporter gene expression overlaps with these associated hindbrain genes (ZFIN gene expression database). These are therefore likely to act as *in vivo* enhancers of these genes. Considering the high degree of sequence identity amongst orthologous CNEs and their size (the range of fragments tested herein is 50-400bp) it is likely that they contain additional motifs that determine their functions aside from Pbx-Hox motifs. If these hindbrain enhancers share some common functionality and are bound by common factors, these should be detectable using motif detection algorithms.

2.6 CONCLUSIONS

In this chapter it was attempted to identify novel hindbrain enhancers through the presence of Pbx-Hox binding motifs. Several novel hindbrain enhancers were identified but enrichment for hindbrain activity in the test set of CNEs was lower than expected based on previously published results. Therefore, the assertion that hundreds of CNEs act as hindbrain enhancers because they contain Pbx-Hox motifs is probably an overestimation.

In summary:

- Pbx-Hox CNEs are not sufficient for hindbrain enhancer activity;
- Using conserved Pbx-Hox motifs, putative Hox target enhancers of *bnc2, dachd*, *foxd3*, *hmx2/hmx3* and the *hoxd* cluster have been identified;
- Differences in selection criteria might explain the disparity in enrichment for hindbrain enhancers in the test sets of this and the previous study[157];
- Sequence context might be important in determining the functionality of Pbx-Hox motifs since these do not always drive hindbrain expression.

The combined number of Pbx-Hox CNEs assayed in the previous study and this chapter is 55, consisting of 22 hindbrain enhancers (hb+ set) and 33 non-hindbrain enhancers (hb- set). These sets can now be used to test hypotheses relating to the sequence basis for hindbrain enhancer activity. Since enhancer activity is thought to be determined by combinatorial binding by many factors, hindbrain enhancers may share some common motifs in addition to Pbx-Hox. In order to increase the accuracy of predictions, it is necessary to understand the underlying basis for hindbrain enhancer activity (hindbrain *cis*-regulatory grammar). Elements of such grammar might include parameters such as the contribution of variable bases in the Pbx-Hox motif (positions 5 and 6, for example), other motifs, and the spacing and orientation of other motifs relative to the Pbx-Hox motif.

CHAPTER 3

Pbx-Hox and Meis/Pknox motifs occur proximally in hindbrain enhancers

3.1 BACKGROUND

The sequences of the 55 Pbx-Hox CNEs for which functional data exist were split in to two sets by virtue of their functions during enhancer assays (hb+ and hb-). Using these sets the sequence basis for hindbrain enhancer activity was scrutinised. The sequence sets possess distinct enhancer capabilities; commonalities within the sets and distinctions between the sets could highlight important and biologically meaningful sequence features relevant to hindbrain enhancer activity.

Since the number of CNEs in each set was quite low, the literature was searched for data from other enhancers. An additional study[226] used an identical functional assay and detected numerous conserved hindbrain enhancers. The zebrafish orthologue of each CNE was assayed in this study, identical to the previously used sets. The data from this study contain 16 hb+ sequences and 120 hb- sequences. These were added to our sets to increase the sample size for subsequent analyses.

It was decided to use the MEME suite[227] as a starting point because these algorithms are well established for the detection and analysis of motifs. The MEME algorithm[228] was originally developed for the discovery of novel motifs from ChIP data, but it is also applicable to the problem presented herein; whether hindbrain enhancers are enriched for certain motifs compared with non-enhancers or enhancers of other tissues. MEME can generate novel motifs from a set of sequences in two ways: *de novo,* where enriched motifs are detected with respect to a background model; or discriminatively, where enriched motifs are detected with respect to a control set. This makes MEME ideal for the comparison of two functionally validated sets of enhancers. MEME generates an output in the form of sequence logos derived from position weight matrices (PWMs). These logos represent the log likelihood of finding a choice of bases at a given position in the motif.

Since 52 of the sequences were originally identified through evolutionary conservation, sequence alignments might be a simple and informative approach to pinpoint functional motifs. CONDOR[142] contains entries for many of the sequences from the sets, and has stored BLAST hits for orthologues from many vertebrate genomes. Multiple orthologous sequences can then be aligned using algorithms such as ClustalW2[220] and searched by eye for conserved motifs (phylogenetic footprinting). This approach is useful because conservation of the motif strongly implies functionality. However, not all the sequences from the sets have entries in CONDOR, so this reduces the sample size.

An alternative and complementary approach is to detect matches by similarity to PWMs. There are two programs from the MEME suite[227] suitable for this purpose. The FIMO[229] algorithm can find significant matches to PWMs whilst also listing information about location and orientation, ideal for testing hypotheses relating to the grammar of multiple motifs. MCAST[230] detects clusters of multiple motifs. The output from MCAST can be manually filtered to list only clusters containing both sites. These approaches will use a larger sample size than phylogenetic footprinting but will detect all motif instances rather than conserved motif instances.

3.2 AIMS AND HYPOTHESES

- **Aim:** to elucidate the enriched motifs from the hb+ and hb- sets, identify likely TF candidates and pursue any significant discoveries, particularly in relation to the grammar of sites in the sequences.
- **Aim:** to assess how often pairs of motifs co-occur in each set using ClustalW2, MCAST and FIMO, and to determine the contribution of site spacing, order and orientation to hindbrain enhancer function.
- **Hypothesis:** hb+ sequences contain distinct motifs distinguishing them from hb- sequences, representing binding sites for accessory factors.
- **Hypothesis:** The hb+ and hb- sets contain distinct grammars of motifs that can distinguish them on the basis of their sequence.

3.3 METHODS

3.3.1 Categorisation of sequence sets

The sequences assayed in chapter 2 and two previous studies[157, 226] were collected and separated in to two files according to their hindbrain enhancer activity. Hindbrain enhancers were placed in one set (hb+) and all other sequences were placed in another (hb-). Enhancer activity or lack thereof in other tissues held no bearing on categorisation. The criteria for whether a CNE was regarded as a hindbrain enhancer are detailed in the methods section of chapter 2. The other studies used similar[157] or more stringent[226] criteria. There were 188 sequences, comprised of 38 hb+ and 150 hb-.

3.3.2 Discovery and analysis of enriched motifs

The hb+ and hb- sets were submitted to the MEME web server (http://meme.nbcr.net/meme/tools/meme) in a series of six experiments using various settings. This was to ensure the enriched motifs discovered were robust to a wide parameter space. The only parameter remaining unchanged throughout all analyses was a motif length of 6-12 nucleotides, the typical length of known eukaryotic monomeric TFBSs. Firstly, *de novo* motif discovery was performed on the hb+ and hb-sets, allowing any number of repetitions per sequence to contribute to the PWM. Secondly, discriminative motif discovery was performed on the hb+ set using the hb-set as a control and *vice versa* allowing zero or one motif occurrence per sequence to contribute to the PWM. Finally, discriminative motif discovery was performed on the hb+ set using the hb- set as a control and *vice versa* forcing one motif occurrence per sequence to contribute to the PWM. In each case, shuffled sequence controls were also performed. Output PWMs were saved for future use.

The discovered motifs were compared to three online databases of known TF binding preferences (TRANSFAC[178], Uniprobe[181] and JASPAR[179]) using two alignment algorithms (TOMTOM[272] and the Smith-Waterman local alignment in STAMP[273]). This was to ensure robustness of the matches to different algorithms and databases. Only matches to factors from vertebrates were considered.

3.3.3 Phylogenetic footprinting with Clustalw2

52 of the sequences from the hb+ and hb- sets have entries in CONDOR[142] (http://condor.nimr.mrc.ac.uk). These entries were accessed and 16-24 orthologous sequences were downloaded. The sequences were trimmed to prevent unaligned ends and aligned using ClustalW2[220] (http://www.ebi.ac.uk/Tools/msa/clustalw2/). The

alignments were then searched for conserved motifs matching the MEME-derived Pbx-Hox (TGATDDATKD) and Meis/Pknox (CTGYCA) motifs, allowing up to one mismatch per motif.

Three parameters assigned to each element were derived from these alignments: 1. Inter-motif distance (a positive integer); 2: Relative site order (categorical variable, preceding or following); 3: relative site orientation (categorical variable, + or -). The inter-motif distance for each element was defined as the number of nucleotides between the closest Pbx-Hox and Meis/Pknox motif pair in the zebrafish orthologue. To assign the other two parameters, the sequences were oriented such that the Pbx-Hox motif appeared in the same ("positive") orientation (TGATDDATKD). These parameters therefore refer to the order and orientation of the Meis/Pknox motif relative to the Pbx-Hox motif.

3.3.4 PWM matching with FIMO

The hb+ and hb- sets were submitted to the FIMO[229] web server (http://meme.nbcr.net/meme/tools/fimo) and searched for motifs matching the MEME-derived Pbx-Hox (TGATDDATKD) and Meis/Pknox (CTGYCA) PWMs with a p-value cut-off of 0.001. The output files were exported as excel spreadsheets. The number of Pbx-Hox and Meis/Pknox motifs per sequence were counted for each set. Motifs were considered to cluster if they contained one significant match to each motif within 100bp. Sequences were considered to have co-occurrences if they contained at least one cluster.

3.3.5 PWM matching with MCAST

The hb+ and the hb- sets were submitted to the MCAST[230] web server (http://meme.nbcr.net/meme/tools/mcast) and searched for pairs of motifs matching the Pbx-Hox (TGATDDATKD) and Meis/Pknox (CTGYCA) PWMs. Motif clusters were considered if they contained one significant match to each motif within 100bp. Sequences were considered to have co-occurrences if they contained at least one cluster.

3.4 RESULTS

3.4.1 Hindbrain CNEs contain Pbx-Hox and Meis/Pknox motifs

MEME was used to derive motifs from the hb+ and hb- sets (see methods). When *de novo* motif discovery is performed, the hb+ set is enriched for two motifs. The first (motif 1, TGATDDATKD, figure 3.1 A) is very similar to the sequence used to select the sequences (TGATNNAT). However, this motif is longer and shows a clear bias against C at positions 5, 6 and 10 and a strong preference for G/T at position 9, suggesting that functional Pbx-Hox sites are more likely to have this composition. Motif 1 aligns to known binding preferences for Pbx and Hox proteins (figure 3.1 B). The second (motif 2, CTGYCA, figure 3.1 C) matches preferences for Meis and Prep proteins, two closely related classes of proteins[231] that preferentially bind the 6mer CTGTCA (figure 3.1 D). These motifs were not found in shuffled sequence controls. Contrastingly, the hb- set is enriched for only a single motif (CTCTCTCTCTCT) that matches known preferences poorly (data not shown), and appears to be derived from repetitive sequence. The motif occurrences contributing to this PWM come from very few of the sample sequences, so this cannot have a significant bearing on function.

Next, to compare and contrast the motif content of the two sets, discriminative motif searches were performed using MEME. The hb+ set is enriched for DTGATKDATK with respect to the hb- set, irrespective of the settings used (whether the number of motif occurrences per sequence contributing to the PWM is set to 1, or 0 or 1). This suggests that Pbx-Hox motifs from the hb+ set have a different composition to those in the hb- set, with distinct preferences at positions -1, 5, 6 and 9 (data not shown). This motif is not enriched in shuffled sequence controls. The only motif found enriched in the hb- set when compared to the hb+ set is CTCTCTCTCTCT, identical to that found in *de novo* searches (data not shown).

The hb+ set, hb- set and the PWMs for motif 1 and motif 2 (derived from the hb+ set) can be found in the appendix (8.2.3).

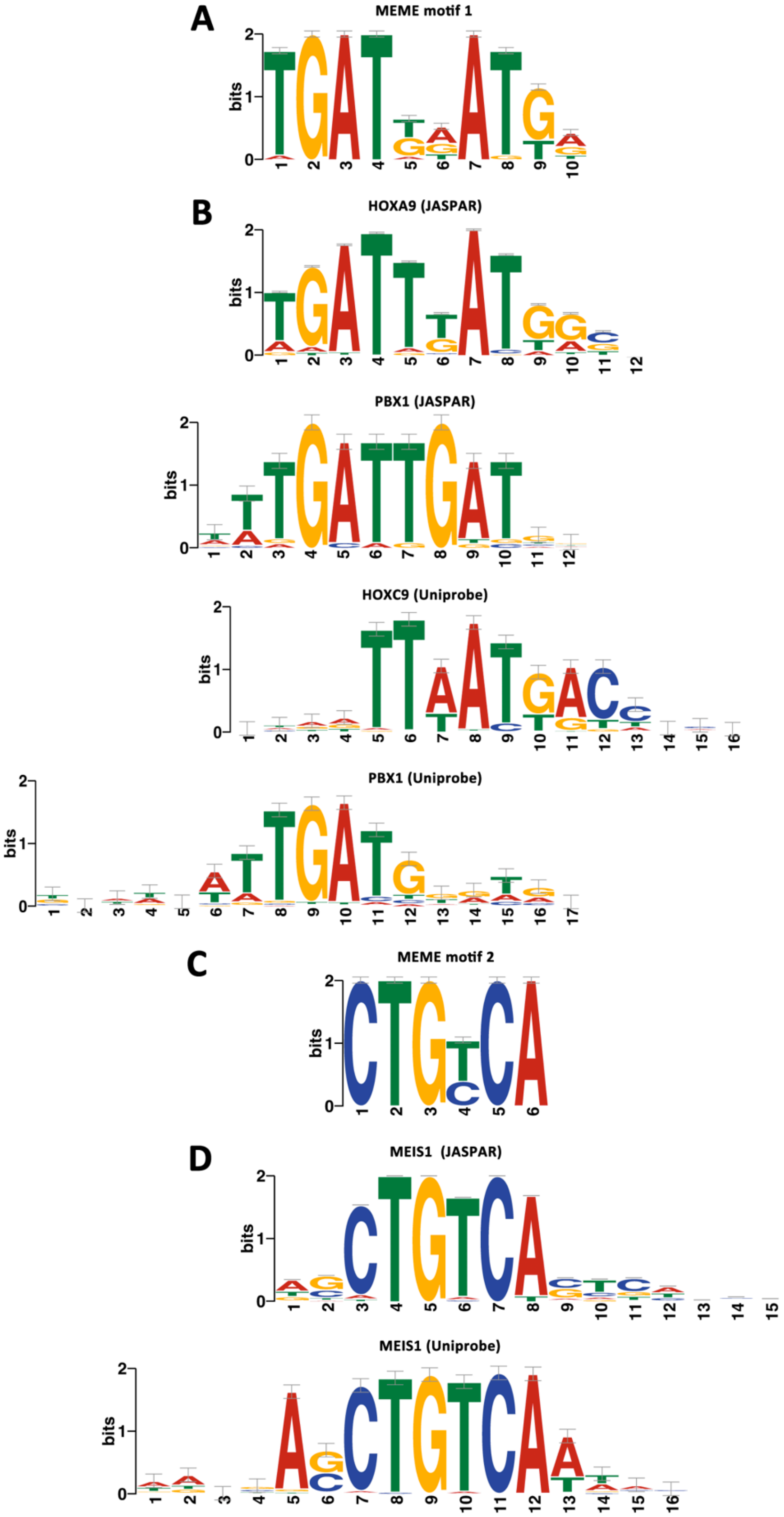
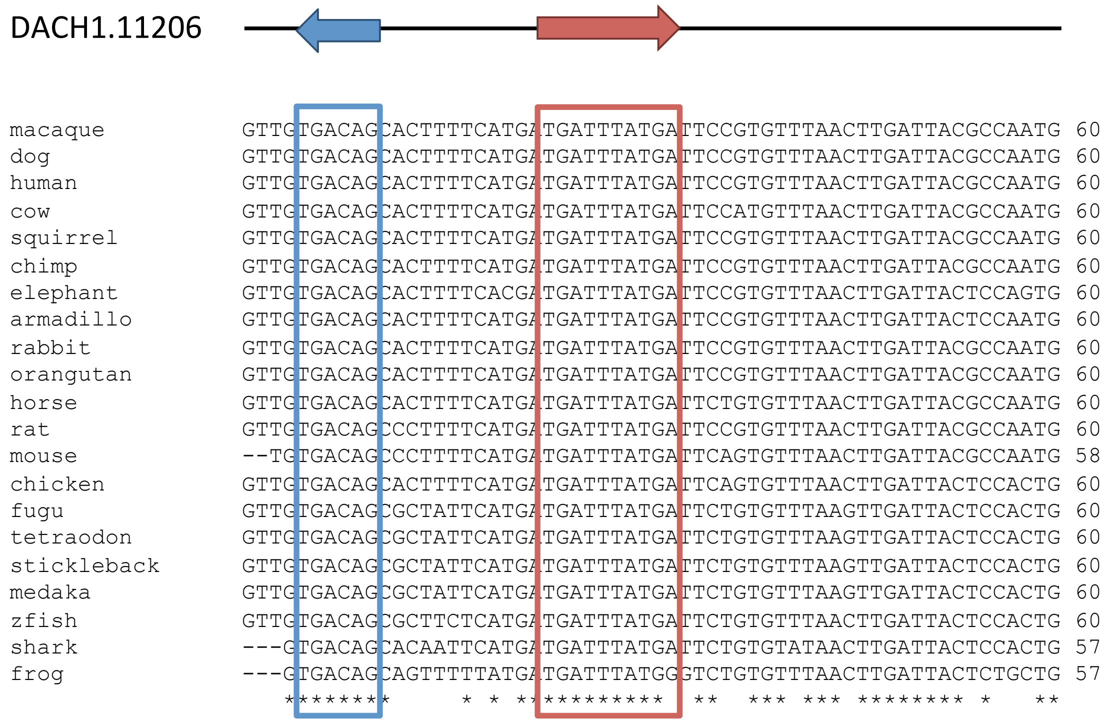


Figure 3.1: Hindbrain enhancers typically contain both Pbx-Hox and Meis/Pknox motifs. Motif discovery on a set of 38 hindbrain enhancers using MEME detects two enriched motifs. The first (A) resembles Pbx and Hox binding preferences (B) and the second (C) resembles Meis and Prep binding preferences (D). Alignments were performed using TOMTOM against JASPAR and Uniprobe. The PWMs for motif 1 and motif 2 can be found in the appendix (8.2.3).

3.4.2 Pbx-Hox and Meis/Pknox motifs occur proximally

In order to investigate the distribution and conservation of motifs in the hb+ and hb- sets, the sub-set of sequences that have entries in CONDOR (hb+ n=22 and hb- n=30) were aligned using ClustalW2. 20/22 hb+ sequences contain conserved instances of both Pbx-Hox and Meis/Pknox motifs (an example is shown in figure 3.2). All 20 of these sequences have an inter-motif distance of 89bp or less (figure 3.3). 17 have an inter-motif distance of less than 50bp and 14 have an inter-motif distance of less than 25bp. Whilst the remaining 2 sequences contain motifs loosely resembling the Meis/Pknox consensus near their Pbx-Hox sites, these were not counted as containing a Meis/Pknox site according to the criteria used. However, these locations may be low-affinity Meis/Pknox binding sites. In contrast to the hb+ set, 5/30 hb- sequences contain conserved instances of both motifs. The likelihood of seeing this distribution of co-occurrences by chance is 9.8×10^{-8} (Fisher's exact test), demonstrating that hindbrain enhancers are significantly more likely to contain both motifs. Furthermore, 3/5 of the hb- sequences containing co-occurrences have an inter-motif distance exceeding 100bp. When considering only cases where the motifs occur within 100bp, the likelihood of seeing this distribution by chance is 2.7×10^{-9} (Fisher's exact test). This suggests that motif proximity also plays a role as well as co-occurrence.

To assess the contribution of motif orientation and order to hindbrain enhancer activity, the motifs were considered in pairs (consisting of one Pbx-Hox and one Meis/Pknox motif) and the incidence of each possible orientation/order combination was counted. The hb+ sequences contain 22 Pbx-Hox and Meis/Pknox motif pairs. Most sequences contain one pair, *meis2a*.1102 contains two pairs of sites[157] and *znf503.1*.10105 contains a cluster of two Meis/Pknox motifs flanking a single a Pbx-Hox motif (figure 3.3), which was considered to be two separate pairs for the purpose of this analysis. With respect to the Pbx-Hox motif in "positive" orientation (TGATDDATKD), there are 12 pairs where the Meis/Pknox motif precedes the Pbx-Hox motif and 8 pairs where the Meis/Pknox motif is in "positive" orientation (CTGYCA). There is no significant preference for any of the four possible combinations of motif orientation/order: forward/preceding (n=4), forward/succeeding (n=5), reverse/preceding (n=10) or reverse/succeeding (n=6). The p-value for this distribution is 0.43 (Fisher's exact test). This suggests that site order and orientation are inconsequential. Annotated alignments of the 22 hb+ sequences can be found in the appendix (8.2.3).



```
macaque         GTTGTGACAGCACTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACGCCAATG 60
dog             GTTGTGACAGCACTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACGCCAATG 60
human           GTTGTGACAGCACTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACGCCAATG 60
cow             GTTGTGACAGCACTTTTCATGATGATTTATGATTCCATGTTTAACTTGATTACGCCAATG 60
squirrel        GTTGTGACAGCACTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACGCCAATG 60
chimp           GTTGTGACAGCACTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACGCCAATG 60
elephant        GTTGTGACAGCACTTTTCACGATGATTTATGATTCCGTGTTTAACTTGATTACTCCAGTG 60
armadillo       GTTGTGACAGCACTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACTCCAATG 60
rabbit          GTTGTGACAGCACTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACGCCAATG 60
orangutan       GTTGTGACAGCACTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACGCCAATG 60
horse           GTTGTGACAGCACTTTTCATGATGATTTATGATTCTGTGTTTAACTTGATTACGCCAATG 60
rat             GTTGTGACAGCCCTTTTCATGATGATTTATGATTCCGTGTTTAACTTGATTACGCCAATG 60
mouse           --TGTGACAGCCCTTTTCATGATGATTTATGATTCAGTGTTTAACTTGATTACGCCAATG 58
chicken         GTTGTGACAGCACTTTTCATGATGATTTATGATTCAGTGTTTAACTTGATTACTCCACTG 60
fugu            GTTGTGACAGCGCTATTCATGATGATTTATGATTCTGTGTTTAAGTTGATTACTCCACTG 60
tetraodon       GTTGTGACAGCGCTATTCATGATGATTTATGATTCTGTGTTTAAGTTGATTACTCCACTG 60
stickleback     GTTGTGACAGCGCTATTCATGATGATTTATGATTCTGTGTTTAAGTTGATTACTCCACTG 60
medaka          GTTGTGACAGCGCTATTCATGATGATTTATGATTCTGTGTTTAAGTTGATTACTCCACTG 60
zfish           GTTGTGACAGCGCTTCTCATGATGATTTATGATTCTGTGTTTAACTTGATTACTCCACTG 60
shark           ---GTGACAGCACAATTCATGATGATTTATGATTCTGTGTATAACTTGATTACTCCACTG 57
frog            ---GTGACAGCAGTTTTTATGATGATTTATGGGTCTGTGTTTAACTTGATTACTCTGCTG 57
                   ********     * * *********** **  *** *** ******** *   **
```

Figure 3.2: Conservation of Pbx-Hox and Meis/Pknox motifs in vertebrate CNEs. A portion of a ClustalW2 alignment of 21 orthologous CNEs (*dach1.*11206) from different vertebrates is shown. The species from which the sequences are derived are indicated to the left. Bases conserved in all aligned species are indicated with an asterisk below the alignment. The conserved Pbx-Hox (red box) and Meis/Pknox (blue box) motifs are indicated. A schematic representation of the element is shown at the top, with conserved motifs represented by arrows (red for Pbx-Hox, blue for Meis/Pknox). This element is also shown in figure 2.1; note the longer Pbx-Hox motif matching the consensus TGATDDATKD (conserved in all species except for the final nucleotide in frog), and the Meis/Pknox motif matching the consensus CTGYCA (conserved in all species).

Returning to the larger sets of sequences (hb+ n=38 and hb- n=150), FIMO was used to detect instances of Pbx-Hox and Meis/Pknox PWMs in each set. The hb+ set contains 73 hits to the Pbx-Hox PWM (1.92 per sequence) and 79 hits to the Meis/Pknox PWM (2.07 per sequence). The hb- set contains 203 hits to the Pbx-Hox PWM (1.27 per sequence) and 190 hits to the Meis/Pknox PWM (1.26 per sequence). Thus, the hb+ has only modest enrichment for both Pbx-Hox and Meis/Pknox motif hits compared to the hb- set (1.5x and 1.6x, respectively). However, 26/38 hb+ sequences and 49/150 hb- sequences contain clusters of motifs according to the chosen criteria (see methods). The likelihood of seeing this distribution by chance is 7.0x10^{-5} (Fisher's exact test). Therefore, the co-occurrence of Pbx-Hox and Meis/Pknox motifs is strongly associated with hindbrain enhancer function, in agreement with the trend seen in the ClustalW2 alignments performed on the smaller set.

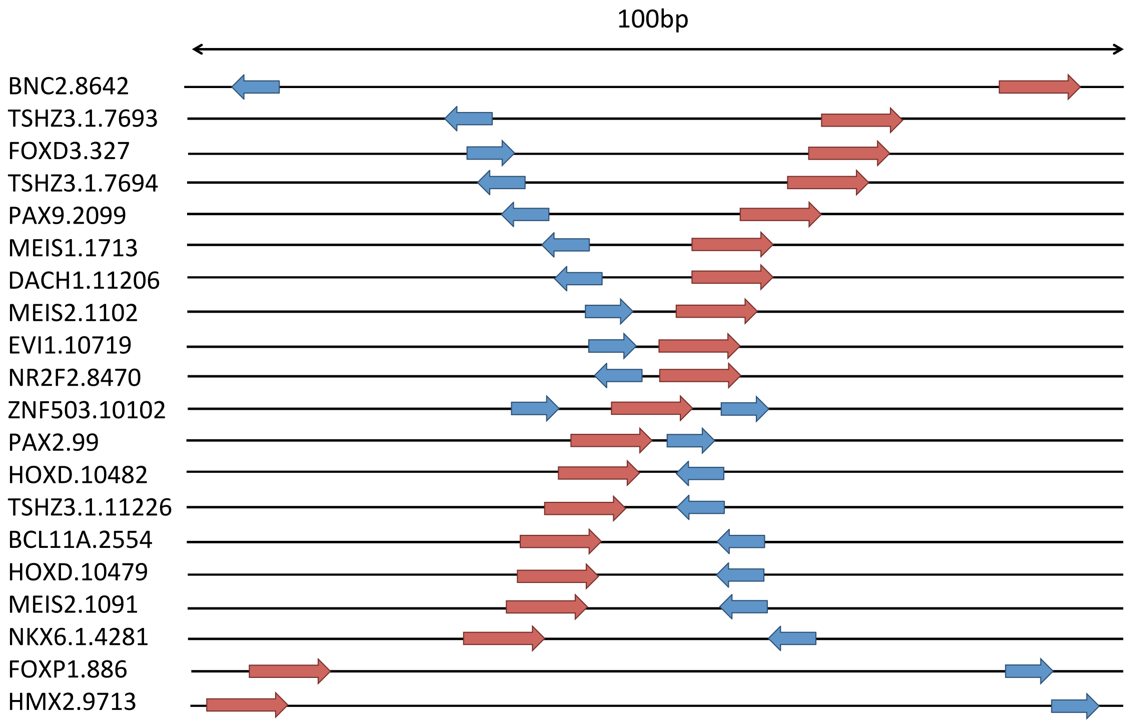


Figure 3.3: Hindbrain enhancers contain a common grammar of Pbx-Hox and Meis/Pknox motifs within 100bp. The figure shows schematic representations of 20 Pbx-Hox and Meis/Pknox motif pairs from functionally validated hindbrain enhancers, showing the locations and relative orientations of Pbx-Hox (red arrows) and Meis/Pknox (blue arrows) motifs. The motifs occur within 100bp in all cases but the distance between motifs varies within this range. The sites occur in variable arrangements; there is no preference for particular orders or orientations of motifs. The name of the element is displayed on the left. The occurrence of this motif structure within 19/21 sequences from the hb+ set suggests that this constitutes a hindbrain enhancer grammar. These schematics are derived from ClustalW2 alignments, available in the appendix (8.2.3)

To test this association further, the hb+ and hb- sets were searched for significant clusters of Pbx-Hox and Meis/Pknox motifs using MCAST. 15/38 hb+ sequences and 7/150 hb- sequences contain significant clusters. The chance of seeing this distribution of significant clusters by chance is 1.8×10^{-7} (Fisher's exact test), again indicating a strong association between Pbx-Hox and Meis co-occurrences and hindbrain enhancer activity. Thus this association is supported by data from three separate analyses, despite the use of different methodologies and match thresholds. These data suggest that conserved Pbx-Hox and Meis/Pknox motifs within 100bp constitute a hindbrain enhancer grammar.

3.5 DISCUSSION

The hb+ set is enriched for two motifs compared to a background model, suggesting that these could be important for hindbrain enhancer activity. These motif occurrences appear to represent binding sites for Pbx-Hox (motif 1) and Meis/Pknox (motif 2) factors. These proteins are known to form heterotrimers at hindbrain enhancers and their binding motifs are required for the function of several hindbrain enhancers[157, 196, 232, 233]. The presence of these motifs in this larger set is consistent with these characterised hindbrain enhancers from the literature. The prediction of hindbrain enhancers using Pbx-Hox sites alone may have been ineffective because hindbrain enhancer activity depends upon Meis/Pknox binding as well as Pbx-Hox binding, and suggests these sites encode hindbrain enhancer function using AND rather than OR logic.

This hypothesis is also supported by results from discriminative searches. Motif 2 is not found when comparing the hb+ set to the hb- set, suggesting that both sets contain instances of this motif. Meis/Pknox family proteins are expressed in and regulate the development of several tissues and organs other than the hindbrain, including the forebrain[234], midbrain[235, 236], heart[237, 238], blood[239] and vasculature[240]. The hb- set contains active enhancers of many tissues, so that some of these sequences contain Meis/Pknox motifs is not surprising. Perhaps, whilst both sets contain occurrences of the Meis/Pknox motif, these do not act as hindbrain enhancers because co-occurring Meis/Pknox and Pbx-Hox motifs are required to direct gene expression in the developing hindbrain.

Motif 1 has strong preferences at positions 5, 6, 9 and 10. This closely resembles published Pbx-Hox binding preferences from EMSA[192] and ChIP[206, 241] experiments, more so than the TGATNNAT motif used to identify hindbrain enhancers previously[157] and in chapter 2. Furthermore, that a similar motif is discovered in discriminative searches suggests that Pbx-Hox motif occurrences in the hb+ set are qualitatively different from those in the hb- set. Searching for instances of this longer and less ambiguous motif may increase the specificity of the model, leading to fewer false positives in the future. There is also some evidence that different combinations of core bases 5/6 can preferentially select particular Hox proteins in the heterodimer[242, 243], and this may explain some of the diversity in segment specificity driven by these enhancers.

Motif 1 matches known binding preferences for Pbx and Hox proteins from Uniprobe. These have preferences for four core bases resembling TGAT and TAAT respectively. Uniprobe contains preferences derived from universal protein-binding microarray (PBM) assays[181], wherein proteins are typically tested as monomers[244]. There is evidence that dimerisation with Pbx modulates the binding preferences of Hox proteins[245], which might make the comparison of *in vivo* dimer sites and monomer preferences difficult or inappropriate. Nevertheless, Pbx and Hox monomeric preferences align to regions of motif 1 suggestive of a Pbx-Hox dimer binding site, where the Pbx protein contacting the first 4 bases and the Hox protein contacting the last 6 bases[242].

In contrast to these monomeric preferences, JASPAR and TRANSFAC contain some preferences derived from ChIP experiments[178, 179], wherein all the available cofactors would be present. This could explain why preferences from these databases match more closely to motif 1 than the preferences from Uniprobe. For example, ChIP-seq peaks for HoxA2 are enriched for the 10mer DTGATDDATD[206], suggesting frequent co-binding with Pbx. Likewise, regions occupied by both Pbx1 and Meis1 are enriched for both TGATKDATKR and CTGTCA[241], suggesting that Hox proteins frequently co-bind with Pbx1. ChIP experiments using an antibody for either Pbx proteins or Hox proteins therefore appear to recover dimeric Pbx-Hox sites; this explains the similarity between the PWMs listed as Pbx and Hox motifs from JASPAR (figure 3.1 B) and motif 1.

The hb- set is enriched for only a single motif resembling a CT dinucleotide repeat. There is some evidence that dinucleotide repeats can contribute to the function of enhancers[246]. However, the sequences from which this motif is derived are very few and, furthermore, these do not act as enhancers of any kind. This suggests that this motif is enriched simply because some non-functional repetitive sequence has been included in cloned PCR products, probably due to the difficulties associated with PCR primer placement. The hb- set is highly heterogeneous, containing enhancers of many tissues as well as non-enhancer sequences. That this set is not enriched for any other motifs is therefore unsurprising. These results suggest that functionally validated hindbrain enhancers are bound *in vivo* by a Pbx-Hox dimer and Meis/Pknox. Furthermore functional Pbx-Hox motifs appear to be less variable than the motif used in chapter 2; this motif defines a more stringent consensus. This also suggests that co-occurrence of these motifs could be important for hindbrain enhancer function, suggesting that the motifs use AND logic.

The data from this chapter define criteria describing the structure of conserved, Hox-dependent hindbrain enhancers. These criteria can now be used to predict additional hindbrain enhancers and inform appropriate mutagenesis experiments to dissect their functions. The approach used is conceptually similar to that used by a previous study which identified motifs associated with forebrain enhancer function[226], but in this case the factors binding to such motifs were not known. Resultantly, this rendered the ascription of a mechanism of action to these enhancers difficult. In contrast the motifs contributing to the hindbrain grammar identified herein support the notion that these elements are bound by Pbx-Hox and Meis/Pknox factors in order to activate the expression of nearby target genes in the hindbrain.

Another study used a machine-learning approach to create a hindbrain enhancer model, generated from a training set of hindbrain enhancers from the VISTA enhancer browser[120]. Since these enhancers were identified in large-scale screens of conserved elements with no other criteria, they are very likely to be functionally heterogeneous. However, the hb+ set used herein were identified mostly by virtue of their Pbx-Hox motifs. Therefore, these enhancers are more likely to use Pbx-Hox and Meis/Pknox motifs as key activating factors, evidenced by their enrichment for co-occurrences of these motifs. Applying this model predictively may therefore identify a group of hindbrain enhancers that operate using a common mechanism. Despite these advantages over previous hindbrain enhancer models, enhancers that drive expression in multiple tissues were present in the hb+ set. The identified motifs have a strong correlation with hindbrain enhancer activity but there is no reason to expect that sequences tested on this basis will not act as enhancers of other tissues.

The presence of and requirement for co-occurring Pbx-Hox and Meis/Pknox motifs has been noted in a number of studies of individual hindbrain enhancers. Two enhancers that have been studied extensively are a pair of enhancers from the mouse *hoxb* cluster active in r4, thought to be *hoxb1* and *hoxb2* CRMs. These elements are both activated by the r4-specific HoxB1, and the requirement for both Pbx-Hox and Meis/Pknox motifs for their functions have been demonstrated using mutagenesis[192, 195, 232, 233, 247]. Analysis of hoxb3 and hoxb4 regulatory regions from zebrafish found that many of the previously identified hindbrain enhancers from mice are also conserved at the sequence and functional levels[248]. A recent study found 28 vertebrate CNEs amongst the duplicated and subsequently fragmented teleost Hox clusters[249]. Indeed, a CRM mediating *hoxa2* expression in r4 is highly conserved amongst gnathostomes. This element also contains proximal Pbx-Hox and Meis/Pknox sites essential for its function, and is also activated by HoxB1[250].

The fact that these three elements use Pbx-HoxB1-Meis/Pknox heterotrimers to activate expression in r4 suggests that the choice of Hox protein contributes to the specificity of the element. Indeed, a study found that the choice of Hox protein can be influenced by the central, variable bases of the Pbx-Hox site; mutating these bases in the *hoxb1* CRM from GG to TA changes the expression pattern from a domain resembling the *labial*/*hoxb1* pattern to one resembling the *deformed*/*hoxb4* pattern in transgenic flies. However, the authors recognised that in most enhancers, which are arrays of binding sites, other sites most probably play an instructive role in the readout of expression as well[243].

Aside from these r4 specific enhancers, there are several other hindbrain enhancers containing Pbx-Hox and Meis/Pknox motifs, each with varying segment specificities, and conserved in various vertebrate genomes. An early comparative study using the human, mouse and fugu *hoxa* clusters discovered a conserved *hoxa3* CRM capable of driving expression in r5-6. A conserved Pbx-Hox and Meis/Pknox motif pair is visible in the alignment of this element but the necessity of these motifs for hindbrain enhancer function was not tested in this study[86]. More recently, a CRM called element C of *egr2* (krox-20) was found to be conserved in human, mouse, chicken, frog and zebrafish. All species used in the alignment contain a conserved Pbx-Hox motif, but the location and number of Meis/Pknox consensus motifs varies across evolution. The contribution of this variation to function was not assessed as only the chicken orthologue was assayed. The wild-type element is active in r3-5. Mutations targeting the Pbx-Hox motif prevent the element from upregulating reporter gene in the hindbrain at all, whereas mutations targeting the Meis/Pknox motifs abolish expression specifically in r3, indicating an ability for this Pbx-Hox motif to operate in both a Meis/Pknox-dependent and independent manner in different domains[196, 251]. Finally, a *meis2* CNE (called CNE 3299 in a previous publication, but called *meis2.*1102 here) is conserved between jawed and jawless vertebrates. In this case, functional analysis was performed on both the zebrafish and lamprey orthologues. The zebrafish element drives reporter expression in r3-4 and the corresponding neural crest, whereas the lamprey orthologue drives expression in r2-4, showing that this element's segment specificity has altered over the course of evolution[157, 162]. Furthermore, the zebrafish element contains two pairs of Pbx-Hox and Meis/Pknox motifs with distinct contributions to the role of the enhancer. Mutating the 5' motif pair causes a loss of neural crest expression whilst generating ectopic expression beyond the anterior (r2/r3 interface) and posterior (r4/r5 interface) boundaries of the wild-type pattern. This suggests that this cluster of sites act as a neural crest enhancer and an r2/r5 repressor. Mutating the 3' motif pair abolishes expression by the element altogether, suggesting these are activatory sites.

As discussed above, numerous hindbrain enhancers from the literature contain matches to the hindbrain grammar identified in this chapter but additional binding sites must underpin the diversity of patterns driven by these elements. Nevertheless, functionally diverse hindbrain enhancers may be able to be identified by applying this hindbrain grammar predictively, but only if these motifs are informative of mechanism, as the previously described association suggests.

The previously published cases discussed above all contain Pbx-Hox and Meis/Pknox sites separated by 20bp or less. However, in the phylogenetic footprinting analysis, hb+ sequences were found to contain motifs gapped by up to 89bp, although these gaps were smaller than 50bp in 17/22 cases. Within this range the size of the gaps are highly variable and there was no observable bias for particular distances between the motifs. This is consistent with data demonstrating that Pbx-Hox sites are inseparable[191] but that Meis proteins can bind proximal, but potentially gapped and/or inverted sites[192]. This suggests that these factors can activate gene expression without an enhanceosome-like grammar that requires a fixed absolute distance, at least in the context of the elements studied thus far.

At odds with this observation, two hb- sequences from a previous study[157] contain motifs gapped by less than 25bp: *gli3*.2152 and *pou3f2*.9802. Both of these genes are expressed in the hindbrain at the time-points studied, so that these sequences do not act as hindbrain enhancers is surprising. Perhaps in these cases the fully functional elements were not isolated due to primer placement, they required additional flanking sequence not classed as conserved by CONDOR, or these clusters of sites may have a repressive function similar to the 5' cluster in *meis2*.1102. Assaying these elements as part of larger fragments might help to distinguish between these possibilities. Alternately, despite the strongly predictive nature of these two co-occurring motifs, these may still not be sufficient to confer hindbrain enhancer activity to an element and additional, as yet identified motifs, form an additional component of the hindbrain grammar.

This loose arrangement of sites is seen in other classes of enhancers, for example those of the notochord[174]. The lack of a bias for particular distances is in contrast to the trend seen with, for example, dimer binding sites in *Drosophila* enhancers, where gaps are subject to periodicity matching the helical turns of DNA[171]. This loose grammar would suggest that vertebrate CNEs operate more in line with the “billboard” rather than the “enhanceosome” model of enhancer function, at least in the case of Pbx-Hox and Meis/Pknox motifs. However, site distance and orientation are remarkably well conserved amongst orthologous sequences, suggesting that these parameters have been maintained by purifying selection. Whether these parameters somehow define the enhancer activity of these enhancers remains to be established.

3.6 CONCLUSIONS

In this chapter the hb+ and hb- sets were searched for enriched motifs. The hb+ set can be distinguished from the hb- set by both the composition of Pbx-Hox motifs and the presence of proximally occurring Meis/Pknox motifs. Furthermore, these sites occur within 100bp of one another in all cases examined. The association between the presence of this grammar and hindbrain enhancer function is supported by data from two distinct approaches: phylogenetic footprinting and PWM matching. This grammar is also present in numerous published examples of hindbrain enhancers, lending support to its potential as a predictive tool.

In summary:

- The hb+ set is enriched for Pbx-Hox and Meis/Pknox binding motifs;
- Functional Pbx-Hox motifs resemble TGATDDATKD rather than TGATNNAT;
- The hb- set is enriched only for a repetitive motif;
- The hb+ set contains only modest enrichment for Pbx-Hox and Meis/Pknox motifs compared to the hb- set;
- The hb+ set contains significantly more Pbx-Hox and Meis/Pknox motif co-occurrences than the hb- set;
- CNEs active in hindbrain typically contain Pbx-Hox and Meis/Pknox motifs within 100bp;
- Site orientation, order and variation in inter-motif distance appear to make no significant contribution to the ability of an element to upregulate hindbrain expression.

The aforementioned findings support the model that these hindbrain enhancers are targeted combinatorially by Pbx-Hox dimers and Meis/Pknox factors. These motifs can now replace the original motif (TGATNNAT) used to select putative hindbrain enhancers when further candidates are selected. The analyses also define criteria that associate strongly with hindbrain enhancer activity: conserved Pbx-Hox (TGATDDATKD) and Meis/Pknox (CTGYCA) motifs occurring within 100bp.

Applying this model predictively to vertebrate genomes may identify additional Hox-target hindbrain enhancers. The predictive power of the model (sufficiency) and the contribution of each motif (necessity) need to be tested experimentally. If the hindbrain enhancer activity of the previously tested CNEs was predicted on the basis of this heuristic model (using a simple binary on/off classification at a 20% cutoff), 48/52 elements would be correctly categorised (2 false positives and 2 false negatives, an accuracy of 92%). These criteria can now be applied to the whole set of CNEs to test its predictive capacity.

CHAPTER 4

Proximal Pbx-Hox and Meis/Pknox motifs predict hindbrain enhancers

4.1 BACKGROUND

In the previous chapter, criteria were identified which could sort the hb+ and hb- sets with 92% accuracy, lending support to the hypothesis that these criteria constitute a hindbrain enhancer grammar. Many studies have identified sequence-level criteria associating with sets of enhancers with shared tissue specificities[153, 174], but subsequently failed to apply these predictively. Those studies that have applied sequence-level models predictively often discover a high proportion of false negatives[154, 176, 177], with a few exceptions[120]. In particular, the use of Pbx-Hox motifs alone to predict hindbrain enhancers has been applied with some success[157], but data from chapter 2 suggest that Pbx-Hox motifs are not sufficient to drive hindbrain expression. Therefore, it was attempted to use the hindbrain grammar to identify additional hindbrain enhancers putatively targeted by Hox proteins and their TALE-class homeodomain cofactors, with the aim to predict hindbrain enhancers with high accuracy.

The CONDOR database[142] contains entries for orthologous CNEs from fugu, mouse, rat and human derived from mLAGAN[98] and sLAGAN[99] alignments. CONDOR also contains putative orthologues from other species derived from BLAST hits. By searching for Pbx-Hox and Meis/Pknox motifs in pre-aligned sequences using MCAST[230] and FIMO[229], conserved instances of both motifs within 100bp can be detected.

Because many *hox* auto- and cross-regulatory elements containing this grammar have been found in the vertebrate *hox* clusters in previous publications[192, 195, 233, 247, 252], it was also decided to search these for additional candidate CNEs that might correspond to novel hindbrain enhancers. However, CONDOR only contains CNEs derived from alignments of the *hoxd* cluster. The other clusters are present in duplicate in the zebrafish genome, but with selective loss of genes[253], making exon-anchored alignments difficult. It was decided to perform mLAGAN[98] alignments using *hoxa* and *hoxb* clusters from human, mouse, chicken, and the two duplicated *hoxaa*/*hoxab* and *hoxba*/*hoxbb* clusters from zebrafish to detect CNEs conserved at either zebrafish cluster. The *hoxc* clusters were not aligned because the chicken genome lacks an assembled *hoxc* cluster.

Based on the strong association between this grammar and hindbrain enhancer activity discovered in chapters 3 and 4, novel elements found to contain this grammar will be predicted to be hindbrain enhancers as well. Subjecting the candidates to the enhancer assay[150] will then assess the predictive power of this grammar.

Once novel hindbrain enhancers have been identified using these criteria, these can be added to the hb+ set. This larger set can then be searched for additional enriched motifs and shared aspects of grammar. Furthermore, there is evidence to suggest that the core bases 5/6 of the Pbx-Hox motif in certain hindbrain enhancers play an instructive role by contributing to the choice of Hox protein used in the heterodimer[242, 243]. With a larger set of hindbrain enhancers, the contribution of these purportedly instructive bases can be investigated by assessing the segment specificity of each enhancer.

