NOTE TO USERS

This reproduction is the best copy available.

UMI®

uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Kyle Martin
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (Biology)
GRADE / DEGREE

Department of Biology
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Genomic Characterization of Cyclostome *Dlx* Gene Family Members:
Insight into the Evolution of the Chordate Genome and Body Plan from the Organizational and Transcriptional Regulatory Properties of *Dlx* Genes in the Petromyzontiformes (Lamprey) and the Hyperotreti (Hagfish)

TITRE DE LA THÈSE / TITLE OF THESIS

Dr. Marc Ekker
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Dr. Tom Moon

Dr. Ashkan Golshani

Dr. Rees Kassen

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

MSc. THESIS

Genomic Characterization of Cyclostome *Dlx* Gene Family Members:

Insight into the Evolution of the Chordate Genome and Body Plan from the Organizational and Transcriptional Regulatory Properties of *Dlx* Genes in the Petromyzontiformes (Lamprey) and the Hyperotreti (Hagfish).

By:

Kyle J. Martin

Submitted to the Department of Biology

University of Ottawa

In partial fulfillment of the requirements for the degree of Master of Science

Supervisor:

Dr. Marc Ekker

Jury:

Dr. Tom Moon

Dr. Ashkan Golshani

Dr. Rees Kassen

© Kyle Martin, Ottawa, Canada, 2009

Library and Archives Canada

Bibliothèque et Archives Canada

Published Heritage Branch

Direction du Patrimoine de l'édition

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-59868-9
Our file Notre référence
ISBN: 978-0-494-59868-9

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Acknowledgements

I thoroughly appreciate the help and advice I have received from the many individuals who gave it during the course of this program. My sincere thanks go to my supervisor Dr. Marc Ekker for his guidance and encouragement to pursue my research interests. Thank you to the members of my committee, Dr. Ashkan Golshani, Dr. Anne-Gaelle Roland-Lagan, and Dr. Catherine Tsilfidis for useful comments and questions. For your extraordinary efforts in reviewing this thesis I offer my thanks to the members of the jury: Dr. Tom Moon, Dr. Rees Kassen, and Dr. Ashkan Golshani.

I would like to extend my extreme gratitude to my many collaborators and colleagues without whom it would have been impossible to attempt the experiments presented herein. Thank you to Dr. Sylvie Rétaux and Dr. Sylvie Mazan for hosting me in your labs, and allowing me access to your facilities, materials, and insight. Thank you also to Dr. Shigeru Kuratani for your supervision during my stay in your lab, to Dr. Shigehiro Kuraku for sharing results and useful discussion, and to Dr. Kinya Ota for all your guidance in experimental studies using hagfish, for sharing data, materials, and laughs. Thank you to Dr. Tom Moon, to Dr. Bruce Cameron, and to Dr. Chris Amemiya for providing materials indispensable for the completion of this work.

I appreciate all the help from past and present members of the Ekker Lab. Thank you to Ashish Maurya, Ryan MacDonald, Gary Hatch, and Vishal Saxena for your direct contributions to this project during my tenure. I am forever grateful for the loving encouragement to pursue my studies from all my family and loved ones, most especially Arleen Martin, Raymond Martin, Kristin Martin and Kirstie Gordon-Loiello.

Quotations

"Embryology is to me by far the strongest single class of facts in favour of change of form..."

- Charles Darwin

"So long and thanks for all the fish"