4.2 AIMS AND HYPOTHESES

- **Aim:** to use the previously identified sequence grammar (conserved Pbx-Hox and Meis/Pknox sites within 100bp) to predict additional hindbrain enhancers.
- **Aim:** to assess the predictive power of this grammar and expand the hb+ sequence set by subjecting candidates to functional assays.
- **Aim:** to assess parameters of this grammar in a larger set of hindbrain enhancers, including the contribution of the variable bases in the Hox binding site (5/6 and 9/10 of the Pbx-Hox motif).
- **Hypothesis:** the hb+ grammar will have greater predictive power than Pbx-Hox motifs alone when applied to full sets of CNEs.
- **Hypothesis:** specific base compositions at the variable bases in the Pbx-Hox motif contribute to the segment specificity of each enhancer.

4.3 METHODS

4.3.1 Detection of conserved mCAST clusters

Full sets of CNEs over 40bp in length from human (n=6347) and zebrafish (n=4259) were downloaded from CONDOR (http://condor.nimr.mrc.ac.uk). These were then searched for clusters as detailed in the methods section of chapter 3 (3.3.5). CNEs containing a significant cluster in both the human and zebrafish orthologues were retained for further analysis. The divergence in inter-motif distance between human and zebrafish was calculated as the difference between the inter-motif distances of each orthologue. Some interesting individual cases were investigated further using ClustalW2[220] alignments using orthologous sequences from CONDOR.

4.3.2 Identification of candidate CNEs

Full sets of CNEs over 100bp from human, mouse, rat/dog and fugu were downloaded from CONDOR (http://condor.nimr.mrc.ac.uk) and merged in to a single file (n=12329). These sequences were then searched for clusters using FIMO[229] as in the methods section of chapter 3. CNEs were considered as potential candidates if all four orthologues of the CNE contained a cluster. Potential candidates were subjected to phylogenetic footprinting as detailed in the methods section of chapter 3 (3.3.3) to confirm the conservation of the Pbx-Hox and Meis/Pknox motifs. Any candidates that did not contain canonical Pbx-Hox (TGATNNAT) and Meis/Pknox (CTGTCA) motifs in at least 90% of aligned species were discarded. The trans-dev genes from the loci with CNEs containing significant clusters were checked for hindbrain expression in the *in situ* hybridization database at ZFIN[254] (http://zfin.org/cgi-bin/webdriver?MIval=aa-xpatselect.apg).

4.3.3 Detection of candidate CNEs at vertebrate *hox* clusters

Orthologous regions containing the *hoxa* and *hoxb* clusters from human, mouse, chicken and zebrafish were downloaded from the Ensembl genome browser[221] with Vista formatted annotations. 5 homologous regions were aligned for each cluster: the single clusters from human, mouse and chicken and the duplicated clusters from zebrafish. Discovered CNEs were downloaded and searched for conserved Pbx-Hox and Meis/Pknox motifs within 100bp. Those containing this grammar were used as BLAST[91] queries against other vertebrate genomes from Ensembl[221] to discover putative orthologues. These sequences were then trimmed to prevent unaligned sequence ends and aligned using ClustalW2[220] to confirm the conservation of Pbx-Hox and Meis/Pknox sites in a total of 21 vertebrate species.

4.3.4 Cloning of candidate CNEs

The cloning of candidate CNEs was performed as detailed in the methods section of chapter 2 (2.3.2).

4.3.5 Enhancer assay in zebrafish embryos

The enhancer assay was performed as described previously[150] with the alterations detailed in the methods section of chapter 2 (2.3.3).

4.4 RESULTS

4.4.1 Inter-motif distance is conserved amongst orthologous CNEs

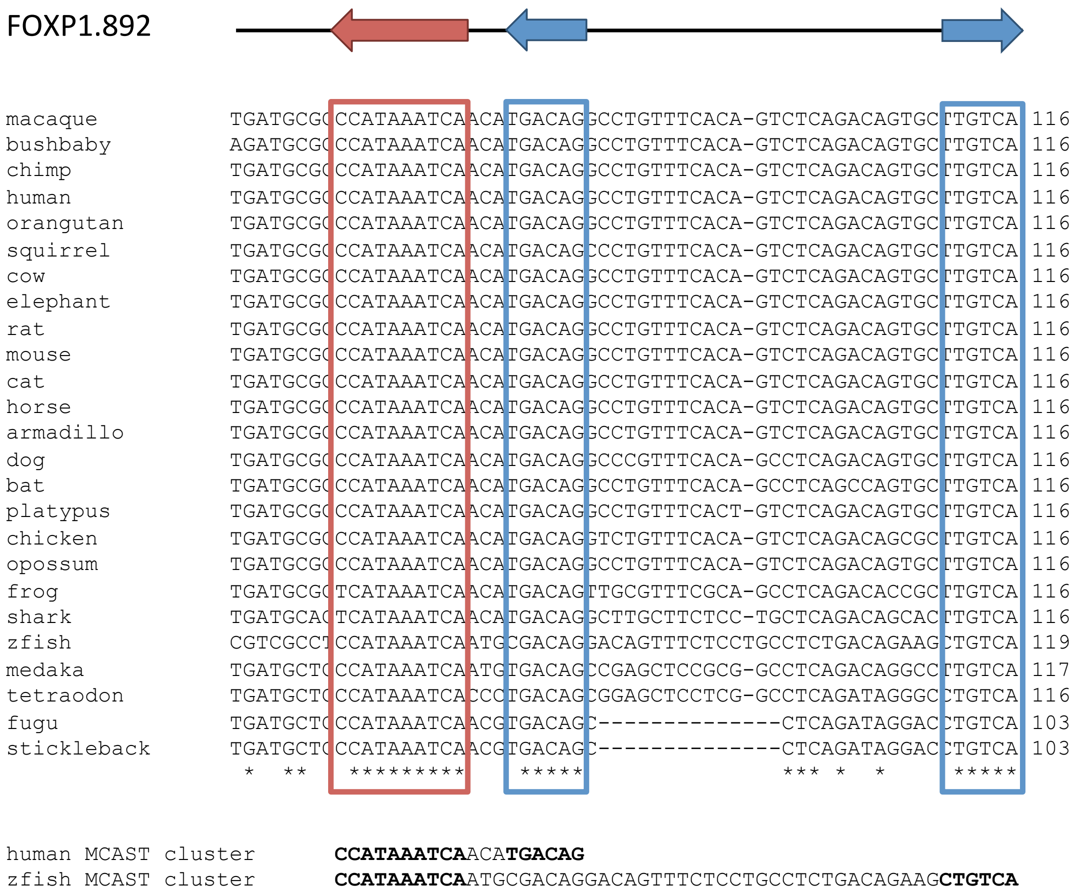
Initially, human and zebrafish CNEs were searched for instances of the hb+ grammar with MCAST[230]. 91/6347 human CNEs and 70/4259 zebrafish CNEs contained significant matches to the grammar. Of these, 39 were orthologous to one another. The inter-motif distance divergence is conserved in 23 cases and varies by 1 or 2 bp in a further 14 cases. This is to be expected as these sequences are subject to strong purifying selection by definition. Unexpectedly, The 2 remaining cases have a larger divergence in distance: one of 16 bp and another of 24 bp (Table 4.1).

The two cases where the distance divergence was (*foxp1*.892 and *zfhx1b*.2928) were examined using ClustalW2 alignments. In these cases, the Meis/Pknox motifs are not positionally conserved. This is apparent in the MCAST results: high-affinity Meis/Pknox sites matching the CTGTCA consensus are discovered at different positions in the zebrafish and human orthologues, resulting in clusters of different sizes. In the case of *foxp1*.892, this might be because the appearance of a second high-affinity site in teleosts has relaxed purifying selection at the other site (figure 4.1). In the case of *zfhx1b*.2928, might be because of a compensatory mutation in the lineage leading to zebrafish (figure 4.2). Both of these cases appear to suggest selection for high-affinity Meis/Pknox motifs proximal to the Pbx-Hox motif, without constraint on absolute distance.

7/39 elements had been confirmed to act as hindbrain enhancers previously; either in chapter 2 (figure 2.2) or a previous study[157]. The remaining element (*pou3f2b*.9802) was not regarded to function as a hindbrain enhancer in the previous study[157].

Table 4.1: Pbx-Hox and Meis/Pknox site spacing is typically conserved amongst orthologous CNEs. 39 CNEs containing significant MCAST clusters in both the human and zebrafish orthologues are shown. The inter-motif distance in human and zebrafish is shown. These were used to calculate the distance divergence. CNEs for which the zebrafish orthologue has been previously assayed for hindbrain enhancer function are highlighted in green (for positive elements) and red (for negative elements).

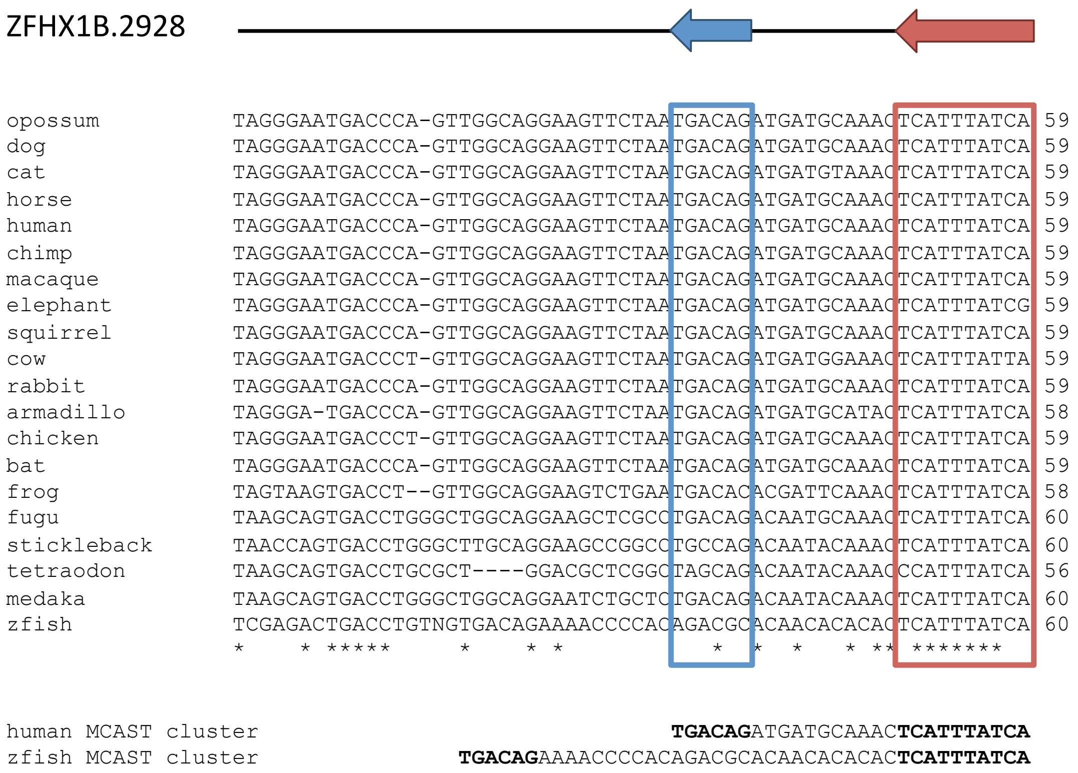
CNE	Human inter-motif distance	Zebrafish inter-motif distance	Distance divergence
pax2.99	2	2	0
dachd.232	6	7	1
foxp1.892	3	27	24
meis2a.1102	5	7	2
meis1.1730	5	6	1
pax9.2099	24	24	0
gli3.2152	10	10	0
bcl11a.2554	15	15	0
zfhx1b.2928	11	27	16
sall3.2991	7	7	0
sall3.3016	25	27	2
pbx3b.3198	31	33	2
pbx3b.3213	20	20	0
foxp2.3505	12	13	1
foxp2.3548	13	14	1
ebf3.3817	7	8	1
barhl2.3939	10	10	0
emx2.4548	11	11	0
emx2.4559	6	6	0
znf503.4953	11	11	0
shox2.5643	3	3	0
zic1.5745	23	24	1
lmo1.6396	5	5	0
irx1/2/4.6677	7	9	2
irx1/2/4.6679	26	26	0
irx1/2/4.6714	6	7	1
irx1/2/4.6731	6	8	2
irx1/2/4.6739	12	12	0
irx3/5/6.7193	18	18	0
tshz3.7673	2	2	0
pou3f1.7785	8	8	0
_nr2f2_.8107	23	22	1
_nr2f2_.8470	2	2	0
tshz1.8804	11	12	1
pou3f2b.9802	15	15	0
znf703.10890	2	2	0
znf703.10897	3	3	0
dachd.11206	12	12	0
evi1.11239	19	19	0



```
macaque           TGATGCGCCCATAAATCAACATGACAGGCCTGTTTCACA-GTCTCAGACAGTGCTTGTCA 116
bushbaby          AGATGCGCCCATAAATCAACATGACAGGCCTGTTTCACA-GTCTCAGACAGTGCTTGTCA 116
chimp             TGATGCGCCCATAAATCAACATGACAGGCCTGTTTCACA-GTCTCAGACAGTGCTTGTCA 116
human             TGATGCGCCCATAAATCAACATGACAGGCCTGTTTCACA-GTCTCAGACAGTGCTTGTCA 116
orangutan         TGATGCGCCCATAAATCAACATGACAGGCCTGTTTCACA-GTCTCAGACAGTGCTTGTCA 116
squirrel          TGATGCGCCCATAAATCAACATGACAGCCCTGTTTCACA-GTCTCAGACAGTGCTTGTCA 116
cow               TGATGCGCCCATAAATCAACATGACAGGCCTGTTTCACA-GTCTCAGACAGTGCTTGTCA 116
elephant          TGATGCGCCCATAAATCAACATGACAGGCCTGTTTCACA-GTCTCAGACAGTGCTTGTCA 116
rat               TGATGCGCCCATAAATCAACATGACAGGCCTGTTTCACA-GTCTCAGACAGTGCTTGTCA 116
mouse             TGATGCGCCCATAAATCAACATGACAGGCCTGTTTCACA-GTCTCAGACAGTGCTTGTCA 116
cat               TGATGCGCCCATAAATCAACATGACAGGCCTGTTTCACA-GTCTCAGACAGTGCTTGTCA 116
horse             TGATGCGCCCATAAATCAACATGACAGGCCTGTTTCACA-GTCTCAGACAGTGCTTGTCA 116
armadillo         TGATGCGCCCATAAATCAACATGACAGGCCTGTTTCACA-GTCTCAGACAGTGCTTGTCA 116
dog               TGATGCGCCCATAAATCAACATGACAGGCCCGTTTCACA-GCCTCAGACAGTGCTTGTCA 116
bat               TGATGCGCCCATAAATCAACATGACAGGCCTGTTTCACA-GCCTCAGCCAGTGCTTGTCA 116
platypus          TGATGCGCCCATAAATCAACATGACAGGCCTGTTTCACT-GTCTCAGACAGTGCTTGTCA 116
chicken           TGATGCGCCCATAAATCAACATGACAGGTCTGTTTCACA-GTCTCAGACAGCGCTTGTCA 116
opossum           TGATGCGCCCATAAATCAACATGACAGGCCTGTTTCACA-GTCTCAGACAGTGCTTGTCA 116
frog              TGATGCGCTCATAAATCAACATGACAGTTGCGTTTCGCA-GCCTCAGACACCGCTTGTCA 116
shark             TGATGCACTCATAAATCAACATGACAGGCTTGCTTCTCC-TGCTCAGACAGCACTTGTCA 116
zfish             CGTCGCCTCCATAAATCAATGCGACAGGACAGTTTCTCCTGCCTCTGACAGAAGCTGTCA 119
medaka            TGATGCTCCCATAAATCAATGTGACAGCCGAGCTCCGCG-GCCTCAGACAGGCCTTGTCA 117
tetraodon         TGATGCTCCCATAAATCACCCTGACAGCGGAGCTCCTCG-GCCTCAGATAGGGCCTGTCA 116
fugu              TGATGCTCCCATAAATCAACGTGACAGC--------------CTCAGATAGGACCTGTCA 103
stickleback       TGATGCTCCCATAAATCAACGTGACAGC--------------CTCAGATAGGACCTGTCA 103
                   *  **   *********    *****               *** *  *     *****

human MCAST cluster       CCATAAATCAACATGACAG
zfish MCAST cluster       CCATAAATCAATGCGACAGGACAGTTTCTCCTGCCTCTGACAGAAGCTGTCA
```

Figure 4.1: The emergence of a compensatory Meis/Pknox motif in teleost *foxp1*.892 may have relaxed selection and caused the observed distance divergence. A portion of the ClustalW2 alignment of 25 orthologues of *foxp1*.892 is shown. The MCAST clusters from human and zebrafish are shown below the alignment. The Pbx-Hox site (red box, red arrow) is conserved in all species. There are two Meis/Pknox motifs with 5/6 bases conserved in all species (blue boxes, blue arrows). In tetrapods and shark, the 3' Meis/Pknox motif deviates from the consensus sequence in the first base. There was a substitution at this position in the lineage leading to teleost fish, changing the motif from TTGTCA>CTGTCA, the best match to the Meis/Pknox consensus. Resultantly, this position was detected by MCAST in the zebrafish, but not the human, orthologue. Subsequently, the lineage leading to zebrafish substituted the last base of the 5' Meis/Pknox motif to deviate from the consensus (CTGTCA>CTGTCG), rendering this motif below MCAST's match threshold. This suggests relaxed selection after the emergence of a compensatory site.



```
opossum          TAGGGAATGACCCA-GTTGGCAGGAAGTTCTAATGACAGATGATGCAAACTCATTTATCA 59
dog              TAGGGAATGACCCA-GTTGGCAGGAAGTTCTAATGACAGATGATGCAAACTCATTTATCA 59
cat              TAGGGAATGACCCA-GTTGGCAGGAAGTTCTAATGACAGATGATGTAAACTCATTTATCA 59
horse            TAGGGAATGACCCA-GTTGGCAGGAAGTTCTAATGACAGATGATGCAAACTCATTTATCA 59
human            TAGGGAATGACCCA-GTTGGCAGGAAGTTCTAATGACAGATGATGCAAACTCATTTATCA 59
chimp            TAGGGAATGACCCA-GTTGGCAGGAAGTTCTAATGACAGATGATGCAAACTCATTTATCA 59
macaque          TAGGGAATGACCCA-GTTGGCAGGAAGTTCTAATGACAGATGATGCAAACTCATTTATCA 59
elephant         TAGGGAATGACCCA-GTTGGCAGGAAGTTCTAATGACAGATGATGCAAACTCATTTATCG 59
squirrel         TAGGGAATGACCCA-GTTGGCAGGAAGTTCTAATGACAGATGATGCAAACTCATTTATCA 59
cow              TAGGGAATGACCCT-GTTGGCAGGAAGTTCTAATGACAGATGATGGAAACTCATTTATTA 59
rabbit           TAGGGAATGACCCA-GTTGGCAGGAAGTTCTAATGACAGATGATGCAAACTCATTTATCA 59
armadillo        TAGGGA-TGACCCA-GTTGGCAGGAAGTTCTAATGACAGATGATGCATACTCATTTATCA 58
chicken          TAGGGAATGACCCT-GTTGGCAGGAAGTTCTAATGACAGATGATGCAAACTCATTTATCA 59
bat              TAGGGAATGACCCA-GTTGGCAGGAAGTTCTAATGACAGATGATGCAAACTCATTTATCA 59
frog             TAGTAAGTGACCT--GTTGGCAGGAAGTCTGAATGACACACGATTCAAACTCATTTATCA 58
fugu             TAAGCAGTGACCTGGGCTGGCAGGAAGCTCGCCTGACAGACAATGCAAACTCATTTATCA 60
stickleback      TAACCAGTGACCTGGGCTTGCAGGAAGCCGGCCTGCCAGACAATACAAACTCATTTATCA 60
tetraodon        TAAGCAGTGACCTGCGCT----GGACGCTCGGCTAGCAGACAATACAAACCCATTTATCA 56
medaka           TAAGCAGTGACCTGGGCTGGCAGGAATCTGCTCTGACAGACAATACAAACTCATTTATCA 60
zfish            TCGAGACTGACCTGTNGTGACAGAAAACCCCACAGACGCACAACACACACTCATTTATCA 60
                 *    * *****     *    * *           *  *  *   * ** *******


human MCAST cluster                               TGACAGATGATGCAAACTCATTTATCA
zfish MCAST cluster               TGACAGAAAACCCCACAGACGCACAACACACACTCATTTATCA
```

Figure 4.2: A compensatory mutation may have preserved a single high-affinity Meis/Pknox motif in zebrafish *zfhxb1*.2928 and caused the observed distance divergence. A portion of the ClustalW2 alignment of 20 orthologues of *zfhxb1*.2928 is shown. The MCAST clusters from human and zebrafish are shown below the alignment. The Pbx-Hox site (red box, red arrow) is conserved in all species except elephant and cow; these may correspond to miscalled bases. The Meis/Pknox site is conserved in most species but has been lost in several teleosts, including zebrafish (blue box, blue arrow). Nevertheless, the zebrafish orthologues contains a good match to the Meis/Pknox consensus in its corresponding MCAST cluster, 10 bp 5' of the ancestral motif pair. This zebrafish-specific Meis/Pknox motif may be the product of compensatory mutations or lineage-specific insertions between the sites.

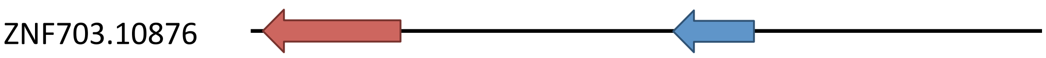
4.4.2 FIMO detects 80 candidate hindbrain enhancers from CONDOR

Many previously assayed hindbrain enhancers were not discovered using MCAST, suggesting that the match threshold used by MCAST to detect significant clusters is too strict. In an attempt to detect more candidate hindbrain enhancers, significant clusters were also detected in the full set of CONDOR CNEs using FIMO (figure 4.3). This method detects 110 mammal-fugu CNEs with conserved hb+ grammar, 94 of which are conserved in zebrafish. 13 of these had been tested previously: 3 (of 7) are positives from chapter 2; and 9 (of 15) are positives from the previous study[157]. Again, the negative element *pou3f2b*.9802 was detected. This demonstrates that FIMO rediscovers more known hindbrain enhancers than does MCAST, but still does not recover them all. The FIMO output file can be found in the appendix (8.3.4).

Of the 81 remaining CNEs discovered by FIMO, there were three pairs of duplicated CNEs, one of each pair associating with the *znf503.1* and *znf503.2* genes. These pairs each correspond to a single CNE in zebrafish, reducing the total to 78 candidates. Additionally, mLAGAN[98] alignments of the *hoxa* and *hoxb* clusters discovered 14 and 10 CNEs, respectively (data not shown). Only 2 of these CNEs contained the hindbrain grammar, both associating with the *hoxa* cluster (*hoxa*.12003 and *hoxa*.12006), bringing the total number of candidates to 80. The conservation of these sites in all 80 orthologues was confirmed using alignments (see methods), which are available in the appendix (8.3.4).

The 80 candidates occur at 44 distinct loci as defined in CONDOR. Trans-dev genes known to be expressed in the hindbrain are found at 36 of these loci (81% of loci). The remaining genes are either known to be expressed in other tissues (3 in somites, 2 in the pharyngeal arches, and 1 in the hatching gland). The 2 remaining loci had no *in situ* hybridisation data stored in ZFIN. This demonstrates that this grammar can locate elements associating with genes known to be expressed in the hindbrain, suggesting that it can identify hindbrain enhancers. The conservation of sites in at least 16

Motif	CNE and species	start	stop	strand	score	p-value	q-value	matched sequence
PBX-HOX	CRCNE00010876_Human	193	202	-	12.8	9.46E-05	1.98E-01	TGATTTATTG
PBX-HOX	CRCNE00010876_Mouse	193	202	-	12.8	9.46E-05	1.98E-01	TGATTTATTG
PBX-HOX	CRCNE00010876_Dog	193	202	-	12.8	9.46E-05	1.98E-01	TGATTTATTG
PBX-HOX	CRCNE00010876_Fugu	197	206	-	12.8	9.46E-05	1.98E-01	TGATTTATAG
MEIS	CRCNE00010876_Human	224	229	-	12.0	2.37E-04	2.63E-01	CTGTCA
MEIS	CRCNE00010876_Mouse	224	229	-	12.0	2.37E-04	2.63E-01	CTGTCA
MEIS	CRCNE00010876_Dog	224	229	-	12.0	2.37E-04	2.63E-01	CTGTCA
MEIS	CRCNE00010876_Fugu	228	233	-	12.0	2.37E-04	2.63E-01	CTGTCA



```
fugu             TCTATAAATCACACTCGTACACATGCCCCCAATGACAGCTAAAT-CAACA-TGGAGTAAT 253
tetraodon        TCTATAAATCATGCTCGTACACATGCCCCCAATGACAGCTAAAT-CAACA-TGGAGTAAT 253
stickleback      TCTATAAATCACGCTCGAACACATGCCCCCAATGACAGCTAAAT-CAACA-TGGAGTAAT 253
medaka           TCTTTAAATCACATCGATCCCCAAGCCGCCAATGACAGCTAAAT-CAACA-TGAAGTACT 253
zfish            TCTATAAATCATGTTGAAACACATGCCCCCAATGACAGCTAAAT-CAACA-TAGAGTAAT 251
dog              TCAATAAATCATGTTCAAACACATGCCTGTAATGACAGGTAAAT-CAACAATGAAGTAAT 250
bushbaby         TCAATAAATCATGTTCAAACACATGCCTGTAATGACAGGTAAAT-CAACAATGAAGTAAT 250
armadillo        TCAATAAATCATGTTCAAACACATGCCTGTAATGACAGGTAAAT-CAACAATGAAGTAAT 250
horse            TCAATAAATCATGTTCAAACACATGCCTGTAATGACAGGTAAAT-CAACAATGAAGTAAT 250
elephant         TCAATAAATCATGTTCAAACACATGCCTGTAATGACAGGTAAAT-CAACAATGAAGTAAT 250
bat              TCAATAAATCATGTTCAAACACATGCCTGTAATGACAAGTAAAT-CAACAATGAAGTAAT 250
orangutan        TCAATAAATCATGTTCAAACACATGCCTGTAATGACAGGTAAAT-CAACAATGAAGTAAT 250
squirrel         TCAATAAATCATGTTCAAACACATGCCTGTAATGACAGGTAAAT-CAACAATGAAGTAAT 250
cat              TCAATAAATCATGTTCAAACACATGCCTGTAATGACAGGTAAAT-CAACAATGAAGTAAT 250
chimp            TCAATAAATCATGTTCAAACACATGCCTGTAATGACAGGTAAAT-CAACAATGAAGTAAT 250
human            TCAATAAATCATGTTCAAACACATGCCTGTAATGACAGGTAAAT-CAACAATGAAGTAAT 250
macaque          TCAATAAATCATGTTCAAACACATGCCTGTAATGACAGGTAAAT-CAACAATGAAGTAAT 250
cow              TCAATAAATCATGTTCAAACACATGCCTGTAATGACAGGTAAAT-CAACAATGAAGTAAT 250
rat              TCAATAAATCATGTTCAAACACATGCCTGCAATGACAGGTAAAT-CAACAATGAAGTAAT 250
mouse            TCAATAAATCATGTTCAAACACATGCCTGTAATGACAGGTAAAT-CAACAATGAAGTAAT 250
opossum          TCAATAAATCATGTCCCAACACATGCCTGTAATGACAGGTAAAT-CAACAATGAAGTAAT 250
chicken          TCAATAAATCATGTTCAAACACATGCCTCTAATGACAGGTAAAT-CAACAATGAAGTAAT 250
shark            CCCATAAATCATAGCACACAACAC-CCACTGGTGACAGATTAATGTAAC--CAAAATAAT 252
frog             GCAATAAATCACACTGGCACACATGCCATTCCTGACAGGCAGGG-CAGCTACGTTGTAAT 271
                  *  *******          **  **     *****         * *       ** *
```

Figure 4.3: FIMO detects significant clusters of Pbx-Hox and Meis/Pknox motifs conserved amongst osteichthyans. The use of prealigned sequences together with the FIMO algorithm from CONDOR allows clusters of conserved sites to be detected. The table (top) shows an example of a significant cluster detected by FIMO. Each line shows a single hit for either the Pbx-Hox or Meis/Pknox PWM (motif) within 4 orthologues of *znf703*.10876 (CNE and species). The location (start/stop), orientation (strand), three parameters describing the quality of the match (score, p-value, q-value) and the composition of the motif (matched sequence). The score indicates the strength of the match and is calculated by summing the values of the matched bases from the PWM. The p-value states the probability that a random motif would match this position in the query with the same or better score. The q-value is the estimated false discovery rate. A ClustalW2 alignment of 24 orthologues of this CNE is shown below, highlighting the conserved Pbx-Hox (red box, red arrow) and Meis/Pknox (blue box), demonstrating the conservation of these sites amongst all gnathostomes (osteichthyans plus elephant shark). The FIMO output and annotated alignments for all 80 candidates can be found in the appendix (8.3.4)

4.4.3 The hindbrain syntax predicts zebrafish hindbrain enhancers

74 of the 80 candidates were cloned in to the expression vector (see methods). 6 of the candidate elements failed to amplify during PCR. Of the 74 elements subjected to the functional assay (see methods), 66 (89%) act as hindbrain enhancers consistently and reproducibly during the first 3 days of zebrafish development (figure 4.4, table 4.2), a 3-fold enrichment compared to the set assayed in chapter 2 (CNEs containing TGATNNAT). None of the elements were considered active in hindbrain at 1 dpf, perhaps reflecting the limitations of transient transgenesis. 62/66 elements were considered to be active in hindbrain at 2 dpf and 43/66 active at 3 dpf, with 29/66 active at both time points. Hindbrain is one of the most common tissues for 48/66 elements at 2 dpf, and for 34/66 elements at 3dpf. Images of representative embryos expressing each construct can be found in the appendix (8.2.4).

These enhancers are highly heterogeneous in their functions. 44/66 hindbrain enhancers generate appreciable expression in at least one other tissue between 1-3 dpf (table 5.2), suggesting that multifunctionality is to be frequently found in these CNEs. Aside from hindbrain, the most common tissues observed are forebrain, midbrain, spinal cord (most commonly rohon-beard cells), peripheral neurons (most commonly cranial ganglia), eye (both retina and lens), trunk muscle cells and heart. Furthermore, 23/66 elements switch their specificity from one tissue to another between the time-points.

Considering the domains of hindbrain expression, there are a number of elements where expression is targeted to subdomains of the hindbrain, as noted in chapter 2 and in previous work[157]. Several enhancers have their anterior limit coincident with a rhombomere boundary, for example *nr2f2*.8394 posterior of the r1r2 boundary (figure 4.4 D), *znf703*.10876 posterior of the r2r3 boundary (figure 4.4 J) and *znf503*.10105 expressed posterior of the r4r5 boundary (figure 4.4 M). More commonly, elements exhibit a peak of expression at a particular anteroposterior position but do not exhibit reporter expression restricted to specific rhombomeres. Examples include *fign*.5158 with its peak in the posterior hindbrain at approximately r6/r7 (figure 4.4 B) *foxp1*.892 with its peak in r2 (figure 4.4 C) and *pax2*.174 with its peak in r4 (figure 4.4 E).

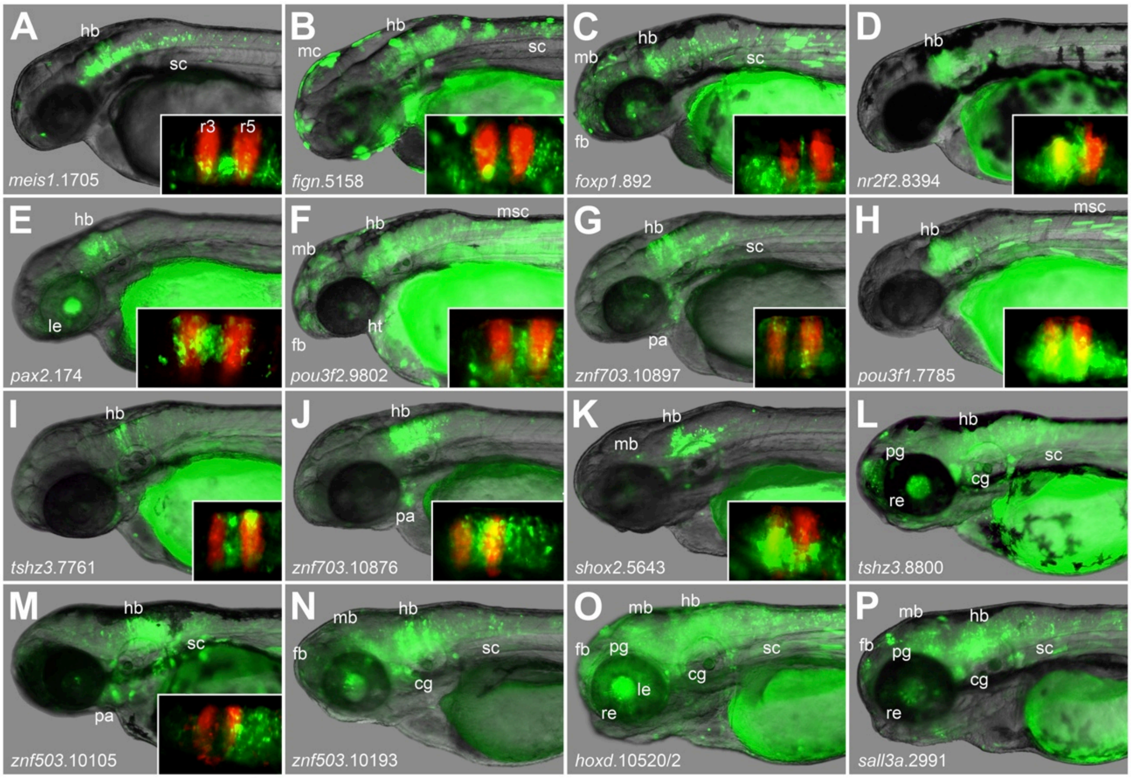


Figure 4.4: Conserved Pbx-Hox and Meis/Pknox motifs predict hindbrain enhancers accurately. Images show F^0 transgenic embryos between 2 and 3 dpf expressing Egfp under the control of CNEs. Insets show comparison with mCherry in rhombomeres 3 and 5. Constructs drive expression in hindbrain and in other tissues *meis1*.1705 (A) in hindbrain and spinal cord; *fign*.5158 (B) in hindbrain, spinal cord and melanocytes; *foxp1*.892 (C) in the central nervous system; *nr2f2*.8394 (D) in hindbrain; *pax2*.174 (E) in hindbrain and lens; *pou3f2*.9802 (F) in the central nervous system, heart and muscle; *znf703*.10897 (G) in hindbrain, spinal cord and pharyngeal arches/neural crest; *pou3f1*.7785 (H) in hindbrain; *tshz3.1*.7761 (I) in hindbrain; *znf703*.10876 (J) in hindbrain and pharyngeal arches/neural crest; *shox2*.5643 (K) in hindbrain; *tshz1*.8800 (L) in hindbrain, spinal cord, retina, pineal gland and cranial ganglia; *znf503*.10105 (M) in hindbrain, spinal cord and pharyngeal arches; *znf503*.10193 (N) in the central nervous system and cranial ganglia; *hoxd*.10520/1 (O) in the central nervous system, lens, retina, pineal gland and cranial ganglia; and *sall3a*.2991 (P) in the central nervous system, retina, pineal gland and cranial ganglia. fb: forebrain; mb: midbrain; hb: hindbrain; sc: spinal cord; pa: pharyngeal arches/neural crest; cg: cranial ganglia; mc: melanocytes; msc: trunk muscle cells; le: lens; re: retina; pg: pineal gland.

Table 4.2: hindbrain expression driven by CNEs containing the hb+ grammar. The table shows the name of CNEs containing the hb+ grammar (CNE) and the percentage of hindbrain positive embryos as a proportion of the total number injected (hb+/total) and as a proportion of GFP positive embryos (hb+/total). The most common regions observed for each element are also displayed (common regions). Sequences considered as hindbrain enhancers (those expressing in at least 20% of the injected embryo's hindbrains) are indicated in green. Other enhancers are indicated in red.

	2 dpf			**3 dpf**		
CNE	**hb+ /total**	**hb+ /GFP+**	**common regions**	**hb+ /total**	**hb+ /GFP+**	**common regions**
auts2.8975	63.3%	100.0%	hb	13.3%	57.1%	none
barhl2.3939	6.3%	30.0%	none	2.8%	33.3%	none
bnc2.8570	73.3%	100.0%	hb	36.7%	100.0%	cns
cst.9947	15.0%	20.0%	ht	3.6%	4.8%	ht
dachd.229	3.3%	100.0%	none	3.3%	100.0%	none
dachd.232	67.1%	94.0%	hb	8.5%	23.8%	none
dachd.240	64.2%	95.6%	hb	41.5%	90.0%	hb
emx2.4548	6.7%	18.2%	sk	6.7%	22.2%	sk
esrrb.9004	55.3%	84.0%	hb	30.3%	100.0%	hb
esrrb.9024	76.9%	98.6%	hb	76.9%	98.6%	hb
esrrb.9032	44.7%	100.0%	mb, hb	29.0%	100.0%	mb, hb
evi1.10711	33.3%	100.0%	hb	13.3%	100.0%	none
evi1.11239	0.0%	0.0%	ht	0.0%	0.0%	ht
fign.5158	38.1%	100.0%	hb, sk	33.3%	100.0%	hb, sk
fign.5161	57.4%	96.9%	hb	17.5%	63.6%	msc
foxd3.330	45.7%	100.0%	cns	52.3%	100.0%	cns
foxp1.892	31.4%	90.0%	hb	12.7%	75.0%	msc
foxp1.906	60.7%	100.0%	hb	20.0%	92.3%	nc, ht
foxp2.3531	36.7%	52.4%	eye	0.0%	0.0%	ht, pf
hmx2.9711/2	68.3%	81.2%	sc	11.1%	47.1%	ne, ey, pa
hmx2.9741	11.5%	47.4%	ht	10.5%	80.0%	ht
hoxa.12003	50.0%	100.0%	hb	4.0%	25.0%	ey, ht, pf
hoxa.12006	48.8%	66.7%	ey, pa	8.1%	27.3%	ey, pa, pf
hoxd.10470	48.6%	100.0%	hb	13.8%	66.7%	none
hoxd.10483	36.6%	82.0%	mb, hb	7.2%	63.6%	msc, ht
hoxd.10498	55.3%	84.0%	fb, hb, sk, pa	15.6%	83.3%	msc, ht
hoxd.10520/1	50.0%	100.0%	cns, ne, ey	56.7%	100.0%	cns, ey, ne
irx1/2/4.6679	2.0%	10.0%	msc	51.1%	100.0%	hb
irx3/5/6.7352	41.5%	48.6%	ht	25.7%	100.0%	fb, hb, sc, ey
lmo1.6396	76.6%	94.4%	fb, msc, ht	35.3%	88.2%	msc, ht
maf.11498	51.5%	87.5%	hb	42.7%	100.0%	hb, sc
meis1.1705	82.1%	100.0%	hb	66.7%	100.0%	hb, sc
meis2a.1042	59.5%	100.0%	hb	51.3%	100.0%	hb
meis2a.1089	58.6%	100.0%	hb	31.5%	100.0%	hb
_nr2f2_.8394	66.7%	100.0%	hb	17.6%	100.0%	none
_nr4a2b_.2390	50.0%	100.0%	hb	23.5%	80.0%	hb

fb: forebrain; mb: midbrain; hb: hindbrain; sc: spinal cord; cns: central nervous system; pa: pharyngeal arches/neural crest; ne: other neuron; ey: eye; msc: trunk muscle cells; ht: heart; sk: skin.

Table 4.2: hindbrain expression driven by CNEs containing the hb+ grammar (continued). Hindbrain enhancers are indicated in green. Other enhancers are indicated in red.

	2 dpf			3 dpf		
CNE	**hb+ /total**	**hb+ /GFP+**	**common regions**	**hb+ /total**	**hb+ /GFP+**	**common regions**
***pax1a*.10260**	43.8%	93.3%	hb	26.7%	72.7%	hb
***pax2a*.135**	50.0%	73.9%	fb	19.4%	54.5%	msc
***pax2a*.174**	96.6%	100.0%	hb, ey	61.7%	100.0%	hb, ey
***pax2a*.90**	47.3%	63.4%	msc	27.5%	37.9%	msc
***pbx3b*.3200**	56.5%	100.0%	hb	33.0%	100.0%	hb
***pbx3b*.3211**	78.8%	100.0%	hb	37.5%	100.0%	hb
***pbx3b*.3212/3**	62.5%	100.0%	mb, hb	37.1%	92.9%	mb, hb
***pbx3b*.3244**	62.2%	100.0%	cns, ey, pa	28.2%	100.0%	cns, pa
***pbx3b*.3253**	52.1%	100.0%	hb, sc	12.5%	100.0%	none
***pou3f1*.7785**	92.5%	100.0%	hb, msc	77.5%	100.0%	hb, msc
***pou3f2b*.9802**	65.1%	100.0%	hb, sc	45.3%	85.0%	msc
***pou6f2*.1568**	12.5%	100.0%	none	35.7%	100.0%	hb
***sall3a*.2991**	60.0%	100.0%	hb, ne	43.3%	92.9%	ne
***sall3a*.3016**	39.6%	95.0%	hb, sc, nc, pf	20.0%	66.7%	msc
***satb1*.5966**	38.4%	71.6%	hb	3.5%	40.0%	ht
***shox*.11103**	56.5%	100.0%	hb, sc	18.3%	100.0%	none
***shox2*.5627**	0.0%	0.0%	none	46.7%	100.0%	hb
***shox2*.5643**	74.5%	100.0%	hb	19.6%	100.0%	none
***sox14*.11250**	58.7%	61.4%	msc	9.9%	77.8%	ey, msc, pa
***sox6*.2298**	0.0%	0.0%	ht	6.7%	10.5%	ht
***tcf7l2*.5398**	45.8%	66.0%	fb	29.9%	57.1%	fb
***tshz1*.8799**	35.7%	40.0%	ey	36.7%	84.6%	ey, msc, ht
***tshz1*.8800**	45.0%	100.0%	hb, ne	51.3%	100.0%	hb
***tshz1*.8804**	56.8%	100.0%	hb	57.1%	100.0%	hb
***tshz2*.8749**	50.0%	80.0%	fb	56.0%	80.0%	fb
***tshz3*.7689**	1.9%	100.0%	none	10.4%	71.4%	none
***tshz3*.7761**	50.8%	100.0%	hb, pa	56.4%	91.2%	hb, pa
***uncx*.9830**	17.0%	88.9%	none	20.5%	100.0%	hb
***zic1*.5745**	61.3%	100.0%	hb, sc, msc	63.3%	100.0%	hb, sc, msc
***zic1*.5763**	31.7%	57.1%	fb	14.0%	92.3%	fb
***znf503*.10049**	50.0%	100.0%	hb	50.0%	100.0%	hb
***znf503*.10105**	81.0%	100.0%	hb, pa	94.1%	100.0%	hb, pa
***znf503*.10147**	52.3%	100.0%	hb	56.0%	100.0%	hb
***znf503*.10156**	47.1%	100.0%	hb	33.3%	100.0%	hb
***znf503*.10193**	39.5%	100.0%	hb	9.0%	29.0%	ey
***znf503*.10196**	28.6%	54.5%	ne, msc, pa	19.7%	41.2%	msc, pa
***znf703*.10876**	48.4%	100.0%	hb, pa	40.0%	100.0%	hb, pa
***znf703*.10897**	31.3%	100.0%	hb, pa	9.7%	100.0%	none

fb: forebrain; mb: midbrain; hb: hindbrain; sc: spinal cord; cns: central nervous system; pa: pharyngeal arches/neural crest; ne: other neuron; ey: eye; msc: muscle; ht: heart; sk: skin.

4.5 DISCUSSION

In this chapter, matches to the hindbrain grammar were detected by applying MCAST and FIMO to prealigned sequences from CONDOR or mLAGAN alignments of the *hox* clusters. MCAST has a high match threshold, and detects only 39 human-zebrafish conserved matches to the hb+ grammar. However, for the most part these instances of the hb+ grammar have been subject to strong purifying selection; this has maintained not only the composition and orientation of the ancestral motifs but also their spacing, typically within narrow limits. In the two cases where the inter-motif distance varies considerably, compensatory mutations appear to preserve high-affinity Meis/Pknox sites well within the 100bp limit suggested in chapter 3. These compensatory mutations explain the apparent divergence in inter-motif distance between the orthologues and suggest selection for high-affinity binding sites rather than absolute distance in these cases. Selection for motif composition in enhancers, rather than spacing, is consistent with the finding that selection for the composition of Bicoid and Krüppel binding sites in *Drosophila* can create the illusion of a conserved grammar during enhancer evolution simulations[172]. Indeed, the apparent selection for spacing/orientation in CNEs could be an artefact caused by the detection of these sequences by identity; elements subject to binding site turnover (as has been commonly noted in enhancers from vertebrates[158, 255-257] and *Drosophila*[53, 67]) would not be classified as CNEs. This could explain why cases where inter-motif distances vary appreciably are uncommon. Nevertheless, CNEs are an interesting group of sequences because in the majority of cases they appear to have resisted such turnover and rearrangement of binding sites.

The hindbrain grammar has been very successful at identifying additional hindbrain enhancers. 74/80 candidate sequences were tested (92.5%), a high proportion of candidates compared with a previous study using only Pbx-Hox sites[157]. 66/74 hb+ grammar CNEs upregulate hindbrain expression consistently during transient transgenesis (89%). A comparable study attempted to identify hindbrain enhancers in the human genome using many motifs using a machine-learning enhancer classifier[120]. Contrastingly, the authors tested only a small fraction of their predicted enhancers (55/40,000, 0.14%) and also tested these transiently. However, the value they provide for the efficacy of their classifier (88%) comes from a subset of 30/34 of these that were subsequently studied in stable transgenic lines. In contrast, the approach used herein appears to have greater specificity: of the 75 candidates tested, 89% are active during transient transgenic assays. The greater specificity of this model might arise from the strict conservation requirements of a specific grammar. However, this approach appears to have sacrificed sensitivity for specificity, though it has identified a relatively small number of elements that can be practically subjected to functional validation.

In a previous publication *pou3f2b*.9802 was not regarded as an enhancer[157]. However, due to its significant match to the hindbrain grammar when using either MCAST (table 4.1) or FIMO it was decided to include this element as a candidate. When a longer version of this element was cloned, it does indeed act as a hindbrain enhancer (figure 4.5 F, table 4.2). This indicates that whilst some CNEs tested in these screens were hindbrain negative, this may be because essential sequence was not contained in the cloned fragment. This suggests that additional sequence is necessary for the function of these hindbrain enhancers and that Pbx-Hox and Meis/Pknox sites alone are not sufficient for hindbrain enhancer activity. Nevertheless, the sequences with this grammar at their functional core may still act as hindbrain enhancers *in vivo*.

Trends observed amongst the full set of hindbrain enhancers containing the hb+ grammar may now be assessed. This set is composed of 6 CNEs from chapter 2 and 66 CNEs from this chapter. In 34/72 cases (47%), the characteristic patterns driven by these elements are highly reproducible between embryos. In these cases it appears that fully functional regulatory elements have been cloned and tested, shown by their ability to upregulate gene expression in a consistent manner irrespective of the site of insertion. This can be seen by comparing the patterns driven by the same element in two different embryos, such as one wild type (main image) and one Tg(*egr2b*:*kalta4*;*uas*:*mCherry*) (inset) embryo (figure 4.4). Whilst all these sequences were considered to act as hindbrain enhancers (i.e. they drive reporter expression in at least 20% of injected embryos), in 38/72 cases these CNEs exhibited high variation in pattern between embryos, high mosaicism, or were simply expressed throughout the hindbrain. These cloned fragments may contain only part of a more specific enhancer, or alternatively these enhancers may just be active in wide domains. The establishment of stable transgenic lines could distinguish between these possibilities: this would overcome mosaicism but it would not circumvent intrinsic properties of the fragments being tested, such as cell-type specificity or sensitivity to endogenous regulatory regions near the insertion site (positional effects). Positional effects could be counterracted by targeted insertion of the reporter cassette, mediated by PhiC31 integrase[258] or CRISPR-Cas9[259]. However, this would introduce a consistent bias that might be considered undesirable when testing sets of sequences for a given tissue specificity.

Many of these enhancers exhibit patterns restricted to certain anteroposterior domains which may or may not be coincident with segment interfaces. Several expression pattens are reminiscent of Hox expression patterns, implying activation by particular Hox proteins or paralogous groups. For example, *nr2f2*.8394 overlaps the expression pattern of *hoxa2b* (figure 4.4 D), *znf703*.10876 of *hoxb2a* (figure 4.4 J) and *znf503*.10105 of *hoxb3a* (figure 4.4 M)[185, 186]. Such enhancers may control the expression of segment specific genes. Since many of these genes are themselves transcription factors, these enhancers may be responsible for the activation of a cascade of transcriptional regulatory events, imparting unique molecular and morphological identities to the segments. However, in many cases segment specific enhancers are discovered near genes which are not expressed in a segment-specific manner, for example *znf503*, *meis2a*, *tshz3* and *foxd3*. Perhaps these enhancers generate quantitative differences in expression of their target genes between rhombomeres. This again may prefigure patterning events and control segment identity. As such, these CNEs may provide a means for the interpretation of the collinear *hox* code by downstream genes by controlling segment specific expression and dosage. This may ensure the appropriate specification and interconnectivity of developing neurons and underlie the phylotypic pattern of hindbrain neuroanatomy shared by all gnathostomes, and to a lesser extent all vertebrates[184, 186].

Aside from enhancers driving expression restricted to particular antero-posterior domains, apico-basal restriction is also noticable for many elements. The basal hindbrain contains a pool of neural progeniters, and as these differentiate they move to the periphery; as such, the more mature neurons are generally located laterally. Small, isolated cells, perhaps neural progenitors, are Egfp positive in embryos injected with many different elements (appendix). There are many neural progenitors in the hindbrain at the time-points being studied. Indeed, HoxB2 has been shown to play a role in preventing the differentiation of neurons, thereby maintaining progenitor populations[205]. Perhaps medially-restricted elements are active in neural progenitors, implicating these enhancers in the maintenance a GRN controlling a proliferative state. Conversely, those elements active in lateral domains (for example, *hoxd*.10482, figure 2.1 C) appear to be expressed in maturing neurons, suggesting possible roles for these elements in activating genes downstream of EGBs for differentiation, specification and axon targeting.

This screen of elements containing Pbx-Hox and Meis/Pknox sites also provides a list of putative *hox* target genes, as has been noted before in a screen of CNEs containing Pbx-Hox sites[157]. Previous work has identified rhombomere-specific genes using microarray experiments[207], but these may be either direct or indirect targets of Hox proteins. The enhancers identified herein contain appropriate sites for Hox proteins and their cofactors suggesting that their associated genes are direct Hox targets. The putative gene-regulatory interactions identified herein require validation by perturbation experiments, perhaps using morpholinos for specific inputs.

Multifunctionality appears to be common in this set of sequences. 66% of the hindbrain enhancers are active in at least one other tissue over the timecourse observed. This multifunctionality appears to reflect the roles of the factors putatively binding to these elements. For example, several Hox proteins are expressed in and positionally identify somites[260]. Additionally Meis1 is expressed in the heart where it regulates angiogenesis[238, 240]. Indeed, both heart and muscle expression were commonly observed when these sequences were tested (table 5.2). Furthermore, Meis, Prep and Pbx family proteins are expressed in the developing zebrafish forebrain and midbrain as well as the hindbrain (ZFIN gene expression database). Forebrain and midbrain expression were also commonly observed during this screen (table 5.2). Therefore, the multifunctionality of these CNEs is consistent with the expression patterns of Pbx, Hox and Meis/Pknox proteins and their roles in the development of multiple tissues. In some cases, multifunctionality may also be noise caused by cloning incomplete enhancers or isolating them from their immediate genomic context. This spatiotemporal heterogeneity suggests that these enhancers share a core functional grammar that is further refined by different mechanisms, generating the observed diversity in expression patterns. Indeed, the assayed elements are often hundreds of bases in length and are conserved at high identity, strongly implying the presence of additional binding sites.

Multifunctionality has previously been hypothesised to account for the startling conservation of vertebrate CNEs[90]. In this hypothesis, multiple overlapping TFBSs cause hundreds of base pairs to be maintained by purifying selection. The data in this chapter are consistent with this hypothesis, and suggest that CNEs are actually conjoined or juxtaposed enhancers of different tissues containing diverse TFBSs. This may contribute to their extensive conservation over hundreds of base pairs.

4.6 CONCLUSIONS

The strong association between conserved and proximal Pbx-Hox and Meis/Pknox motifs has been further confirmed by applying this grammar predictively. This grammar identifies hindbrain enhancers with high accuracy. However, the functions of these enhancers are highly heterogeneous and suggest that they each contain unique binding sites to specifiy their characteristic expression patterns.

In summary:

- CNEs containing the hb+ grammar are sufficient to generate hindbrain enhancer activity in 89% of cases;
- Most enhancers are active in multiple tissues and at multiple time-points, suggesting that they regulate a variety of developmental processes;
- The set of enhancers display a high degree of spatiotemporal heterogeneity in expression pattern and may have their patterns refined by distinct mechanisms.

This grammar has been shown to be predictive of hindbrain enhancer activity in the context of the functional assay. Elements containing this grammar are sufficient to activate transgene expression in the developing zebrafish hindbrain. However, whether these motifs are necessary for hindbrain enhancer function is not clear. To demonstrate that Pbx-Hox and Meis/Pknox sites within 100bp constitutes a hindbrain enhancer grammar, mutagenesis should be carried out on these motifs in a variety of enhancers.

CHAPTER 5

Pbx-Hox and Meis/Pknox motifs are necessary for enhancer function

5.1 BACKGROUND

The previous chapters identified a grammar that was predictive of hindbrain activity both from the sample from which it was derived (20/22 hb+ sequences, 91%) and from a new sample identified using phylogenetic footprinting and motif searches (66/74 hb+ sequences, 89%). Whilst this grammar has strong predictive power, the necessity of these motifs has not yet been demonstrated in the hindbrain enhancers discovered in the previous chapters. Previous studies have demonstrated the necessity of Pbx-Hox and Meis/Pknox motifs for hindbrain enhancer activity in several independent elements[157, 192, 195, 232, 250, 252]. It was decided to test for the necessity of the motifs comprising the hindbrain grammar using a number of CNEs.

The traditional approach for testing the necessity of motifs in an enhancer assay is to perform site-directed mutagenesis on the sequence of interest, and compare this to a wild-type control. 4 enhancers were selected to represent a range of loci and antero-posterior specificities. This was to ensure the sample was functionally heterogeneous to test the necessity of these motifs in different contexts. Two constructs were generated for each enhancer: one with a mutant Pbx-Hox motif and another with a mutant Meis/Pknox motif. This was to test the contribution of the motifs to hindbrain enhancer activity independently.

As an additional test of this grammar, it was decided to assay CNEs that contained only one or the other motif. This was to test the requirement for motif co-occurrence in naturally occurring enhancers. There are 6 CNEs from the *meis2a* locus containing Pbx-Hox and Meis/Pknox motifs that are known to act as hindbrain enhancers, 4 identified in a previous study[157] and 2 identified in chapter 5, indicating the correlation between the hb+ grammar and hindbrain activity at this locus. CNEs containing either Pbx-Hox or Meis/Pknox motifs from the zebrafish *meis2a* locus were selected. The expression of nearby genes is likely to affect the results of the assays; *meis2a* was chosen because it is strongly expressed in the hindbrain and would bias towards the identification of hindbrain enhancers.

5.2 AIMS AND HYPOTHESES

- **Aim:** to assess the functionality of Pbx-Hox and Meis/Pknox motifs in hindbrain enhancers using site-directed mutagenesis.
- **Aim:** to test *meis2a* elements containing only one of the two sites (either Pbx-Hox or Meis/Pknox)
- **Hypothesis:** mutations disrupting the composition of the Pbx-Hox or Meis/Pknox motif lead to a significant reduction in hindbrain expression.
- **Hypothesis:** CNEs from *meis2a* containing either motif do not drive hindbrain enhancer activity.

5.3 METHODS

5.3.1 Site directed mutagenesis

Mutations were introduced to plasmids using the QuikChange approach (Agilent). Pairs of long (35-45bp) complementary primers containing the desired mutations (mutagenesis primers) were designed for four CNEs: *pax2*.174, *foxd3*.327, *meis2a*.1042 and *meis1*.1705. Oligonucleotides were synthesised and PAGE purified by Sigma-Aldrich. pGW_*cfos*GFP vectors containing the relevant CNE inserts were used as templates for PCRs with mutagenesis primers, dNTPs, Pfu Fusion high-fidelity DNA polymerase and buffers (Agilent). Reaction mixes were subjected to 12 of the following thermal cycles: 95°C for 30 seconds (melting); 55°C for 1 minute (annealing); 68°C for 13 minutes (extension). PCR products were incubated with DPNI for 60min at 37°C to digest the wild-type template before transformation in to Oneshot TOP10 chemically competent cells (Invitrogen). Outgrown cultures were spread on ampicillin agar plates and grown at 37°C overnight. Colonies were picked and grown in 3ml LB plus ampicillin and grown at 37°C overnight with agitation. Plasmids were obtained from 2ml of culture using the QIAprep Spin Miniprep Kit (qiagen) according to the manufacturer's guidelines and eluted from the column with double-distilled water. Inserts were sanger sequenced (Source Bioscience) to confirm the substituted bases.

5.3.2 Comparison of wild-type and mutant zebrafish enhancers

The approach for the enhancer assay was performed as described previously[150] and the methods section of chapter 2 (2.3.2), with the following modifications.

The concentrations of each set of three constructs (the wild-type element and two mutant constructs) were normalised. To ensure minimum variation in experimental conditions, wild-type and mutant constructs were injected in to embryos collected from the same tank of wild-type zebrafish and injected on the same morning. The same aliquot of *tol2* mRNA was used for the preparation of microinjection mixes. 3 replicates of 15-70 embryos were injected for each of the 12 constructs. In order to discount cases where transient transgenesis was not successful, only embryos that exhibited reporter expression in any tissue were counted. Total samples comprised at least 40 GFP+ embryos for each construct.

The total number of embryos exhibiting GFP positive cells in the forebrain, midbrain, hindbrain and spinal cord were counted at a predetermined time point for each element, corresponding to the time when most GFP positive embryos were observed in previous experiments (56h for *pax2*.174, *meis2a*.1042 and *meis1*.1705; 72h for *foxd3*.327). To test for reduction in the proportion of embryos exhibiting hindbrain expression, wild-type and mutant hb+ embryo counts were compared with a one-tailed paired t-test with a p-value cutoff of 0.05.

For elements driving substantial (>20% of injected embryos) levels of hindbrain expression, six GFP-positive embryos from the same replicate were selected at random and imaged using stereotypic zoom and exposure. These six images were then rotated and aligned using Adobe Photoshop and stacked in ImageJ before performing a Z-projection to generate an average of the six images. Averaged images for corresponding wild-type and mutant constructs were artificially coloured and overlaid in Adobe Photoshop for comparative purposes.

5.3.3 Identification of candidate *meis2a* CNEs

The human, mouse, chicken *meis2* loci and the zebrafish *meis2a* locus were aligned using mLAGAN[98] and sLAGAN[99] to detect and define CNEs from zebrafish. These were used as BLAST queries against CONDOR to check they gave the best hit to the correct CNE. These were cross-referenced with the FIMO search detailed in the methods section of chapter 5, but in this case only CNEs associating with either a hit to the Pbx-Hox motif or the Meis/Pknox motif were considered as candidates. 16-24 orthologues of each CNE were aligned and checked for putative binding sites by using conserved sequence blocks as queries against the Uniprobe database[181].

5.3.4 Cloning of candidate CNEs

The cloning of candidate CNEs was performed as detailed in the methods section of chapter 2 (2.3.2).

5.3.5 Enhancer assay in zebrafish embryos

The enhancer assay was performed as described previously[150] with the alterations detailed in the methods section of chapter 2 (2.3.3).

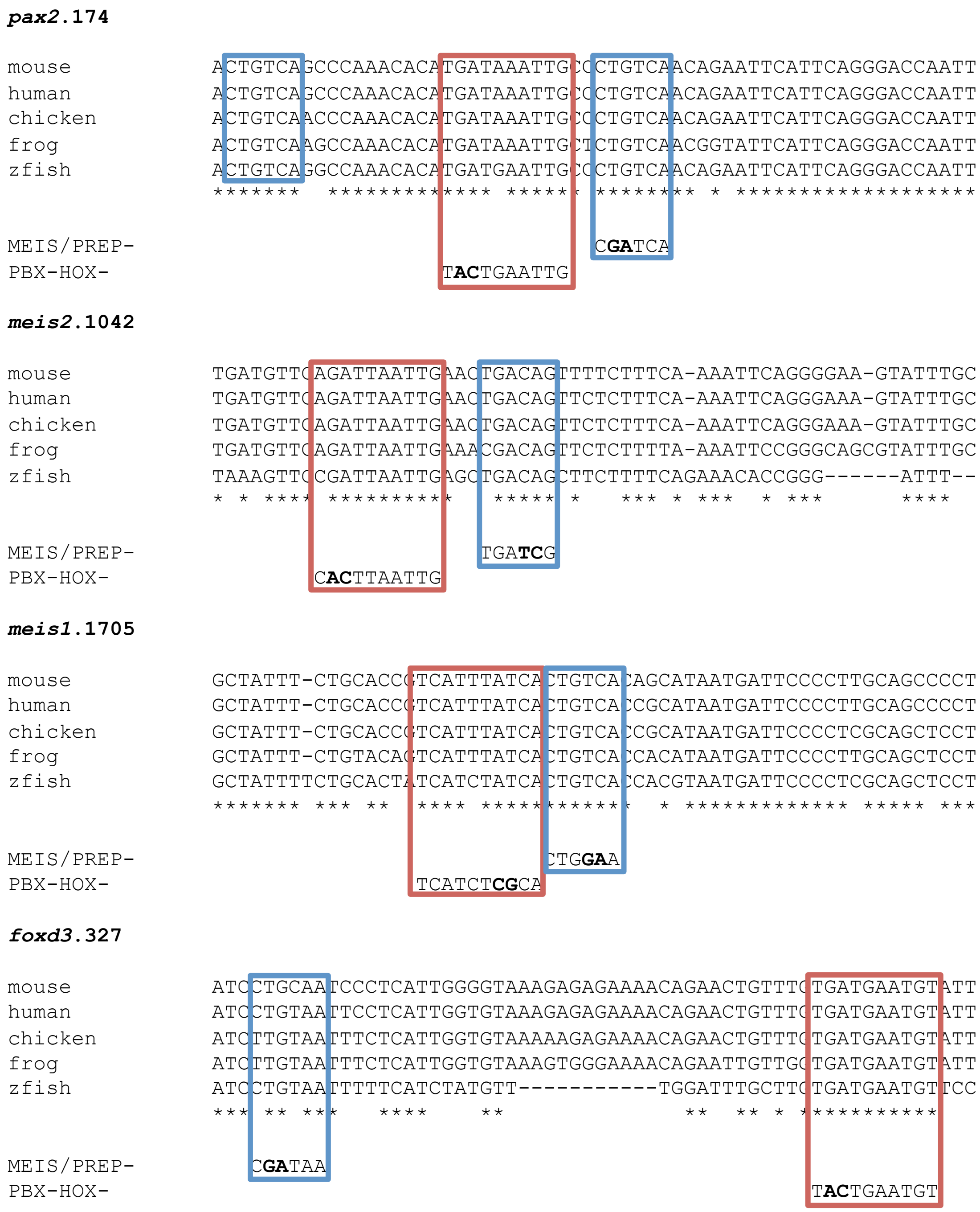
5.4 RESULTS

5.4.1 Mutation of motifs abolishes hindbrain activation or specificity

First, site-directed mutagenesis was performed to test the necessity of Pbx-Hox and Meis/Pknox motifs. The most proximal Pbx-Hox and Meis/Pknox motif pair for each CNE was identified in the chapter 4 (figure 4.3). For each construct, two separate dinucleotide site-directed substitutions were performed: one targeting the Pbx-Hox motif and one targeting the Meis/Pknox motif, termed Pbx-Hox- and Meis/Pknox- respectively (figure 5.1).

In all four cases studied, a dinucleotide substitution disrupting the composition of either the Pbx-Hox or Meis/Pknox motif leads to a statistically significant reduction in the proportion of hindbrain positive embryos (student's t-test, $p < 0.05$). The magnitude of this effect is seen most strikingly in three of the four enhancers (*pax2*.174, *meis2a*.1042 and *meis1.*1705), where mutation of either motif leads to reduction of hindbrain positive embryos (measured as a proportion of total Egfp positive embryos) from ~90% to ~35% (figure 5.2 A, B, C). A much more modest effect is seen for *foxd3*.327, where a reduction from ~90% to ~80% is observed (figure 5.2 D). Because the mutants targeting each site are not significantly different from one another, these motifs appear to make approximately equal contributions to the activation of hindbrain expression. Furthermore, mutant embryos frequently express mosaically in very few hindbrain cells, in contrast to the reproducible and segment-targeted expression seen in the wild-type constructs.

Interestingly, when the Pbx-Hox motif or the Meis/Pknox motif of *foxd3*.327 is mutated, the proportion of embryos expressing GFP in forebrain, midbrain and spinal cord increases. The increase in forebrain, midbrain and spinal cord expression can also be observed when comparing average expression patters driven by the constructs (figure 5.2 H). Other than a Pbx-Hox and Meis/Pknox motif, this element also contains two conserved Hox consensus sites. The zebrafish orthologue of *foxd3*.327 also contains a Mafb consensus site (figure 5.3). Mafb is expressed in r5r6, matching the expression pattern of *foxd3*.327 (figure 2.1 G, figure 5.2 H). This suggests that Hox proteins or Mafb could activate this enhancer in the posterior hindbrain.



***pax2*.174**

```
mouse           ACTGTCAGCCCAAACACATGATAAATTGCCCTGTCAACAGAATTCATTCAGGGACCAATT
human           ACTGTCAGCCCAAACACATGATAAATTGCCCTGTCAACAGAATTCATTCAGGGACCAATT
chicken         ACTGTCAACCCAAACACATGATAAATTGCCCTGTCAACAGAATTCATTCAGGGACCAATT
frog            ACTGTCAAGCCAAACACATGATAAATTGCTCTGTCAACGGTATTCATTCAGGGACCAATT
zfish           ACTGTCAGGCCAAACACATGATGAATTGCCCTGTCAACAGAATTCATTCAGGGACCAATT
                *******  ************* ****** ******** * *******************

MEIS/PREP-                                    CGATCA
PBX-HOX-                          TACTGAATTG
```

***meis2*.1042**

```
mouse           TGATGTTCAGATTAATTGAACTGACAGTTTTCTTTCA-AAATTCAGGGGAA-GTATTTGC
human           TGATGTTCAGATTAATTGAACTGACAGTTCTCTTTCA-AAATTCAGGGAAA-GTATTTGC
chicken         TGATGTTCAGATTAATTGAACTGACAGTTCTCTTTCA-AAATTCAGGGAAA-GTATTTGC
frog            TGATGTTCAGATTAATTGAAACGACAGTTCTCTTTTA-AAATTCCGGGCAGCGTATTTGC
zfish           TAAAGTTCCGATTAATTGAGCTGACAGCTTCTTTTCAGAAACACCGGG------ATTT--
                * * **** **********   ***** *   *** * ***  * ***      ****

MEIS/PREP-                             TGATCG
PBX-HOX-                ACTTAATTG
```

***meis1*.1705**

```
mouse           GCTATTT-CTGCACCGTCATTTATCACTGTCACAGCATAATGATTCCCCTTGCAGCCCCT
human           GCTATTT-CTGCACCGTCATTTATCACTGTCACCGCATAATGATTCCCCTTGCAGCCCCT
chicken         GCTATTT-CTGCACCGTCATTTATCACTGTCACCGCATAATGATTCCCCTCGCAGCTCCT
frog            GCTATTT-CTGTACAGTCATTTATCACTGTCACCACATAATGATTCCCCTTGCAGCTCCT
zfish           GCTATTTTCTGCACTATCATCTATCACTGTCACCACGTAATGATTCCCCTCGCAGCTCCT
                ******* *** **  **** ************  * ************* ***** ***

MEIS/PREP-                                CTGGAA
PBX-HOX-                         TCATCTCGCA
```

***foxd3*.327**

```
mouse           ATCCTGCAATCCCTCATTGGGGTAAAGAGAGAAAACAGAACTGTTTCTGATGAATGTATT
human           ATCCTGTAATTCCTCATTGGTGTAAAGAGAGAAAACAGAACTGTTTCTGATGAATGTATT
chicken         ATCTTGTAATTTCTCATTGGTGTAAAAAGAGAAAACAGAACTGTTTCTGATGAATGTATT
frog            ATCTTGTAATTTCTCATTGGTGTAAAGTGGGAAAACAGAATTGTTGCTGATGAATGTATT
zfish           ATCCTGTAATTTTTCATCTATGTT-----------TGGATTTGCTTCTGATGAATGTTCC
                *** ** ***   ****    **              **  ** * ***********

MEIS/PREP-         CGATAA
PBX-HOX-                                                       TACTGAATGT
```

Figure 5.1: Introduction of targeted substitutions to Pbx-Hox and Meis/Pknox motifs. Sections of alignments for 4 CNEs containing conserved Pbx-Hox (red boxes) and Meis/Pknox (blue boxes) motifs are shown. Mutations were designed to target either the Meis/Pknox site (Meis/Pknox-) or Pbx-Hox site (Pbx-Hox-) in the zebrafish orthologues which had been confirmed to act as hindbrain enhancer in chapter 2 (*foxd3*.327, figure 2.2 G) or in chapter 4 (*pax2*.174, figure 4.4, *meis2*.1042, figure 4.4, meis1.1705, figure 4.4). The mutated sites are shown below each alignment with the substituted bases highlighted in bold.

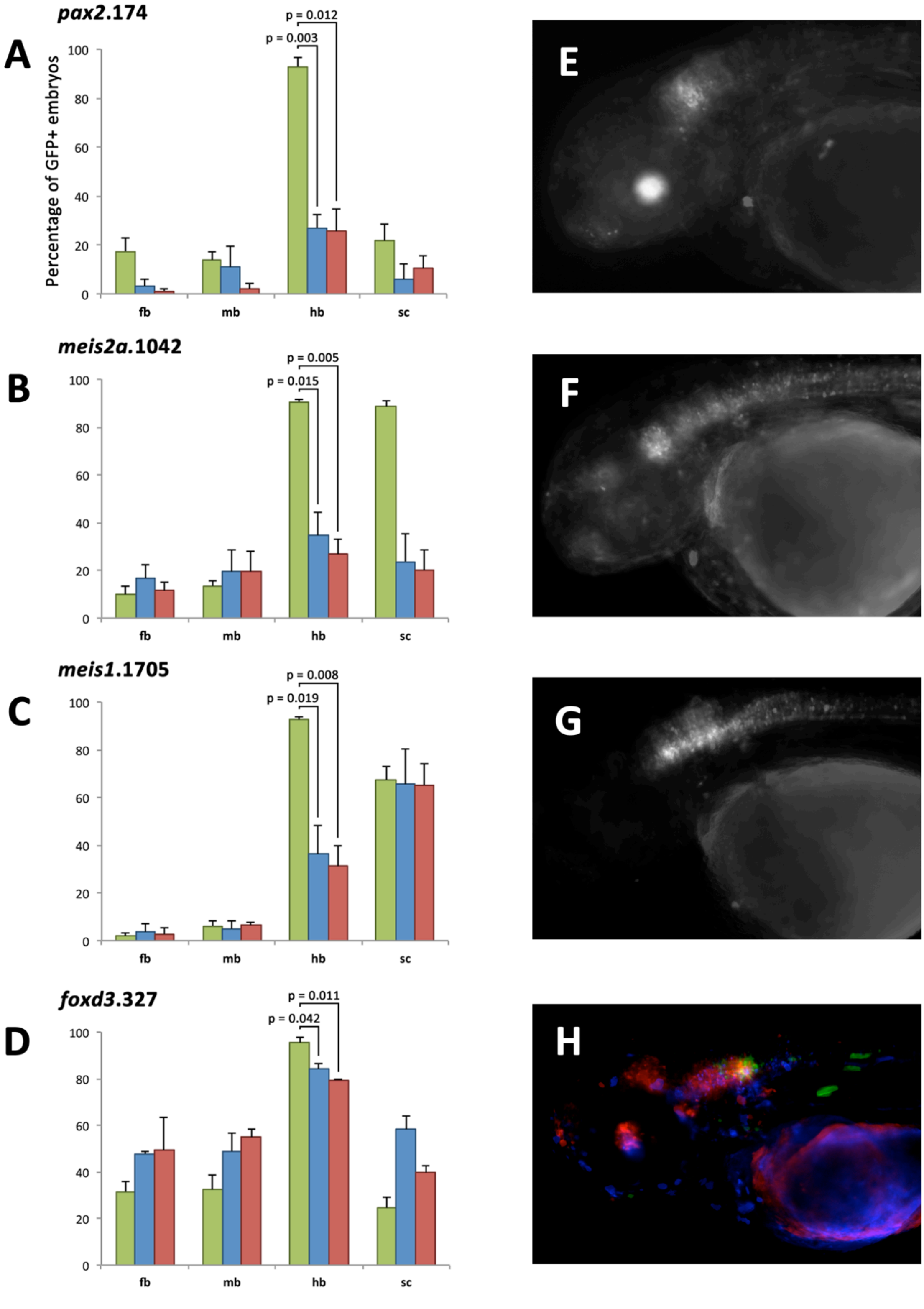
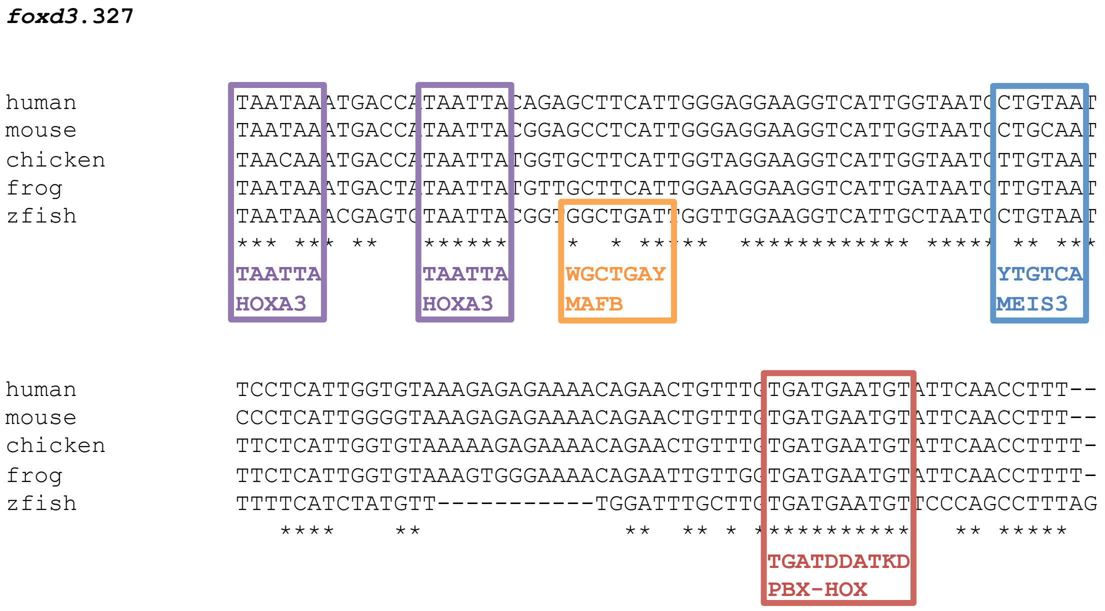


Figure 5.2: Mutation of Pbx-Hox or Meis/Pknox motifs abrogates hindbrain enhancer activity or generates aberrant reporter gene expression (caption overleaf).

Figure 5.2: mutation of Pbx-Hox or Meis/Pknox motifs abrogates hindbrain enhancer activity or generates aberrant reporter gene expression Histograms for four elements, *pax2*.174 (A), *meis2a*.1042 (B), *meis1*.1705 (C) and *foxd3*.327 (D), showing the number of embryos with GFP positive cells in forebrain (fb), midbrain (mb), hindbrain (hb) and spinal cord (sc) when expressing wild-type (green), Meis/Pknox site mutant (blue) or Pbx-Hox site mutant (red) constructs. Annotation displays p values for one-tailed paired t tests. All mutations result in a significant (t test p = <0.05) reduction in the number of embryos positive for hindbrain. Wild-type *pax2*.174 (E) drives expression in hindbrain and lens (green), whereas mutant constructs do not drive this pattern. Wild-type *meis2a*.1042 (F) drives expression in the central nervous system particularly the anterior hindbrain. Mutant constructs fail to recapitulate this expression. Wild-type *meis1*.1705 (G) drives expression in the hindbrain and spinal cord, but mutant constructs drive expression only in spinal cord. Wild-type *foxd3*.327 (H) drives expression in posterior hindbrain (green), but expression driven by constructs where either the Meis/Pknox (blue) or the Pbx-Hox (red) motif is mutated is frequently ectopic in midbrain, anterior hindbrain and spinal cord.



```
foxd3.327

human           TAATAAATGACCATAATTACAGAGCTTCATTGGGAGGAAGGTCATTGGTAATCCTGTAAT
mouse           TAATAAATGACCATAATTACGGAGCCTCATTGGGAGGAAGGTCATTGGTAATCCTGCAAT
chicken         TAACAAATGACCATAATTATGGTGCTTCATTGGTAGGAAGGTCATTGGTAATCTTGTAAT
frog            TAATAAATGACTATAATTATGTTGCTTCATTGGAAGGAAGGTCATTGATAATCTTGTAAT
zfish           TAATAAACGAGTCTAATTACGGTGGCTGATTGGTTGGAAGGTCATTGCTAATCCTGTAAT
                *** *** **   ******    *  * *****  ************ ***** ** ***
                TAATTA       TAATTA    WGCTGAY                       YTGTCA
                HOXA3        HOXA3     MAFB                          MEIS3

human           TCCTCATTGGTGTAAAGAGAGAAAACAGAACTGTTTGTGATGAATGTATTCAACCTTT--
mouse           CCCTCATTGGGGTAAAGAGAGAAAACAGAACTGTTTGTGATGAATGTATTCAACCTTT--
chicken         TTCTCATTGGTGTAAAAAGAGAAAACAGAACTGTTTGTGATGAATGTATTCAACCTTTT-
frog            TTCTCATTGGTGTAAAGTGGGAAAACAGAATTGTTGGTGATGAATGTATTCAACCTTTT-
zfish           TTTTCATCTATGTT-----------TGGATTTGCTTGTGATGAATGTTCCCAGCCTTTAG
                   ****    **              **  ** * ***********   ** *****
                                                     TGATDDATKD
                                                     PBX-HOX
```

Figure 5.3: the hindbrain enhancer foxd3*.327* contains Hox, Pbx-Hox, Meis and Mafb sites. The figure shows a section of the alignment of *foxd3*.327 from different vertebrates. The boxes highlight matches to TF binding preferences from Uniprobe or chapter 3. The name of the most significant match is displayed but all Pbx, Meis/Pknox and Hox proteins have very similar binding preferences. This sequence contains the hindbrain grammar, with a Pbx-Hox (red box) and a Meis/Pknox (blue box) motif. Ectopic expression is observed when these motifs are mutated (figure 5.2 D and H), suggesting that these are binding sites for repressive factors. The sequence also contains two conserved Hox consensus motifs (purple boxes) and a Mafb consensus motif in the zebrafish (orange box), which might be responsible for activation of this element in r5r6.

5.4.2 Pbx-Hox or Meis/Pknox motifs are insufficient for hindbrain activity

The mutagenesis results suggested that both motifs were required for hindbrain enhancer activity. However, there remains the possibility that either motif could be sufficient for hindbrain activity in the context of endogenous enhancers. To test this, elements containing either Pbx-Hox (*meis2a*.960, *meis2a*.984/5/6, *meis2a*.1043/4) or Meis/Pknox (*meis2a*.957/8, *meis2a*.962/3, *meis2a*.965, *meis2a*.974, *meis2a*.982, *meis2a*.983) motifs were subjected to the enhancer assay. All of these sequences drove reporter expression during the assay (table 5.1, figure 5.4).

3 of these were considered to act as hindbrain enhancers. Two of the hindbrain enhancers, *meis2a*.962/3 and *meis2a*.974, generate mosaic but appreciable hindbrain expression at 2 dpf (table 5.1). However, each element has its own specific and reproducible pattern, in the midbrain for *meis2a*.962/3 (figure 5.4 C) and in the forebrain and eye for *meis2a*.974 (figure 5.4 E). In contrast, *meis2a*.965 generates a specific and reproducible pattern in r7, the anterior spinal cord and associated neural crest (figure 5.4 D). This demonstrates that other motif combinations can also lead to robust and specific hindbrain activity; *meis2a*.965 contains motifs resembling monomeric Pbx, Hox and Meis/Pknox binding sites, and contains a Mafb consensus overlapping the Meis/Pknox site (figure 5.5).

Table 5.1: *meis2a* CNEs containing either Pbx-Hox or Meis/Pknox motifs do not typically drive hindbrain expression. The table shows the name of the element (CNE) and the percentage of hindbrain positive embryos as a proportion of the total number injected (hb+/total) and as a proportion of GFP positive embryos (hb+/total). The most common tissues driven by the elements are also displayed. Hindbrain enhancers are indicated in green. Other enhancers are indicated in red. The source data for this table can be found in appendix 1.

	2 dpf			3 dpf		
CNE	**hb+ /total**	**hb+ /GFP+**	**common tissues**	**hb+ /total**	**hb+ /GFP+**	**common tissues**
meis2a_957/8	5.7%	6.4%	mb	3.3%	4.5%	mb
meis2a_960	0.0%	0.0%	sc, ne	0.0%	0.0%	sc, ne
meis2a_962/3	33.3%	45.7%	mb	9.8%	36.4%	mb
meis2a_965	48.6%	71.8%	hb	89.1%	96.5%	hb
meis2a_974	35.3%	50.0%	fb, ne, ey	12.9%	33.3%	msc
meis2a_982	6.7%	18.2%	fb, msc	3.3%	16.7%	fb, msc, ht
meis2a_983	3.1%	13.3%	msc	10.0%	25.0%	ne, msc
meis2a_984/5/6	2.8%	18.2%	ea, msc, sk	7.8%	18.5%	ea, msc
meis2a_1043/4	14.6%	31.8%	msc	15.0%	54.5%	ne, msc

fb: forebrain; mb: midbrain; hb: hindbrain; sc: spinal cord; ne: other neuron; ey: eye; ea: ear; msc: trunk muscle cells; ht: heart; sk: skin.

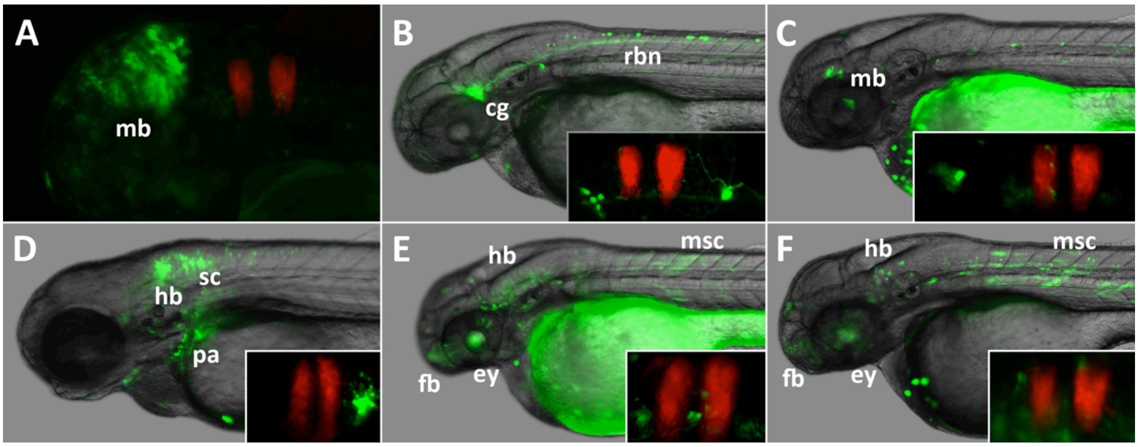
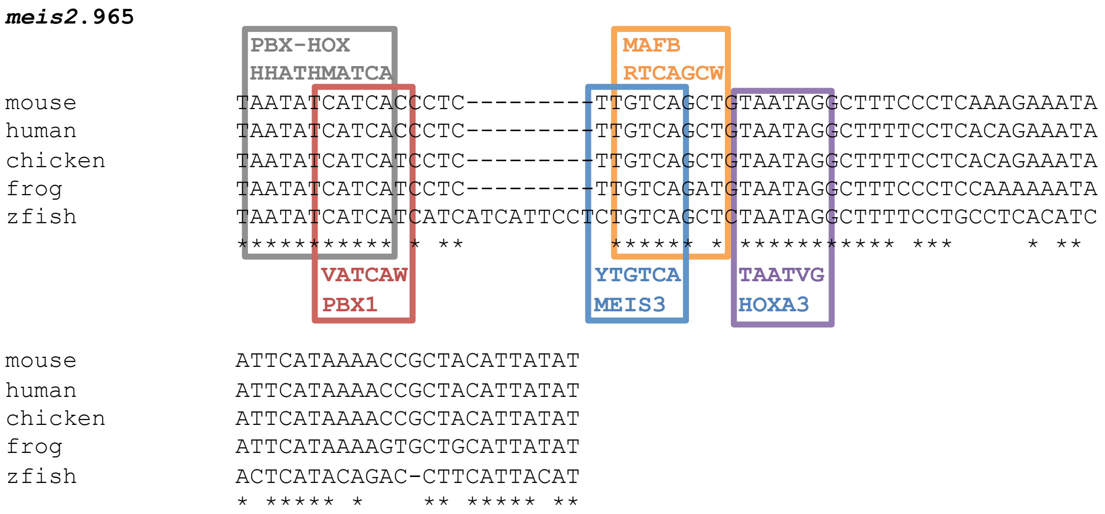


Figure 5.4: *meis2a* CNEs with incomplete hindbrain grammar can act as enhancers of midbrain, hindbrain and other tissues. *meis2a*.957/8 (A) is a midbrain enhancer. *meis2a*.960 (B) is active in cranial ganglia and rohon-beard neurons. *meis2a*.962/3 (C) drives expression in small population of midbrain cells, but also generates mosaic expression in hindbrain. *meis2a*.965 (D) is active in r7, anterior spinal cord and neural crest. *meis2a*.974 (E) drives expression in forebrain, eye, hindbrain and trunk muscle cells. *meis2a*.982 (F) is a muscle enhancer but brain and eye expression are observed in a small proportion of embryos. fb: forebrain; mb: midbrain; hb: hindbrain; sc: spinal cord; pa: pharyngeal arches/neural crest; cg: cranial ganglia; rbn: rohon-beard neurons; ey: eye; msc: trunk muscle cells.



```
mouse      TAATATCATCACCCTC---------TTGTCAGCTGTAATAGGCTTTCCCTCAAAGAAATA
human      TAATATCATCACCCTC---------TTGTCAGCTGTAATAGGCTTTTCCTCACAGAAATA
chicken    TAATATCATCATCCTC---------TTGTCAGCTGTAATAGGCTTTTCCTCACAGAAATA
frog       TAATATCATCATCCTC---------TTGTCAGATGTAATAGGCTTTCCCTCCAAAAAATA
zfish      TAATATCATCATCATCATCATTCCTCTGTCAGCTCTAATAGGCTTTTCCTGCCTCACATC
           *********** *  **         ****** * *********** ***       * **

mouse      ATTCATAAAACCGCTACATTATAT
human      ATTCATAAAACCGCTACATTATAT
chicken    ATTCATAAAACCGCTACATTATAT
frog       ATTCATAAAAGTGCTGCATTATAT
zfish      ACTCATACAGAC-CTTCATTACAT
           * ***** *    ** ***** **
```

Figure 5.5: the hindbrain enhancer *meis2a*.965 does not contain the hindbrain grammar. The figure shows a section of the alignment of *meis2*.965 from different vertebrates. The boxes highlight matches to TF binding preferences from Uniprobe. Whilst this sequence was not considered to have the hindbrain grammar due to two mismatches in the Pbx-Hox motif (grey box), it nevertheless acts as a robust and specific hindbrain enhancer (figure 5.3 D, table 5.1). However, the sequence does contain consensus sites for Pbx (red box), Meis/Pknox (blue box) and Hox (purple box) binding preferences. There is also a Mafb site overlapping the Meis/Pknox site (orange box).

5.5 DISCUSSION

In this chapter, the requirement for the previously defined grammar was assessed. The results from the mutagenesis experiments suggest that both motifs are required for the normal function of these enhancers. Embryo counts for mutations targeting each site are equivalent, suggesting that each site makes an additive and synergistic contribution to the observed activity. This is consistent with these enhancers operating by either: i) the formation of Pbx-Hox-Meis/Pknox heterotrimers; or ii) a Pbx-Hox dimer and Meis/Pknox factor binding combinatorially to the enhancer.

In the first three cases (*pax2*.174, *meis2a*.1042 and *meis1*.1705) mutating either motif leads to a reduction of expressing embryos to around 1/3 of that seen with the wild type constructs. Embryos expressing mutant constructs do so mosaically and in few cells, but do not lose hindbrain expression completely. Unfortunately, the construction of overlays for mutants proved difficult and uninformative due to the small number of Egfp-positive cells in each embryo. In some cases, mutations abolish other patterns of expression as well. For example, the *pax2*.174 mutant constructs abolish lens as well as hindbrain expression (data not shown), and the *meis2a*.1042 mutant constructs abolish both spinal cord and hindbrain expression (figure 5.2). This suggests either multifunctional Pbx and Meis/Pknox sites or the existence of overlapping sites at the mutated positions. Contrastingly, the spinal cord expression driven by *meis1*.1705 remains unchanged. This indicates that, in this case, the activity in hindbrain and spinal cord are encoded at distinct positions within the same element. This highlights the diversity of the mechanistic functions of CNEs as CRMs, even when these elements have been identified using a grammar that is strongly predictive of tissue specificity. This is consistent with their unique sequences and independent evolution.

A contrasting case is observed for *foxd3*.327 and its mutants. The patterns observed in functional assays demonstrate that mutation of either motif leads to ectopic expression throughout the major subdivisions of the CNS (figure 5.2 G). Furthermore, overlays of 6 randomly selected embryos demonstrate that the mutant constructs more frequently express in anterior hindbrain. Because the number of hindbrain-positive embryos was calculated as a proportion of Egfp-positive embryos, slight reduction in the proportion of hindbrain positive embryos is probably caused by an increase in the number of embryos exhibiting ectopic expression.

This supports a model where these Pbx-Hox and Meis/Pknox motifs act as a platform for a repressive complex. While repressive Hox sites are rare, there are two notable examples. First, a mouse TBX5 CRM generates expression in the forelimb, but not hindlimb, forming region of mouse embryos and is activated by HoxC5[261].