- The dolphins from Douglas N. Adams' The Hitchhiker's Guide to the Galaxy

Table of Contents

Acknowledgements 2

Quotations 3

List of Figures 6

Abstract 8

Résumé 9

Section 1: Introduction 10

- 1.1 Evolutionary Developmental Biology, or "Evo-Devo" 10
- 1.2 Cyclostomata and the Choice of Model Organisms in Evo-Devo 12
  - 1.2.1 Phylogeny of the Cyclostomes 12
  - 1.2.2 Cyclostomes: An Outgroup to Gnathostome Vertebrates 20
- 1.3 Gene and Genome Duplication 24
  - 1.3.1 The Fates of Duplicated Genes 24
  - 1.3.2 Genome Duplication 26
  - 1.3.3 The Timing of Genome Duplication Events 28
  - 1.3.4 Gene Clusters 31
- 1.4 The *Dlx* Gene Family 36
  - 1.4.1 Introduction to the *Dll/Dlx* Gene Family 36
  - 1.4.2 Cephalochordate and Ascidian *Dll* Expression 39
  - 1.4.3 Vertebrate *Dlx* Gene Expression 41
  - 1.4.4 Cyclostome *Dlx* Gene Expression 48
- 1.5 *Dlx* Gene Regulation 51
- 1.6 Major Questions 57

Section 2: Materials and Methods 57

- 2.1 Genomic DNA Extraction from Hagfish Tissue 57
- 2.2 Isolation of Novel Hagfish *Dlx* Gene Fragments 58
- 2.3 Isolation of Lamprey *Dlx* Gene Fragments 60
- 2.4 Isolation of BACs Containing Hagfish and Lamprey *Dlx* Genes 61
- 2.5 PCR Screening BAC Clones 62
- 2.6 Southern Hybridization to *Dlx* BAC Clones 63
- 2.7 Cloning BAC Fragments in pBluescript 64

2.8 Fluorescent *in situ* Hybridization to *E. burgeri* Chromosomes 64
2.9 Microinjection of DNA into Zebrafish Embryos 67
2.10 Screening Zebrafish for Transgene Transmission 67
2.11 Microinjection of DNA into *L. fluviatilis* Embryos 68
2.12 Total RNA Isolation from Adult Hagfish Brain 68
2.13 Reverse Transcription of Adult Hagfish Brain Total RNA 68
2.14 Amplification of Hagfish *Dlx* cDNA 69
Section 3: Results 70
3.1 Identification of *Dlx* Genes in the Lamprey Genome 70
3.2 Identification of *Dlx* Genes in the Hagfish Genome 77
3.3 Physical Mapping of Hagfish *Dlx* Genes using FISH 89
3.4 Phylogenetic Analysis of Cyclostome *Dlx* Genes 96
3.5 Regulation of Cyclostome *Dlx* Genes 102
3.6 Transgenic Analysis of Noncoding DNA Surrounding Lamprey *Dlx* Loci 108
3.6.1 Zebrafish Transgenic Assay 109
3.6.1a *LfDlxA-eGFP* Activity in Zebrafish Embryos 110
3.6.1b *LfDlxC/D-eGFP* Activity in Zebrafish Embryos 115
3.6.1c *LfDlxC/D*-intergenic-β-globin-*eGFP* Activity in Zebrafish Embryos 119
3.6.2 Lamprey Transgenic Assay 122
3.6.2a Activity of the CMV-*eGFP* Transgene in *L. fluviatilis* Embryos 123
3.6.2b *LfDlxA-eGFP* Activity in Lamprey Embryos 126
3.6.2c Activity of the *LfDlxC/D-eGFP* Transgene in Lamprey Embryos 128
3.6.2d Expression of Zebrafish *Dlx1a/Dlx2a*-intergenic-β-globin-*eGFP* in Lamprey Embryos 128
Section 4: Discussion 131
Section 5: Conclusions 141
Section 6: References 144

List of Figures

Figure 1: Theories on the Phylogenetic Relationship Between Hagfishes, Lamprey, and Gnathostomes 17

Figure 2: Molecular Phylogeny of the Deuterostomes 18

Figure 3: Time Scale for the Divergence of Cyclostomes and Gnathostomes 19

Figure 4: Morphological Novelties in the Chordate Lineage are Associated With the Expression of *Dlx* Genes 23

Figure 5: The Timing of Whole Genome Duplication Events in the Chordate Lineage 30

Figure 6: Duplication of the *Dll/Dlx* Gene Clusters and *Hox* Complexes can be Used to Resolve the Timing of Ancient Whole Genome Duplications 35