This CRM also contains a Hox binding site which, when mutated, causes expansion of expression to the posterior. The authors then demonstrated that this site is occupied by HoxC9 which mediates repressive activity in the posterior lateral plate mesoderm[262]. Secondly, a case from *Drosophila* demonstrated that a *sal* CRM is bound and repressed by UBX (homologous to vertebrate paralogous groups Hox6-Hox8) in the haltere[263]. These examples show that in some cases, Hox proteins can mediate repression as well as activation, though in both cases this is without the requirement for Pbx, Meis or Prep cofactors (or their *Drosophila* orthologues EXD, HTH and VIS/ACHI). There is one further, albeit artificial, case from *Drosophila* where three copies of a GRH binding site can drive expression uniformly in the ectoderm[264]; three copies of an EXD-UBX/ABDA site can repress the activity of this element in abdominal segments[264]. Due to the patterns driven by foxd3.327 and its mutants, the Hox paralogs involved here are likely to be those expressed in the hindbrain and anterior spinal cord (paralogous groups 1-5), and these are not known to contain repressor domains. The HoxC9 proteins of limb-forming vertebrates possess a repressive domain[265], but this is the only Hox protein for which a repressive domain has been characterised. It therefore appears likely that repressive action by anterior proteins is not possible because they lack repressor domains.

In *Drosophila*, EXD/HTH have been shown to cooperate with EN (a homeodomain-containing transcriptional repressor) at a *slp* CRM to mediate repression[266]. This may be occurring at this CRM with the zebrafish Pbx, Meis/Pknox and Eng proteins. Such an interaction has not been noted in vertebrates, but the zebrafish *eng1b* gene is expressed in the hindbrain during the high-pec stage, prior to the activation of this element (zfin gene expression database). Pbx and Meis proteins can also co-bind with Klf4, which can both activate and repress gene expression[267]. There are many possibilities of how a repressive complex could form at this CNE, and further experimentation is required to investigate these possibilities. This result also hints at the possibility of two distinct modes for the Pbx-Hox-Meis/Pknox trimeric complex: one activatory and one repressive. Further work on the library of hindbrain enhancers described here could gather more evidence to test these hypotheses.

A number of consensus binding motifs are present in *foxd3*.327, including those for Hox monomers and Mafb. These could be activatory sites, whilst the Pbx-Hox and Meis/Pknox motifs are responsible for repression, limiting expression to the posterior hindbrain. Indeed, Mafb is expressed in r5r6, colocalised with the expression pattern of the wild-type enhancer. Further experiments on *foxd3*.327 could determine whether these Hox monomer sites or the Mafb site are activatory using mutagenesis. There exists a zebrafish *mafb* null mutant line known as *valentino* (*val*). Injection of the expression construct in to *val-/-* embryos could determine whether this element is activated by Mafb. However, *val-/-* embryos also have altered *hox* gene expression[215] so this would also need to be taken in to account. Injection of this construct in to mutant lines for various *hox* genes could test the requirement for Hox binding as well. Alternatively, co-injection of the construct with morpholinos targeting Mafb, Meis, Pbx and Hox family members could test the requirement for these proteins. These experiments could decrypt the mechanism of action at this enhancer and assess the contribution of each site and input to its function, as detailed work on individual enhancers has done before[53].

Following these findings, elements containing incomplete grammar (either Pbx-Hox or Meis/Pknox sites) were tested. Only 3/9 (33%) of elements tested on this basis were regarded as hindbrain enhancers. It is not surprising that we find some hindbrain enhancers at *meis2a* locus since it is strongly expressed in the hindbrain (ZFIN gene expression database). However, the proportion of hindbrain enhancers identified in this smaller screen is much smaller than the proportion identified in chapter 5 (33% rather than 89%). 2/3 of the hindbrain enhancers generate mosaic expression in low numbers of cells. Low levels of expression are by no means irrelevant to *in vivo* function; indeed, some of the hindbrain enhancers identified in chapter 5 generated faint and mosaic, but visible, levels of hindbrain expression. These enhancers are just as likely to be important for function as those identified in chapter 5. There are two possibilities that might explain this activity: these are low-level hindbrain enhancers, but they are activated by mechanism other than a Pbx-Hox-Meis/Pknox trimer; or, the complete enhancers have not been cloned in to the constructs. In the second case, these enhancers could exhibit leaky expression in the hindbrain because the full elements have not been cloned. An example of this is *meis2a*.962/3, which is most commonly a midbrain enhancer (figure 5.4 C), but nevertheless drives some hindbrain expression. Alternatively, they could be part of more specific hindbrain enhancers. Cloning larger elements could distinguish between these possibilities.

The other element, *meis2a*.965, acts as one of the most robust hindbrain enhancers from any of the screens performed (table 5.1). It also displays striking segment specificity (figure 5.4). This is clear evidence that the hindbrain grammar misses a proportion of conserved hindbrain enhancers, consistent with its failure to identify all the known hindbrain enhancers from CONDOR in chapter 5. Upon closer inspection, the sequence does contain a Pbx-Hox-like motif, but this contained two mismatches (TA instead of AT at position 7/8), so was not considered a valid site by our criteria (figure 5.5, grey box. This could still be an in-vivo Pbx-Hox site although it was excluded by the criteria for the FIMO search (chapter 3). When conserved sequence blocks from this sequence are used as queries against Uniprobe, there are matches to the binding consensus of Pbx, Hox, Meis/Pknox and Mafb monomers (figure 5.5). Perhaps Pbx, Hox and Meis/Pknox are binding in a noncanonical formation at this enhancer. It is also known that Pbx and Meis proteins can cooperate with bound Hox proteins without binding to DNA themselves[268], suggesting that perhaps close arrangement of all three binding sites is not necessary in all circumstances. However, the hindbrain grammar is still strongly predictive of hindbrain enhancers and appears to have identified a group of hindbrain enhancers that are mechanistically similar.

5.6 CONCLUSIONS

The mutagenesis results are consistent with a model where the motifs comprising the hindbrain grammar make equal contributions to hindbrain enhancer activity. This suggests that the enhancers are bound either by Pbx-Hox-Meis/Pknox trimeric complex, or that a Pbx-Hox dimer and Meis/Prep monomer are both required to bind to generate hindbrain expression. These motifs are therefore interdependent, using AND logic to achieve an appropriate level of expression in the developing hindbrain. This is consistent with, and complementary to, previously published results. Furthermore, in most cases CNEs containing either Pbx-Hox or Meis/Pknox motifs are not hindbrain enhancers. There are a few cases where CNEs containing isolated sites can still activate hindbrain expression. Because of the large number of transcription factors in the hindbrain GRN, this is not surprising. It seems likely, however, that hindbrain grammar elements appear to be activated directly by Hox proteins because of their motif content. Hindbrain grammar CNEs are therefore candidates for the CRMs that mediate the establishment of the GRN directly downstream of Hox proteins.

In summary:

- In four different cases, the components of the hindbrain syntax are required for the ordinary function of the enhancer;
- Three elements (*pax2*.174, *meis2a*.1042 and *meis1*.1705) require these sites for an appropriate level of activation;
- The remaining element (*foxd3*.327) requires these sites for restriction to r5r6;
- In most cases, CNEs containing either motif are not hindbrain enhancers, but those that do must either contain cryptic sites or use distinct mechanisms;

- The hindbrain grammar has failed to identify all segment-specific hindbrain enhancers (such as *meis2a*.965), reflective of the choice to detect a particular mechanistic class of elements and specificity over sensitivity.

The finding that most CNEs containing the hindbrain grammar drive hindbrain expression provides information on how these CNEs might operate at the molecular level: targeting by Hox proteins and their cofactors. However, their multifunctionality and functional heterogeneity suggest a great diversity in activating mechanisms, even for CNEs that utilise a recurring binding syntax. However, since the majority of vertebrate CNEs containing this signature are hindbrain enhancers, it seems reasonable to assume that the presence of this signature should be indicative of hindbrain/anterior neural enhancers targeted by Hox proteins. Therefore, it was decided to try and track the evolution of this signature across the chordates, to test the hypothesis that these kinds of elements have been gained in vertebrates.

CHAPTER 6

Gnathostome hindbrain enhancers are rarely conserved in cyclostomes

6.1 BACKGROUND

In a pair of reviews, Vavouri and Lehne[135] and Nelson and Wardle[219] have independently proposed that CNEs coordinate morphogenesis during the phylotypic stage of the group concerned based on both morphological and molecular evidence. However, which processes CNEs control and the mechanisms by which they perform these have been limited to few examples. Previous work drew the connection between CNEs and hindbrain patterning for the first time via the presence of Pbx-Hox motifs[157]. This body of work has been expanded in the previous chapters to identify a predictive hindbrain enhancer syntax that is necessary for the proper function of these CNEs.

Related to this, enhancers have been postulated to arise *de novo* in genomes by the gradual accumulation of clustered binding sites by chance; these act as a platform for the recruitment of additional selectively advantageous sites over evolutionary time, and eventually become fixed as CNEs[90, 269]. In previous hypotheses, Pbx-Hox motifs were thought to have acted as such platforms[157]. A related observation is that CNEs appear to get larger over the course of evolution; whether this 'extended' conservation is neutral or adaptive is not, as yet, clear.

In the majority of cases analysed in chapter 2, the hindbrain enhancers identified from this study also have high-affinity Meis/Pknox sites. Now the model that numerous colocalised Pbx-Hox and Meis/Pknox sites might have arisen at the base of the vertebrate lineage to couple the Hox code to novel downstream genes in the evolving hindbrain will be investigated.

There are several sets of CNEs from previously published work identified using various criteria (table 6.1). Presumably, these CNEs are descended from sequences present in the common ancestors of the species analysed, and thus represent approximations of regulatory elements from various nodes on the chordate phylogenetic tree (figure 6.1). The aim was to use these CNE sets to assess the correlation between enrichment for the hindbrain syntax (corresponding to putative Hox target enhancers) and the acquisition of hindbrain segmentation during chordate evolution. Comparative study of orthologous sequences from different vertebrates may determine when hindbrain activity was acquired and whether this correlates with the presence of the hindbrain grammar.

Table 6.1: CNE sets used for the analysis. Table shows the lineage in which the elements are conserved (CNE set), The number of individual elements (CNEs), the combined length of the sequences (Length) and the reference for the database or publication from which the set was derived (Citation).

CNE set	CNEs	Length (bp)	Citation	Abbreviation
Actinopterygii	54533	6428822	Hiller *at al.* 2013[144]	AcCNEs
Osteicthyes	7065	830109	Woolfe *et al.* 2007[142]	OsCNEs
Gnathostomata	4782	1000528	Venkatesh *et al.* 2006[143]	GnCNEs
Vertebrata	476	37651	Smith *et al.* 2013[35]	VeCNEs
Olfactores	183	8116	Sanges *et al.* 2013[138]	OlCNEs
Chordata	55	3295	Putnam *et al.* 2008[137]	ChCNEs
Ciona	2336	424229	Doglio *et al.* 2013[97]	CiCNEs

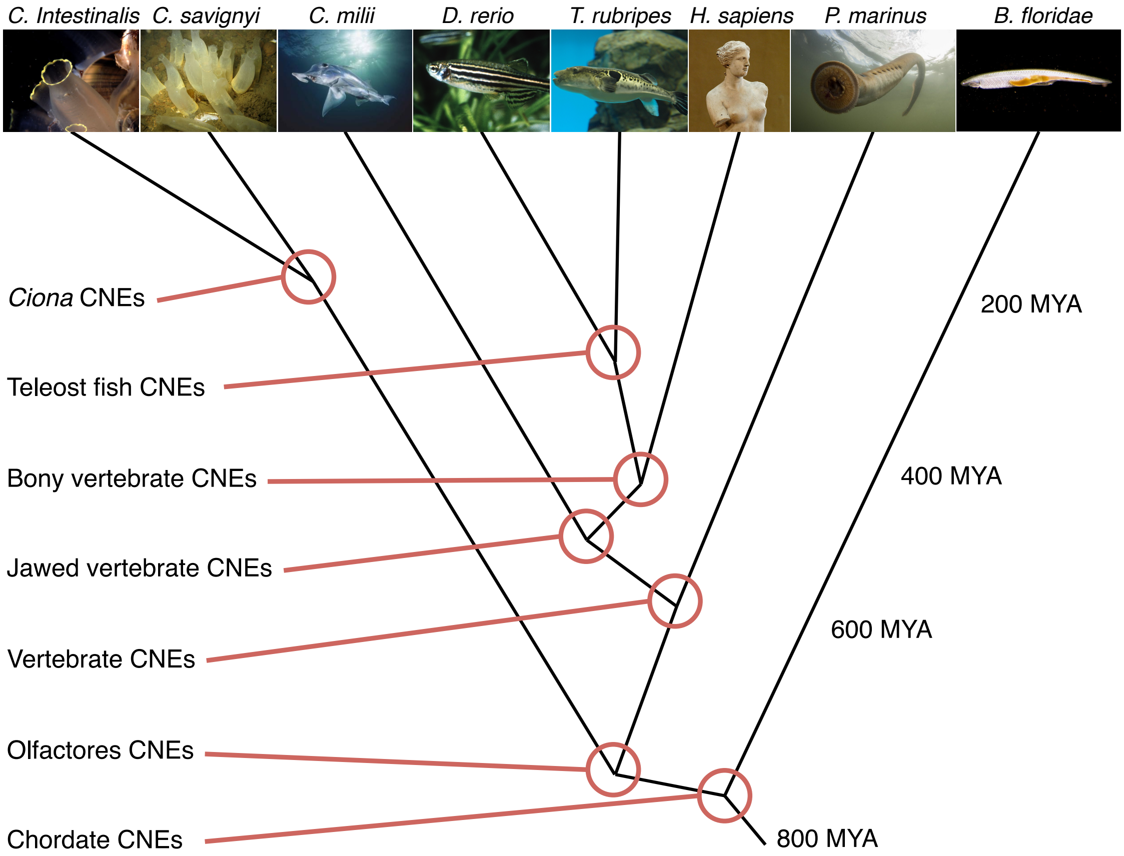


Figure 6.1: Approximate age of and clade represented by published CNE sets. The figure shows the level of conservation of the different CNE sets superimposed on a phylogeny of chordates.

Enriched motifs in large sequence sets can be discovered using CisFinder[270]. Cisfinder first counts kmers, then creates a 1bp substitution matrix for each kmer and sums the columns to create a PWM. The same is performed for a control set (usually shuffled sequences) and the control matrix is subtracted from the test matrix to give the 'elementary motif'. Cisfinder then aligns similar elementary motifs to form 'clustered motifs'. Enrichment of the elementary or clustered motifs versus the control set can then be calculated. As such Cisfinder detects motifs *de novo* and can provide information on the most common binding signatures present in each set, though it does not count individual occurrences in the set. Similarity between motifs can also be determined by STAMP, and the output tree used to group motifs. By analysing the number of motifs derived from each set appearing in each clade, changes in motif content over the course of evolution can be tracked.

Comparative genomic studies have typically only identified large sets of CNEs when comparing gnathostome (jawed vertebrates) genomes. However, this could be because more distant genomes pass some threshold of identity imposed by current methods. Indeed, many of these genomes are phylogenetically isolated; there are no closely related genomes to compare them to because they have yet to be sequenced or their relatives are largely extinct. However, recent comparative genomic data from the zebrafish suggests that additional conserved elements can be detected by using a number of approaches and a variety of reference sequences[144], and hints at the prospect of detecting lineage-specific elements. One such approach is ancestral reconstruction, which aims to infer the sequence of a common ancestor to reduce the evolutionary distance being spanned by similarity searches. Ancestors v1.1 (http://ancestors.bioinfo.uqam.ca/ancestorWeb/) uses this approach[271, 272] and combines algorithms for 1. sequence alignment 2. the inference of insertions/deletions and 3. the inference of substitutions.

The amount of sequence orthologous to gnathostome CNEs identifiable in the sea lamprey genome by BLAST (~38kb in the non-redundant set)[35] is far smaller the full sets from mammal-fugu (~0.8mb) and human-chimaera (~1.0Mb) comparisons. This suggests two possibilities: that the divergence of the cyclostome and gnathostomes lineages predates the evolution and fixation of most gnathostome CNEs; or that the orthologous sequences in the lamprey have diverged beyond what is recognisable by the previously used methods. Ancestral reconstruction may aid in the identification of putative orthologues beyond that recognisable by using modern orthologues as BLAST queries. Previous work has shown that 4 *meis2* CNEs act as hindbrain enhancers from both lamprey and zebrafish, and later used the lamprey genome as a reference for identifying some hindbrain enhancers[157].

Whilst the lamprey hindbrain possesses genetic and morphological rhombomeres, the boundaries between segments are not as overt as in gnathostome hindbrains. Therefore, whether this control network targets the same enhancers and genes in lamprey is not clear. The expression patterns of several components of the hindbrain GRN, including lamprey *hox1*, *hox2*, *hox3, egr2,* and *mafb,* are largely conserved in lamprey[184], but there are some discrepancies. The lamprey *hox* code may not have been fully integrated to downstream factors which, in gnathostomes, control the formation of segment boundaries[273]. Lamprey embryos also sometimes interpret gnathostome hindbrain enhancers differently to zebrafish[162, 184], suggesting that not all components of the gnathostome hindbrain GRN are present in lamprey. The hindbrain enhancers discovered here are, with a few exceptions, not present in the BLAST set of lamprey CNEs, but more putative orthologues may be detectable using ancestral reconstruction. Even fewer CNEs are conserved in amphioxus and *Ciona*, none of which match known hindbrain enhancers; ancestral reconstruction may therefore identify the regulatory sequences from which these hindbrain enhancers first evolved in stem vertebrates.

6.2 AIMS AND HYPOTHESES

- **Aim:** Find the most enriched motifs in these CNE sets to query their likely functions on a global scale.
- **Aim:** Attempt to identify orthologues of known hindbrain enhancers in the lamprey, *Ciona* and amphioxus genomes to assess the likelihood that these elements played a role in hindbrain evolution.
- **Hypothesis:** Vertebrate, but not invertebrate, CNEs contain detectable signatures of the hindbrain grammar.
- **Hypothesis:** Using ancestral reconstruction, orthologues of hindbrain enhancers can be detected amongst vertebrates (i.e. in lamprey) but not in invertebrate chordates.

6.3 METHODS

6.3.1 Acquisition and preparation of CNE sets

FASTA files of CNE sets were downloaded from the supplementary data files of publications and online databases. For CNEs from lamprey, there are two sets of CNEs: those derived from BLAST hits to osteichthyan CNEs; and those derived from BLAST hits to Gnathostome CNEs. These were combined in to a non-redundant set, with overlapping CNEs becoming a single, longer sequence. All files were softmasked for regions of low complexity using the fastnucleo algorithm from USEARCH (http://www.drive5.com/usearch/manual/masking.html).

6.3.2 Estimating PWMs and calculating enrichment using cisfinder

Each set was shuffled using EMBOSS Shuffleseq (http://emboss.open-bio.org/). The text files containing the CNE sets and the corresponding shuffled sets were uploaded to Cisfinder[270] (http://lgsun.grc.nia.nih.gov/CisFinder/bin/cisfinder.cgi) in order to estimate PWMs and calculate enrichment. Various match thresholds were used in order to obtain informative motifs from each set. The discovered motifs were aligned to JASPAR[179] and Uniprobe[181] using TOMTOM[274], and to TRANSFAC[178] using STAMP[275].

In order to compare the motifs discovered from all sets, the top 50 elementary motifs from each set were submitted to STAMP, which aligns the motifs and generates a newick tree. The newick tree was then visualised in MEGA 6[276]. Where clustered motifs matched the same family of transcription factor binding preferences, these were grouped in to a clade. Each clade was named according to its matched family/families. Individual motif identifiers were then used to establish the presence or absence of motifs derived from each set in each clade, indicative of enrichment for different transcription factor families in each CNE set.

6.3.3 Pipeline for the identification of putative lamprey orthologues

Candidate CNEs were chosen from this thesis as well as publications where they had been confirmed to act as hindbrain enhancers. 4 orthologues of CNEs from representative species (zebrafish, frog, chicken and human) were downloaded from CONDOR (or ensembl for *egr2*_C) and aligned using clustalw2.

To perform ancestral reconstruction, this alignment was fed in to Ancestors v. 1.1 (rooted using the centre point of the tree and using the "best heuristic scenario" setting) to generate a theoretical sequence for the osteichthyan ancestor (OA).

The OA sequence was then used as a BLAST query against the elephant shark (*C. milii*) genome using default settings. The top hit was extended to the full length of the osteichthyan alignment, downloaded and added to the alignment. The alignment was trimmed to prevent unaligned sequence ends and a subsequent round of ancestral reconstruction was performed to generate a theoretical sequence for the gnathostome ancestor (GA).

The GA sequence was then used as a BLAST query against the sea lamprey (*P. marinus*) genome at ensembl using the "distant homologies" setting. The region containing the top hit was downloaded and aligned to the other five putative orthologues in order to generate a sequence of similar length. This sequence was then used as a BLAST query against CONDOR to confirm it matched to the correct CNE (best reciprocal BLAST hit).

The BLAST hit from lamprey was added to the alignment and a further round of ancestral reconstruction was performed as above to infer the sequence of the vertebrate ancestor (VA). This sequence was used as a BLAST query to attempt to find putative orthologues in the tunicate (*Ciona intestinalis*) and amphioxus (*Branchiostoma floridae*) genomes.

The presence of Pbx-Hox (TGATNNAT) and Meis/Pknox (CTGTCA) motifs in the elephant shark and lamprey orthologues was then assessed by scanning for conserved co-occurrences of these motifs in the clustalw2 alignments.

6.3.4 Cloning of putative lamprey orthologues

Lamprey genomic DNA was prepared from frozen liver tissue from a single adult male lamprey using the ISOLATE genomic DNA kit (Bioline). Primers were designed such that the product contained the lamprey BLAST hits and was as similar as possible in size to the zebrafish insert. Thereafter, the cloning of candidate CNEs was performed as detailed in the methods section of chapter 2 (2.3.2) using lamprey genomic DNA as a template for the PCR.

6.3.5 Comparison of orthologous zebrafish and lamprey CNEs

The approach for the enhancer assay was performed as described previously[150] and the methods section of chapter 2 (2.3.2). Experimental conditions for zebrafish and lamprey orthologues were maintained as for the wild-type and mutant zebrafish enhancers in the methods section of chapter 5 (5.3.2).

6.4 RESULTS

6.4.1 Pbx-Hox and Meis/Pknox motifs are enriched in vertebrate CNEs

In order to assess the global changes in conserved element motif content, a range of CNE sets were analysed for enriched motifs (table 6.1, figure 6.1) using CisFinder. These sets were chosen to represent elements that evolved in a range of vertebrate lineages, as well as the few elements conserved between vertebrates and invertebrate chordates. A parallel set of elements from *Ciona* was also analysed. Initially, The top ten clustered motifs discovered by Cisfinder (see methods) were aligned to known binding preferences from online databases.

The most evolutionarily recent set of elements from vertebrates, the AcCNEs, are enriched for motifs resembling diverse TFBSs (table 6.2). The highest scoring motif (CWGCWCWG) resembles binding preferences for some C2H2 zinc fingers such as ZIC1 and ZIC2, and is by far the most enriched motif in the set, with an enrichment of 15x versus the shuffled sequence control. The second most highly scoring motif, an Oct motif (ATTWGCAT) is also the most frequent in the set, with 46,503 occurrences (7.23 occurrences per kilobase). A Pbx-Hox-like motif is detected, but Meis/Pknox are apparently absent.

Table 6.2: The most highly scoring motifs in AcCNEs. The table shows the motif as discovered by cisfinder (Motif), the number of ocurrences (Freq), the enrichment versus the shuffled sequence control (Ratio), the information content of the motif (Info), and the score (Score). Matches to the motif from online databases are also shown: the matches to JASPAR and Uniprobe using TOMTOM, and the preferences to Transfac using STAMP. Matches refer to groups of factors with similar binding preferences.

Actinopterygian CNEs						
Motif	**Freq**	**Ratio**	**Info**	**Score**	**TOMTOM vs Jaspar, Uniprobe**	**STAMP vs Transfac**
B**CWGCWCWG**H…	4250	15.0	6.0	1012.2	Zic, Zfp691	None
…M**ATTWGCAT**D…	4.6k	3.6	12.3	779.9	Oct	Oct
AAAKGDCAS	1656	3.6	4.5	405.0	Nr2, Nr4, Esrr	None
TRATWAATKD	1759	3.5	5.4	389.3	Hox, Pbx	Pbx
TYKCMATKGMRA	501	4.5	6.8	387.5	Rfx	Rfx, Oct, Nkx
…S**WGSAAAT**	2160	3.6	3.8	335.6	Ets, Oct	Ets, Oct
YARASAG…	700	3.5	7.0	317.6	Fox, Smad	Smad
…D**VARWCAVWD**	2176	3.5	4.3	254.2	Sox	Fox, Smad
S**CTBTGAWR**	1254	3.1	5.1	234.8	Tcf/Lef	Tcf/Lef
…W**WWWCWGC**	2031	3.5	5.1	226.7	Hox	Homeodomain*

* Motifs matching various homeodomain binding preferences.

Contrastingly, there appears to be a clear enrichment for the components of the hindbrain grammar in OsCNEs (table 7.3) and GnCNEs (table 7.4). These include Pbx-Hox (TGATDWATKR) and Meis/Pknox (CTGWCA) motifs. There is one motif matching Maf preferences (DWCWGYH): *mafb* is a transcription factor expressed in r5r6. Note that Maf and Meis/Pknox preferences are very similar due to a potential 4bp overlap (CTGT), so this motif might be a variant of the Meis/Pknox consensus and may not carry functional relevance. Several motifs matching canonical homeodomain TFBSs (approximating TAATTA) are enriched, as well as motifs resembling binding preferences for POU domain factors (ATTTGCAT).

Table 6.3: The most highly scoring motifs in OsCNEs and GnCNEs. Cisfinder output for OsCNEs (A) and GaCNEs (B). Matches to the motif from online databases are also shown: the matches to JASPAR and Uniprobe using TOMTOM; and the preferences to Transfac using STAMP. Matches refer to groups of factors with similar binding preferences. Pbx-Hox, Meis, Oct and Maf motifs are displayed in red, blue, green and orange, respectively.

A. Osteichthyan CNEs (OsCNEs)

Motif	Freq	Ratio	Info	Score	TOMTOM vs Jaspar, Uniprobe	STAMP vs Transfac
…R**DGCTGWC**A…	242	4.6	7.4	303.5	Maf, Meis	Tgif
…M**ATTWGCAT**D…	1553	4.4	8.3	295.0	Oct	Oct
R**TGATHAATKR**B…	109	4.1	11.4	251.7	Homeodomain*	Homeodomain*
…Y**TNATTA**R…	137	4.6	7.9	228.0	Homeodomain*	Lhx
…K**TAATTA**V…	1485	3.5	5.6	226.7	Homeodomain*	Lhx, Otx, Hox
…G**CTGWCA**V	226	3.5	6.5	185.7	Meis, Prep, Tgif	Tgif
…T**TWATSA**B…	215	3.9	6.5	185.4	Homeodomain*	Otx, Oct
…N**TGATKWATGA**M…	1388	3.9	7.6	184.2	Hox, Pbx	Pbx
…B**CTGWCA**G…	1659	3.9	6.2	170.8	Meis	Meis
…T**THATTA**R…	265	3.2	7.7	149.0	Homeodomain*	Lhx

B. Gnathostome CNEs (GnCNEs)

Motif	Freq	Ratio	Info	Score	TOMTOM vs Jaspar, Uniprobe	STAMP vs Transfac
…K**CTGWCA**R…	2771	4.1	6.5	449.5	Meis	Tgif
…H**ATTTKMWT**D…	1020	3.8	6.9	278.8	Oct, Hox, Nkx, Sox	Oct
G**CTGKMA**W…	160	4.7	6.7	252.9	Prep, Tgif, Meis	Tgif
A**TGATWAATK**	618	4.6	5.6	242.4	Hox, Pbx	Pbx
…B**CTGWCA**R…	271	3.9	5.9	239.4	Meis, Tgif, Maf	Tgif
…T**WAATKR**A…	336	3.9	7.4	182.8	Homeodomain*	Otx, Oct
…B**TAWTKR**M…	379	3.6	4.6	175.6	Homeodomain*	Lhx, Hox, Oct
…G**CTGTCA**D…	226	3.3	7.5	146.6	Tgif, Meis, Prep	Tgif
…W**DHTGHAA**	253	3.4	6.5	143.3	Oct, Sox	Otx
…T**WRCWGWH**W…	480	3.6	5.6	132.0	Maf, Sox	None

* Motifs matching various homeodomain binding preferences.

In VeCNEs, the enriched motifs are again distinct from OsCNEs and GnCNEs. The most frequently occurring motifs are a Sox/Hox/Fox motif (AATRAAAT), Oct motifs (ATTTKCAT), Pbx-Hox motifs (TGATNAATKW) and a canonical homeodomain motif (WAATKA). There are no Maf or Meis/Pknox-like motifs in the top ten (table 7.5). Since this set of CNEs is very small, all the motifs discovered have an FDR of 1, precluding their inclusion in further analysis.

Table 6.4: The most highly scoring motifs in vertebrate CNEs. Cisfinder output for VeCNEs. Matches to the motif from online databases are also shown: the matches to JASPAR and Uniprobe using TOMTOM; and the preferences to Transfac using STAMP. Matches refer to groups of factors with similar binding preferences.

Vertebrate CNEs (VeCNEs)						
Motif	**Freq**	**Ratio**	**Info**	**Score**	**TOMTOM vs Jaspar, Uniprobe**	**STAMP vs Transfac**
…W**AATRAAAT**W…	53	5.3	7.1	156.2	Sox, Hox	Hox, Fox
W**TGATNAATGW**M	36	4.0	8.5	108.5	Hox, Pbx	Pbx
…M**TGATRAATKW**	15	3.4	10.3	88.8	Hox, Pbx	Pbx
…W**AWTSDMAW**T…	21	3.6	9.0	85.8	Sox, Oct	Oct
…M**ATTTGCAT**W…	53	2.9	9.2	77.8	Oct	Fox, Oct
…T**WAATKA**H…	47	2.9	7.0	74.4	Homeodomain*	Otx, Oct, Lhx
…H**ATTTKCAT**D…	18	3.9	9.8	73.5	Oct	Oct
…A**TGAMWTGC**Y…	9	2.5	15.3	68.8	bZip**	p54, bZip
…S**WAATARS**V…	17	2.4	15.0	66.0	Gata	Plzf
…W**WCATTTW**W…	27	3.0	7.7	53.0	Irx	Nkx, Pit1

* Motifs matching various homeodomain binding preferences.

* * Motifs matching various basic leucine zipper binding preferences.

The results are very different for elements from urochordates. The olfactores set contains some enriched motifs, but these are long, nebulous and have a false discovery rate of 1, indicating that they could have been discovered by chance (data not shown). This suggests that these elements are not bound by TFs. Interestingly, the CiCNEs are enriched for motifs matching TFBSs, but there are no Oct, Pbx-Hox, Meis/Pknox or Maf motifs. Instead, basic leucine zipper (bZip) domain dimer binding preferences (TGACGTCA) occupy nine of the ten top motifs, along with some motifs which match both leucine zipper and other preferences. The tenth motif is a Sox/Hmx motif (table 7.6).

Table 6.5: The most highly scoring motifs in tunicate CNEs. The table shows the motif as discovered by cisfinder (Motif), the number of ocurrences (Freq), the enrichment versus the shuffled sequence control (Ratio), the information content of the motif (Info), and the score (Score). Matches to the motif from online databases are also shown: the matches to JASPAR and Uniprobe using TOMTOM, and the preferences to Transfac using STAMP. Matches refer to groups of factors with similar binding preferences.

Ciona CNEs						
Motif	**Freq**	**Ratio**	**Info**	**Score**	**TOMTOM vs Jaspar, Uniprobe**	**STAMP vs Transfac**
…R**TGACGTCA**Y…	4298	5.4	11.0	1454.1	bZip*	bZip *
…S**TGACBTCA**Y…	424	3.8	8.8	399.9	bZip*, Nr4a2	bZip* Nr4a2
…R**TGAMSKMA**Y…	87	5.1	11.7	374.1	bZip*	Creb
…G**YKACGTHA**C…	147	4.5	8.7	358.9	ATF, Sox	Pax, Creb, Sox
…R**TGABRYGA**Y…	175	4.6	8.4	271.9	Sreb, Cdx, Atf	Six, Sox
…R**TKRMSKYA**Y…	85	4.7	10.2	215.6	bZip*, Cdx	bZip*
…C**GWMACAAT**R…	47	4.1	14.9	210.8	Sox, Hox, bZip*	Creb, Sox
…G**TGWCGTCW**M…	79	4.4	12.6	209.5	bZip*, Nr4a2	bZip*
…R**TGAMBKCA**Y…	86	3.4	11.9	176.2	bZip*, Rxr	Creb
…C**KKAACAAT**W…	63	3.5	14.0	171.8	Sox	Hmx

* Motifs matching various basic leucine zipper binding preferences.

Finally, two sets of elements from amphioxus could not be analysed; the first[137] was too short for Cisfinder to process, and the second[149] gave spurious results, suggestive that the sequences were derived from unannotated or unmasked coding sequence (data not shown). To summarise, the vertebrate sets are enriched for Pbx-Hox and Meis/Pknox motifs The only exception is the AcCNE set, which are not enriched for Meis/Pknox motifs (table 6.6). The CiCNE set does not have a Pbx-Hox or a Meis/Pknox motif in its top ten enriched motifs.

Table 6.6: Cisfinder discovers Pbx-Hox and Meis/Pknox motifs from vertebrate, but not chordate, CNE sets. Table shows the motif pattern discovered by cisfinder (Motif), the number of times the motif occurs in the set (Frequency), the number of motifs per kb in the set (Motifs/kb), and the enrichment ratio vs a shuffled sequence control set (Enrichment). Motifs resembling Pbx-Hox (red) and Meis/Pknox (blue) discovered from each set are shown. These motifs were not discovered in elements from invertebrate chordate genomes.

CNE set	Motif	Frequency	Motifs/kb	Enrichment
Actinopterygii (Ray-finned fish)	**TRATWAATKD**	1759	0.27	3.5
	N/A	N/A	N/A	N/A
Osteicthyes (Bony vertebrates)	…N**TGATKWATGA**M…	1388	1.67	3.9
	…B**CTGWCA**G…	1659	2.00	3.9
Gnathostomata (Jawed vertebrates)	A**TGATWAATK**	618	0.61	4.6
	…K**CTGWCA**R…	2771	2.77	4.1
Vertebrata (Vertebrates)	…A**TGATNAATGW**M	26	0.69	3.7
	…A**CTGTCA**A…	12	0.32	2.6

6.4.2 Binding site motifs are non-uniformly represented in the CNE sets

The majority of clustered motifs from the sets resemble known TF binding preferences, consistent with the known roles of CNEs as enhancers composed of TF binding sites. Next, attempts were made to address how similar the CNE sets are with regards to their enriched motifs. The top 50 elementary motifs derived from each set were aligned to one another using STAMP (see methods). The elementary motifs were used in order to ensure each was identified using the same criteria. Since Cisfinder discovers multiple variants of the same motif,

The motifs cluster in to thirteen clades, eight of which match the following transcription factor binding preferences in a 1:1 fashion: Pou, Pbx-Hox, Pbx, canonical homeodomain, Pitx/Otx, Err/Nr, leucine zipper and Meis/Pknox (figure 6.2). The remaining five clades do not match known monomer sites unambiguously. These match either multiple monomeric sites or do not significantly match any known TF binding preferences, and were excluded from further analysis.

Elementary motifs derived from each vertebrate sets are similar, but some clades lack representatives from all sets. Interestingly, none of the eight clades possess representatives from all four sets (figure 6.3 A). The vertebrate sets (AcCNEs, OsCNEs and GnCNEs) are most similar, with representatives in the Pou, Pbx-Hox, Pbx, Homeodomain and Pitx/Otx clades (figure 6.3 B). Notably, the AcCNEs are not enriched for any motif falling within the Meis/Pknox clade. Motifs derived from OsCNEs and GnCNEs fall within both the Pbx-Hox and Meis/Pknox clades, consistent with their previously characterised roles as hindbrain enhancers.

48 of the top 50 motifs derived from CiCNEs cluster in to a single clade, resembling dimeric sites for bZIP transcription factors (figure 6.3 A, B). One remaining motif falls in to the Meis/Pknox clade and the other in to an ambiguous clade. This suggests that these evolutionarily recent elements are mainly regulatory elements bound by bZIP proteins.

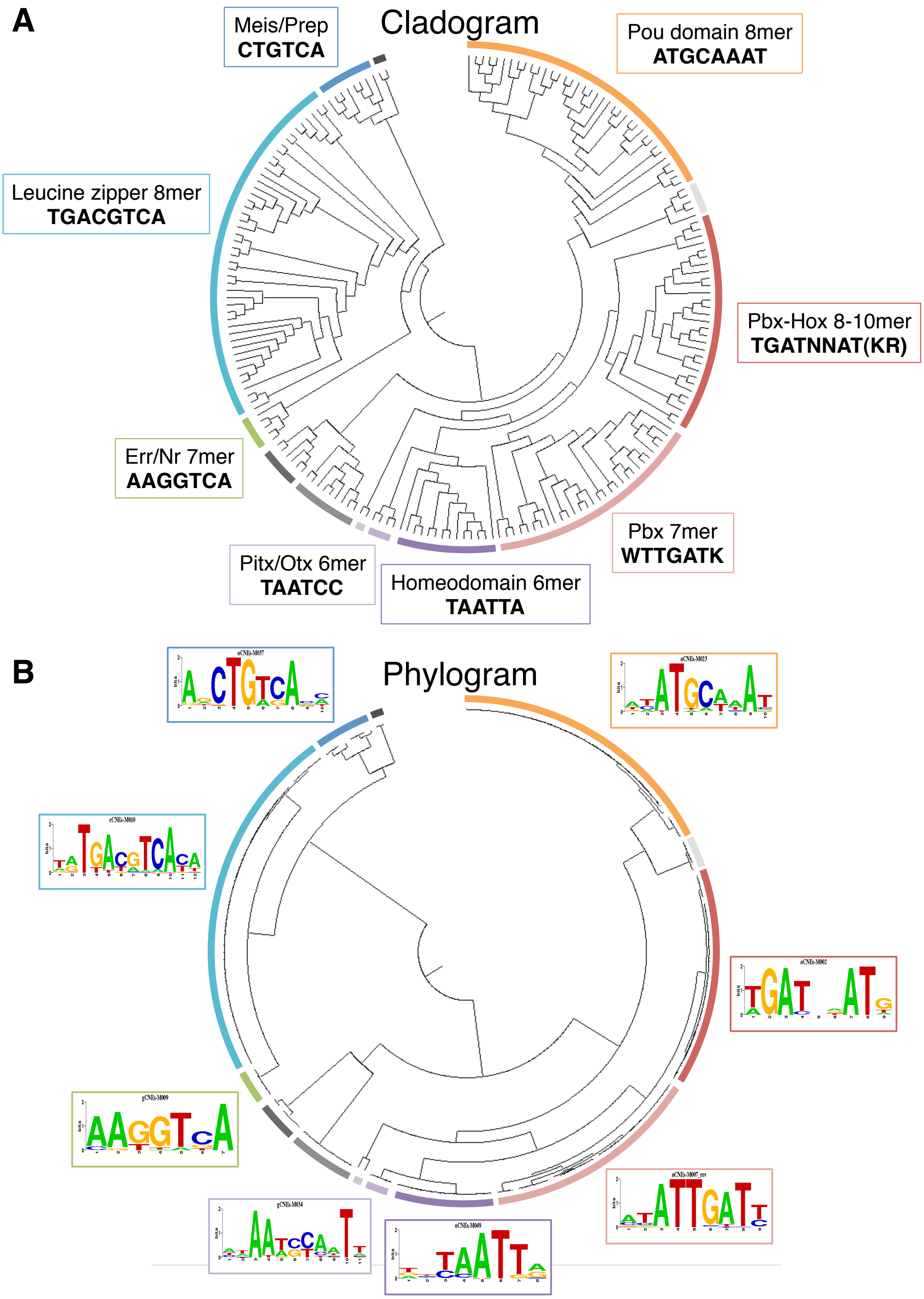


Figure 6.2: Elementary motifs derived from CNE sets cluster in to eight clades matching known TF preferences. Alignment of the top 50 motifs derived from each of the four sets, aligned using STAMP. The output tree was visualised with MEGA (see methods). The motifs are clustered hierarchically with equalised branch lengths (A) or using the Smith-Waterman distance (B). 176/200 motifs fall in to one of eight clades that resemble known TF binding preferences (coloured bars, clockwise from top: Pou, Pbx-Hox, Pbx, Homeodomain, Pitx/Otx, Err/Nr, leucine zipper, Meis/Pknox). 24 motifs fall in to one of five ambigious clades (grey bars). Each clade is annotated with its and consensus (A) and an example of a representative PWM (B).

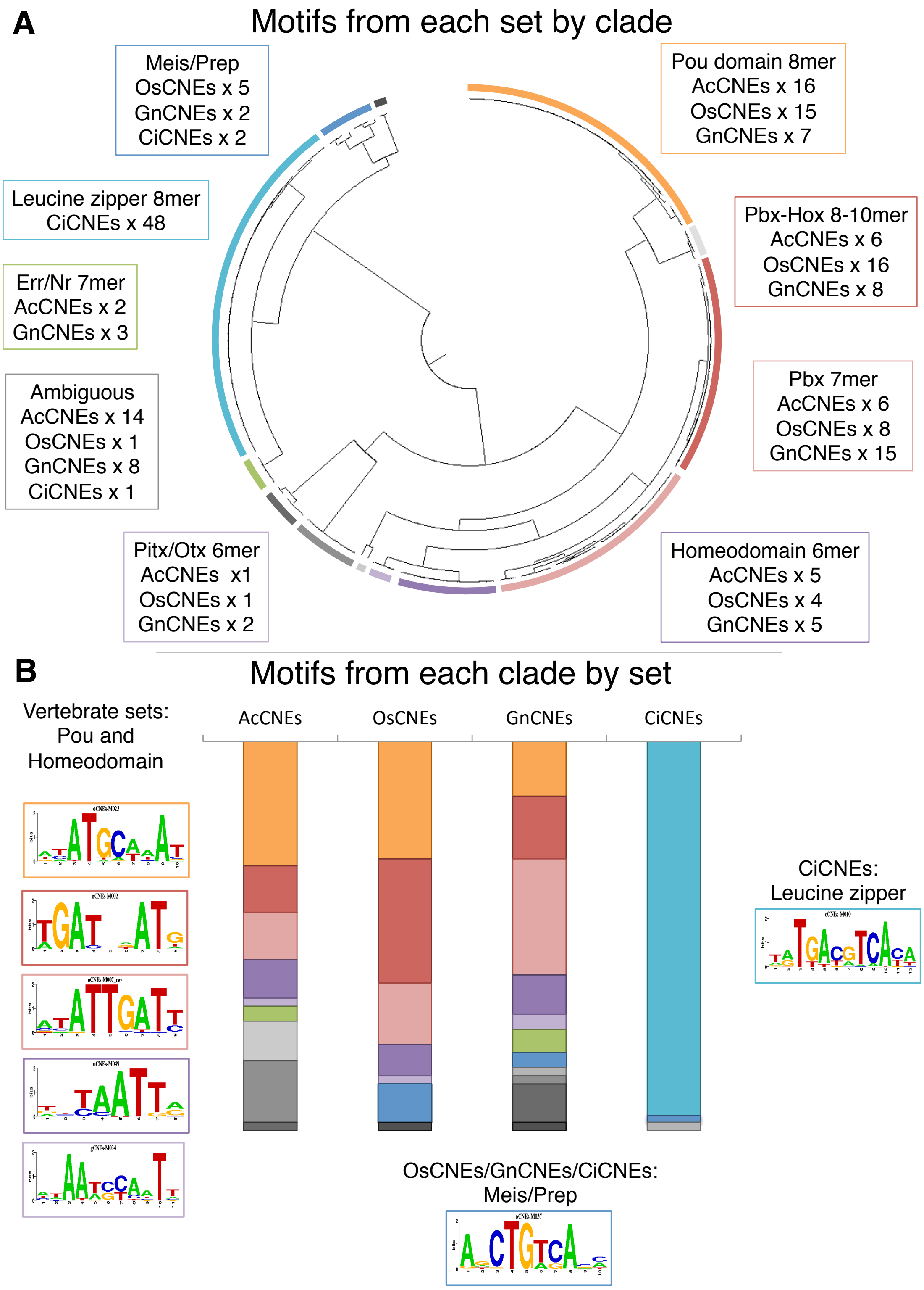


Figure 6.3: Only GnCNEs and OsCNEs contain elementary motifs clustering in both the Pbx-Hox and Meis/Pknox clades. The number of motifs derived from each set falling within each clade (A) and the number of motifs falling within each clade derived from each set (B) are shown. No clade contains motifs derived from all four sets. Consistent with their phylogenetic position, the vertebrate sets are more similar to one another in terms of motif content than they are to CiCNEs. The top 50 motifs derived from each set are distributed amongst unique selections of clades, suggesting gradual shifts in function over the course of evolution.

6.4.3 Lamprey orthologues typically lack complete hindbrain grammar

In an attempt to identify elements orthologous to these gnathostome hindbrain enhancers in the sea lamprey, an ancestral reconstruction approach was taken (see methods). A sample of CNE hindbrain enhancers was chosen from the previous chapters (x14) and Parker *et al.* 2011[157] (x7). Because of its interesting role as an initiatior element of *egr2/krox20*, *krox20* element C (henceforth *egr2b.*C) from Wassef et al. 2008[196] was also investigated.

Of the 22 elements for which an osteichthyan ancestor (OA) was constructed, 21 provided BLAST hits against the chimaera genome (table 7.8). The OA sequences had lengths of 71-481bp (mean 221bp) and gave BLAST hits in the chimaera genome with lengths of 32-490bp (mean 163bp). The osteichthyan hindbrain enhancer *foxd3*.327 could not be detected in the chimaera genome, and could therefore not be included in the analysis. The adjacent CNEs (*foxd3*.326 and *foxd3*.328) are present in the chimaera genome, but aligning this whole region with the more sensitive mLAGAN[98] cannot detect this element either (data not shown). This element may have: first arisen in osteichthyans; become deleted or diverged beyond high identity in chimaera; or the region has not been assembled correctly.

A further round of ancestral reconstruction (see methods) generated the gnathostome ancestor (GA) sequence. The GA sequences are 62%-100% of the length of their OA counterparts (mean 94%). Surprisingly, all of these GA sequences provide BLAST hits in the lamprey genome. The lamprey BLAST hits also provide best reciprocal hits to the correct CNE. The percentage ID for the regions corresponding to the BLAST hit from lamprey are 66-96% (mean 78%) between zebrafish and chimaera, 45%-76% (mean 56%) between zebrafish and lamprey and 50%-82% (mean 60%) between chimaera and lamprey (table 6.7).

Table 6.7: Sequence identity between zebrafish, chimaera and lamprey orthologues of zebrafish hindbrain enhancers. The table shows the identity of each element in pairwise comparisons for each species, calculated with ClustalW2. The name of the element being compared is displayed on the left. In accordance with the interrelationships of the species, chimaera and zebrafish share higher sequence identity with each other than they do with lamprey. Dr: zebrafish; Cm: chimaera; Pm: lamprey.

CNE name	Dr vs Cm	Dr vs Pm	Cm vs Pm
egr2b.C	67.5%	51.2%	72.5%
evi1.10719	70.4%	61.1%	71.8%
hoxd.10479	78.7%	45.9%	51.9%
meis1.1705	75.0%	59.1%	49.6%
meis2.1042	73.2%	52.6%	53.5%
meis2.1089	79.0%	56.2%	59.7%
meis2a.1090/1	68.6%	54.4%	65.0%
meis2a.1102	77.6%	62.6%	66.4%
nkx6-1.4281/2	71.5%	48.5%	57.0%
nr2f2.8394	82.8%	52.4%	58.7%
nr2f2.8470	71.9%	67.5%	66.7%
pax2a.174	96.1%	76.3%	72.4%
pou3f1.7785	84.4%	59.6%	50.9%
pou3f2b.9802	82.5%	52.3%	59.7%
shox2.5643	86.8%	44.6%	61.3%
tshz3.11226	94.6%	76.3%	82.8%
tshz3.7761	71.1%	56.5%	51.1%
znf503.10102	87.8%	54.7%	52.9%
znf503.10105	86.6%	49.5%	50.2%
znf703.10876	70.2%	57.1%	57.1%
znf703.10897	65.7%	50.8%	55.6%
mean	78.2%	56.6%	60.3%

🟥 <50%
🟥
🟩
🟩 >70%

When the alignments are extended to encompass the region containing the Pbx-Hox and Meis/Pknox motifs from the original osteichthyan alignment, a striking pattern is observed. The majority (20/21) of the chimaera orthologues contain the conserved Pbx-Hox and Meis/Pknox motif pair (table 6.8), but few of the putative lamprey orthologues contain these motif pairs (9/21), with the others missing one (4/13) or both (9/13) motifs (table 6.9).

Table 6.8: Chimaera orthologues of zebrafish hindbrain enhancers typically contain conserved hindbrain grammar. Table shows the name of the CNE for which the OA sequence was reconstructed (CNE name), the length of the OA sequence (OA length), the length of the BLAST hit from the chimaera genome (Cm BLAST), the E-value, and whether the Pbx-Hox and Meis/Pknox motifs are conserved in the chimaera orthologue. 20/21 chimaera orthologues contain conserved Pbx-Hox and Meis/Pknox motif pairs.

CNE name	OA length	Cm BLAST	E-value	Pbx-Hox	Meis/Pknox
hoxd.10479	139	120	2.00E-30	Present	Present
meis1.1705	190	32	4.00E-04	Present	Present
meis2.1042	481	490	2.00E-92	Present	Present
meis2.1089	201	186	3.00E-58	Present	Present
nr2f2.8394	111	69	2.00E-21	Present	Present
pax2a.174	76	77	1.00E-30	Present	Present
pou3f1.7785	230	218	3.00E-43	Present	Present
pou3f2b.9802	276	207	8.00E-75	Present	Present
shox2.5643	288	288	1.00E-123	Present	Present
tshz3.7761	195	45	1.00E-04	Present	Present
znf503.10105	221	210	4.00E-76	Present	Present
znf703.10876	254	53	1.90E-01	Present	Present
znf703.10897	145	24	2.00E-03	Absent	Absent
evi1.10719	71	69	5.00E-17	Present	Present
meis2a.1090/1	405	375	1.00E-133	Present	Present
meis2a.1102	275	271	7.00E-97	Present	Present
nkx6-1.4281/2	365	72	2.00E-08	Present	Present
nr2f2.8470	121	67	2.00E-15	Present	Present
tshz3.11226	93	93	2.00E-38	Present	Present
znf503.10102	366	316	1.00E-128	Present	Present
egr2b.C	133	132	3.00E-04	Present	Present

Inspecting the alignments more closely, it is evident that a few elements have retained their hindbrain grammar since the divergence of the jawless and jawed vertebrate lineages. For the most part, though, the lamprey orthologues are missing the Pbx-Hox motif, the Meis/Pknox motif, or both. Additionally, the lamprey orthologues do not contain the motifs anywhere in the CNE, indicating that there have been no compensatory sites introduced, or that the sites have been moved, perhaps by the action of insertions or deletions. This suggests that these sites have either been recruited to these regions on the gnathostome stem or have been secondarily lost in the lineage leading to lamprey. There is also one case (*znf703*.10897) where the Pbx-Hox motif is conserved in lamprey but absent from chimaera, indicating secondary loss in the lineage leading to the chimaera.

Table 6.9: Lamprey orthologues of zebrafish hindbrain enhancers do not typically contain hindbrain grammar. Table shows the name of the CNE for which the GA sequence was reconstructed (CNE name), the length of the GA sequence (GA length), the length of the BLAST hit from the chimaera genome (Pm BLAST), the E-value, and whether the Pbx-Hox and Meis/Pknox motifs are conserved in the lamprey orthologue. 9/21 lamprey orthologues contain conserved Pbx-Hox and Meis/Pknox motif pairs.

CNE name	GA length	Pm BLAST	E-value	Pbx-Hox	Meis/Pknox
hoxd.10479	122	50	4.00E-01	Absent	Absent
meis1.1705	156	42	1.70E+00	Absent	Absent
meis2.1042	481	88	5.60E-02	Absent	Absent
meis2.1089	187	125	6.20E+00	Present	Present
nr2f2.8394	69	26	3.40E+00	Absent	Present
pax2a.174	76	75	2.10E-02	Present	Present
pou3f1.7785	218	55	7.70E+00	Absent	Absent
pou3f2b.9802	207	71	1.40E-01	Absent	Absent
shox2.5643	288	45	1.60E+00	Absent	Present
tshz3.7761	196	58	9.20E-01	Absent	Absent
znf503.10105	215	35	2.80E+00	Present	Present
znf703.10876	247	65	1.80E-01	Absent	Absent
znf703.10897	117	43	7.30E+00	Present	Absent
evi1.10719	71	47	3.00E-08	Present	Present
meis2a.1090/1	403	354	8.00E-19	Present	Present
meis2a.1102	274	245	5.00E-19	Present	Present
nkx6-1.4281/2	354	56	2.00E+00	Absent	Absent
nr2f2.8470	121	113	5.00E-08	Present	Present
tshz3.11226	93	95	1.00E-15	Present	Present
znf503.10102	365	93	2.10E+00	Absent	Absent
egr2b.C	134	43	8.70E-02	Present	Present

Despite a further round of ancestral reconstruction (see methods), no BLAST hits could be obtained against urochordate or cephalochordate genomes. Taken together, these results demonstrate that whilst hindbrain enhancers shared amongst gnathostomes may have orthologues in lamprey, these do not always contain the hindbrain grammar. These traces of gnathostome hindbrain enhancer CNEs are not detectable in invertebrate chordates.

6.4.4 Hindbrain activity is dependent on the conservation of motifs

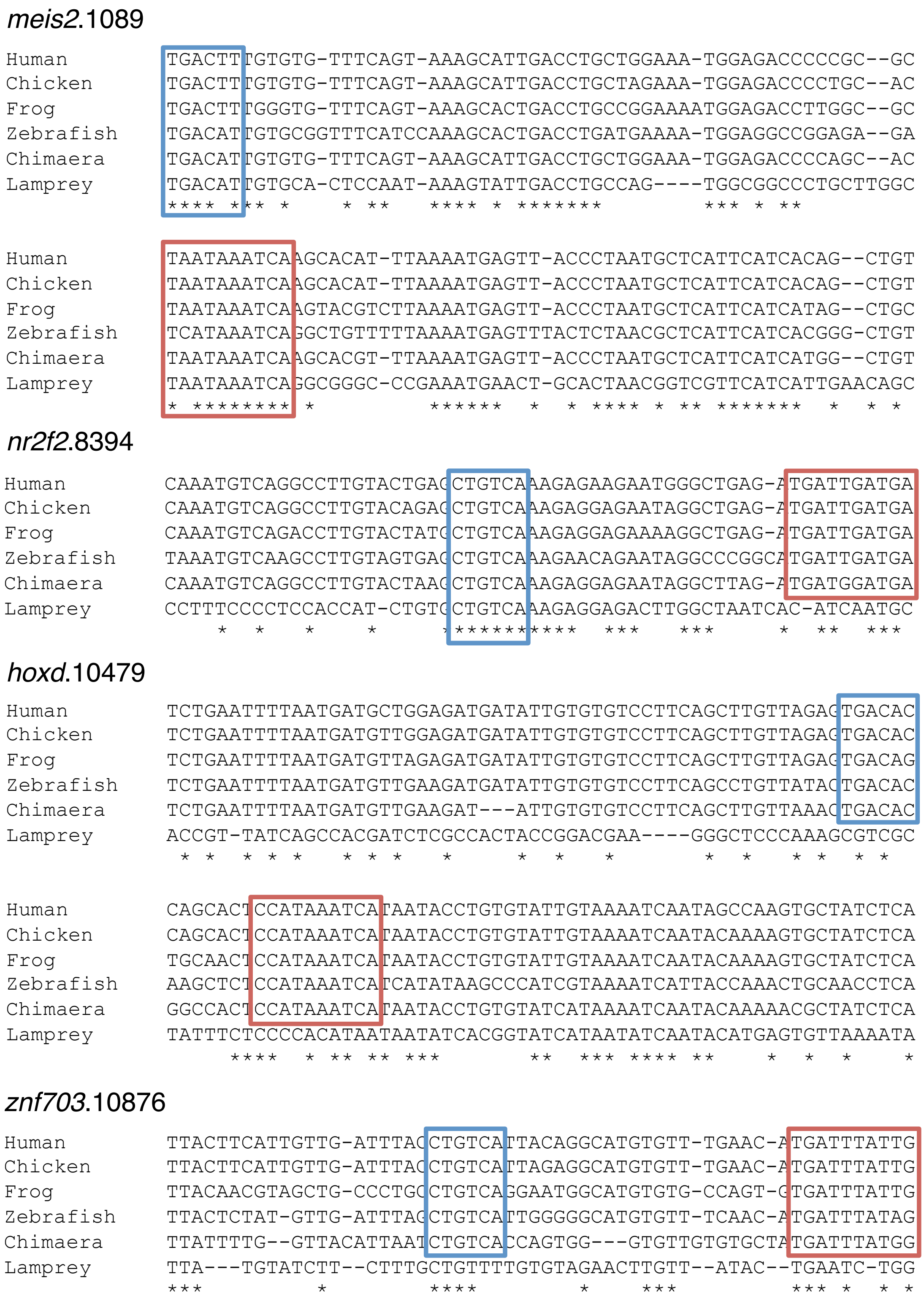
To compare the function of gnathostome hindbrain enhancers with their lamprey orthologues, four candidates were cloned and subjected to functional assay (see methods). These four elements share high identity (59-77%) with the full length of the reconstructed GA CNE. The cloned inserts share lower identity (44-57%) with the full length of the zebrafish insert, for two reasons: the zebrafish CNEs are more diverged from the gnathostome ancestor than non-actinopterygian gnathostome CNEs, and the inserts contain low-identity flanks (averaging a length of 115bp per element). Only one of these elements, Pm.*meis2*.1089, contains the hindbrain grammar, indicating the conservation of these sites to the base of vertebrates. For the other three elements, complete grammar is lacking: Pm.*nr2f2*.8394 contains only a Meis/Pknox motif, whereas Pm.*hoxd*.10479 and Pm.*znf703*.10876 contain neither motif (figure 6.4).

Table 6.10: The lamprey elements share high identity with the gnathostome ancestor CNEs. The table summarises the identity of lamprey clones across the full length of the corresponding GA CNE (Pm vs GA CNE) and the cloned zebrafish insert (Pm vs Dr insert).

	Pm vs GA CNE		Pm vs Dr insert	
Lamprey CNE	**Length (bp)**	**Identity**	**Length (bp)**	**Identity**
Pm.*meis2*.1089	187	66.3%	225	56.9%
Pm.*nr2f2*.8394	69	76.8%	220	50.0%
Pm.*hoxd*.10479	122	62.3%	346	44.0%
Pm.*znf703*.10876	251	59.0%	298	53.0%

🟥 <50%
🟥
🟩
🟩 >70%

Akin to the zebrafish orthologue, Pm.*meis2*.1089 acts as a robust hindbrain enhancer in r5r6. It appears from the expression patterns that these elements are functionally equivalent, consistent with the conservation of the hindbrain grammar (figure 6.5 A-F). Conversely, for the elements where the hindbrain is lacking, robust and segment-restricted enhancer activity is not observed. Pm.*nr2f2*.8394 is active in cranial ganglia, heart and blood. No embryos were observed that recapitulated the expression of the zebrafish enhancer in r2-r5 (figure 6.5 G-L). Pm.*hoxd*.10479 is active in mainly in trunk muscle cells, unlike the zebrafish enhancer which reproducibly drives expression in r5r6 (figure 6.5 M-R). Pm.*znf703*.10876 is also mainly active in trunk muscle cells. No embryos exhibiting specific expression posterior of the r2/3 interface were observed (figure 6.5 S-X). There are a few hindbrain positive embryos in the samples injected with Pm.*hoxd*.10479 and Pm.*znf703*.10876. However, these lack the brightness and specificity of embryos expressing zebrafish constructs.



meis2.1089

```
Human           TGACTTTGTGTG-TTTCAGT-AAAGCATTGACCTGCTGGAAA-TGGAGACCCCCGC--GC
Chicken         TGACTTTGTGTG-TTTCAGT-AAAGCATTGACCTGCTAGAAA-TGGAGACCCCTGC--AC
Frog            TGACTTTGGGTG-TTTCAGT-AAAGCACTGACCTGCCGGAAAATGGAGACCTTGGC--GC
Zebrafish       TGACATTGTGCGGTTTCATCCAAAGCACTGACCTGATGAAAA-TGGAGGCCGGAGA--GA
Chimaera        TGACATTGTGTG-TTTCAGT-AAAGCATTGACCTGCTGGAAA-TGGAGACCCCAGC--AC
Lamprey         TGACATTGTGCA-CTCCAAT-AAAGTATTGACCTGCCAG----TGGCGGCCCTGCTTGGC
                **** *** *    * **   **** * *******        *** * **

Human           TAATAAATCAAGCACAT-TTAAAATGAGTT-ACCCTAATGCTCATTCATCACAG--CTGT
Chicken         TAATAAATCAAGCACAT-TTAAAATGAGTT-ACCCTAATGCTCATTCATCACAG--CTGT
Frog            TAATAAATCAAGTACGTCTTAAAATGAGTT-ACCCTAATGCTCATTCATCATAG--CTGC
Zebrafish       TCATAAATCAGGCTGTTTTTAAAATGAGTTTACTCTAACGCTCATTCATCACGGG-CTGT
Chimaera        TAATAAATCAAGCACGT-TTAAAATGAGTT-ACCCTAATGCTCATTCATCATGG--CTGT
Lamprey         TAATAAATCAGGCGGGC-CCGAAATGAACT-GCACTAACGGTCGTTCATCATTGAACAGC
                * ******** *         ******  *  * **** * ** *******  *  * *
```

nr2f2.8394

```
Human           CAAATGTCAGGCCTTGTACTGAGCTGTCAAAGAGAAGAATGGGCTGAG-ATGATTGATGA
Chicken         CAAATGTCAGGCCTTGTACAGAGCTGTCAAAGAGGAGAATAGGCTGAG-ATGATTGATGA
Frog            CAAATGTCAGACCTTGTACTATGCTGTCAAAGAGGAGAAAAGGCTGAG-ATGATTGATGA
Zebrafish       TAAATGTCAAGCCTTGTAGTGAGCTGTCAAAGAACAGAATAGGCCCGGCATGATTGATGA
Chimaera        CAAATGTCAGGCCTTGTACTAAGCTGTCAAAGAGGAGAATAGGCTTAG-ATGATGGATGA
Lamprey         CCTTTCCCCTCCACCAT-CTGTGCTGTCAAAGAGGAGACTTGGCTAATCAC-ATCAATGC
                    *  *   *    *     ***********  ***   ***     *  **  ***
```

hoxd.10479

```
Human           TCTGAATTTTAATGATGCTGGAGATGATATTGTGTGTCCTTCAGCTTGTTAGAGTGACAC
Chicken         TCTGAATTTTAATGATGTTGGAGATGATATTGTGTGTCCTTCAGCTTGTTAGAGTGACAC
Frog            TCTGAATTTTAATGATGTTAGAGATGATATTGTGTGTCCTTCAGCTTGTTAGAGTGACAG
Zebrafish       TCTGAATTTTAATGATGTTGAAGATGATATTGTGTGTCCTTCAGCCTGTTATACTGACAC
Chimaera        TCTGAATTTTAATGATGTTGAAGAT---ATTGTGTGTCCTTCAGCTTGTTAAACTGACAC
Lamprey         ACCGT-TATCAGCCACGATCTCGCCACTACCGGACGAA----GGGCTCCCAAAGCGTCGC
                 * *  * * *   * * *   *     *  *   *       *  *   * *  * *

Human           CAGCACTCCATAAATCATAATACCTGTGTATTGTAAAATCAATAGCCAAGTGCTATCTCA
Chicken         CAGCACTCCATAAATCATAATACCTGTGTATTGTAAAATCAATACAAAAGTGCTATCTCA
Frog            TGCAACTCCATAAATCATAATACCTGTGTATTGTAAAATCAATACAAAAGTGCTATCTCA
Zebrafish       AAGCTCTCCATAAATCATCATATAAGCCCATCGTAAAATCATTACCAAACTGCAACCTCA
Chimaera        GGCCACTCCATAAATCATAATACCTGTGTATCATAAAATCAATACAAAAACGCTATCTCA
Lamprey         TATTTCTCCCCACATAATAATATCACGGTATCATAATATCAATACATGAGTGTTAAAATA
                     **** * ** ** ***       **  *** **** **    *  *  *    *
```

znf703.10876

```
Human           TTACTTCATTGTTG-ATTTACCTGTCATTACAGGCATGTGTT-TGAAC-ATGATTTATTG
Chicken         TTACTTCATTGTTG-ATTTACCTGTCATTAGAGGCATGTGTT-TGAAC-ATGATTTATTG
Frog            TTACAACGTAGCTG-CCCTGCCTGTCAGGAATGGCATGTGTG-CCAGT-GTGATTTATTG
Zebrafish       TTACTCTAT-GTTG-ATTTAGCTGTCATTGGGGGCATGTGTT-TCAAC-ATGATTTATAG
Chimaera        TTATTTTG--GTTACATTAATCTGTCACCAGTGG---GTGTTGTGTGCTATGATTTATGG
Lamprey         TTA---TGTATCTT--CTTTGCTGTTTTGTGTAGAACTTGTT--ATAC--TGAATC-TGG
                ***         *        ****        *    ***         *** *  * *
```

Figure 6.4: Hindbrain grammar became fixed in CNEs on either the vertebrate or gnathostome stems. The relevant regions of lamprey CNEs identified using an ancestral reconstruction approach are shown. Some CNEs, such as *meis2*.1089, contain Pbx-Hox (red boxes) and Meis/Pknox (blue boxes) motifs in all extant vertebrate orthologues. Typically, however, the lamprey orthologues of the CNE lack one (e.g. *nr2f2*.8394) or both (e.g. *hoxd*.10479 and *znf703*.10876) motifs.

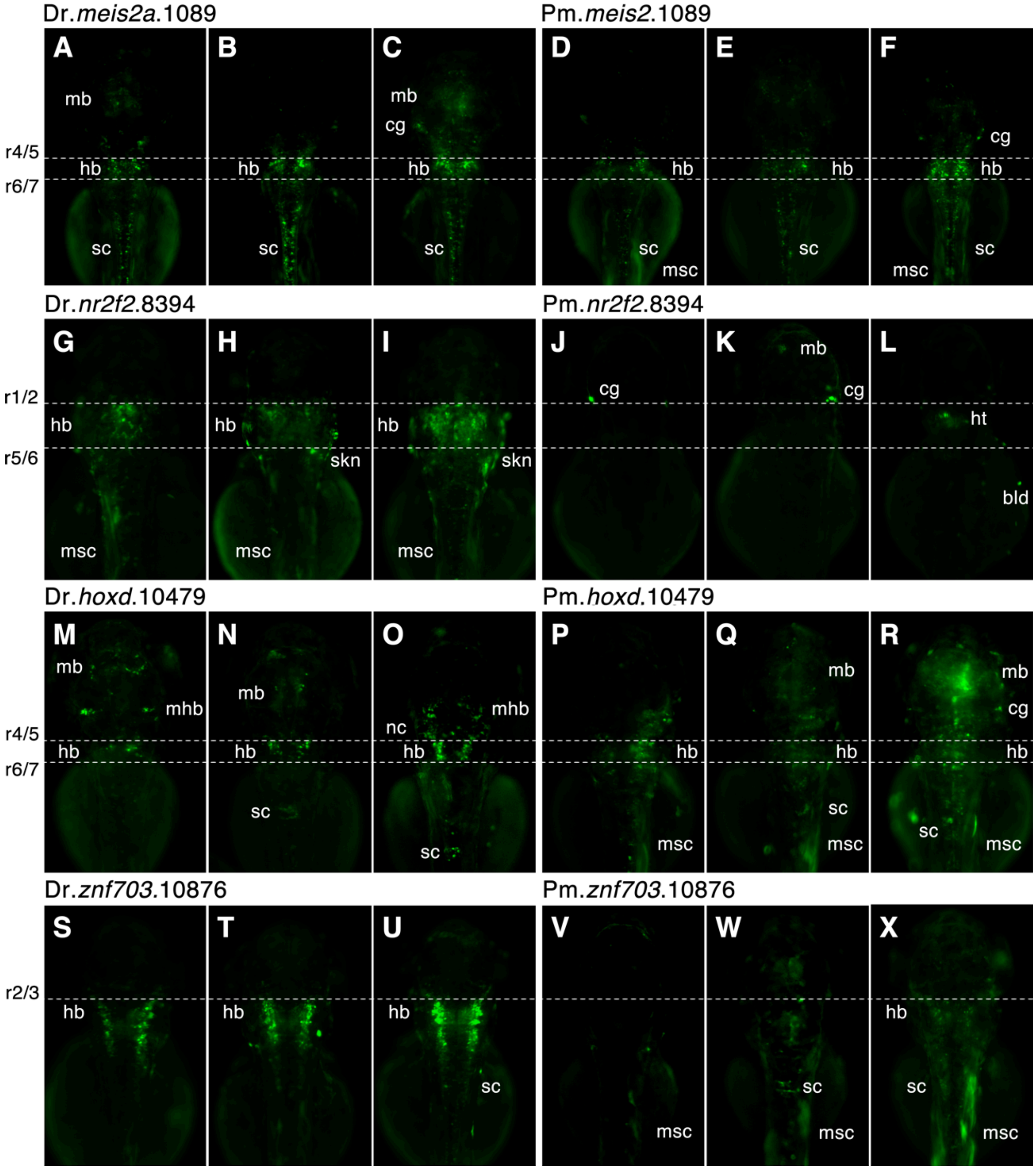


Figure 6.5: Conservation of hindbrain grammar correlates with conservation of function (caption overleaf).

mb: midbrain; hb: hindbrain; mhb: midbrain-hindbrain boundary; sc; spinal cord cg: cranial ganglia; ht: heart; bld: blood; msc: muscle; skn: skin.

Figure 6.5: Conservation of hindbrain grammar correlates with conservation of function. Zebrafish (Dr) hindbrain enhancers (left side) were compared with their lamprey (Pm) orthologues (right side). Three representative embryos are shown for each construct, demonstrating variation amongst the sample of independent insertions. Dr.*meis2a*.1089 drives expression reproducibly in r5/6 and spinal cord at ~56hpf (A, B, C). This function is conserved in Pm.*meis2*.1089 (D, E, F), correlating with the conservation of hindbrain grammar. Dr.*nr2f2*.8394 acts as an enhancer of r2-r5 at ~56hpf (G, H, I). Pm.*nr2f2*.8394, which lacks the consensus Pbx-Hox site, is typically active in cranial ganglia, heart and blood (J, K, L). Dr.*hoxd*.10479 consistently upregulates expression in r5r6 at 72 hpf, with a proportion of embryos also expressing in neural crest, midbrain, midbrain-hindbrain boundary and spinal cord (M, N, O). For Pm.*hoxd*.10479, expression in the brain is rarely observed and, where present, lacks r5r6 specificity (P, Q, R). Dr.*znf703*.10876 drives expression posterior of the r2/r3 boundary at ~72 hpf (S, T, U). Contrastingly, Pm.*znf703*.10876 drives little or no expression in hindbrain, and instead in muscle and/or spinal cord (V, W, X).

It should be noted that the only lamprey element that acts as a hindbrain enhancer is Pm.*meis2*.1089, according to the criteria used to define hindbrain enhancers first introduced in chapter 2 (Egfp positive cells in r2-r7 in at least 20% of injected embryos). This demonstrates that lamprey elements lacking one, or both, motifs do not act as hindbrain enhancers, despite sharing sequence identity with the zebrafish element outside of these motifs.

6.5 DISCUSSION

In this chapter, the evolution of binding site motifs within CNEs were investigated by elucidating the enriched motifs from different chordate CNE sets. Broadly speaking, the sets of vertebrate elements should be nested within one another according to the binary tree that describes the interrelationships of the groups concerned; the tree (((((cyprinid fishes (percomorph fishes)) lobe-finned fishes) cartilaginous fishes) jawless fishes) leads to the inverse nested pattern of the CNE sets: (actinopterygian CNEs (osteichthyan CNEs (gnathostome CNEs (vertebrate CNEs)))). This is not strictly true, since lineages can both gain and lose CNEs[277]. For example, *foxd3*.327 appears to have been lost in the chimaera genome since there are no significant BLAST hits when using the OA sequence as a query. The situation is further complicated in ray-finned fishes: they have both gained and lost CNEs with respect to the chimaera/human set; and many CNEs are retained in duplicate in ray-finned fish following their own WGD[163, 164]. A further complication is that the orthologues from the out-group were used each time, rather than, for example, sets of human orthologues being compared to one another. This was a result of the relevant publications usually publishing the orthologues from the out-group.

Even considering these facts, the overlap between the different CNE sets is likely to be substantial. The AcCNE set is likely to contain the zebrafish orthologues of all of the OsCNEs and GnCNEs, if indeed they are present in the zebrafish genome. However, since this set is >6.4 Mb, these elements probably comprise only a small fraction of the total set. For OsCNEs and GnCNEs, the overlap is likely to be more substantial, since the factor providing high specificity to the OsCNEs is the rapidly evolving teleost out-group. The overlap between the sets has been previously calculated to be 271kb; this overlap corresponds to 26.5% and 34.9% of the total length of GnCNEs and OsCNEs, respectively[157]. The discovery that similar motifs are enriched in OsCNEs and GnCNEs is therefore not surprising. Indeed, five of the top ten scoring motifs from these two sets are identical to one another, although they understandably appear with different frequencies and levels of enrichment.

For all the vertebrate CNE sets, the ten most high-scoring motifs contain both Oct motifs and Pbx-Hox motifs. POU domain factors regulate pluripotency in numerous contexts, most famously Pou5f3 (OCT4 in mouse/chicken) in regulating pluripotency during early embryonic development[278, 279]. The enrichment for Oct motifs in all the CNE sets is therefore not surprising, but it is not particularly informative since it is not known which family members are binding to these motifs in each element. Contrastingly, only OsCNEs and GnCNEs are enriched for both Pbx-Hox and Meis/Pknox (and, possibly, Maf) motifs. This reflects the functions of many individual elements conserved to this level as hindbrain enhancers remarkably well, and suggests that many more gnathostome CNEs may coordinate gene expression in the developing hindbrain and remain to be identified.

Since OsCNEs and GnCNEs sets are presumably orthologous to a subset of the AcCNEs, it may appear strange that Meis/Pknox and Maf motifs are not also enriched in the actinopterygian set. This could be because these evolutionarily older hindbrain CNEs have been ‘diluted’ by the fixation of lineage-specific elements controlling the development of lineage-specific characteristics. For example, the highest scoring motif from actinopterygian CNEs matches ZIC preferences, which are expressed in and control neurogenesis in the embryonic and adult cerebellum [214, 280]. Whilst the cerebellum is considered anatomically to be part of the hindbrain, r1, from which the cerebellum develops, is genetically distinct from the other rhombomeres; it does not express *hox* genes. One possible function for the elements from the actinopterygian set might be to control actinopterygian-specific cerebellar development. There is also a motif resembling Tcf/Lef binding preferences in the top ten motifs from actinopterygian CNEs; these factors are the mediators of Wnt signalling and this pathway has been redeployed numerous times over the course of vertebrate evolution. In zebrafish, wnt signalling activates Tcf/Lef family TFs in the skin, paired and medial fins, sensory organs and gills at various time points during development[281]. It is conceivable that these structures have a higher degree of morphological similarity amongst actinopterygians than between actinopterygians and other groups: chondrichthyans (e.g. gills and fins) or sarcopterygians (e.g. limbs). This enrichment for Tcf/Lef motifs may reflect enrichment for enhancers controlling the formation of the blood, fins, gills and digestive tract in this set. Functional validation of AcCNEs containing these motifs could test these hypotheses.

In accordance with the large number of hindbrain enhancers discovered in previous chapters, OsCNEs/GnCNEs are enriched for motifs resembling the binding sites of transcription factors involved in hindbrain development. Pbx-Hox and Meis/Pknox motifs are some of the most highly enriched motifs in CNEs, reflecting the known roles of Hox paralogs 1-4, Pbx, Meis and Prep upstream of both hindbrain segmentation and the identification of rhombomeres. Taken together with the results from the previous chapters, this could indicate that the gnathostome ancestor had a large complement of hindbrain enhancers, which are mostly retained in all gnathostome lineages today. 19 hindbrain enhancers were identified by Parker *et al.*[157], 7 hindbrain enhancers were identified in chapter 2, 66 were identified in chapter 4 by virtue of the hindbrain grammar and 3 were identified in chapter 5 by virtue of associating with *meis2a*. Of these 95 hindbrain enhancers, most are conserved in the shark genome but only a handful are detectable (by BLAST) in the lamprey genome when using extant orthologues as a query.

To investigate whether these elements were present in lamprey, an ancestral reconstruction approach was taken with a sample of validated hindbrain enhancers from zebrafish. Putative orthologues of all the zebrafish hindbrain enhancers were discovered in the lamprey, but the majority of these lack the hindbrain grammar. Even though these elements share long stretches of high identity (they would be classified as CNEs according to, for example, the CONDOR definition), they do not necessarily have conserved functions; the origins of sequence identity predate the origin of hindbrain enhancer function. Where these have been functionally validated in zebrafish, conservation of grammar correlates with conservation of function (*meis2*.1089). Conversely, those putative lamprey orthologues lacking the hindbrain grammar display divergent functions. Lamprey CNEs lacking the hindbrain grammar either do not express in hindbrain (*nr2f2*.8394) or, in the few embryos that exhibit hindbrain expression, they lack the robustness or specificity of their zebrafish counterparts (*hoxd*.10479, *znf703.10876*). This mirrors the mutagenesis results detailed in chapter 5 (figure 5.1, figure 5.2), where zebrafish elements lost their robustness or specificity after the introduction of substitutions within either the Pbx-Hox or the Meis/Pknox motif.

There are some issues that remain to be resolved. While there is as yet no evidence that the lamprey regulatory links between the *hox* code and these putative Hox targets from gnathostomes, analogous divergent or independently evolved enhancers may mediate such links. Alternatively, the lamprey regulatory state may be capable of interpreting these lamprey elements as hindbrain enhancers, although these cannot be interpreted by the zebrafish regulatory state. Where cross-species validation of lamprey and gnathostome elements has been performed, these are largely interpretable by the lamprey hindbrain GRN[157,162,183] Investigating further elements from these regions in the lamprey genome and testing these enhancers in the lamprey could distinguish between these different eventualities.

Taking in to account the results of the motif analysis, conservation of hindbrain grammar and the functional assay results, it appears that only part of the Hox target hindbrain enhancer complement present in gnathostomes evolved in stem vertebrates. This indicates that most of these Hox target enhancers arose on the gnathostome stem. Alternatively, they could have been present in the vertebrate ancestor, and secondarily lost in lamprey. Whichever the case, the fact that lamprey orthologues frequently lack the hindbrain grammar sheds doubt on the extent to which lamprey and gnathostomes share their Hox target genes. The similarity of the cyclostome and gnathostome hindbrain GRN is therefore still unclear. The results from this chapter suggest a model where some elements mediating the activation of members of the segmentation network (such as *egr2*.C and *meis2* hindbrain enhancers) arose in stem vertebrates; subsequently, in stem gnathostomes, this network was wired to target new genes through the co-option of Pbx-Hox and Meis/Pknox motifs in to established regulatory elements (such as *nr2f2*.8394, *hoxd*.10479 and *znf703*.10876). In this model, the lamprey-like vertebrate ancestor is an ecological, morphological and gene-regulatory intermediate between the amphioxus-like chordate ancestor and the gnathostome ancestor.

This model requires thorough investigation. More comprehensive sets of CNEs can be identified in lamprey using ancestral reconstruction, and the full complement of CNEs needs to be assessed for their motif content. Additional lamprey orthologues lacking the hindbrain grammar must be assayed to determine whether they can still act as hindbrain enhancers; they may do so using parallel or compensatory sites. If the lamprey orthologues lacking the hindbrain grammar do not act as hindbrain enhancers, mutations could be performed to introduce them. These modified elements could then be assayed to see if the introduced sites have activated the element as a hindbrain enhancer. This experiment would test the hypothesis that the hindbrain grammar is sufficient to generate a hindbrain enhancer from a pre-existing regulatory sequence.

This chapter has aimed to address two related, but distinct, questions: 1. what are the functions of CNE sets conserved at different phylogenetic depth; and 2. have any lineage specific changes been effected within individual CNEs? The main shortfall of this analysis is that the sets of CNEs are not directly comparable in terms of the species from which they were derived, the criteria for their inclusion or the regions of the ancestral genome from which they are descended. Depending upon the question being addressed, it would be ideal to slightly alter the sorts of sets to be analysed.

Firstly, if the aim is to assess the function of CNEs with different evolutionary ages, nested sets of sequences from a reference species should be identified using a series of progressively more distantly related species. As an example, human could be compared first with mouse; then ancestral sequences reconstructed, and these sequences used as queries against chicken; and so on with frog, zebrafish, shark and lamprey. Then, in order to assess the global functions of these nested sets, the motif content of inclusive or exclusive sets could be compared, to test various hypotheses. Given the preliminary analysis presented here, the hypothesis that human-shark elements are more highly enriched for the hindbrain grammar than human-lamprey elements could be tested. This would assess whether a larger proportion of hindbrain regulatory elements became fixed on the vertebrate or gnathostome stem. In general, methods such as these could be applied to elucidate the timing of fixation of regulatory elements with different functions, enriching our knowledge of the evolution of GRNs underlying development. Another experiment could compare selection in different predefined regions of the same CNEs using human population data; for example, do regions of the CNEs conserved in shark show stronger signatures of selection than the peripheries of the elements conserved in chicken? This would test the hypothesis that sequence gets gradually recruited and fixed at the borders of CNEs over evolution.

Secondly, if the aim is to track lineage-specific changes in individual CNEs, sets of orthologous CNEs should be defined according to the out-group, as is common in traditional comparative genomics. It would be useful to have multiple species-specific orthologues of the CNEs conserved between, for example, human and lamprey. This would define a minimal set. These minimal orthologous sets can then be compared for motif content, and orthologous CNEs from multiple species can be assayed in order to assess functional changes. This approach has been attempted within, for example, CONDOR, but complete sets are only available for fugu and human; near-complete sets from mouse and rat are available. The orthologous from other species are derived from BLAST hits, which generally works well. However, to make the sets directly comparable, these orthologous sets would need to be filtered to exclude any CNEs that cannot be found in all species of interest. If these sorts of minimal sets from different species were well defined, orthologous CNEs from any extant species (or ancestrally reconstructed sequences representing any lineage) may be analysed to track the introduction of sequence and functional novelty.

6.6 CONCLUSIONS

In this chapter, the conservation of hindbrain grammar was investigated across vertebrate and chordate evolution. The hindbrain grammar appears to be fairly common amongst gnathostome CNEs, far less common in the lamprey CNEs and totally lacking in invertebrate chordate CNEs.

In summary:

- The motif content of vertebrate and invertebrate CNE sets suggests functional distinctions between vertebrate and *Ciona* CNEs, the former containing mostly homeodomain motifs and the latter containing bZip motifs;
- Gnathostome CNEs contain clear signatures of hindbrain regulatory potential in the form of Pbx-Hox, Meis/Pknox and Maf motifs, lending support to the functional assay data;
- The lack of complete hindbrain grammar in the lamprey CNEs suggests that the lamprey either contains fewer anterior Hox target enhancers or that these remain to be identified;
- Conservation of hindbrain grammar correlates with conservation of reproducible hindbrain enhancer activity, suggesting that lamprey orthologues of gnathostome hindbrain enhancers are not typically hindbrain enhancers;
- In the elements studied, the origin of the element often predates the origin of hindbrain enhancer function, suggesting that these regions were elaborated with Pbx-Hox and Meis/Pknox sites on the gnathostome stem.

The data suggest a model where elaboration of enhancers contributed to the evolution of the hindbrain from the corresponding region in the chordate ancestor. The gradual acquisition of these enhancers appears to have elaborated the hindbrain GRN during early vertebrate evolution and contributed to the evolution of the existing overtly segmented gnathostome hindbrain via a lamprey-like intermediate possessing more subtle segmentation.

CHAPTER 7

Discussion and conclusions

7.1 Heterogeneous hindbrain enhancers share common grammar

In this thesis, the role of CNEs in the patterning and evolution of the hindbrain has been investigated. An approach combining comparative genomics, motif identification algorithms and functional assays for enhancer activity in zebrafish embryos has identified a hindbrain enhancer syntax that is both predictive of and essential for hindbrain enhancer activity.

Pbx-Hox and Meis/Pknox sites have been demonstrated to have essential roles in the activity of several well-studied hindbrain enhancers, many of which were identified in work by Robb Krumlauf and his colleagues[185]. Furthermore, because of the fascinating developmental roles of the Hox proteins in specifying segmental identity along the primary (A-P) axis in all bilaterians, Hox proteins have been subjected to much biochemical interrogation. This work led to the identification of the Hox cofactors Pbx and Meis/Pknox. Dimer- and trimerisation with these cofactors extends and modulates the binding specificities of Hox proteins to what are now well-characterised motifs. The extensive literature on the biochemistry of Hox proteins and Hox-dependent enhancers, together with novel bioinformatic data on motif content, have been used to apply binding sites for these trimeric complexes predictively, with striking results. 66/74 (89%) of the candidate sequences were shown to be active in the developing zebrafish hindbrain: a 3.7x enrichment when using either Pbx-Hox sites alone (7/29, 24%); or 2.6x enrichment using either Pbx-Hox or Meis/Pknox sites within CNEs from a known hindbrain gene (3/9, 33%). The methods outlined in this thesis could be used to identify grammars predictive of other conserved, tissue-specific enhancers, for example the majority of CNEs that remain to be functionally characterised. However, it is unclear how applicable this approach will be for TFs other than Hox proteins, since most TFs do not have such well-characterised cofactor dependence.

Closely associated Pbx-Hox and Meis/Pknox motifs are present in over 90 conserved vertebrate hindbrain enhancers identified both from previously published work[157] and this thesis (chapter 2, chapter 4). In all cases where the requirement of these sites has been assessed (both previously documented and novel), these sites are co-dependently required for the ordinary function of these enhancers, resulting in the abolition of hindbrain enhancer activity in most cases. In one exceptional case, *foxd3*.327, these sites appear to contribute to the specificity of the enhancer; mutation of either site causes ectopic reporter gene expression. The mechanism underlying this interesting result remains to be identified, but could be mediated by a number of repressive homeodomain transcription factors (chapter 5, discussion).

The *de novo* discovered motifs presented in chapter 3 are identical to the known preferences of Pbx, Hox and Meis/Pknox factors derived from EMSA, PBM and ChIP experiments. For example, Penkov *et al*. performed ChIP-seq for Pbx1, Meis1 and Prep1 on E11.5 mouse embryos[241]. Motifs matching TGATKDATD and CTGTCA are enriched under these ChIP-seq peaks. ChIP-seq data for HoxA2 from the mouse E11.5 second pharyngeal arch also detects a significant enrichment for Pbx-Hox (TGATNNAT) and Meis/Pknox (CTGTCA) binding motifs. This suggests that these motifs are likely to be bound by Pbx, Hox and Meis/Pknox proteins *in vivo*. Since the Hox, Pbx and Meis/Pknox proteins have near identical preferences, the particular proteins likely to bind to each enhancer cannot be determined with any confidence.

Despite their common binding site motifs, these enhancers generate a range of tisse-specific expression patterns (table 7.1). The final set of hindbrain grammar enhancers contains no enriched motifs aside from Pbx-Hox and Meis/Pknox. This suggests that, aside from this core syntax specifying targeting by Hox proteins, the hindbrain enhancers use a variety of mechanisms to generate their own unique expression patterns. Preliminary attempts were made to assess sets of hindbrain enhancers that generated expression restricted to subdomains of the hindbrain, but again there were no other significantly enriched motifs. This could be because the sets were very small or because segment-specific enhancers can use a variety of mechanisms to achieve the same specificity. Indeed, Burzynski *et al.* failed to find significant differences in motif content between anterior and posterior hindbrain enhancers and attributed this to functional heterogeneity[120].

This grammar identifies a proportion of, but by no means all, hindbrain enhancers. Some highly robust and segment specific enhancers were identified that lack this grammar, at least at the thresholds of significance that were selected. Examples include *foxd3*.365 from chapter 2 (figure 2.2 H), which contains a Pbx-Hox motif but only a weak Meis/Pknox consensus, and *meis2a*.965 from chapter 5 (figure 5.4 D and figure 5.5), which contains a Meis/Pknox motif but only a weak Pbx-Hox consensus. This indicates that the model favoured specificity over sensitivity, but this is perhaps preferable to identifying large numbers of false positives, as occurred in chapter 2. In order to identify more enhancers lying immediately downstream of the conserved hindbrain control network[184], future models could incorporate other transcription factor Egr2 and Mafb sites in addition to Hox sites.

Table 7.1: Heterogeneity of function amongst hindbrain grammar CNEs. Schematic representations of ISH for hindbrain TFs (A, blue) and Egfp reporter gene expression driven by CNEs (B, green) are shown. Segment specificity was determined by comparison with stable mCherry in r3 and r5 (B, red, or co-expression with Egfp, yellow). Solid shading denotes high levels of expression and dotted shading denotes low levels of expression. The expression of segment-specific TFs (A) determines downstream genes and subsequently segmentation and patterning. CNEs interpret these inputs to generate unique, complex and informative expression patterns (B). Very few of these patterns correspond to the expression patterns of single TFs (C), suggesting that most of the patterns are determined combinatorially by multiple factors.