Figure 7: Branchial Arches in the Mouse Embryo 46

Figure 8: The *Dlx* Code in the Branchial Arches of the Mouse Embryo 46

Figure 9: A *Dlx* Code Patterns the Dorso-ventral Axis of the Branchial Arches in Gnathostomes 47

Figure 10: Loss of Maxillary-Mandibular Asymmetry in $Dlx5^{-/-}$ $Dlx6^{-/-}$ Double Muntant Mice 47

Figure 11: *in situ* Analysis of *Petromyzon marinus Dlx* Expression 50

Figure 12: Conserved Noncoding Elements (CNEs) are Found in the Intergenic Region of *Dlx* Gene Pairs 55

Figure 13: Activity of Intergenic Enhancers Associated with the *Dlx5/Dlx6* Bigene Clusters in the Mouse Embryo 56

Figure 14: Activity of Intergenic Enhancers Associated with the Zebrafish *dlx1a/dlx2a* Cluster 56

Figure 15: Schematic Representation of the Arrangement of *Dlx* Genes in *Petromyzon marinus* 76

Figure 16: General Schematic of the Organization of *Dlx* Genes in the Hagfish *Eptatretus stoutii* 86

Figure 17: Detailed Structure of the *E. stoutii DlxW/DlxX* Gene Cluster ____ 86
Figure 18: cDNA and Amino Acid Sequence of *E. stoutii Dlx* Genes ____ 87, 88
Figure 19: *E. burgeri* Metaphase Chromosome Preparations ____ 92
Figure 20: FISH Localization of *E. burgeri* 18S rDNA Cluster ____ 93
Figure 21: FISH Localization of *E. burgeri EbDlx35B* ____ 94
Figure 22: FISH Colocalization of *E. burgeri EbDlx16A* and *Eb18S* ____ *95*
Figure 23: Amino Acid Level Alignment of Chordate *Dlx* Homeodomains ____ 99
Figure 24: Protein Neighbour-Joining Gene Tree of Chordate *Dlx* Genes ____ 100, 101
Figure 25: Activity of *LfDlxA-eGFP* Cosmid Transgene in the Zebrafish Embryo ____ 113, 114
Figure 26: Activity of the *LfDlxC/DlxD-eGFP* Cosmid Transgene in the Zebrafish Embryo ____ 117, 118
Figure 27: Activity of the 14kb *LfDlxC/DlxD*-intergenic-β-globin-*eGFP* Transgene in the Zebrafish Embryo ____ 121
Figure 28: Activity of the CMV-*eGFP* Ubiquitous Transgene in the *Lampetra fluviatilis* Embryo ____ 125
Figure 29: Activity of the *LfDlxA-eGFP* Cosmid Transgene in the *Lampetra Fluviatilis* Embryo ____ 127
Figure 30: Activity of the *LfDlxC/DlxD-eGFP* Cosmid Transgene in the *Lampetra fluviatilis* Embryo ____ 127
Figure 31: Activity of the Zebrafish 6kb *dlx1a/dlx2a*-intergenic-β-globin-*eGFP* Transgene in the *Lampetra fluviatilis* Embryo ____ 130
Figure 32: Summary of *Dlx* Gene Organization and Gene/Genome Duplication Events Across Chordate Phylogeny ____ 139
Figure 33: Summary of Major Properties of *Dlx* Gene Organization and Regulation in the Chordate Lineage ____ 140

Abstract

Gnathostome novelties include jaws, paired appendages, and true teeth. *Dlx* genes encode transcription factors indispensable for embryonic development of these novelties. Gnathostomes possess at least 6 *Dlx* genes organized in 3 bi-gene clusters, a physical arrangement which is proposed to affect their expression though shared enhancer elements. I studied the *Dlx* genes of Cyclostomes, the evolutionary sister group of Gnathostomes. I identified 4 novel members of the *Dlx* gene family in hagfish *(Eptatretus stoutii)*, and confirmed the presence of 6 *Dlx* genes in lamprey *(Petromyzon marinus)*. I found that Cyclostomes have only 1 gene cluster and several orphan genes. This lack of conserved arrangement is coincident with an absence of conserved *Dlx* enhancers. Irregardless, some regulatory conservation is still apparent as lamprey non-coding DNA is able to drive *dlx* specific expression patterns of reporter genes in zebrafish. Therefore Cyclostome and Gnathostome *Dlx* are both organized and regulated differently. The causes and consequences of these changes in Chordate evolution are discussed.