A: Hindbrain TFs	r2	r3	r4	r5	r6	r7
hoxb1a			🟦			
hoxa2b	🟦	🟦	░🟦	🟦	░🟦	░🟦
hoxb2a		🟦	🟦	░🟦	░🟦	░🟦
hoxb3a				🟦	🟦	░🟦
hoxb4a						🟦
egr2b		🟦		🟦		
tcf2/mafba				🟦	🟦	

B: Hindbrain CNEs	r2	r3	r4	r5	r6	r7
Dr.meis2a.1042	🟩	░🟥	░🟩	░🟥	░🟩	░🟩
Dr.meis1.1705		🟨	🟩	░🟥	░🟩	░🟩
Dr.nr2f2.8394	░🟩	🟨	🟩	░🟥	░🟩	
Dr.pou3f1.7785	░🟩	🟨	🟩	🟨	░🟩	░🟩
Dr.znf703.10876		🟨	🟩	🟨	░🟩	░🟩
Dr.tshz.7761		🟥	🟩	🟨		
Dr.shox2.5643	░🟩	🟨	🟩	🟨	░🟩	░🟩
Dr.pax2.174		░🟥	🟩	░🟥		
Dr.znf503.10105		🟥		🟨	🟩	🟩
Dr.meis2.1089		🟥		🟨	🟩	░🟩
Dr.hoxd.10479		🟥		🟨	🟩	
Dr.foxd3.327		🟥		🟨	🟩	

C: Correspondance	r2	r3	r4	r5	r6	r7
hoxb2a		🟦	🟦	░🟦	░🟦	░🟦
Dr.meis1.1705		🟨	🟩	░🟥	░🟩	░🟩
hoxb3a				🟦	🟦	░🟦
Dr.meis2.1089		🟥		🟨	🟩	░🟩
tcf2/mafba				🟦	🟦	
Dr.foxd3.327		🟥		🟨	🟩	

In relation to this observation, some of the members of the hindbrain segmentation network have similar or overlapping binding preferences. For example, there is a possible 4bp overlap between the 6bp Meis/Pknox consensus (CTGTCA) and the 7bp Mafb consensus (RTCAGCW). This kind of overlap is evident in the r7/anterior spinal-cord-specific enhancer *meis2a*.965 (figure 5.4 D). This suggests that a proportion of nucleotides within these CNEs could be bound by multiple factors in different contexts. Perhaps the sites are not always occupied by Pbx-Hox or Meis/Pknox, but might be occupied by other members of the segmentation network, such as Mafb. This could add an additional layer of complexity to these enhancers and further explain their heterogeneity. As a general principle, this could be one reason why CNEs are conserved to such a high degree across their length; they contain multiple, overlapping binding sites.

The identification of this hindbrain enhancer grammar furnishes our understanding of the function of many CNEs as hindbrain enhancers and the mechanism by which these might operate. This allows these sites to be assessed with regards to the evolution of hindbrain patterning, and may enable lineage-specific Hox-target hindbrain enhancers to be identified. Furthermore, that these CNEs function as hindbrain/neural crest enhancers suggests that they are conserved because mutations within them would be disastrous for hindbrain and craniofacial patterning. Resultantly, mutations within sequences such as these could underlie hindbrain and craniofacial malformations in humans. The human genome may also contain many more Hox-target hindbrain enhancers vital for the proper development of the hindbrain and neural crest, and this grammar may help to identify a proportion of these. The requirement of these sequences for the ordinary development of the hindbrain could be tested by the introduction of targeted mutations to model organism genomes using the CRISPR-cas9 system or TALE-nucleases[282, 283]. The developmental consequences of mutations at these positions may then be fully assessed.

7.2 There is no simple relationship between sequence and specificity

Previously, detailed *in vitro* studies on the exact binding specificities of different Hox paralogs have shown that each has distinct binding preferences when forming Pbx-Hox heterodimers; as such, the central bases of the Pbx-Hox motif (positions 5/6) had been demonstrated to contribute to the choice of Hox paralog which binds the sequence[242]. One striking study demonstrated that substituting these two bases could change the pattern of a Hox-dependant enhancer from an ANTP-like pattern to an UBX-like pattern when controlling a reporter gene in *Drosophila* embryos[243]. However the authors did note that the magnitude of the effect was small, and likely due to the fact that specificity is somehow programmed in to the enhancer by the presence of other binding sites.

Nevertheless, this finding led Parker *et al.*[157] to posit that CNEs achieve segment specificity by selectively binding particular Hox proteins. Whilst this might be considered an attractive model because of its simplicity, there was no observable correlation between the bases 5/6 of the Pbx-Hox motif and segment specificity in the sample tested in this thesis (figure 7.1), and very few elements simply recapitulate Hox expression patterns (table 7.1 C). This could be because the sample of hindbrain enhancers is simply too small to detect such a correlation, or that the contribution of these bases is completely obscured by the effect of other binding sites on the elements' specificity. The second case seems plausible since CNEs are likely to be long arrays of TFBSs that have evolved their specificities by any or all means available. Regionalisation of gene expression patterns can be bought about by a number of dynamic mechanisms involving any and all factors expressed in the hindbrain during this time. These include activation by segment-specific factors, repression by factors expressed in adjacent segments, positive or negative feed-back or feed-forward, et cetera. Additionally, many factors can act as repressors or activators in different contexts; different elements bound by the same factor may not therefore be enhancers in a given segment.

Another possibility is that the long half-life of EGFP, combined with the alteration of *hox* expression patterns over the course of development, could cause some elements to generate widespread expression even if they bind particular Hox proteins. This could also explain why there appears to be a 'peak' of expression in certain segments; the lower level of expression observed in other segments may be driven at a time-point before the borders of Hox patterns have become well defined. Considering these facts, it is unsurprising that, in the full sample, there is no simple correlation between segment specificity and sequence. The development of more temporally sensitive assays, such as using a fluorescent protein with a proteasome degradation tag, may be necessary to fully assess how the segment specificity of these enhancers alters over the course of development.

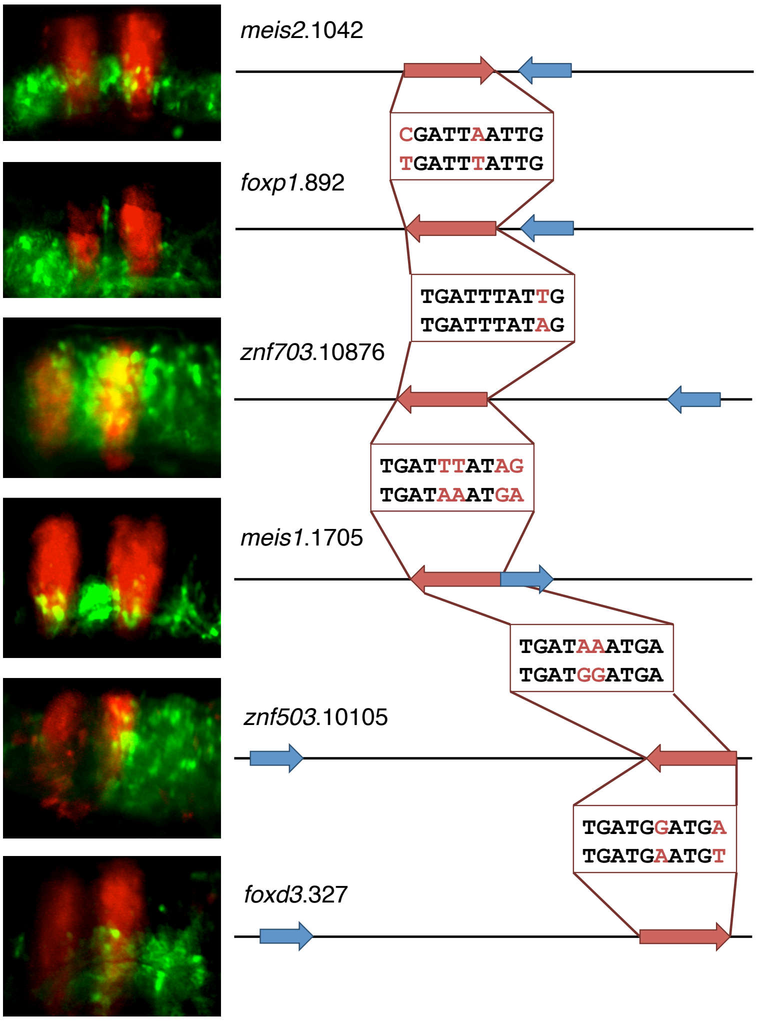


Figure 7.1: Base composition at position 5/6 of the Pbx-Hox motif does not correlate with segment specificity. Egfp expression driven by CNEs and compared with mCherry expression in r3r5 is shown on the left. Each element harbours a Pbx-Hox motif, which is shown aligned to the motif from wother elements (red boxes). Red bases denote mismatches. Elements sharing the same segment specificity (for example, *znf703*.10876 and *meis1*.1705 in r3 and posterior) have different bases at position 5/6 of the Pbx-Hox motif (TT and AA). Coversely, elements with different segment specificities (for example, *foxp1*.892, active in r2, and *znf703*.10876, inactive in r2) can possess the same bases at these positions (both have TT). Therefore, a model where these bases are informative and contribute to the choice of Hox protein in order to achieve specificity is not sufficient to explain their functions. A model of combinatorial binding appears to be more applicable.

In the full set of hindbrain grammar enhancers, there is no preference for any combination of site order or orientation. In addition, there is no observable preference for particular distances between the Pbx-Hox and Meis/Pknox motifs; the frequency of different distances is very high for the lowest distances but then decays exponentially, with no gaps in the distribution (figure 7.2). It appears that there are no unfavourable distances for determining the functionality of these sites, and that absolute distance between the sites is irrelevant in determining hindbrain enhancer activity. This distribution suggests that these enhancers require co-occurrence of the motifs within a certain loosely defined genomic space, and is what one might expect if there was selection only for site co-occurrence.

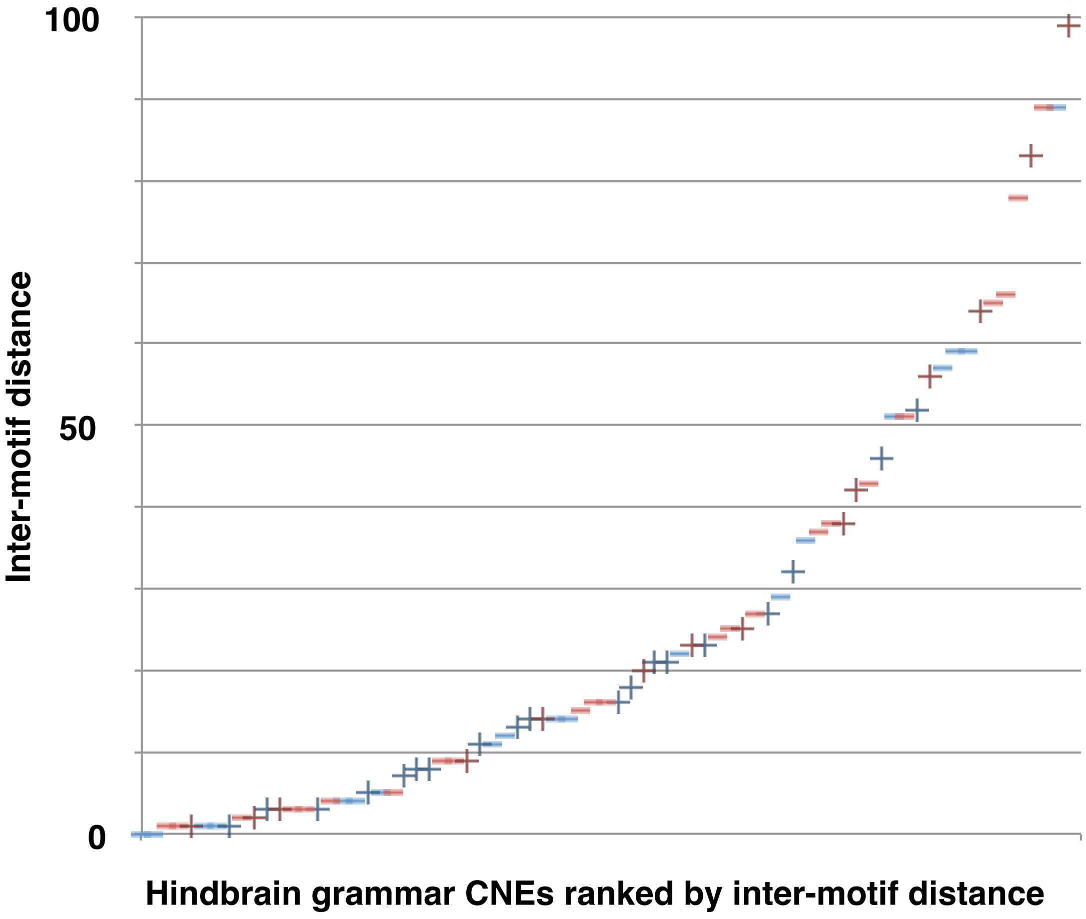


Figure 7.2: Pbx-Hox and Meis/Pknox motifs do not form higher-order structures within hindbrain enhancer CNEs. 75 hindbrain grammar CNEs are displayed, ranked by inter-motif distance in bp. Symbols denote the element's structure. All the sequences are oriented with respect to the Pbx-Hox motif (TGATDDATKD). A blue symbol denotes that the Meis/Pknox motif precedes the Pbx-Hox motif; a red symbol, follows. The orientation of the Meis/Pknox motif is also indicated (+ denotes CTGTCA, - denotes TGACAG). There is no preferred motif order or orientation. There appear to be no particularly favourable inter-motif distances.

These parameters could affect the function of individual enhancers in more subtle ways, such as the level of expression generated. The data in this thesis were insufficient to determine any such correlation; to determine any such differences would require intensive investigation. One possible experiment could be to introduce single targeted insertions of CNE-*cfos*:*egfp* transgenes to independent zebrafish lines, allowing direct comparison of Egfp expression levels by either: 1. assessing Egfp brightness with identical camera settings; or 2. By performing comparative qRT-PCRs for *egfp* on batches of embryos. Furthermore, it is difficult to assess the likely parameters that could determine hindbrain enhancer activity because very few hindbrain grammar-containing elements are hindbrain negative (n=8). This may be because the grammar of these enhancers was for some reason insufficient, or equally these enhancers may not have had their boundaries appropriately delineated leading to the cloning and testing of incomplete elements. One case that supports this is *pou3f2*.9802; a shorter version of this element was assayed previously and was shown to have no significant activity[157], whereas a longer version of this element has significant hindbrain enhancer activity (figure 4.4 F and table 4.2). There are other published cases where even small changes in the length of the CNEs drastically alters the observed activity[163], presumably because of the inclusion of poorly conserved or lineage-specific, but nevertheless functional, sites.

Despite this variation when considering the set as a whole, individual enhancers usually conserve the distance between and orientation of the sites. There are three possible explanations for this observation:

1. The distance and orientation are important to the particular function of individual enhancers, such that distance is maintained by purifying selection and the DNA between the sites is uninformative;
2. The distance is conserved because other important binding sites lie between and/or overlap the Pbx-Hox and Meis/Pknox sites;
3. The apparent maintenance of distance and orientation may be an artefact caused by the selection of only very highly conserved sequences for this analysis; sequences capable of tolerating such mutations would not be classed as CNEs.

Given that CNEs have been maintained at very high sequence identity for at least 450 million years, the first eventuality seems unlikely. The second and third possibilities both seem applicable but are not mutually exclusive. Since the majority of the hindbrain enhancers from this thesis express in at least one additional tissue, CNEs might not be "enhanceosome" elements, but rather several abutting, juxtaposed or nested "billboard" enhancers. If this is the case, we might expect that CNEs are active in multiple tissues (they are), and that these enhancer activities should be separable. However, the results of the mutagenesis experiments are mixed in this regard. There are cases where the generation of two tissue-specific expression patterns rely on distinct positions in the sequence. For example, *meis1*.1705 is an enhancer of the ventral hindbrain and spinal cord (figure 5.2 G). Mutating either its Pbx-Hox or Meis/Pknox motif abolishes hindbrain expression but does not affect the number of embryos positive for spinal cord expression (figure 5.2 C). In a contrasting case, *pax2*.174 is an enhancer of hindbrain (r3-r5) but also contains a lens enhancer (figure 5.2 E). Mutating either its Pbx-Hox or one of its two Meis/Pknox motifs ablates both hindbrain and lens expression. This demonstrates that these positions within the CNE are required for expression in multiple tissues, or perhaps they perform some necessary prerequisite function, for example by marking the sequence as accessible with epigenetic marks at some earlier time-point during development. Experiments to test these different eventualities e.g. insertions/deletions of interstitial DNA, mutations to meaningless sequence or inversions would probe this grammar and test the two models further. Detailed studies of individual enhancers[52, 53] have complemented large screens for enhancer activity and experiments of this sort must be performed to rigorously test this grammar.

The properties of these enhancers as a set suggest they are not structured enhanceosomes. They achieve diverse functions within the hindbrain and are commonly multifunctional (i.e. active in other tissues); this is better accounted for with a model of combinatorial binding. The set contains no additional enriched motifs other than Pbx-Hox and Meis/Pknox; this suggests the formation of analogous protein complexes at these elements cannot explain their extensive conservation. I suggest a model whereby CNEs are long arrays of loosely arranged, but closely associated, abutting and/or overlapping binding sites. Their functions are mediated by combinatorial binding in a billboard-type fashion, but fixation of multiple sites in the same genomic space prevents insertions, deletions or rearrangement from taking place. The concept of overlapping or abutting binding sites better accounts for the extensive conservation of these enhancers, the complex expression patterns they encode in developing tissues, and their activity in multiple tissues. With regards to the hindbrain, both activatory and repressive sites may program segment specificity; an enhancer of one segment is superimposed with a repressor of another. Enhancers of different tissues may also be superimposed upon one another. I refer to this model as the “inflexible billboard” to distinguish it from the “billboard” model, because whilst these elements could theoretically reconfigure, their densely-packed binding sites make this practically impossible, due to constraint on tightly regulated segmental expression or trade-off between activity in multiple tissues (figure 7.3).

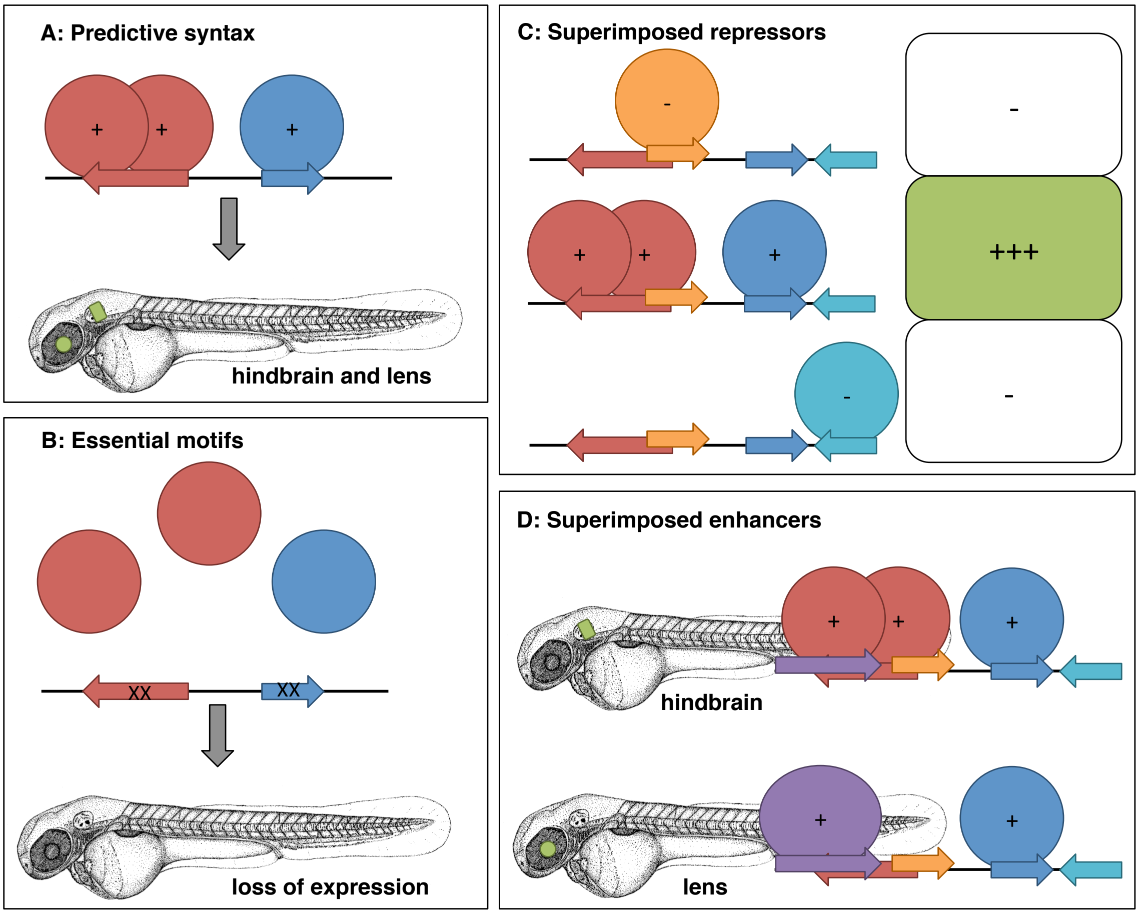


Figure 7.3: CNEs as multifunctional enhancers and repressors (inflexible billboards). A hypothetical CNE, which acts as a regulatory element in the lens and hindbrain, is shown to illustrate the model. A: Although Pbx-Hox and Meis/Pknox sites predict the enhancer and correctly identify one of its functions (hindbrain), lens expression is also observed. B: The syntax is shown to be essential for both hindbrain and lens enhancer function. C: segment-specific activation in the hindbrain is achieved through activation of the element in the hindbrain and repression of the element in adjacent segments, integrated through the unique regulatory states of each segment and overlapping binding sites for repressive factors. D: Two combinatorial codes, one for hindbrain and one for lens, are juxtaposed in the same genomic space. Thus, the sequence acts as a bifunctional enhancer. These two concepts (the superimposition of enhancers and repressors) are consistent with the known functions of the hindbrain enhancer CNEs and go some way to explaining their heterogeneity in function and extensive conservation. Mutations within overlapping or multifunctional sites will be strongly selected against, consistent with the maintenance of CNEs by purifying selection; therefore, whilst theoretically mutations and rearrangements are permitted, these are highly likely to disrupt other functions or binding sites. I suggest that these mechanisms, and not selection for precise protein-protein interactions, can account for the conservation of CNEs and suggest that they are inflexible billboards.

Indeed, in more general terms, using evolutionary conservation detects only elements that have resisted extensive changes (sequence divergence, insertions, deletions and rearrangements). Therefore, studying only CNEs to distinguish between the "billboard" and "enhanceosome" models is not appropriate: if we study only the most highly conserved sequences, the enhanceosome model will be invariably found to be more applicable. To address this issue, sets of regulatory elements identified using various methods must be compared and contrasted. For example, CNEs at different taxonomic levels could be compared to one another as was attempted in chapter 6, or elements identified through other, motif-based approaches could be studied in parallel. If common rules can be found in the organization of regulatory elements identified by multiple methods, then these rules can truly be said to constitute syntax.

The characterisation of the Pbx-Hox and Meis/Pknox hindbrain enhancer grammar in a large number of examples more easily allows us to assess the functional consequences of mutations at these sites with regards to evolution, development and human disease. This grammar could be used to discover putative human hindbrain/neural crest enhancers, which might be the kinds of regulatory sequences mutated in human developmental disorders. Targeted deletion or mutation at these sites, using for example TALE-nuclease or the CRISPR/cas9 system, could also be used to test their requirement for normal hindbrain/neural crest development in model organisms. Furthermore, the generic method outlined in this thesis could be used to generate enhancer grammars for other tissues. There remains the caveat that this approach may not be applicable to other factors that have little or no characterised cofactor dependence.

The lack of an easily comprehensible relationship between sequence and function precludes the design of bespoke enhancers with desired tissue specificities. However, since individual tissues and unique cell types must, during development, be determined genetically, enhancer design, whilst theoretically possible, should be unnecessary. Enhancer prediction using conservation, motif clustering and systems-level biochemical approaches, alongside enhancer trap assays, can increase the arsenal of known enhancer sequences available to the molecular biologist. These enhancers can then be used to drive reporter genes for the study of cell behaviour or other biological molecules for the study of the molecular control of development.

7.3 The acquisition of Hox-dependant CRMs in early vertebrates

Previous work suggested that the acquisition of Pbx-Hox sites within CNEs contributed to the elaboration of the hindbrain and pharyngeal region. This indicated the first direct mechanistic connection between CNEs and hindbrain evolution[157]. This study estimated that 400-500 such hindbrain enhancers might exist, by virtue of their containing conserved Pbx-Hox motifs. However, the data presented in chapter one cast doubt on the sufficiency of Pbx-Hox sites alone to act as hindbrain enhancers. It must be stated that this changes only the numbers of elements estimated to perform a hindbrain regulatory function. The principal that CNEs contain a grammar specifying hindbrain expression, and thus coordinate patterning events in the hindbrain/pharyngeal region is still valid, and indeed is supported by much of the data in this thesis. Furthermore, these are not the only elements that could act as conserved hindbrain enhancers, as shown by the expression pattern of *meis2a*.965, which does not contain a Pbx-Hox site.

Vertebrate CNEs have few orthologues detectable in invertebrate chordates[137, 138]. This is consistent with the idea that they control aspects of development that are vertebrate novelties (vertebrate phylotypic development). The hindbrain enhancers detected here are correspondingly absent from invertebrate chordates. Putative orthologues of these enhancers were also identified in the sea lamprey genome (chapter 6), many of which lack the hindbrain grammar. Where the hindbrain grammar is conserved in these orthologues, segment specific function is also conserved. Conversely, orthologues lacking Pbx-Hox and Meis/Pknox motif pairs do not drive gene expression in the zebrafish hindbrain (table 7.2). It seems that some hindbrain grammar CNEs evolved on the vertebrate stem but most evolved later on the gnathostome stem, possibly by incorporating Pbx-Hox and Meis/Pknox sites at the periphery of preexisting regulatory elements. This suggests a model where the lamprey represents an ecological, morphological and gene-regulatory intermediate between amphioxus and gnathostomes, consistent with its phylogenetic position.

Table 7.2: Conservation or absence of hindbrain enhancer activity in lamprey orthologues of gnathostome hindbrain enhancers. Putative orthologues of gnathostome hindbrain enhancers, whilst detectable in lampreys, do not act as segment-specific enhancers in the zebrafish hindbrain, unless the hindbrain enhancer grammar is also conserved.

	r2	r3	r4	r5	r6	r7
Dr.meis2.1089*		🟥		🟨	🟩	▨
Pm.meis2.1089*		🟥		🟨	🟩	▨
Dr.nr2f2.8394*	▨	🟨	🟩	▧	▨	
Pm.nr2f2.8394		🟥		🟥		
Dr.hoxd.10479*		🟥		🟨	🟩	
Pm.hoxd.10479		🟥		🟥		
Dr.znf703.10876*		🟨	🟩	🟨	▨	▨
Dr.znf703.10876		🟥		🟥		

*elements containing Pbx-Hox and Meis/Pknox motifs.

CNEs, which are maintained by purifying selection[133, 284], have long been hypothesised to control aspects of development phylotypic for the group concerned[135]. Several lines of evidence support this hypothesis. First, CNEs exist in lineages that share developmental similarities (i.e. they share a body plan). Diptera, *Caenorhabditis* sp., *Ciona* sp. and vertebrates all have their own sets of CNEs that have evolved independently[97, 134]. This demonstrates a correlation between large amounts of non-coding sequence conservation within, but not between, phyla. This has been thought to indicate the similarity of GRN topology amongst different phyla during mid-development.

Second, there is biochemical evidence that supports the notion that CNEs play a role in regulating genes expressed at the phylotypic stage. For example, RNA-seq data from zebrafish suggests that genes expressed at phylotypic stage typically associate with more CNEs than those expressed during early or late development[285]. Furthermore, histones associated with conserved regions of the genome become increasingly marked with H3K27Ac, thought to indicate regulatory element use, as the phylotypic stage approaches[116]. These studies suggest that more CNEs, and thus their associated genes, are activated during mid-embryogenesis, again suggesting that regulatory links are highly constrained during this time.

Indeed, a recent computational model simulating the evolution of a GRN controlling development found that, by selecting against excessive perturbation to the GRN, that the network naturally acquires an hourglass-like shape, with fewer genes at the waist[286]. Together, these lines of evidence indicate that, through the action of purifying selection, both regulatory sequences and network topologies are most constrained during mid-embryogenesis.

That CNEs act to constrain such topologies seems plausible but lacks strong support. The data described herein are consistent with the idea that CNEs control phylotypic development. Firstly, a large number of CNEs act as enhancers of the hindbrain, a vertebrate phylotypic structure. Further to this, the number of hindbrain grammar elements is consistent with the acquisition (in lamprey) and elaboration (in gnathostomes) of the hindbrain GRN. I hypothesise that these enhancers contributed to the evolution of hindbrain patterning and segmentation during early vertebrate evolution and are conserved because they specify GRN topologies required for normal hindbrain development in extant vertebrates.

These hypotheses could be tested with a variety of experiments. There are identifiable putative orthologues of the zebrafish hindbrain enhancers in the shark and lamprey genomes, but these do not necessarily contain the hindbrain grammar. Indeed, the results presented in chapter 6 demonstrate that lamprey orthologues lacking the hindbrain grammar cannot function as hindbrain enhancers in zebrafish. Previously, it was hypothesised that Pbx-Hox motifs acted as platforms for clustered binding sites to become fixed as CNEs[157]. In the light of new expression data (table 7.2), the reverse appears true: these motifs were apparently introduced in to pre-established CNEs in gnathostomes, leading to the acquisition of novel function. Perhaps introduction of these sites in to the lamprey orthologues by site-directed mutagenesis might activate some latent hindbrain regulatory potential in these sequences. This would test the model that this hindbrain grammar was introduced to pre-existing regulatory elements during early vertebrate evolution to generate novel gene expression patterns in the hindbrain.

CNEs have often been referred to as DNA "fossils", implying that they give us insight in to the regulatory nature of the ancestral vertebrate. However, the observation that CNEs are "frozen in time" may be because, akin the incompleteness of the fossil record, very few animals have full genome sequences. Even fewer of these genomes have been probed fully for conserved regulatory sequences. Constrained elements have been detected frequently amongst vertebrates (table 1.1, table 6.1, figure 6.1) and other model genera[134, 136]. Numerous conservation tracks (generated by, for example, GERP[287]) have also been incorporated in to the Ensembl and UCSC genome browsers. However, these published sets of conserved elements identified between different vertebrate species are not directly comparable; each was generated using different approaches, algorithms and identity thresholds.

As a further complication, a number of groups, such as lampreys and lancelets, are phylogenetically isolated due to extinction events, making the identification of their regulatory sequences by conservation difficult[144]. It is also evident that CNEs 'grow' over the course of evolution, perhaps because lineage-specific modules arise and become fixed at the peripheries of existing elements[90, 269]. In order to determine the evolutionary significance of CNEs, comprehensive, pairwise, hierarchical sets of conserved elements must be determined for a range of vertebrate species occupying a phylogenetic tree with symmetrical topology and similar branch lengths. Once pairwise sets have been established, ancestral reconstruction can be performed to attempt to bridge large evolutionary distances and identify elements from phylogenetically isolated genomes such as the lampreys, *ciona* and lancelets. In this way, we can begin to understand how novel functions have been incorporated in to essential regulatory elements in spite of architectural constraint on their ancestral functions.

7.4 Implications for hindbrain evolution and its underlying GRN

The patterns of expression driven by *hox* genes are, with some exceptions, conserved amongst all bilaterian animals, and expressed in a collinear fashion from the anterior to the posterior of the organism. Amongst chordates, *hox* expression patterns in the neural tube have been inherited and modified from the chordate ancestor. There are many examples of modification to the ancestral code observed in modern chordate lineages. In amphioxus the *hox2* gene has lost its expression pattern in the nerve cord[216] and in vertebrates *hoxb1* breaks collinearity by expressing in r4 instead of the most anterior segments; this is in response to feedback from the downstream GRN[185]. In urochordates, the larvae is thought to be the stage most representative the ancestral chordate body plan; however, *Ciona* sp. have lost numerous *hox* genes and the expression patterns of the remaining *hox* genes at the larval stage appear to be highly derived[147], correlating with the degenerate morphology seen amongst urochordates. Below, I discuss the similarities and differences in the development of the anterior neural tube amongst chordates, i.e. the region expressing *hox* paralogs 1-4, the region most closely corresponding to the vertebrate hindbrain.

Gnathostomes possess a GRN for hindbrain development that is largely conserved. Correlating with the conservation of this control network, all gnathostomes possess a stereotypic pattern of cranial nerves that originate from the hindbrain[204]. Correspondingly, gnathostomes retain thousands of CNEs and ~90 of these are hindbrain enhancers associating with ~40 genes[157] (chapter 2, chapter 4). This suggests that these CNEs and their associated genes are conserved *hox* targets controlling phylotypic hindbrain development. These genes are usually transcription factors themselves, and perhaps they activate specific programs of gene expression in each segment. These *hox* target genes may be part of the hindbrain control network themselves and/or might mediate links (directly or indirectly) to the segmentation and identity EGBs.

Unlike gnathostomes, cyclostomes lack overt morphological segmentation[184], and therefore the level of similarity between the hindbrain development of lampreys and gnathostomes is a matter of contention[184, 273]. Lampreys express patterns of *hox*, *egr2* and *mafb* in the hindbrain very similar to those of gnathostomes[184, 273]. Furthermore, the first rhombomere boundary to form in gnathostomes, the presumptive 4/5 boundary, is positioned by *irx* and *hnf1* genes; these genes are also expressed in the corresponding region of lampreys[288]. These genes do not lead to overt segmentation in the lamprey hindbrain; whilst the lamprey possesses similar expression patterns, these do not lead to the same morphological output. The targets of these TFs must be somehow different in gnathostomes and cyclostomes. Some aspects of hindbrain patterning and regional identity also differ. The timing of development and the resulting morphology of the cranial nerves is largely conserved amongst vertebrates, but there are some differences in cranial nerve organisation between cyclostomes and gnathostomes, reflecting their distinct craniofacial morphology[289, 290]. The patterning of reticular neurons arising from the hindbrain is largely conserved between lamprey and zebrafish, with characteristic neurons arising from each segment[273]. Contrastingly, the lamprey *hox3* expression pattern determines branchiomotor neuron identity (as *hoxb3* does in gnathostomes), but the identity of these neurons can be altered following retinoic acid (RA) treatment without altering the position of the segment boundary, determined by the expression of a lamprey *eph* gene[273]. This is unlike the simultaneous homeosis and boundary movement seen in RA treatment in gnathostomes, and could indicate that the *hox* code does not control the positioning of morphological boundaries in lampreys. However, more recent evidence suggests that the *hox* code is fully coupled to hindbrain segmentation because lamprey hindbrains can interpret gnathostome *hox*/*egr2*/*mafb* target enhancers, for the most part, appropriately[162, 184]. The discrepancy between these two studies suggests that, whilst the control network evolved on the vertebrate stem, the genes targeted by this network, and therefore the segmentation and identity EGBs, could potentially be very different, and the two processes might not be fully integrated.

Two further lines of molecular evidence from lamprey embryos suggest that cyclostomes might not have not fully coupled the control network to the segmentation EGBs. Firstly, *in situ* hybridisation data for *eph* genes suggests both similarities and differences in expression patterns between cyclostomes and gnathostomes. In gnathostomes, *egr2* directly activates *epha4a* in r3r5. The Japanese lamprey expresses *ephc*, a divergent *eph* gene, in a similar pattern to gnathostome *epha4* in the hindbrain[273, 291], suggesting that *egr2* activates *ephc* in the lamprey. Gnathostomes also express *ephb4* in a segmentally restricted fashion; in zebrafish, *ephb4a* is expressed in r2 and r5r6[204]. However, a Japanese lamprey *ephb* gene is not segmentally restricted and is expressed throughout the hindbrain[205]. This suggests that cyclostomes diverged from gnathostomes before this segmental expression of *ephb* was established in the hindbrain or it has been secondarily lost in lampreys. Secondly, the mouse EphA4 enhancer, which is activated by *egr2* in r3 and r5 in gnathostomes, is only active in r3 in the lamprey[184], suggesting that the lamprey lacks some aspects of trans-regulatory control for the activation of this enhancer in r5.

The regions most likely to correspond to the vertebrate hindbrain in invertebrate chordates (the region of the nerve cord expressing *hox* paralogs 1-4) do not exhibit overt morphological segmentation, although in amphioxus the expression patterns of some genes, such as *islet*, suggest some cryptic form of metamerism[292]. This region also lacks expression patterns of *egr* or *maf* genes[6, 212, 213]. In addition, there are no precise boundaries between *hox* expression patterns in amphioxus[146, 216], presumably because the divergence of vertebrates and amphioxus predates the evolution of key members of the hindbrain control network and its linkage to the segmentation network. CRMs from the amphioxus and *Ciona hox* clusters can generate segment-specific expression patterns in vertebrate embryos[293, 294], but in these cases the elements are being interpreted in the context of the vertebrate hindbrain control network, where morphological boundaries are strongly delineated. As such this does not demonstrate that these elements program sharp expression borders in their host species. Amphioxus shares very few CNEs with gnathostomes and none of these contain the Pbx-Hox-Meis/Pknox grammar, suggesting that Hox-target enhancers are not conserved amongst chordates. If Hox target genes are conserved, these interactions are mediated by divergent elements that have yet to be identified.

Together, these observations suggest, rather than the dramatic and rapid gain of a complete hindbrain development network in stem vertebrates, a gradual addition of new components to the hindbrain GRN on both the vertebrate and gnathostome stems. To summarise, these facts indicate the following chronology of events in the evolution of the hindbrain segmentation network:

1. **The ancestral chordate possessed a collinear *hox* code with nebulous expression boundaries**, suggesting that it lacked both the CRMs responsible

for mediating the establishment of sharp patterns and a mechanism of border sharpening (segmentation).

2. **Stem vertebrates evolved sharp boundaries between *hox* expression domains** using three complementary mechanisms:
   - **i.** by the **co-option of new genes** (*egr2*, *mafb* et c.);
   - **ii.** by **gaining new regulatory elements** mediating cross-repressive interactions between these transcription factors;
   - **iii. by the targeting of some eph/ephrin genes** by the segmentation network, such as *ephc*/*epha4* by *egr2*.
3. **Stem gnathostomes fully coupled the *hox* code to morphological segmentation**, with two effects:
   - **i. reinforcing the sharp boundaries** between segments;
   - **ii. solidifying the link to segment identity**, i.e the hindbrain control network determining neuron morphology and connectivity.

After the EGB mediating segmentation had become coupled to the control network, it is simple to envision how novel expression patterns of segment-specific genes might arise: this could occur by generating new enhancers targeted by members of the established network, co-opting further genes in to downstream networks involved in determining the identity of neurons and guiding their axons. This could have lead to the establishment of unique identities for each rhombomere, which are then disrupted when the function of these segmentation genes are lost.

One example where a hindbrain activatory element appears to be conserved in all vertebrates is that of *egr2.C*[196, 251, 295] (chapter 7, figure 7). *egr2.C* is one of the elements responsible for the initiation of *egr2* expression in r3-5, and contains Pbx-Hox and Meis/Pknox motifs. It is possible that this element was the first to evolve in order to co-opt *egr2* in to the hindbrain control network. Indeed, there is a correlation between the presence of this conserved element and the characteristic expression of egr2 in r3/r5 in the embryo. There is no detectable orthologue of e*gr2.C* in amphioxus, which also lacks expression of any *egr* gene in the region expression *hox* 1-4. The lamprey possesses an orthologue of *egr2.C* (chapter 7) and displays r3/r5 restricted expression of *egr2*[184]. However orthologues of *egr2.A*, an r3 initiator element, and *egr2.B*, an autoregulatory element, cannot be found using the same approach**.** This suggests that the addition of Hox-responsive enhancers might have been a way to co-opt additional TFs in to the hindbrain control network in early vertebrates (figure 8.2 A). The gain of Pbx-Hox and Meis/Pknox sites to innocuous sequence (or pre-existing CRMs) might have led to the *de novo* generation (or neofunctionalisation) of enhancers in order to co-opt additional downstream genes to the hindbrain GRN. Novel expression patterns of regulatory genes in the hindbrain could have gradually built the control network and linked them to the segmentation and identity EGBs.

One speculative case for such co-option is *znf703* (figure 8.2 B). There are at least two *hox*-responsive hindbrain enhancers at this locus in zebrafish, *znf703.*10876 and *znf703.*10897 (chapter 4, figure 4.4 G and J), suggesting that at least one of the anterior Hox proteins (possibly HoxB2) activates *znf703* in the hindbrain. Orthologues of these elements exist in the chimaera and lamprey genomes (table 6.7), but the extent to which the hindbrain grammar is conserved varies (table 6.8, 6.9). In wild-type zebrafish embryos, *znf703* represses *tcf2* in r4, which acts as a repressor of *hoxb1a* in r5 and posterior. This gene is required for the normal formation of r4 because it acts as a repressor of non-r4 genes (such as the repressor *tcf2*) to permit the expression of r4 determinants (such as *hoxb1*a)[202, 203]. Little is known about this sub-network in the lamprey or chimaera. However, the lack of the hindbrain grammar in lamprey and chimaera orthologues of these elements suggests that Hox proteins might not activate these enhancers, and perhaps these species do not express *znf703* in the hindbrain, or they do so through some distinct mechanism. Indeed, the lamprey orthologue of this sequence is not active in the hindbrain (table 7.2). Cross-species comparisons for the enhancer activity of these orthologous elements and in situ hybridisation for *znf703* in zebrafish, chimaera and lamprey embryos could distinguish between these possibilities. Investigating individual cases such as this could determine whether the trends observed in chapter 6 carry functional significance.

These CNEs appear to program the links between the *hox* code and the hindbrain control network, and subsequently the segmentation and identity EGBs. Several of these elements appear to be auto- or cross-regulatory elements of *hoxa* and *hoxd* genes or their cofactors *meis1*, *meis2* and *pbx3*, due to their close association with these genes. Some of these elements activate members of the downstream control network, evidenced by the presence of Hox-target enhancers at genes such as *egr2* and *znf703*, which are required for the proper formation of r3/r5 and r4, respectively.

These facts support the model that a large number of CNEs operate as Hox dependant enhancers programming essential regulatory links in the gnathostome hindbrain, regulating the expression of *hox* genes and their cofactors, a control network of downstream TFs and EGBs for segmenting and identifying rhombomeres. The findings presented in this thesis, in the context of what is known about the GRNs underlying hindbrain development, provide evidence that numerous gnathostome CNEs act to program hindbrain segmentation and patterning in response to the *hox* code. Some, but by no means all, of these CNEs from the lamprey have conserved regulatory functions (of the orthologues discovered in chapter 6, only 43% contain conserved hindbrain grammar). This evidence suggests that the lamprey is a gene-regulatory, as well as a morphological, intermediate with regards to hindbrain development. None of these elements are detectable in invertebrate chordates, correlating with the lack of a true hindbrain in these organisms. This demonstrates a correlation between the gradual acquisition of hindbrain morphology and gene expression patterns and the number of hindbrain grammar-containing CNEs. These facts suggest a scheme where hindbrain grammar CNEs linked the *hox* code to novel downstream targets during chordate evolution to contribute to the acquisition and elaboration of hindbrain development to generate adaptive morphology, explaining their extensive conservation in modern vertebrates. This appears to have occurred by the introduction of sites to pre-existing CNEs, rather than *de novo* generation. This model is summarised in figure 7.4.

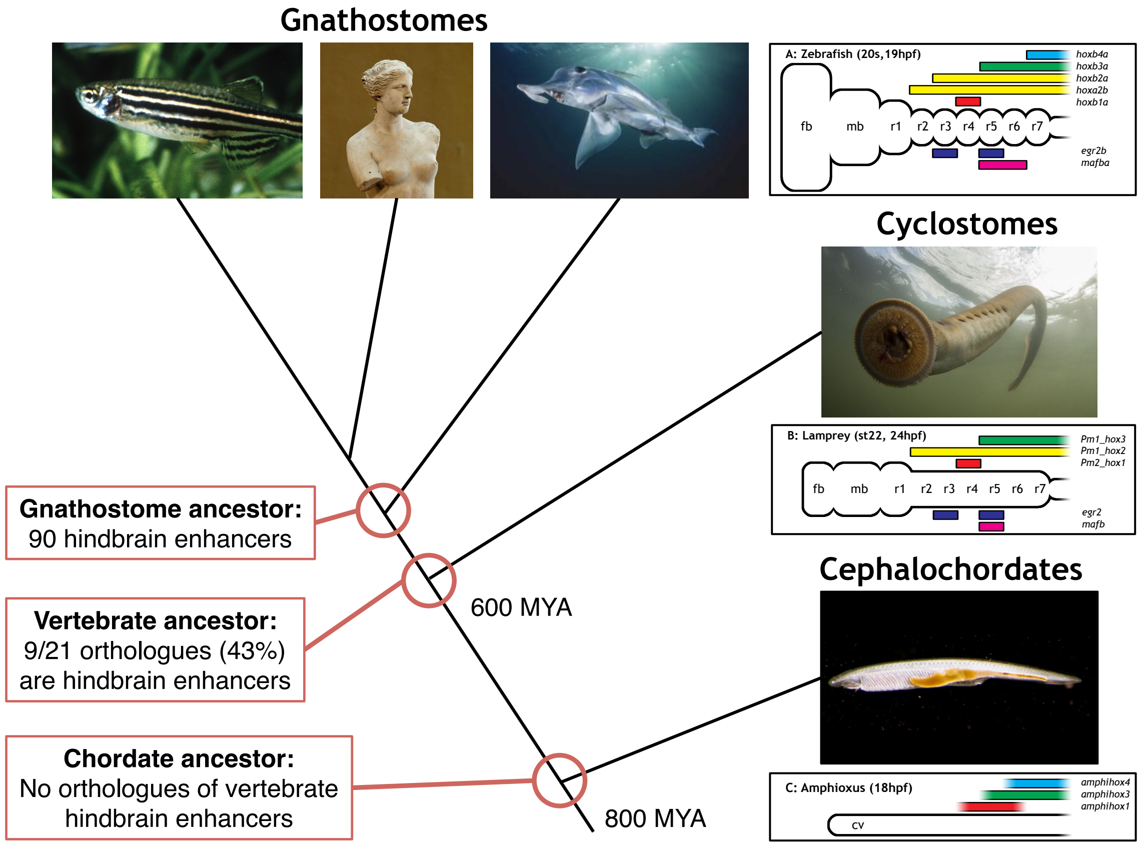


Figure 7.4: CNEs as contributors to the establishment and elaboration of the hindbrain GRN. A complement of 90 hindbrain enhancers containing Pbx-Hox and Meis/Pknox motifs are conserved amongst gnathostomes, Of the orthologues that have been found in lamprey, hindbrain activity is only conserved when these Pbx-Hox and Meis/Pknox sites are conserved; fewer than half contain the hindbrain grammar. No orthologues of these elements can be found in amphioxus. This suggests a model whereby the acquisition of hindbrain grammar programmed new regulatory links between the Hox code and downstream regulatory genes.