Résumé

Les nouveautés évolutives des Gnathostomes comprennent les mâchoires, les membres pairés et les dents vraies. Les gènes *Dlx* codent pour des facteurs de transcription indispensables au développement des ces structures. Les Gnathostomes possèdent au moins 6 gènes *Dlx*, organisés en trois bigènes, une organisation génomique dont on pense qu'elle affecte l'expression génique grâce au partage d'éléments régulateurs de type intensificateur. J'ai étudié les gènes *Dlx* des cyclostomes, le groupe sœur des Gnathostomes d'un point de vue évolutif. J'ai identifié 4 nouveaux gènes *Dlx* chez la myxine *(Eptatretus stoutii)*, et confirmé l'existence de 6 gènes *Dlx* chez la lamproie *(Petromyzon marinus)*. J'ai démontré que les cyclostomes ne possèdent qu'un seul bigène *Dlx* et plusieurs gènes isolés. L'absence de conservation de l'organisation génomique coïncide avec une absence de conservation des séquences des éléments intensificateurs. Il semble toutefois qu'une conservation des mécanismes de régulation subsiste puisque que de l'ADN non-codant est capable de diriger l'expression de gènes rapporteurs avec des profils semblables à ceux des gènes *dlx* lorsque testés chez le poisson-zèbre. Il semble donc que les gènes *Gnathostomes* des cyclostomes et des Gnathostomes sont organisés et régulées différemment. Les causes et conséquences de ces changements survenus au cours de l'évolution des cordés sont discutées.

Section 1: Introduction

1.1 Evolutionary Developmental Biology, or "Evo-Devo"

Cross-disciplinary studies in science combining two relatively disparate fields can yield extremely valuable information that otherwise would be unreachable by single conventional method alone. Modern Evolutionary Developmental Biology, or "Evo-Devo" as it has come to be known, is the result of combining many of the tools and knowledge wielded by conventional developmental biologists, with evolutionary biology. Evo-Devo, however, is far from a new science. In Darwin's era the terms "evolution" and "development" were intimately connected. Much of the best evidence for classical Darwinian evolution came from the comparative study of animals in their embryonic stages, and Darwin stated in The Origin "Embryology is to me by far the strongest single class of facts in favour of change of form..." (Darwin 1859). Today we are not witnessing the birth of a new highly specific field of science applicable to only the most remote academic applications. We are witnessing the reunion of a single field which for over a century had been separated by creative conflicts and which have sought the answer to essentially the same question from different points of view. That question is: how does a complex organism emerge from comparatively simple beginnings. The embryologist will ask this question in terms of how a multicellular organism develops from a single cell, while the evolutionary biologist has traditionally been occupied with how complex organisms emerged from simpler ancestors over time. It was to the great credit of several modern developmental geneticists that today we are witnessing the reunion of these estranged sister-sciences. The combined efforts of Edward B. Lewis, Christiane Nüsslein-Volhard, and Eric F. Wieschaus, for which they shared a the Nobel Prize in

Medicine in 1995, demonstrated, among other things, that organism as disparate as arthropods and vertebrates use a common set of developmental genes to pattern their embryological development. This ignited interest in the genetic bonds that diverse organisms shared. This breakthrough led the field in a series of discoveries which clearly re-established the tight link between animal development, and evolutionary change. .

The similarities in developmental genes across disparate organisms sets up what some have described as the central paradox of evolutionary developmental biology. That is, given that we now know that disparate organisms use a common set of developmental genes to control their embryogenesis how does one account for the massive morphological differences observed across different clades? The modern field of evolutionary developmental biology attempts to resolve this paradox by applying the powerful tools of molecular developmental biology, originally designed in model species such as the fly and the mouse, to diverse taxa in key phylogenetic positions. The answer that is emerging is that the striking conservation of developmental genes is mirrored by highly divergent mechanisms of deployment of these genes, and that apparently minor change in the coding sequence or non-coding sequences of these developmental genes can have powerful impacts on morphology.