7.5 Concluding remarks and future directions

In summary, a library of conserved, vertebrate hindbrain enhancers has been generated. This provides a resource for those interested in studying hindbrain development, particularly for the labelling of specific hindbrain substructures. Transgenic animals harbouring stable insertions for these constructs could be grown in order to study cell sorting in particular segments. These lines could also serve as fluorescent reporters to provide readout for knockdown experiments on members of the hindbrain GRN, in the same manner as segment-specific *in situ* hybridisations or other reporter genes have been used. Such transgenic animals could also be used to assess the inputs to each enhancer, again through knockdown experiments. Failure of the enhancer to activate in the absence of a particular TF would indicate that this factor lies upstream in the GRN.

The discovery of a hindbrain enhancer grammar within vertebrate CNEs is a further step towards understanding the relationship between enhancer sequence and expression pattern. The sequences of these enhancers and their accompanying expression data generated here provide a platform for future work in investigating this relationship further. These enhancers can be subjected to further functional dissection using molecular biology approaches to introduce substitutions, deletions, insertions, part-inversions, or to generate cross-species chimearic elements.

Together with previous work, these data suggest that at least 90 gnathostome CNEs choreograph *hox* dependant gene regulatory interactions in the vertebrate hindbrain, supporting the theory that CNEs contributed to hindbrain evolution in early vertebrates. This indicates a function contributing to their conservation in gnathostomes. The data herein also suggest that CNEs program phylotypic development by maintaining essential GRN topologies. Further studies of this nature might identify links between the emergence of CNEs and other aspects of vertebrate-specific development, such as the forebrain, midbrain, cranial nerves, paired sense organs and neural crest.

The methods used in this study (comparative genomics, motif searches and enhancer assays in zebrafish) are individually well established. However, their arrangement in to a novel pipeline, informed by relevant biochemical and embryological data, has lead to the accurate prediction of tissue-specific enhancer activity in this study. The pipeline outlined here could be applied in order to learn *cis*-regulatory grammars for other tissues or those encoded in other categories of non-coding sequence.

APPENDIX

8.1 Relevant publication

The data from chapter 1-5 were reported in the following publication:

Grice, J., Noyvert, B., Doglio, L. and Elgar, G. A simple predictive enhancer syntax for hindbrain patterning is conserved in vertebrate genomes.
PLoS One. 2015 Jul 1;10(7):e0130413. doi: 10.1371/journal.pone.0130413. eCollection 2015.

This publication can be found at the following URL:
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130413

8.2 Supplementary information

8.2.1 Coordinates of assayed CNEs

The coordinates of all the zebrafish CNEs assayed in this study, and their orthologues from human, can be found in a .xlsx file available at the following URL:
http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0130413.s009

8.2.2 Chapter 2

Expression data for 29 CNEs containing Pbx-Hox motifs (TGATNNAT) can be found in a .xlsx file available at the following URL:
http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0130413.s006

8.2.3 Chapter 3

The hb+ set (n=38) used for the MEME analysis can be found in a .txt file available at the following URL:
http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0130413.s013

The hb- set (n=150) used for the MEME analysis can be found in a .txt file available at the following URL:
http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0130413.s014

The PWMs for motif 1 and motif 2 derived from the hb+ set can be fount in a .txt file available at the following URL:
http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0130413.s015

Alignments of 22 hb+ sequences, annotated with the position of conserved Pbx-Hox (TGATDDATKD) and Meis/Pknox (CTGTCA) motifs can be found in a .pptx file available at:
http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0130413.s004

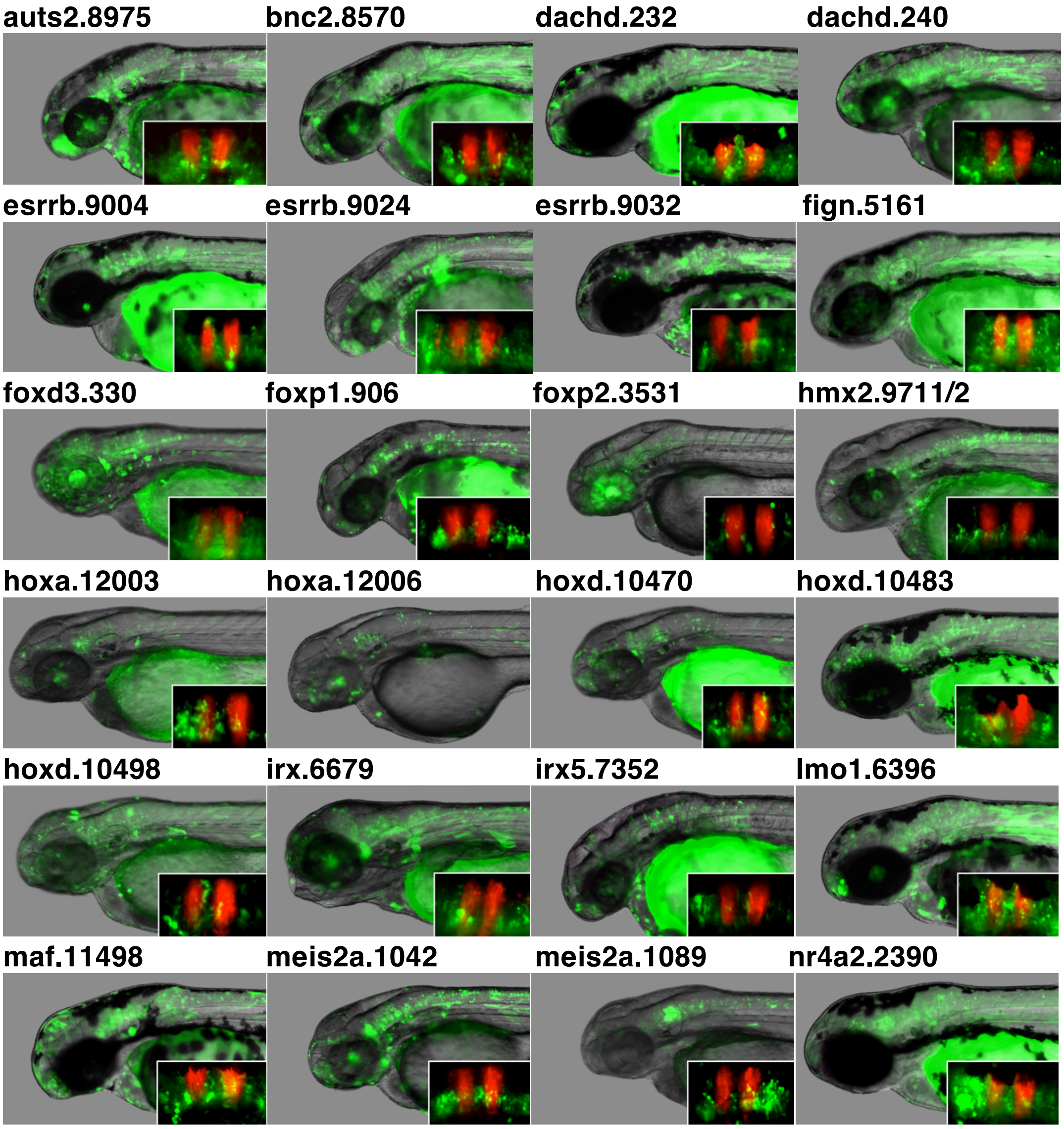
8.2.4 Chapter 4

The output of the FIMO search using human, mouse, rat/dog and fugu CNEs can be found in a .xlsx file available at:
http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0130413.s007

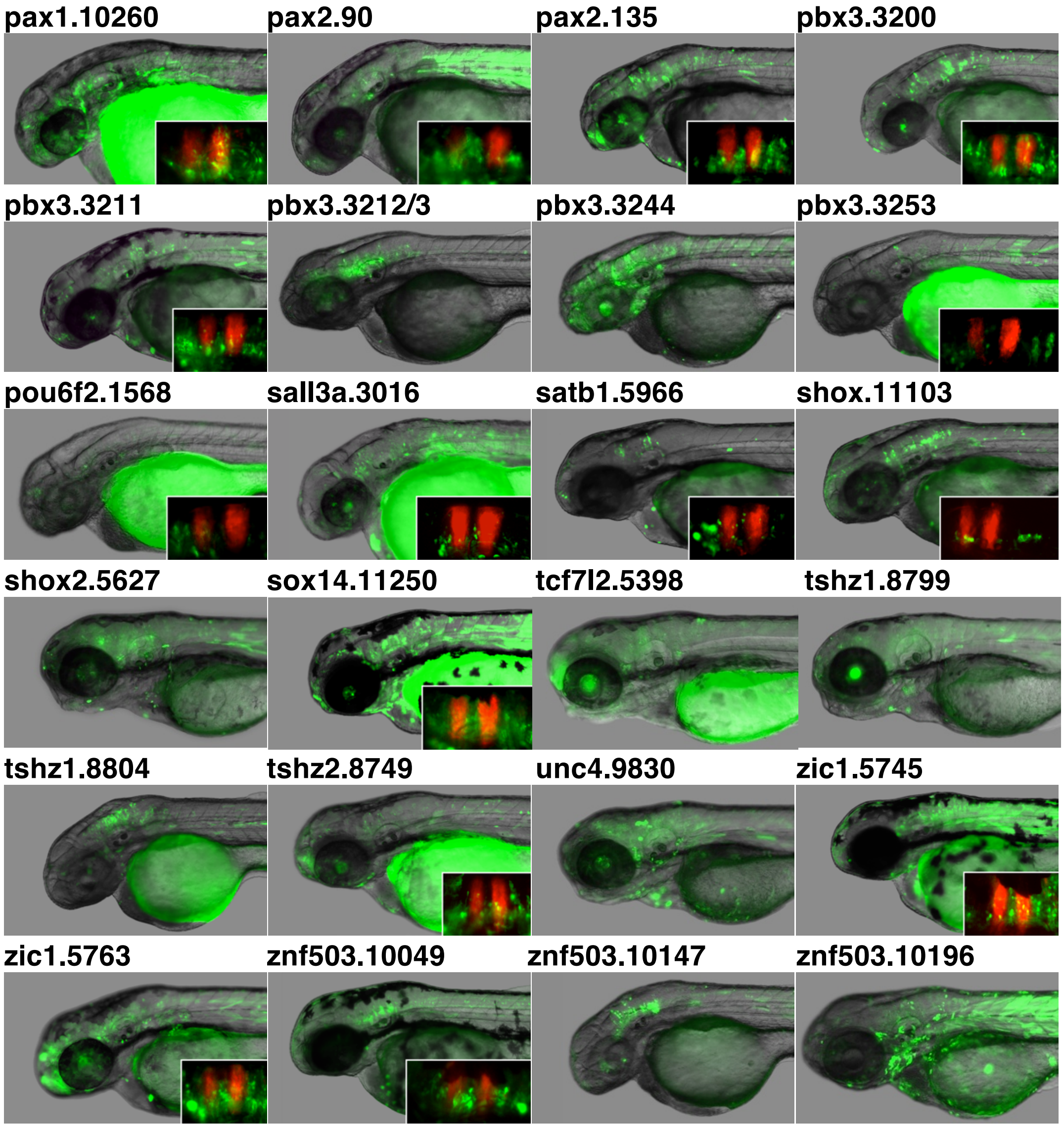
Alignments of 74 hindbrain enhancer candidate sequences, annotated with the position of conserved Pbx-Hox (TGATDDATKD) and Meis/Pknox (CTGTCA) motifs can be found in a .pptx file available at:
http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0130413.s005

Expression data for 74 CNEs containing Pbx-Hox (TGATDDATKD) and Meis/Pknox (CTGYCA) motifs can be found in a .xlsx file available at:
http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0130413.s006

Supplementary figures: reporter gene expression patterns driven by zebrafish hindbrain enhancers.



Supplementary figures: reporter gene expression patterns driven by zebrafish hindbrain enhancers.



8.2.5 Chapter 5

Expression data for 9 CNEs containing Pbx-Hox (TGATDDATKD) or Meis/Pknox (CTGYCA) from the *meis2a* locus can be found in a .xlsx file available at:

http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0130413.s010

REFERENCES

1. Irie N, Kuratani S. The developmental hourglass model: a predictor of the basic body plan? Development. 2014;141(24):4649-55. Epub 2014/12/04. doi: 10.1242/dev.107318. PubMed PMID: 25468934.
2. Irie N, Sehara-Fujisawa A. The vertebrate phylotypic stage and an early bilaterian-related stage in mouse embryogenesis defined by genomic information. Bmc Biology. 2007;5. doi: 1 10.1186/1741-7007-5-1. PubMed PMID: WOS:000243921100001.
3. Irie N, Kuratani S. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. Nat Commun. 2011;2. doi: 248 10.1038/ncomms1248. PubMed PMID: WOS:000289982600038.
4. Tena JJ, Gonzalez-Aguilera C, Fernandez-Minan A, Vazquez-Marin J, Parra-Acero H, Cross JW, et al. Comparative epigenomics in distantly related teleost species identifies conserved cis-regulatory nodes active during the vertebrate phylotypic period. Genome Res. 2014;24(7):1075-85. Epub 2014/04/09. doi: 10.1101/gr.163915.113. PubMed PMID: 24709821; PubMed Central PMCID: PMC4079964.
5. Holland PWH, Garcia-Fernandez J, Williams NA, Sidow A. Gene duplications and the origins of vertebrate development. Development (Cambridge). 1994;0(SUPPL.):125-33. PubMed PMID: BCI:BCI199598138944.
6. Shimeld SM, Holland PWH. Vertebrate innovations. Proc Natl Acad Sci U S A. 2000;97(9):4449-52. doi: 10.1073/pnas.97.9.4449. PubMed PMID: WOS:000086703000013.
7. Holland LZ, Holland ND. Chordate origins of the vertebrate central nervous system. Current opinion in neurobiology. 1999;9(5):596-602.
8. Mazet F, Shimeld SM. The evolution of chordate neural segmentation. Developmental Biology. 2002;251(2):258-70. doi: 10.1006/dbio.2002.0831. PubMed PMID: WOS:000179377900005.
9. Holland LZ, Short S. Gene Duplication, Co-Option and Recruitment during the Origin of the Vertebrate Brain from the Invertebrate Chordate Brain. Brain Behav Evol. 2008;72(2):91-105. doi: 10.1159/000151470. PubMed PMID: WOS:000259875900002.
10. Holland LZ. Chordate roots of the vertebrate nervous system: expanding the molecular toolkit. Nature Reviews Neuroscience. 2009;10(10):736-46. doi: 10.1038/nrn2703. PubMed PMID: WOS:000269978800012.

11. Butts T, Holland PWH, Ferrier DEK. Ancient homeobox gene loss and the evolution of chordate brain and pharynx development: deductions from amphioxus gene expression. Proc R Soc B-Biol Sci. 2010;277(1699):3381-9. doi: 10.1098/rspb.2010.0647. PubMed PMID: WOS:000283448800002.

12. Hedges SB, Dudley J, Kumar S. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics. 2006;22(23):2971-2. doi: 10.1093/bioinformatics/btl505. PubMed PMID: WOS:000242246300027.

13. Kumar S, Hedges SB. TimeTree2: species divergence times on the iPhone. Bioinformatics. 2011;27(14):2023-4. doi: 10.1093/bioinformatics/btr315. PubMed PMID: WOS:000292554900027.

14. Holland ND, Chen JY. Origin and early evolution of the vertebrates: new insights from advances in molecular biology, anatomy, and palaeontology. Bioessays. 2001;23(2):142-51. doi: 10.1002/1521-1878(200102)23:2<142::aid-bies1021>3.0.co;2-5. PubMed PMID: WOS:000166619500006.

15. Janvier P. Vertebrate characters and the Cambrian vertebrates. C R Palevol. 2003;2(6-7):523-31. doi: 10.1016/j.crpv.2003.09.002. PubMed PMID: WOS:000188512200017.

16. Hedges SB. Molecular evidence for the early history of living vertebrates. Ahlberg PE, editor. London: Taylor & Francis Ltd; 2001. 119-34 p.

17. Near TJ. Conflict and Resolution Between Phylogenies Inferred From Molecular and Phenotypic Data sets for Hagfish, Lampreys, and Gnathostomes. J Exp Zool Part B. 2009;312B(7):749-61. doi: 10.1002/jez.b.21293. PubMed PMID: WOS:000271108600009.

18. Shimeld SM, Donoghue PCJ. Evolutionary crossroads in developmental biology: cyclostomes (lamprey and hagfish). Development. 2012;139(12):2091-9. doi: 10.1242/dev.074716. PubMed PMID: WOS:000304398400002.

19. Nikitina N, Bronner-Fraser M, Sauka-Spengler T. The sea lamprey Petromyzon marinus: a model for evolutionary and developmental biology. Cold Spring Harbor protocols. 2009;2009(1). PubMed PMID: MEDLINE:20147008.

20. Northcutt RG, Gans C. The Genesis of Neural Crest and Epidermal Placodes - a Reinterpretation of Vertebrate Origins. Quarterly Review of Biology. 1983;58(1):1-28. doi: 10.1086/413055. PubMed PMID: WOS:A1983QK24200001.

21. Gans C, Northcutt RG. Neural Crest and the Origin of Vertebrates: A New Head. Science. 1983;220(4594):268-73. doi: 10.1126/science.220.4594.268.

22. Nikitina NV, Bronner-Fraser M. Gene regulatory networks that control the specification of neural-crest cells in the lamprey. Biochim Biophys Acta-Gene

Regul Mech. 2009;1789(4):274-8. doi: 10.1016/j.bbagrm.2008.03.006. PubMed PMID: WOS:000265729800006.

23. Sauka-Spengler T, Bronner-Fraser M. Insights from a sea lamprey into the evolution of neural crest gene regulatory network. Biological Bulletin. 2008;214(3):303-14. PubMed PMID: WOS:000257187200009.

24. Le Douarin NM, Creuzet S. Neural crest and Vertebrate evolution. Biologie Aujourd hui. 2011;205(2):87-94. doi: 10.1051/jbio/2011009. PubMed PMID: BCI:BCI201100562342.

25. Bronner ME, LeDouarin NM. Development and evolution of the neural crest: An overview. Developmental Biology. 2012;366(1):2-9. doi: 10.1016/j.ydbio.2011.12.042. PubMed PMID: WOS:000304501600002.

26. Lours-Calet C, Alvares LE, El-Hanfy AS, Gandesha S, Walters EH, Sobreira DR, et al. Evolutionarily conserved morphogenetic movements at the vertebrate head-trunk interface coordinate the transport and assembly of hypopharyngeal structures. Dev Biol. 2014;390(2):231-46. Epub 2014/03/26. doi: 10.1016/j.ydbio.2014.03.003. PubMed PMID: 24662046; PubMed Central PMCID: PMC4010675.

27. Schlosser G. Do vertebrate neural crest and cranial placodes have a common evolutionary origin? Bioessays. 2008;30(7):659-72. doi: 10.1002/bies.20775. PubMed PMID: WOS:000257186900008.

28. Graham A, Shimeld SM. The origin and evolution of the ectodermal placodes. J Anat. 2013;222(1):32-40. doi: 10.1111/j.1469-7580.2012.01506.x. PubMed PMID: WOS:000312648500004.

29. Cavodeassi F, Houart C. Brain regionalization: of signaling centers and boundaries. Dev Neurobiol. 2012;72(3):218-33. Epub 2011/06/22. doi: 10.1002/dneu.20938. PubMed PMID: 21692189.

30. Robertshaw E, Kiecker C. Phylogenetic origins of brain organisers. Scientifica (Cairo). 2012;2012:475017. Epub 2012/01/01. doi: 10.6064/2012/475017. PubMed PMID: 24278699; PubMed Central PMCID: PMC3820451.

31. Ohno S, Wolf U, Atkin NB. Evolution from fish to mammals by gene duplication. Hereditas. 1968;59(1):169-87. Epub 1968/01/01. PubMed PMID: 5662632.

32. Holland LZ. Evolution of new characters after whole genome duplications: Insights from amphioxus. Semin Cell Dev Biol. 2013. Epub 2013/01/08. doi: 10.1016/j.semcdb.2012.12.007. PubMed PMID: 23291260.

33. Soshnikova N, Dewaele R, Janvier P, Krumlauf R, Duboule D. Duplications of hox gene clusters and the emergence of vertebrates. Developmental Biology.

2013;378(2):194-9. doi: 10.1016/j.ydbio.2013.03.004. PubMed PMID: MEDLINE:23501471.

34. Escriva H, Manzon L, Youson J, Laudet V. Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. Mol Biol Evol. 2002;19(9):1440-50. Epub 2002/08/30. PubMed PMID: 12200472.

35. Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, Campbell MS, et al. Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. Nature Genet. 2013;45:415-21. Epub 2013/02/26. doi: 10.1038/ng.2568. PubMed PMID: 23435085.

36. Tusscher KH, Hogeweg P. Evolution of Networks for Body Plan Patterning; Interplay of Modularity, Robustness and Evolvability. PLoS Comput Biol. 2011;7(10). doi: e1002208 10.1371/journal.pcbi.1002208. PubMed PMID: WOS:000297262700032.

37. Pougach K, Voet A, Kondrashov FA, Voordeckers K, Christiaens JF, Baying B, et al. Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network. Nat Commun. 2014;5:4868. Epub 2014/09/11. doi: 10.1038/ncomms5868. PubMed PMID: 25204769; PubMed Central PMCID: PMC4172970.

38. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, et al. Ultraconserved Elements in the Human Genome. Science. 2004;304(5675):1321-5.

39. Sandelin A, Bailey P, Bruce S, Engstrom P, Klos J, Wasserman W, et al. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics. 2004;5(1):99. PubMed PMID: doi:10.1186/1471-2164-5-99.

40. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, et al. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 2004;3(1):116-30. doi: e7 10.1371/journal.pbio.0030007. PubMed PMID: WOS:000227169800015.

41. McEwen GK, Goode DK, Parker HJ, Woolfe A, Callaway H, Elgar G. Early Evolution of Conserved Regulatory Sequences Associated with Development in Vertebrates. Plos Genetics. 2009;5(12). doi: e1000762 10.1371/journal.pgen.1000762. PubMed PMID: WOS:000273469700014.

42. Ravi V, Venkatesh B. Rapidly evolving fish genomes and teleost diversity. Current Opinion in Genetics & Development. 2008;18(6):544-50. doi: 10.1016/j.gde.2008.11.001. PubMed PMID: WOS:000263399400011.

43. Blair JE, Hedges SB. Molecular phylogeny and divergence times of deuterostome animals. Mol Biol Evol. 2005;22(11):2275-84. doi: 10.1093/molbev/msi225. PubMed PMID: WOS:000232426500017.

44. Kumar S, Hedges SB. A molecular timescale for vertebrate evolution. Nature. 1998;392(6679):917-20.

45. Morange M. Evolutionary developmental biology its roots and characteristics. Developmental Biology. 2011;357(1):13-6. doi: 10.1016/j.ydbio.2011.03.013. PubMed PMID: WOS:000294144800003.

46. Carroll SB. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. Cell. 2008;134(1):25-36.

47. Lieberman DE, Hall BK. The evolutionary developmental biology of tinkering: an introduction to the challenge. Novartis Foundation symposium. 2007;284:1-19; discussion 110-5. PubMed PMID: MEDLINE:17710844.

48. Davidson EH. Evolutionary bioscience as regulatory systems biology. Developmental Biology. 2011;357(1):35-40.

49. Nam JM, Dong P, Tarpine R, Istrail S, Davidson EH. Functional cis-regulatory genomics for systems biology. Proc Natl Acad Sci U S A. 2010;107(8):3930-5. doi: 10.1073/pnas.1000147107. PubMed PMID: WOS:000275130900116.

50. Howard ML, Davidson EH. cis-regulatory control circuits in development. Developmental Biology. 2004;271(1):109-18. doi: 10.1016/j.ydbio.2004.03.031. PubMed PMID: WOS:000222318100010.

51. Peter IS, Davidson EH. Evolution of Gene Regulatory Networks Controlling Body Plan Development. Cell. 2011;144(6):970-85. doi: 10.1016/j.cell.2011.02.017. PubMed PMID: WOS:000288543500014.

52. Evans NC, Swanson CI, Barolo S. Sparkling insights into enhancer structure, function, and evolution. Current topics in developmental biology. 2012;98:97-120. PubMed PMID: MEDLINE:22305160.

53. Swanson CI, Schwimmer DB, Barolo S. Rapid Evolutionary Rewiring of a Structurally Constrained Eye Enhancer. Curr Biol. 2011;21(14):1186-96. doi: 10.1016/j.cub.2011.05.056. PubMed PMID: WOS:000293320000019.

54. Arnosti DN, Kulkarni MM. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? Journal of cellular biochemistry. 2005;94(5):890-8. Epub 2005/02/08. doi: 10.1002/jcb.20352. PubMed PMID: 15696541.

55. Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet. 2011;13:59-69. Epub 2011/12/07. doi: 10.1038/nrg3095. PubMed PMID: 22143240.

56. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012;484(7392):55-61. doi: 10.1038/nature10944. PubMed PMID: WOS:000302343400033.

57. Rebeiz M, Jikomes N, Kassner VA, Carroll SB. Evolutionary origin of a novel gene expression pattern through co-option of the latent activities of existing regulatory sequences. Proc Natl Acad Sci U S A. 2011;108(25):10036-43. doi: 10.1073/pnas.1105937108. PubMed PMID: WOS:000291857500010.

58. Wray GA. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 2007;8(3):206-16. doi: 10.1038/nrg2063. PubMed PMID: WOS:000244274400014.

59. Razeto-Barry P, Maldonado K. Adaptive *cis*-regulatory changes may involve few mutations. Evolution. 2011;65(11):3332-5. doi: 10.1111/j.1558-5646.2011.01412.x. PubMed PMID: WOS:000296702800026.

60. Clune J, Mouret JB, Lipson H. The evolutionary origins of modularity. Proc R Soc B-Biol Sci. 2013;280(1755). doi: 20122863 10.1098/rspb.2012.2863. PubMed PMID: WOS:000314357600008.

61. Frankel N, Wang S, Stern DL. Conserved regulatory architecture underlies parallel genetic changes and convergent phenotypic evolution. Proc Natl Acad Sci U S A. 2012;109(51):20975-9. Epub 2012/12/01. doi: 10.1073/pnas.1207715109. PubMed PMID: 23197832; PubMed Central PMCID: PMC3529038.

62. Frankel N, Erezyilmaz DF, McGregor AP, Wang S, Payre F, Stern DL. Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. Nature. 2011;474(7353):598-603. doi: 10.1038/nature10200. PubMed PMID: WOS:000292204300030.

63. Rogers WA, Salomone JR, Tacy DJ, Camino EM, Davis KA, Rebeiz M, et al. Recurrent modification of a conserved cis-regulatory element underlies fruit fly pigmentation diversity. PLoS Genet. 2013;9(8):e1003740. Epub 2013/09/07. doi: 10.1371/journal.pgen.1003740. PubMed PMID: 24009528; PubMed Central PMCID: PMC3757066.

64. Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, et al. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science (New York, NY). 2010;327(5963):302-5. doi: 10.1126/science.1182213. PubMed PMID: PMC3109066.

65. Cleves PA, Ellis NA, Jimenez MT, Nunez SM, Schluter D, Kingsley DM, et al. Evolved tooth gain in sticklebacks is associated with a cis-regulatory allele of

Bmp6. Proc Natl Acad Sci U S A. 2014;111(38):13912-7. Epub 2014/09/11. doi: 10.1073/pnas.1407567111. PubMed PMID: 25205810; PubMed Central PMCID: PMC4183278.

66. Domyan ET, Guernsey MW, Kronenberg Z, Krishnan S, Boissy RE, Vickrey AI, et al. Epistatic and combinatorial effects of pigmentary gene mutations in the domestic pigeon. Curr Biol. 2014;24(4):459-64. Epub 2014/02/11. doi: 10.1016/j.cub.2014.01.020. PubMed PMID: 24508169; PubMed Central PMCID: PMC3990261.

67. Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, et al. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. Nat Genet. 2014;46(7):685-92. Epub 2014/06/09. doi: 10.1038/ng.3009. PubMed PMID: 24908250.

68. Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jónsson B, et al. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature. 2004;428(6984):717-23.

69. Colosimo PF, Peichel CL, Nereng K, Blackman BK, Shapiro MD, Schluter D, et al. The genetic architecture of parallel armor plate reduction in threespine sticklebacks. PLoS Biol. 2004;2(5):e109.

70. Fraser HB, Babak T, Tsang J, Zhou YQ, Zhang B, Mehrabian M, et al. Systematic Detection of Polygenic cis-Regulatory Evolution. Plos Genetics. 2011;7(3). doi: e1002023 10.1371/journal.pgen.1002023. PubMed PMID: WOS:000288996600053.

71. Louis A, Roest Crollius H, Robinson-Rechavi M. How much does the amphioxus genome represent the ancestor of chordates? Brief Funct Genomics. 2012;11(2):89-95. doi: 10.1093/bfgp/els003. PubMed PMID: 22373648.

72. O'Brien KP, Remm M, Sonnhammer EL. Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. 2005;33(Database issue):D476-80. Epub 2004/12/21. doi: 10.1093/nar/gki107. PubMed PMID: 15608241; PubMed Central PMCID: PMC540061.

73. Epstein DJ. Cis-regulatory mutations in human disease. Briefings in Functional Genomics & Proteomics. 2009;8(4):310-6. doi: 10.1093/bfgp/elp021. PubMed PMID: BCI:BCI200900595337.

74. Vandermeer JE, Ahituv N. cis-regulatory mutations are a genetic cause of human limb malformations. Developmental Dynamics. 2011;240(5):920-30.

75. Bhatia S, Kleinjan DA. Disruption of long-range gene regulation in human genetic disease: a kaleidoscope of general principles, diverse mechanisms and unique

phenotypic consequences. Hum Genet. 2014;133(7):815-45. Epub 2014/02/06. doi: 10.1007/s00439-014-1424-6. PubMed PMID: 24496500.

76. Benko S. Cis-ruptions of highly conserved non-coding genomic elements distant from the SOX9 gene in the Pierre Robin sequence. Biol Aujourdhui. 2011;205(2):111.

77. Smith E, Shilatifard A. Enhancer biology and enhanceropathies. Nat Struct Mol Biol. 2014;21(3):210-9. Epub 2014/03/07. doi: 10.1038/nsmb.2784. PubMed PMID: 24599251.

78. Gordon CT, Attanasio C, Bhatia S, Benko S, Ansari M, Tan TY, et al. Identification of novel craniofacial regulatory domains located far upstream of SOX9 and disrupted in Pierre Robin sequence. Hum Mutat. 2014;35(8):1011-20. Epub 2014/06/18. doi: 10.1002/humu.22606. PubMed PMID: 24934569.

79. Benko S, Fantes JA, Amiel J, Kleinjan DJ, Thomas S, Ramsay J, et al. Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. Nature Genet. 2009;41(3):359-64. doi: 10.1038/ng.329. PubMed PMID: WOS:000263640200019.

80. Lettice LA, Daniels S, Sweeney E, Venkataraman S, Devenney PS, Gautier P, et al. Enhancer-adoption as a mechanism of human developmental disease. Hum Mutat. 2011;32(12):1492-9. doi: 10.1002/humu.21615. PubMed PMID: WOS:000297246800021.

81. Anderson E, Peluso S, Lettice LA, Hill RE. Human limb abnormalities caused by disruption of hedgehog signaling. Trends in Genetics. 2012;28(8):364-73. doi: 10.1016/j.tig.2012.03.012. PubMed PMID: WOS:000307157700002.

82. Laurell T, Vandermeer JE, Wenger AM, Grigelioniene G, Nordenskjold A, Arner M, et al. A novel 13 base pair insertion in the sonic hedgehog ZRS limb enhancer (ZRS/LMBR1) causes preaxial polydactyly with triphalangeal thumb. Hum Mutat. 2012;33(7):1063-6. Epub 2012/04/13. doi: 10.1002/humu.22097. PubMed PMID: 22495965; PubMed Central PMCID: PMC3370115.

83. Smemo S, Campos LC, Moskowitz IP, Krieger JE, Pereira AC, Nobrega MA. Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. Hum Mol Genet. 2012;21(14):3255-63. Epub 2012/05/01. doi: 10.1093/hmg/dds165. PubMed PMID: 22543974; PubMed Central PMCID: PMCPmc3384386.

84. Fantes J, Redeker B, Breen M, Boyle S, Brown J, Fletcher J, et al. Aniridia-associated cytogenetic rearrangements suggest that a position effect may cause the mutant phenotype. Hum Mol Genet. 1995;4(3):415-22. Epub 1995/03/01. PubMed PMID: 7795596.

85. Kleinjan DA, Seawright A, Mella S, Carr CB, Tyas DA, Simpson TI, et al. Long-range downstream enhancers are essential for Pax6 expression. Dev Biol. 2006;299(2):563-81. Epub 2006/10/04. doi: 10.1016/j.ydbio.2006.08.060. PubMed PMID: 17014839; PubMed Central PMCID: PMC2386664.

86. Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, et al. Detecting Conserved Regulatory Elements with the Model Genome of the Japanese Puffer Fish, Fugu rubripes. Proc Natl Acad Sci U S A. 1995;92(5):1684-8. doi: 10.1073/pnas.92.5.1684. PubMed PMID: WOS:A1995QK07700090.

87. Poulin F, Nobrega MA, Plajzer-Frick I, Holt A, Afzal V, Rubin EM, et al. In vivo characterization of a vertebrate ultraconserved enhancer. Genomics. 2005;85(6):774-81. Epub 2005/05/12. doi: 10.1016/j.ygeno.2005.03.003. PubMed PMID: 15885503.

88. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, et al. In vivo enhancer analysis of human conserved non-coding sequences. Nature. 2006;444(7118):499-502.

89. Elgar G, Vavouri T. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends in Genetics. 2008;24(7):344-52. doi: 10.1016/j.tig.2008.04.005. PubMed PMID: WOS:000258008400007.

90. Elgar G. Pan-vertebrate conserved non-coding sequences associated with developmental regulation. Briefings in Functional Genomics & Proteomics. 2009;8(4):256-65. doi: 10.1093/bfgp/elp033. PubMed PMID: BCI:BCI200900595332.

91. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. Journal of Molecular Biology. 1990;215(3):403-10. doi: 10.1006/jmbi.1990.9999. PubMed PMID: WOS:A1990ED16700008.

92. Kent WJ. BLAT - The BLAST-like alignment tool. Genome Research. 2002;12(4):656-64. doi: 10.1101/gr.229202. PubMed PMID: WOS:000174971100015.

93. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000;7(1-2):203-14. Epub 2000/07/13. doi: 10.1089/10665270050081478. PubMed PMID: 10890397.

94. Buhler J. Efficient large-scale sequence comparison by locality-sensitive hashing. Bioinformatics. 2001;17(5):419-28. Epub 2001/05/02. PubMed PMID: 11331236.

95. Ma B, Tromp J, Li M. PatternHunter: faster and more sensitive homology search. Bioinformatics. 2002;18(3):440-5. Epub 2002/04/06. PubMed PMID: 11934743.

96. Lee AP, Kerk SY, Tan YY, Brenner S, Venkatesh B. Ancient Vertebrate Conserved Noncoding Elements Have Been Evolving Rapidly in Teleost Fishes. Mol Biol Evol. 2011;28(3):1205-15. doi: 10.1093/molbev/msq304. PubMed PMID: WOS:000287745200008.

97. Doglio L, Goode DK, Pelleri MC, Pauls S, Frabetti F, Shimeld SM, et al. Parallel evolution of chordate cis-regulatory code for development. PLoS Genet. 2013;9(11):e1003904. Epub 2013/11/28. doi: 10.1371/journal.pgen.1003904. PubMed PMID: 24282393; PubMed Central PMCID: PMC3836708.

98. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Research. 2003;13(4):721-31.

99. Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, et al. Glocal alignment: finding rearrangements during alignment. Bioinformatics. 2003;19(suppl 1):i54-i62.

100. Sosinsky A, Honig B, Mann RS, Califano A. Discovering transcriptional regulatory regions in Drosophila by a nonalignment method for phylogenetic footprinting. Proc Natl Acad Sci U S A. 2007;104(15):6305-10. doi: 10.1073/pnas.0701614104. PubMed PMID: WOS:000245737500039.

101. Khan MA, Soto-Jimenez LM, Howe T, Streit A, Sosinsky A, Stern CD. Computational tools and resources for prediction and analysis of gene regulatory regions in the chick genome. Genesis. 2013;51(5):311-24. Epub 2013/01/29. doi: 10.1002/dvg.22375. PubMed PMID: 23355428; PubMed Central PMCID: PMC3664090.

102. Nikulova AA, Favorov AV, Sutormin RA, Makeev VJ, Mironov AA. CORECLUST: identification of the conserved CRM grammar together with prediction of gene regulation. Nucleic Acids Research. 2012;40(12). doi: e93 10.1093/nar/gks235. PubMed PMID: WOS:000305829000006.

103. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 2013. Epub 2013/02/23. doi: 10.1093/gbe/evt028. PubMed PMID: 23431001.

104. Doolittle WF. Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci U S A. 2013;110(14):5294-300. Epub 2013/03/13. doi: 10.1073/pnas.1221376110. PubMed PMID: 23479647; PubMed Central PMCID: PMC3619371.

105. Gould SJ, Lewontin RC. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proceedings of the Royal

Society of London Series B, Biological sciences. 1979;205(1161):581-98. Epub 1979/09/21. PubMed PMID: 42062.

106. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489(7414):75-82. doi: 10.1038/nature11232. PubMed PMID: WOS:000308347000040.

107. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012;489(7414):83-90. doi: 10.1038/nature11212. PubMed PMID: WOS:000308347000041.

108. Cui K, Zhao K. Genome-wide approaches to determining nucleosome occupancy in metazoans using MNase-Seq. Methods Mol Biol. 2012;833:413-9. Epub 2011/12/21. doi: 10.1007/978-1-61779-477-3_24. PubMed PMID: 22183607; PubMed Central PMCID: PMC3541821.

109. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Current protocols in molecular biology / edited by Frederick M Ausubel [et al]. 2015;109:21 9 1-9. Epub 2015/01/07. doi: 10.1002/0471142727.mb2129s109. PubMed PMID: 25559105.

110. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10(12):1213-8. Epub 2013/10/08. doi: 10.1038/nmeth.2688. PubMed PMID: 24097267; PubMed Central PMCID: PMC3959825.

111. Niu W, Lu ZJ, Zhong M, Sarov M, Murray JI, Brdlik CM, et al. Diverse transcription factor binding features revealed by genome-wide ChIP-seq in C. elegans. Genome Research. 2011;21(2):245-54. doi: 10.1101/gr.114587.110. PubMed PMID: WOS:000286804100010.

112. Jin H, Stojnic R, Adryan B, Ozdemir A, Stathopoulos A, Frasch M. Genome-wide screens for in vivo Tinman binding sites identify cardiac enhancers with diverse functional architectures. Plos Genetics. 2013;9(1):e1003195. doi: 10.1371/journal.pgen.1003195. PubMed PMID: MEDLINE:23326246.

113. Visel A, Taher L, Girgis H, May D, Golonzhka O, Hoch RV, et al. A high-resolution enhancer atlas of the developing telencephalon. Cell. 2013;152(4):895-908. doi: 10.1016/j.cell.2012.12.041. PubMed PMID: MEDLINE:23375746.

114. Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EE. Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature. 2009;462(7269):65-70. Epub 2009/11/06. doi: 10.1038/nature08531. PubMed PMID: 19890324.

115. Aday AW, Zhu LJ, Lakshmanan A, Wang J, Lawson ND. Identification of cis regulatory features in the embryonic zebrafish genome through large-scale profiling of H3K4me1 and H3K4me3 binding sites. Developmental Biology. 2011;357(2):450-62.

116. Bogdanovic O, Fernandez-Minan A, Tena JJ, de la Calle-Mustienes E, Hidalgo C, van Kruysbergen I, et al. Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. Genome Res. 2012;22(10):2043-53. Epub 2012/05/18. doi: 10.1101/gr.134833.111. PubMed PMID: 22593555; PubMed Central PMCID: PMC3460198.

117. Chen CY, Morris Q, Mitchell JA. Enhancer identification in mouse embryonic stem cells using integrative modeling of chromatin and genomic features. BMC Genomics. 2012;13. doi: 152 10.1186/1471-2164-13-152. PubMed PMID: WOS:000306855200001.

118. Wilczynski B, Liu YH, Yeo ZX, Furlong EE. Predicting spatial and temporal gene expression using an integrative model of transcription factor occupancy and chromatin state. PLoS Comput Biol. 2012;8(12):e1002798. Epub 2012/12/14. doi: 10.1371/journal.pcbi.1002798. PubMed PMID: 23236268; PubMed Central PMCID: PMC3516547.

119. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A. 2010;107(50):21931-6. Epub 2010/11/26. doi: 10.1073/pnas.1016071107. PubMed PMID: 21106759; PubMed Central PMCID: PMC3003124.

120. Burzynski GM, Reed X, Taher L, Stine ZE, Matsui T, Ovcharenko I, et al. Systematic elucidation and in vivo validation of sequences enriched in hindbrain transcriptional control. Genome Research. 2012;22(11):2278-89. doi: 10.1101/gr.139717.112. PubMed PMID: MEDLINE:22759862.

121. Birnbaum RY, Clowney EJ, Agamy O, Kim MJ, Zhao JJ, Yamanaka T, et al. Coding exons function as tissue-specific enhancers of nearby genes. Genome Research. 2012;22(6):1059-68. doi: 10.1101/gr.133546.111. PubMed PMID: WOS:000304728100007.

122. Cadiz-Rivera B, Fromm G, de Vries C, Fields J, McGrath KE, Fiering S, et al. The chromatin "landscape" of a murine adult beta-globin gene is unaffected by deletion of either the gene promoter or a downstream enhancer. PLoS ONE. 2014;9(5):e92947. Epub 2014/05/13. doi: 10.1371/journal.pone.0092947. PubMed PMID: 24817273; PubMed Central PMCID: PMC4015891.

123. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013;339(6123):1074-7. Epub 2013/01/19. doi: 10.1126/science.1232542. PubMed PMID: 23328393.

124. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis C, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74. doi: 10.1038/nature11247. PubMed PMID: WOS:000308347000039.

125. Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, et al. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol. 2011;9(4). doi: e1001046 10.1371/journal.pbio.1001046. PubMed PMID: WOS:000289938900014.

126. Aerts S. Computational Strategies for the Genome-Wide Identification of cis-Regulatory Elements and Transcriptional Targets. In: Serge P, Francois P, editors. Current topics in developmental biology. Volume 98: Academic Press; 2012. p. 121-45.

127. Vavouri T, McEwen GK, Woolfe A, Gilks WR, Elgar G. Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key. Trends in Genetics. 2006;22(1):5-10. doi: 10.1016/j.tig.2005.10.005.

128. Pregizer S, Mortlock DP. Control of BMP gene expression by long-range regulatory elements. Cytokine Growth Factor Rev. 2009;20(5-6):509-15. doi: 10.1016/j.cytogfr.2009.10.011. PubMed PMID: WOS:000273235700020.

129. Hong JW, Hendrix DA, Levine MS. Shadow enhancers as a source of evolutionary novelty. Science. 2008;321(5894):1314-. doi: 10.1126/science.1160631. PubMed PMID: WOS:000258914300039.

130. Perry MW, Boettiger AN, Bothma JP, Levine M. Shadow Enhancers Foster Robustness of Drosophila Gastrulation. Curr Biol. 2010;20(17):1562-7. doi: 10.1016/j.cub.2010.07.043. PubMed PMID: WOS:000281941100034.

131. Lampe X, Samad OA, Guiguen A, Matis C, Remacle S, Picard JJ, et al. An ultraconserved Hox-Pbx responsive element resides in the coding sequence of Hoxa2 and is active in rhombomere 4. Nucleic Acids Research. 2008;36(10):3214-25. doi: 10.1093/nar/gkn148. PubMed PMID: WOS:000257183200006.

132. Dong X, Navratilova P, Fredman D, Drivenes ò, Becker TS, Lenhard B. Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons. Nucleic Acids Research. 2010;38(4):1071-85.

133. De Silva DR, Nichols R, Elgar G. Purifying selection in deeply conserved human enhancers is more consistent than in coding sequences. PLoS ONE. 2014;9(7):e103357. Epub 2014/07/26. doi: 10.1371/journal.pone.0103357. PubMed PMID: 25062004; PubMed Central PMCID: PMC4111549.

134. Vavouri T, Walter K, Gilks WR, Lehner B, Elgar G. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. Genome Biology. 2007;8(2). doi: R15 10.1186/gb-2007-8-2-r15. PubMed PMID: WOS:000246076300005.

135. Vavouri T, Lehne B. Conserved noncoding elements and the evolution of animal body plans. Bioessays. 2009;31(7):727-35. doi: 10.1002/bies.200900014. PubMed PMID: WOS:000267610500006.

136. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034-50. Epub 2005/07/19. doi: 10.1101/gr.3715005. PubMed PMID: 16024819; PubMed Central PMCID: PMC1182216.

137. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, et al. The amphioxus genome and the evolution of the chordate karyotype. Nature. 2008;453(7198):1064-71.

138. Sanges R, Hadzhiev Y, Gueroult-Bellone M, Roure A, Ferg M, Meola N, et al. Highly conserved elements discovered in vertebrates are present in non-syntenic loci of tunicates, act as enhancers and can be transcribed during development. Nucleic Acids Res. 2013;41(6):3600-18. Epub 2013/02/09. doi: 10.1093/nar/gkt030. PubMed PMID: 23393190.

139. Nonchev S, Maconochie M, Vesque C, Aparicio S, Ariza-McNaughton L, Manzanares M, et al. The conserved role of Krox-20 in directing Hox gene expression during vertebrate hindbrain segmentation. Proc Natl Acad Sci U S A. 1996;93(18):9339-45. Epub 1996/09/03. PubMed PMID: 8790331; PubMed Central PMCID: PMC38429.

140. Rowitch DH, Echelard Y, Danielian PS, Gellner K, Brenner S, McMahon AP. Identification of an evolutionarily conserved 110 base-pair cis-acting regulatory sequence that governs Wnt-1 expression in the murine neural plate. Development. 1998;125(14):2735-46. Epub 1998/06/24. PubMed PMID: 9636087.

141. Hemberg M, Gray JM, Cloonan N, Kuersten S, Grimmond S, Greenberg ME, et al. Integrated genome analysis suggests that most conserved non-coding sequences are regulatory factor binding sites. Nucleic Acids Res.

2012;40(16):7858-69. Epub 2012/06/12. doi: 10.1093/nar/gks477. PubMed PMID: 22684627; PubMed Central PMCID: PMC3439890.

142. Woolfe A, Goode DK, Cooke J, Callaway H, Smith S, Snell P, et al. CONDOR: a database resource of developmentally associated conserved non-coding elements. Bmc Developmental Biology. 2007;7(100). doi: 100 10.1186/1471-213x-7-100. PubMed PMID: WOS:000250195300001.

143. Venkatesh B, Kirkness EF, Loh YH, Halpern AL, Lee AP, Johnson J, et al. Ancient noncoding elements conserved in the human genome. Science. 2006;314(5807):1892-. doi: 10.1126/science.1130708. PubMed PMID: WOS:000242996800037.

144. Hiller M, Agarwal S, Notwell JH, Parikh R, Guturu H, Wenger AM, et al. Computational methods to detect conserved non-genic elements in phylogenetically isolated genomes: application to zebrafish. Nucleic Acids Res. 2013;41(15):e151. Epub 2013/07/03. doi: 10.1093/nar/gkt557. PubMed PMID: 23814184; PubMed Central PMCID: PMC3753653.

145. Berna L, Alvarez-Valin F, D'Onofrio G. How Fast Is the Sessile Ciona? Compar Funct Genom. 2009. doi: 875901 10.1155/2009/875901. PubMed PMID: WOS:000273277900001.

146. Pascual-Anaya J, Adachi N, Alvarez S, Kuratani S, D'Aniello S, Garcia-Fernandez J. Broken colinearity of the amphioxus Hox cluster. EvoDevo. 2012;3. doi: 28 10.1186/2041-9139-3-28. PubMed PMID: WOS:000312981400001.

147. Ikuta T, Yoshida N, Satoh N, Saiga H. Ciona intestinalis Hox gene cluster: Its dispersed structure and residual colinear expression in development. Proc Natl Acad Sci U S A. 2004;101(42):15118-23. doi: 10.1073/pnas.0401389101. PubMed PMID: WOS:000224688700027.

148. Royo JL, Maeso I, Irimia M, Gao F, Peter IS, Lopes CS, et al. Transphyletic conservation of developmental regulatory state in animal evolution. Proc Natl Acad Sci U S A. 2011;108(34):14186-91. doi: 10.1073/pnas.1109037108. PubMed PMID: WOS:000294163500065.

149. Hufton AL, Mathia S, Braun H, Georgi U, Lehrach H, Vingron M, et al. Deeply conserved chordate noncoding sequences preserve genome synteny but do not drive gene duplicate retention. Genome Research. 2009;19(11):2036-51. doi: 10.1101/gr.093237.109.

150. Fisher S, Grice EA, Vinton RM, Bessling SL, Urasaki A, Kawakami K, et al. Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. Nature Protocols. 2006;1(3):1297-305. doi: 10.1038/nprot.2006.230. PubMed PMID: WOS:000251155400028.

151. Kawakami K. Tol2: a versatile gene transfer vector in vertebrates. Genome Biology. 2007;8. doi: S7 10.1186/gb-2007-8-S1-S7. PubMed PMID: WOS:000207571900007.

152. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser—a database of tissue-specific human enhancers. Nucleic Acids Research. 2007;35(suppl 1):D88-D92.

153. Davies SR, Chang L-W, Patra D, Xing X, Posey K, Hecht J, et al. Computational identification and functional validation of regulatory motifs in cartilage-expressed genes. Genome Research. 2007;17(10):1438-47. doi: 10.1101/gr.6224007. PubMed PMID: WOS:000249869200005.

154. Kwon AT, Chou AY, Arenillas DJ, Wasserman WW. Validation of Skeletal Muscle cis-Regulatory Module Predictions Reveals Nucleotide Composition Bias in Functional Enhancers. PLoS Comput Biol. 2011;7(12). doi: e1002256 10.1371/journal.pcbi.1002256. PubMed PMID: WOS:000299167800002.

155. Haeussler M, Joly JS. When needles look like hay: How to find tissue-specific enhancers in model organism genomes. Developmental Biology. 2011;350(2):239-54. doi: 10.1016/j.ydbio.2010.11.026. PubMed PMID: WOS:000287117900001.

156. MacDonald RB, Debiais-Thibaud M, Martin K, Poitras L, Tay BH, Venkatesh B, et al. Functional conservation of a forebrain enhancer from the elephant shark (Callorhinchus milii) in zebrafish and mice. BMC Evolutionary Biology. 2010;10. doi: 157

10.1186/1471-2148-10-157. PubMed PMID: WOS:000279829100002.

157. Parker H, Piccinelli P, Sauka-Spengler T, Bronner M, Elgar G. Ancient Pbx-Hox Signatures Define Hundreds of Vertebrate Developmental Enhancers. BMC Genomics. 2011;12(1):637. PubMed PMID: doi:10.1186/1471-2164-12-637.

158. Domene S, Bumaschny VF, de Souza FS, Franchini LF, Nasif S, Low MJ, et al. Enhancer turnover and conserved regulatory function in vertebrate evolution. Philos Trans R Soc Lond B Biol Sci. 2013;368(1632):20130027. Epub 2013/11/13. doi: 10.1098/rstb.2013.0027. PubMed PMID: 24218639; PubMed Central PMCID: PMC3826500.

159. Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science. 2006;312(5771):276-9. doi: 10.1126/science.1124070. PubMed PMID: WOS:000236765300052.

160. Ariza-Cosano A, Visel A, Pennacchio LA, Fraser HB, Luis Gomez-Skarmeta J, Irimia M, et al. Differences in enhancer activity in mouse and zebrafish reporter

assays are often associated with changes in gene expression. BMC Genomics. 2012;13. doi: 713 10.1186/1471-2164-13-713. PubMed PMID: WOS:000313248200001.

161. Ritter DI, Li QA, Kostka D, Pollard KS, Guo S, Chuang JH. The Importance of Being Cis: Evolution of Orthologous Fish and Mammalian Enhancer Activity. Mol Biol Evol. 2010;27(10):2322-32. doi: 10.1093/molbev/msq128. PubMed PMID: WOS:000282174600012.

162. Parker HJ, Sauka-Spengler T, Bronner M, Elgar G. A reporter assay in lamprey embryos reveals both functional conservation and elaboration of vertebrate enhancers. PLoS ONE. 2014;9(1):e85492. Epub 2014/01/15. doi: 10.1371/journal.pone.0085492. PubMed PMID: 24416417; PubMed Central PMCID: PMC3887057.

163. Goode DK, Callaway HA, Cerda GA, Lewis KE, Elgar G. Minor change, major difference: divergent functions of highly conserved cis-regulatory elements subsequent to whole genome duplication events. Development. 2011;138(5):879-84. doi: 10.1242/dev.055996. PubMed PMID: WOS:000287576100009.

164. McEwen GK, Woolfe A, Goode D, Vavouri T, Callaway H, Elgar G. Ancient duplicated conserved noncoding elements in vertebrates: A genomic and functional analysis. Genome Research. 2006;16(4):451-65. doi: 10.1101/gr.4143406. PubMed PMID: WOS:000236700000001.

165. Woolfe A, Elgar G. Comparative genomics using Fugu reveals insights into regulatory subfunctionalization. Genome Biology. 2007;8(4). doi: R53 10.1186/gb-2007-8-4-r53. PubMed PMID: WOS:000246982900012.

166. Pauls S, Smith SF, Elgar G. Lens development depends on a pair of highly conserved Sox21 regulatory elements. Developmental Biology. 2012;365(1):310-08.

167. Quina LA, Kuramoto T, Luquetti DV, Cox TC, Serikawa T, Turner EE. Deletion of a conserved regulatory element required for Hmx1 expression in craniofacial mesenchyme in the dumbo rat: a newly identified cause of congenital ear malformation. Dis Model Mech. 2012;5(6):812-22. doi: 10.1242/dmm.009910. PubMed PMID: WOS:000311859700014.

168. Liska F, Snajdr P, Sedova L, Seda O, Chylikova B, Slamova P, et al. Deletion of a Conserved Noncoding Sequence in Plzf Intron Leads to Plzf Down-regulation in Limb Bud and Polydactyly in the Rat. Developmental Dynamics. 2009;238(3):673-84. doi: 10.1002/dvdy.21859. PubMed PMID: WOS:000264001300016.

169. Matsunami M, Saitou N. Vertebrate paralogous conserved noncoding sequences may be related to gene expressions in brain. Genome biology and evolution. 2013;5(1):140-50. doi: 10.1093/gbe/evs128. PubMed PMID: MEDLINE:23267051.

170. Chen Y, Ding Y, Zhang ZM, Wang W, Chen JY, Ueno NT, et al. Evolution of vertebrate central nervous system is accompanied by novel expression changes of duplicate genes. J Genet Genomics. 2011;38(12):577-84. doi: 10.1016/j.jgg.2011.10.004. PubMed PMID: WOS:000298477100002.

171. Papatsenko D, Goltsev Y, Levine M. Organization of developmental enhancers in the Drosophila embryo. Nucleic Acids Research. 2009;37(17):5665-77. doi: 10.1093/nar/gkp619. PubMed PMID: WOS:000271569100008.

172. Lusk RW, Eisen MB. Evolutionary Mirages: Selection on Binding Site Composition Creates the Illusion of Conserved Grammars in Drosophila Enhancers. Plos Genetics. 2010;6(1). doi: e1000829 10.1371/journal.pgen.1000829. PubMed PMID: WOS:000274194300035.

173. Vavouri T, Elgar G. Prediction of cis-regulatory elements using binding site matrices - the successes, the failures and the reasons for both. Current Opinion in Genetics & Development. 2005;15(4):395-402. doi: 10.1016/j.gde.2005.05.002. PubMed PMID: WOS:000231205400007.

174. Rastegar S, Hess I, Dickmeis T, Nicod JC, Ertzer R, Hadzhiev Y, et al. The words of the regulatory code are arranged in a variable manner in highly conserved enhancers. Developmental Biology. 2008;318(2):366-77. doi: 10.1016/j.ydbio.2008.03.034. PubMed PMID: WOS:000256651500015.

175. Kantorovitz MR, Kazemian M, Kinston S, Miranda-Saavedra D, Zhu QY, Robinson GE, et al. Motif-Blind, Genome-Wide Discovery of cis-Regulatory Modules in Drosophila and Mouse. Developmental Cell. 2009;17(4):568-79. doi: 10.1016/j.devcel.2009.09.002. PubMed PMID: WOS:000271181400016.

176. Narlikar L, Sakabe NJ, Blanski AA, Arimura FE, Westlund JM, Nobrega MA, et al. Genome-wide discovery of human heart enhancers. Genome Research. 2010;20(3):381-92. doi: 10.1101/gr.098657.109. PubMed PMID: WOS:000275124600010.

177. Haeussler M, Jaszczyszyn Y, Christiaen L, Joly JS. A cis-Regulatory Signature for Chordate Anterior Neuroectodermal Genes. Plos Genetics. 2010;6(4). doi: e1000912 10.1371/journal.pgen.1000912. PubMed PMID: WOS:000277354200029.

178. Matys V, Fricke E, Geffers R, Goessling E, Haubrock M, Hehl R, et al. TRANSFAC®: transcriptional regulation, from patterns to profiles. Nucleic Acids Research. 2003;31(1):374-8.

179. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 2013;42(Database issue):D142–D7. Epub 2013/11/07. doi: 10.1093/nar/gkt997. PubMed PMID: 24194598.

180. Mongin E, Auer TO, Bourrat F, Gruhl F, Dewar K, Blanchette M, et al. Combining Computational Prediction of Cis-Regulatory Elements with a New Enhancer Assay to Efficiently Label Neuronal Structures in the Medaka Fish. PLoS ONE. 2011;6(5). doi: e19747 10.1371/journal.pone.0019747. PubMed PMID: WOS:000291052500011.

181. Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res. 2015;43(Database issue):D117-22. Epub 2014/11/08. doi: 10.1093/nar/gku1045. PubMed PMID: 25378322.

182. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Research. 2012;22(9):1798-812. doi: 10.1101/gr.139105.112. PubMed PMID: WOS:000308272800020.

183. Yanez-Cuna JO, Kvon EZ, Stark A. Deciphering the transcriptional cis-regulatory code. Trends in Genetics. 2013;29(1):11-22. doi: 10.1016/j.tig.2012.09.007. PubMed PMID: WOS:000314014700004.

184. Parker HJ, Bronner ME, Krumlauf R. A Hox regulatory network of hindbrain segmentation is conserved to the base of vertebrates. Nature. 2014;514:490–3. Epub 2014/09/16. doi: 10.1038/nature13723. PubMed PMID: 25219855.

185. Tumpel S, Wiedemann LM, Krumlauf R. Hox Genes and Segmentation of the Vertebrate Hindbrain. In: Olivier P, editor. Current topics in developmental biology. Volume 88: Academic Press; 2009. p. 103-37.

186. Philippidou P, Dasen JS. Hox genes: choreographers in neural development, architects of circuit organization. Neuron. 2013;80(1):12-34. Epub 2013/10/08. doi: 10.1016/j.neuron.2013.09.020. PubMed PMID: 24094100; PubMed Central PMCID: PMC3835187.

187. Choe SK, Vlachakis N, Sagerstrom CG. Meis family proteins are required for hindbrain development in the zebrafish. Development. 2002;129(3):585-95. PubMed PMID: WOS:000174360000004.

188. Vlachakis N, Choe SK, Sagerstrom CG. Meis3 synergizes with Pbx4 and Hoxb1b in promoting hindbrain fates in the zebrafish. Development. 2001;128(8):1299-312. PubMed PMID: WOS:000168498900008.

189. Waskiewicz AJ, Rikhof HA, Moens CB. Eliminating zebrafish pbx proteins reveals a hindbrain ground state. Dev Cell. 2002;3(5):723-33. Epub 2002/11/15. PubMed PMID: 12431378.

190. Prin F, Serpente P, Itasaki N, Gould AP. Hox proteins drive cell segregation and non-autonomous apical remodelling during hindbrain segmentation. Development. 2014;141(7):1492-502. Epub 2014/02/28. doi: 10.1242/dev.098954. PubMed PMID: 24574009.

191. Knoepfler PS, Lu Q, Kamps MP. Pbx1-Hox heterodimers bind DNA on inseparable half-sites that permit intrinsic DNA binding specificity of the Hox partner at nucleotides 3' to a TAAT motif. Nucleic Acids Research. 1996;24(12):2288-94. doi: 10.1093/nar/24.12.2288. PubMed PMID: WOS:A1996UW12300013.

192. Jacobs Y, Schnabel CA, Cleary ML. Trimeric association of hox and TALE homeodomain proteins mediates Hoxb2 hindbrain enhancer activity. Mol Cell Biol. 1999;19(7):5134-42. PubMed PMID: WOS:000080952300060.

193. DiRocco G, Mavilio F, Zappavigna V. Functional dissection of a transcriptionally active, target-specific Hox-Pbx complex. Embo Journal. 1997;16(12):3644-54. doi: 10.1093/emboj/16.12.3644. PubMed PMID: WOS:A1997XG52000026.

194. Asahara H, Dutta S, Kao HY, Evans RM, Montminy M. Pbx-hox heterodimers recruit coactivator-corepressor complexes in an isoform-specific manner. Mol Cell Biol. 1999;19(12):8219-25. PubMed PMID: WOS:000083781300032.

195. Ferretti E, Marshall H, Popperl H, Maconochie M, Krumlauf R, Blasi F. Segmental expression of Hoxb2 in r4 requires two separate sites that integrate cooperative interactions between Prep1, Pbx and Hox proteins. Development. 2000;127(1):155-66. PubMed PMID: WOS:000085159700015.

196. Wassef MA, Chomette D, Pouilhe M, Stedman A, Havis E, Dinh CDT, et al. Rostral hindbrain patterning involves the direct activation of a Krox20 transcriptional enhancer by Hox/Pbx and Meis factors. Development. 2008;135(20):3369-78. doi: 10.1242/dev.023614. PubMed PMID: WOS:000259568600006.

197. Hudry B, Thomas-Chollier M, Volovik Y, Duffraisse M, Dard A, Frank D, et al. Molecular insights into the origin of the Hox-TALE patterning system. Elife. 2014;3:e01939. Epub 2014/03/20. PubMed PMID: 24642410; PubMed Central PMCID: PMC3957477.

198. Bel-Vialar S, Itasaki N, Krumlauf R. Initiating Hox gene expression: in the early chick neural tube differential sensitivity to FGF and RA signaling subdivides the HoxB genes in two distinct groups. Development. 2002;129(22):5103-15. Epub 2002/10/26. PubMed PMID: 12399303.

199. Skromne I, Thorsen D, Hale M, Prince VE, Ho RK. Repression of the hindbrain developmental program by Cdx factors is required for the specification of the vertebrate spinal cord. Development. 2007;134(11):2147-58. Epub 2007/05/18. doi: 10.1242/dev.002980. PubMed PMID: 17507415; PubMed Central PMCID: PMC2804982.

200. Sham MH, Vesque C, Nonchev S, Marshall H, Frain M, Gupta RD, et al. The zinc finger gene Krox20 regulates HoxB2 (Hox2.8) during hindbrain segmentation. Cell. 1993;72(2):183-96. Epub 1993/01/29. PubMed PMID: 8093858.

201. Voiculescu O, Taillebourg E, Pujades C, Kress C, Buart S, Charnay P, et al. Hindbrain patterning: Krox20 couples segmentation and specification of regional identity. Development. 2001;128(24):4967-78. PubMed PMID: WOS:000173434700004.

202. Hoyle J, Tang YP, Wiellette EL, Wardle FC, Sive H. nlz gene family is required for hindbrain patterning in the zebrafish. Dev Dyn. 2004;229(4):835-46. Epub 2004/03/26. doi: 10.1002/dvdy.20001. PubMed PMID: 15042707.

203. Nakamura M, Choe SK, Runko AP, Gardner PD, Sagerstrom CG. Nlz1/Znf703 acts as a repressor of transcription. BMC Dev Biol. 2008;8:108. Epub 2008/11/19. doi: 10.1186/1471-213x-8-108. PubMed PMID: 19014486; PubMed Central PMCID: PMCPmc2588584.

204. Xu Q, Wilkinson DG. Boundary formation in the development of the vertebrate hindbrain. Wiley Interdiscip Rev Dev Biol. 2013;2(5):735-45. Epub 2013/09/10. doi: 10.1002/wdev.106. PubMed PMID: 24014457.

205. Bami M, Episkopou V, Gavalas A, Gouti M. Directed neural differentiation of mouse embryonic stem cells is a sensitive system for the identification of novel Hox gene effectors. PLoS ONE. 2011;6(5):e20197. Epub 2011/06/04. doi: 10.1371/journal.pone.0020197. PubMed PMID: 21637844; PubMed Central PMCID: PMC3102681.

206. Donaldson IJ, Amin S, Hensman JJ, Kutejova E, Rattray M, Lawrence N, et al. Genome-wide occupancy links Hoxa2 to Wnt-beta-catenin signaling in mouse embryonic development. Nucleic Acids Res. 2012;40(9):3990-4001. Epub 2012/01/10. doi: 10.1093/nar/gkr1240. PubMed PMID: 22223247; PubMed Central PMCID: PMC3351182.

207. Chambers D, Wilson LJ, Alfonsi F, Hunter E, Saxena U, Blanc E, et al. Rhombomere-specific analysis reveals the repertoire of genetic cues expressed across the developing hindbrain. Neural Dev. 2009;4:6. Epub 2009/02/12. doi: 10.1186/1749-8104-4-6. PubMed PMID: 19208226; PubMed Central PMCID: PMC2649922.

208. Semmelhack JL, Donovan JC, Thiele TR, Kuehn E, Laurell E, Baier H. A dedicated visual pathway for prey detection in larval zebrafish. Elife. 2014;3. Epub 2014/12/10. doi: 10.7554/eLife.04878. PubMed PMID: 25490154; PubMed Central PMCID: PMC4281881.

209. Ruffault PL, D'Autreaux F, Hayes JA, Nomaksteinsky M, Autran S, Fujiyama T, et al. The retrotrapezoid nucleus neurons expressing and are essential for the respiratory response to CO. Elife. 2015;4. Epub 2015/04/14. doi: 10.7554/eLife.07051. PubMed PMID: 25866925.

210. Hagglund M, Borgius L, Dougherty KJ, Kiehn O. Activation of groups of excitatory neurons in the mammalian spinal cord or hindbrain evokes locomotion. Nat Neurosci. 2010;13(2):246-52. Epub 2010/01/19. doi: 10.1038/nn.2482. PubMed PMID: 20081850.

211. Bautista TG, Dutschmann M. Ponto-medullary nuclei involved in the generation of sequential pharyngeal swallowing and concomitant protective laryngeal adduction in situ. The Journal of physiology. 2014;592(Pt 12):2605-23. Epub 2014/03/19. doi: 10.1113/jphysiol.2014.272468. PubMed PMID: 24639482; PubMed Central PMCID: PMC4080941.

212. Schilling TF, Knight RD. Origins of anteroposterior patterning and Hox gene regulation during chordate evolution. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences. 2001;356(1414):1599-613. doi: 10.1098/rstb.2001.0918. PubMed PMID: WOS:000171850000008.

213. Knight RD, Panopoulou GD, Holland PW, Shimeld SM. An amphioxus Krox gene: insights into vertebrate hindbrain evolution. Dev Genes Evol. 2000;210(10):518-21. Epub 2001/02/17. doi: 10.1007/s004270050341. PubMed PMID: 11180801.

214. Longabaugh WJ. BioTapestry: a tool to visualize the dynamic properties of gene regulatory networks. Methods Mol Biol. 2012;786:359-94. Epub 2011/09/23. doi: 10.1007/978-1-61779-292-2_21. PubMed PMID: 21938637.

215. Prince VE, Moens CB, Kimmel CB, Ho RK. Zebrafish hox genes: expression in the hindbrain region of wild-type and mutants of the segmentation gene, valentino. Development. 1998;125(3):393-406. PubMed PMID: WOS:000072350300006.

216. Wada H, Garcia-Fernandez J, Holland PW. Colinear and segmental expression of amphioxus Hox genes. Dev Biol. 1999;213(1):131-41. Epub 1999/08/24. doi: 10.1006/dbio.1999.9369. PubMed PMID: 10452851.

217. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 2013;496(7446):498-503. doi: 10.1038/nature12111. PubMed PMID: MEDLINE:23594743.

218. Satoh N, Satou Y, Davidson B, Levine M. Ciona intestinalis: an emerging model for whole-genome analyses. Trends in Genetics. 2003;19(7):376-81. doi: 10.1016/s0168-9525(03)00144-6. PubMed PMID: WOS:000184538400008.

219. Nelson AC, Wardle FC. Conserved non-coding elements and cis regulation: actions speak louder than words. Development (Cambridge, England). 2013;140(7):1385-95. doi: 10.1242/dev.084459. PubMed PMID: MEDLINE:23482485.

220. Larkin M, Blackshields G, Brown N, Chenna R, McGettigan P, McWilliam H, et al. ClustalW and ClustalX version 2. Bioinformatics. 2007;23(21):2947-8.

221. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res. 2015;43(Database issue):D662-9. Epub 2014/10/30. doi: 10.1093/nar/gku1010. PubMed PMID: 25352552.

222. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods in molecular biology (Clifton, NJ). 2000;132:365-86. PubMed PMID: MEDLINE:10547847.

223. Nikolaou N, Watanabe-Asaka T, Gerety S, Distel M, Koster RW, Wilkinson DG. Lunatic fringe promotes the lateral inhibition of neurogenesis. Development. 2009;136(15):2523-33. Epub 2009/06/26. doi: 10.1242/dev.034736. PubMed PMID: 19553285; PubMed Central PMCID: PMC2709061.

224. Distel M, Wullimann MF, Koster RW. Optimized Gal4 genetics for permanent gene expression mapping in zebrafish. Proc Natl Acad Sci U S A. 2009;106(32):13365-70. Epub 2009/07/25. doi: 10.1073/pnas.0903060106. PubMed PMID: 19628697; PubMed Central PMCID: PMC2726396.

225. Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Research. 2010;20(5):565-77. doi: 10.1101/gr.104471.109. PubMed PMID: WOS:000277244800003.

226. Li Q, Ritter D, Yang N, Dong ZQ, Li H, Chuang JH, et al. A systematic approach to identify functional motifs within vertebrate developmental enhancers.

Developmental Biology. 2010;337(2):484-95. doi: 10.1016/j.ydbio.2009.10.019. PubMed PMID: WOS:000273948300027.

227. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research. 2009;37:W202-W8. doi: 10.1093/nar/gkp335. PubMed PMID: WOS:000267889100037.

228. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology. 1994;2:28-36. PubMed PMID: MEDLINE:7584402.

229. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27(7):1017-8. doi: 10.1093/bioinformatics/btr064. PubMed PMID: WOS:000289162000022.

230. Bailey TL, Noble WS. Searching for statistically significant regulatory modules. Bioinformatics. 2003;19 Suppl 2:ii16-25. Epub 2003/10/10. PubMed PMID: 14534166.

231. Moens CB, Selleri L. Hox cofactors in vertebrate development. Developmental Biology. 2006;291(2):193-206.

232. Ferretti E, Cambronero F, Tumpel S, Longobardi E, Wiedemann LM, Blasi F, et al. Hoxb1 enhancer and control of rhombomere 4 expression: complex interplay between PREP1-PBX1-HOXB1 binding sites. Mol Cell Biol. 2005;25(19):8541-52. Epub 2005/09/17. doi: 10.1128/mcb.25.19.8541-8552.2005. PubMed PMID: 16166636; PubMed Central PMCID: PMC1265741.

233. Maconochie MK, Nonchev S, Studer M, Chan SK, Popperl H, Sham MH, et al. Cross-regulation in the mouse HoxB complex: the expression of Hoxb2 in rhombomere 4 is regulated by Hoxb1. Genes & Development. 1997;11(14):1885-95. doi: 10.1101/gad.11.14.1885. PubMed PMID: WOS:A1997XM83500011.

234. Agoston Z, Heine P, Brill MS, Grebbin BM, Hau AC, Kallenborn-Gerhardt W, et al. Meis2 is a Pax6 co-factor in neurogenesis and dopaminergic periglomerular fate specification in the adult olfactory bulb. Development. 2014;141(1):28-38. Epub 2013/11/29. doi: 10.1242/dev.097295. PubMed PMID: 24284204.

235. Shim S, Kim Y, Shin J, Kim J, Park S. Regulation of EphA8 gene expression by TALE homeobox transcription factors during development of the mesencephalon. Mol Cell Biol. 2007;27(5):1614-30. Epub 2006/12/21. doi: 10.1128/mcb.01429-06. PubMed PMID: 17178831; PubMed Central PMCID: PMC1820445.

236. Agoston Z, Li N, Haslinger A, Wizenmann A, Schulte D. Genetic and physical interaction of Meis2, Pax3 and Pax7 during dorsal midbrain development. Bmc Developmental Biology. 2012;12. doi: 10 10.1186/1471-213x-12-10. PubMed PMID: WOS:000302313300001.

237. Yuan X, Braun T. An unexpected switch: regulation of cardiomyocyte proliferation by the homeobox gene meis1. Circulation research. 2013;113(3):245-8. Epub 2013/07/23. doi: 10.1161/circresaha.113.302023. PubMed PMID: 23868827.

238. Mahmoud AI, Kocabas F, Muralidhar SA, Kimura W, Koura AS, Thet S, et al. Meis1 regulates postnatal cardiomyocyte cell cycle arrest. Nature. 2013;497(7448):249-53. doi: 10.1038/nature12054. PubMed PMID: WOS:000318558200038.

239. Cvejic A, Serbanovic-Canic J, Stemple DL, Ouwehand WH. The role of meis1 in primitive and definitive hematopoiesis during zebrafish development. Haematologica. 2011;96(2):190-8. Epub 2010/11/05. doi: 10.3324/haematol.2010.027698. PubMed PMID: 21048033; PubMed Central PMCID: PMC3031685.

240. Amali AA, Sie L, Winkler C, Featherstone M. Zebrafish hoxd4a Acts Upstream of meis1.1 to Direct Vasculogenesis, Angiogenesis and Hematopoiesis. PLoS ONE. 2013;8(3). doi: e58857 10.1371/journal.pone.0058857. PubMed PMID: WOS:000316409800037.

241. Penkov D, San Martin DM, Fernandez-Diaz LC, Rossello CA, Torroja C, Sanchez-Cabo F, et al. Analysis of the DNA-Binding Profile and Function of TALE Homeoproteins Reveals Their Specialization and Specific Interactions with Hox Genes/Proteins. Cell Reports. 2013;3(4):1321-33. doi: 10.1016/j.celrep.2013.03.029. PubMed PMID: WOS:000321897100033.

242. Chan SK, Mann RS. A structural model for a homeotic protein-extradenticle-DNA complex accounts for the choice of HOX protein in the heterodimer. Proc Natl Acad Sci U S A. 1996;93(11):5223-8. doi: 10.1073/pnas.93.11.5223. PubMed PMID: WOS:A1996UN25300010.

243. Chan SK, Ryoo HD, Gould A, Krumlauf R, Mann RS. Switching the in vivo specificity of a minimal Hox-responsive element. Development. 1997;124(10):2007-14. PubMed PMID: WOS:A1997XD82500016.

244. Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Pena-Castillo L, et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell. 2008;133(7):1266-76. doi: 10.1016/j.cell.2008.05.024. PubMed PMID: WOS:000257144600022.

245. Chang CP, Brocchieri L, Shen WF, Largman C, Cleary ML. Pbx modulation of Hox homeodomain amino-terminal arms establishes different DNA-binding specificities across the Hox locus. Mol Cell Biol. 1996;16(4):1734-45. PubMed PMID: WOS:A1996UB56200049.

246. Yanez-Cuna JO, Arnold CD, Stampfel G, Boryn LM, Gerlach D, Rath M, et al. Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. Genome Res. 2014. Epub 2014/04/10. doi: 10.1101/gr.169243.113. PubMed PMID: 24714811.

247. Popperl H, Bienz M, Studer M, Chan SK, Aparicio S, Brenner S, et al. Segmental Expression of Hoxb-1 Is Controlled by a Highly Conserved Autoregulatory Loop Dependent Upon EXD/PBX. Cell. 1995;81(7):1031-42. doi: 10.1016/s0092-8674(05)80008-x. PubMed PMID: WOS:A1995RG91000008.

248. Hadrys T, Prince V, Hunter M, Baker R, Rinkwitz S. Comparative genomic analysis of vertebrate Hox3 and Hox4 genes. J Exp Zool B Mol Dev Evol. 2004;302(2):147-64. Epub 2004/04/01. doi: 10.1002/jez.b.20012. PubMed PMID: 15054858.

249. Matsunami M, Sumiyama K, Saitou N. Evolution of Conserved Non-Coding Sequences Within the Vertebrate Hox Clusters Through the Two-Round Whole Genome Duplications Revealed by Phylogenetic Footprinting Analysis. J Mol Evol. 2010;71(5-6):427-36. doi: 10.1007/s00239-010-9396-1. PubMed PMID: WOS:000284549800010.

250. Tumpel S, Cambronero F, Ferretti E, Blasi F, Wiedemann LM, Krumlauf R. Expression of Hoxa2 in rhombomere 4 is regulated by a conserved cross-regulatory mechanism dependent upon Hoxb1. Developmental Biology. 2007;302(2):646-60. doi: 10.1016/j.ydbio.2006.10.029. PubMed PMID: WOS:000244433100022.

251. Chomette D, Frain M, Cereghini S, Charnay P, Ghislain J. Krox20 hindbrain cis-regulatory landscape: interplay between multiple long-range initiation and autoregulatory elements. Development. 2006;133(7):1253-62. Epub 2006/02/24. doi: 10.1242/dev.02289. PubMed PMID: 16495311.

252. Manzanares M, Bel-Vialar S, Ariza-McNaughton L, Ferretti E, Marshall H, Maconochie MM, et al. Independent regulation of initiation and maintenance phases of Hoxa3 expression in the vertebrate hindbrain involve auto- and cross-regulatory mechanisms. Development. 2001;128(18):3595-607. Epub 2001/09/22. PubMed PMID: 11566863.

253. Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, et al. Zebrafish hox clusters and vertebrate genome evolution. Science. 1998;282(5394):1711-4. doi: 10.1126/science.282.5394.1711. PubMed PMID: WOS:000077246600048.

254. Thisse C, Thisse B. High Throughput Expression Analysis of ZF-Models Consortium Clones. ZFIN Direct Data Submission. 2005.

255. Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, et al. Conservation of trans-acting circuitry during mammalian regulatory evolution. Nature. 2014;515(7527):365-70. Epub 2014/11/21. doi: 10.1038/nature13972. PubMed PMID: 25409825.

256. O'Quin KE, Smith D, Naseer Z, Schulte J, Engel SD, Loh YE, et al. Divergence in cis-regulatory sequences surrounding the opsin gene arrays of African cichlid fishes. BMC Evolutionary Biology. 2011.

257. Gottgens B, Ferreira R, Sanchez MJ, Ishibashi S, Li JA, Spensberger D, et al. cis-Regulatory Remodeling of the SCL Locus during Vertebrate Evolution. Mol Cell Biol. 2010;30(24):5741-51. doi: 10.1128/mcb.00870-10. PubMed PMID: WOS:000284429300012.

258. Roberts JA, Miguel-Escalada I, Slovik KJ, Walsh KT, Hadzhiev Y, Sanges R, et al. Targeted transgene integration overcomes variability of position effects in zebrafish. Development. 2014;141(3):715-24. Epub 2014/01/23. doi: 10.1242/dev.100347. PubMed PMID: 24449846; PubMed Central PMCID: PMC3899822.

259. Hisano Y, Sakuma T, Nakade S, Ohga R, Ota S, Okamoto H, et al. Precise in-frame integration of exogenous DNA mediated by CRISPR/Cas9 system in zebrafish. Sci Rep. 2015;5:8841. Epub 2015/03/06. doi: 10.1038/srep08841. PubMed PMID: 25740433; PubMed Central PMCID: PMCPmc4350073.

260. Dias AS, de Almeida I, Belmonte JM, Glazier JA, Stern CD. Somites without a clock. Science. 2014;343(6172):791-5. Epub 2014/01/11. doi: 10.1126/science.1247575. PubMed PMID: 24407478; PubMed Central PMCID: PMCPmc3992919.

261. Minguillon C, Nishimoto S, Wood S, Vendrell E, Gibson-Brown JJ, Logan MP. Hox genes regulate the onset of Tbx5 expression in the forelimb. Development. 2012;139(17):3180-8. Epub 2012/08/09. doi: 10.1242/dev.084814. PubMed PMID: 22872086; PubMed Central PMCID: PMC3413163.

262. Nishimoto S, Minguillon C, Wood S, Logan MP. A combination of activation and repression by a colinear hox code controls forelimb-restricted expression of tbx5 and reveals hox protein specificity. PLoS Genet. 2014;10(3):e1004245. Epub

2014/03/22. doi: 10.1371/journal.pgen.1004245. PubMed PMID: 24651482; PubMed Central PMCID: PMC3961185.

263. Galant R, Walsh CM, Carroll SB. Hox repression of a target gene: extradenticle-independent, additive action through multiple monomer binding sites. Development. 2002;129(13):3115-26. Epub 2002/06/19. PubMed PMID: 12070087.

264. Uv AE, Harrison EJ, Bray SJ. Tissue-specific splicing and functions of the Drosophila transcription factor Grainyhead. Mol Cell Biol. 1997;17(11):6727-35. Epub 1997/10/29. PubMed PMID: 9343437; PubMed Central PMCID: PMCPmc232527.

265. Jung H, Mazzoni EO, Soshnikova N, Hanley O, Venkatesh B, Duboule D, et al. Evolving Hox activity profiles govern diversity in locomotor systems. Dev Cell. 2014;29(2):171-87. Epub 2014/04/22. doi: 10.1016/j.devcel.2014.03.008. PubMed PMID: 24746670; PubMed Central PMCID: PMC4024207.

266. Fujioka M, Gebelein B, Cofer ZC, Mann RS, Jaynes JB. Engrailed cooperates directly with Extradenticle and Homothorax on a distinct class of homeodomain binding sites to repress sloppy paired. Dev Biol. 2012;366(2):382-92. Epub 2012/04/28. doi: 10.1016/j.ydbio.2012.04.004. PubMed PMID: 22537495; PubMed Central PMCID: PMCPmc3362656.

267. Bjerke GA, Hyman-Walsh C, Wotton D. Cooperative transcriptional activation by Klf4, Meis2, and Pbx1. Mol Cell Biol. 2011;31(18):3723-33. Epub 2011/07/13. doi: MCB.01456-10 [pii] 10.1128/MCB.01456-10. PubMed PMID: 21746878; PubMed Central PMCID: PMC3165729.

268. Shanmugam K, Green NC, Rambaldi I, Saragovi HU, Featherstone MS. PBX and MEIS as non-DNA-binding partners in trimeric complexes with HOX proteins. Mol Cell Biol. 1999;19(11):7577-88.

269. Ishibashi M, Noda AO, Sakate R, Imanishi T. Evolutionary growth process of highly conserved sequences in vertebrate genomes. Gene. 2012;504(1):1-5. doi: 10.1016/j.gene.2012.05.003. PubMed PMID: WOS:000306205500001.

270. Sharov AA, Ko MSH. Exhaustive Search for Over-represented DNA Sequence Motifs with CisFinder. DNA Research. 2009;16(5):261-73. doi: 10.1093/dnares/dsp014. PubMed PMID: WOS:000271610400002.

271. Diallo AB, Makarenkov V, Blanchette M. Exact and heuristic algorithms for the Indel Maximum Likelihood Problem. J Comput Biol. 2007;14(4):446-61. Epub 2007/06/19. doi: 10.1089/cmb.2007.A006. PubMed PMID: 17572023.

272. Diallo AB, Makarenkov V, Blanchette M. Ancestors 1.0: a web server for ancestral sequence reconstruction. Bioinformatics. 2010;26(1):130-1. Epub 2009/10/24. doi: 10.1093/bioinformatics/btp600. PubMed PMID: 19850756.

273. Murakami Y, Pasqualetti M, Takio Y, Hirano S, Rijli FM, Kuratani S. Segmental development of reticulospinal and branchiomotor neurons in lamprey: insights into the evolution of the vertebrate hindbrain. Development. 2004;131(5):983-95. Epub 2004/02/20. doi: 10.1242/dev.00986. PubMed PMID: 14973269.

274. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biology. 2007;8(2).

275. Mahony S, Benos PV. STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Research. 2007;35:W253-W8. doi: 10.1093/nar/gkm272. PubMed PMID: WOS:000255311500047.

276. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 2013;30(12):2725-9. Epub 2013/10/18. doi: 10.1093/molbev/mst197. PubMed PMID: 24132122; PubMed Central PMCID: PMCPmc3840312.

277. Hiller M, Schaar BT, Bejerano G. Hundreds of conserved non-coding genomic regions are independently lost in mammals. Nucleic Acids Res. 2012;40(22):11463-76. Epub 2012/10/09. doi: 10.1093/nar/gks905. PubMed PMID: 23042682; PubMed Central PMCID: PMC3526296.

278. Frankenberg SR, Frank D, Harland R, Johnson AD, Nichols J, Niwa H, et al. The POU-er of gene nomenclature. Development. 2014;141(15):2921-3. Epub 2014/07/24. doi: 10.1242/dev.108407. PubMed PMID: 25053425.

279. Jean C, Oliveira NM, Intarapat S, Fuet A, Mazoyer C, De Almeida I, et al. Transcriptome analysis of chicken ES, blastodermal and germ cells reveals that chick ES cells are equivalent to mouse ES cells rather than EpiSC. Stem Cell Res. 2015;14(1):54-67. Epub 2014/12/17. doi: 10.1016/j.scr.2014.11.005. PubMed PMID: 25514344; PubMed Central PMCID: PMC4305369.

280. Frank CL, Liu F, Wijayatunge R, Song L, Biegler MT, Yang MG, et al. Regulation of chromatin accessibility and Zic binding at enhancers in the developing cerebellum. Nat Neurosci. 2015;18(5):647-56. Epub 2015/04/08. doi: 10.1038/nn.3995. PubMed PMID: 25849986; PubMed Central PMCID: PMC4414887.

281. Shimizu N, Kawakami K, Ishitani T. Visualization and exploration of Tcf/Lef function using a highly responsive Wnt/beta-catenin signaling-reporter transgenic zebrafish. Dev Biol. 2012;370(1):71-85. Epub 2012/07/31. doi: 10.1016/j.ydbio.2012.07.016. PubMed PMID: 22842099.

282. Bedell VM, Wang Y, Campbell JM, Poshusta TL, Starker CG, Krug RG, 2nd, et al. In vivo genome editing using a high-efficiency TALEN system. Nature. 2012;491(7422):114-8. doi: 10.1038/nature11537. PubMed PMID: MEDLINE:23000899.

283. Dahlem TJ, Hoshijima K, Jurynec MJ, Gunther D, Starker CG, Locke AS, et al. Simple Methods for Generating and Detecting Locus-Specific Mutations Induced with TALENs in the Zebrafish Genome. Plos Genetics. 2012;8(8). doi: e1002861 10.1371/journal.pgen.1002861. PubMed PMID: WOS:000308529300021.

284. Casillas S, Barbadilla A, Bergman CM. Purifying selection maintains highly conserved noncoding sequences in Drosophila. Mol Biol Evol. 2007;24(10):2222-34. Epub 2007/07/25. doi: 10.1093/molbev/msm150. PubMed PMID: 17646256.

285. Piasecka B, Lichocki P, Moretti S, Bergmann S, Robinson-Rechavi M. The Hourglass and the Early Conservation Models-Co-Existing Patterns of Developmental Constraints in Vertebrates. Plos Genetics. 2013;9(4). doi: e1003476 10.1371/journal.pgen.1003476. PubMed PMID: WOS:000318073300067.

286. Akhshabi S, Sarda S, Dovrolis C, Yi S. An explanatory evo-devo model for the developmental hourglass. F1000Res. 2014;3:156. Epub 2014/09/12. doi: 10.12688/f1000research.4583.1. PubMed PMID: 25210617; PubMed Central PMCID: PMC4156030.

287. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15(7):901-13. Epub 2005/06/21. doi: 10.1101/gr.3577405. PubMed PMID: 15965027; PubMed Central PMCID: PMC1172034.

288. Jimenez-Guri E, Pujades C. An ancient mechanism of hindbrain patterning has been conserved in vertebrate evolution. Evol Dev. 2011;13(1):38-46. doi: 10.1111/j.1525-142X.2010.00454.x. PubMed PMID: WOS:000286007500004.

289. Kuratani S, Ueki T, Aizawa S, Hirano S. Peripheral development of cranial nerves in a cyclostome, Lampetra japonica: morphological distribution of nerve branches and the vertebrate body plan. J Comp Neurol. 1997;384(4):483-500. Epub 1997/08/11. PubMed PMID: 9259485.

290. Barreiro-Iglesias A, Gómez-López MP, Anadón R, Rodicio MC. Early Development of the Cranial Nerves in a Primitive Vertebrate, the Sea Lamprey, Petromyzon Marinus L. . The Open Zoology Journal. 2008;1:37-43.

291. Suzuki DG, Murakami Y, Yamazaki Y, Wada H. Expression patterns of Eph genes in the "dual visual development" of the lamprey and their significance in

the evolution of vision in vertebrates. Evol Dev. 2015;17(2):139-47. Epub 2015/03/25. doi: 10.1111/ede.12119. PubMed PMID: 25801221.

292. Jackman WR, Langeland JA, Kimmel CB. islet reveals segmentation in the Amphioxus hindbrain homolog. Dev Biol. 2000;220(1):16-26. Epub 2000/03/18. doi: 10.1006/dbio.2000.9630. PubMed PMID: 10720427.

293. Manzanares M, Wada H, Itasaki N, Trainor PA, Krumlauf R, Holland PWH. Conservation and elaboration of Hox gene regulation during evolution of the vertebrate head. Nature. 2000;408(6814):854-7. PubMed PMID: WOS:000165831300049.

294. Natale A, Sims C, Chiusano ML, Amoroso A, D'Aniello E, Fucci L, et al. Evolution of anterior Hox regulatory elements among chordates. BMC Evolutionary Biology. 2011;11(330). doi: 330 10.1186/1471-2148-11-330. PubMed PMID: WOS:000297651300001.

295. Stedman A, Lecaudey V, Havis E, Anselme I, Wassef M, Gilardi-Hebenstret P, et al. A functional interaction between Irx and Meis patterns the anterior hindbrain and activates krox20 expression in rhombomere 3. Developmental Biology. 2009;327(2):566-77. doi: 10.1016/j.ydbio.2008.12.018. PubMed PMID: WOS:000264060700028.