1.2 Cyclostomata and the Choice of Model Organisms in Evo-Devo

The evolution of the chordate body plan has occurred largely by the stepwise addition of shared derived characters that all have their origins in embryonic development. For evolutionary developmental biologists, the correct choice of model organism used to elucidate the pathways important in the evolution of novel phenotypes is vital. In cases where the goal is to determine the molecular developmental mechanisms responsible for the appearance of an evolutionary novelty it is useful to compare organisms with and without the morphological innovation in question.

Unfortunately, the true ancestral condition of any feature is only preserved in fossils, if at all, and modern biologists are restricted to using living representatives of taxa which diverged from the common ancestor prior to the appearance of the character. A valuable example of such taxa are the only living group of jawless fish variously known as either Cyclostomes (round-mouth) or Agnathans (literally, jawless), which diverged from the chordate lineage prior to the evolution of a suite of massively important morphological characteristics including paired appendages, teeth, and jaws.

1.2.1 Phylogeny of the Cyclostomes

One of the most fundamental principles of evolutionary developmental biology is ancestry. In order to make any kind of reasonable deductions about the origin of characters, it is necessary to be reasonably certain of the evolutionary succession of the species being considered in the comparison. In other words, assurances must be made that the character in question is truly a novelty in the derived species and is absent or primitive in the ancestor, and not secondarily lost. Various methods of phylogenetic

reconstruction are employed in order to determine the evolutionary relationship amongst different organisms.

All chordates are defined by possession of a notochord, but only the craniates possess a true cranium with a tripartite brain, and only animals possessing cartilaginous or ossified vertebrae protecting the spinal cord are considered true vertebrates. While current morphological phylogenies take advantage of several new and powerful developments, the principle is constant: the phenotype of the animal is used to determine its relative phylogenetic position. Animal phylogeny has recently been revisited in the context of molecular data. Molecular phylogenies are constructed by comparing information from the genotypes of different animals and do not take into consideration the actual appearance of the organism. For the most-part, morphological and molecular phylogenies will support each other, mutually validating each method. Recently however, strong molecular phylogenetic data has overturned a long standing view concerning the relationship of some of the most primitive chordates by demonstrating that Ascidians are more closely related to Vertebrates than Cephalochordates are despite the overt differences present in adult body plan which result in Ascidians looking somewhat more like an underwater vegetable than our closest non-vertebrate relative (Delsuc et al 2006). Some important nodes in chordate evolution still remain relatively unresolved. Perhaps one of the most significant conflicts between molecular and morphological phylogenetic classification lies close to the base of the chordates. The group of organisms known as either Cyclostomes (round-mouth), or Agnathans (without jaws) include the hagfishes (Order *Hyperotrei*, Family *Myxinoidea*) and the lamprey (Order *Petromyzontiformes*, Family *Petromyzontoidea*), and lies in a key phylogenetic

position near the base of the chordate phylogenetic tree. Jawless fishes were once a diverse clade and fossil species, especially from the armored Ostracoderms, have been described from nearly all corners of the world (Janvier 2006). The fossil record of hagfishes and lampreys are relatively sparse. The oldest lamprey fossil found to date, *Priscomyzon*, dates to the Devonian period and looks strikingly similar to modern lampreys (Gess et al. 2006), and only a handful of other fossil species have been described (Chang et al. 2006, Janvier and Lund 1983). The record of hagfishes is even sparser but reported fossil forms from the Pennsylvanian period are also similar to modern forms (Bardack 1991). It was from a related but separate lineage of these jawless fishes that Gnathostomes and all extant vertebrates eventually evolved. It is therefore of significant interest to study the only two living representatives of this once great clade in order to understand more about the changes that accompanied the transition to the modern vertebrate phylotype.

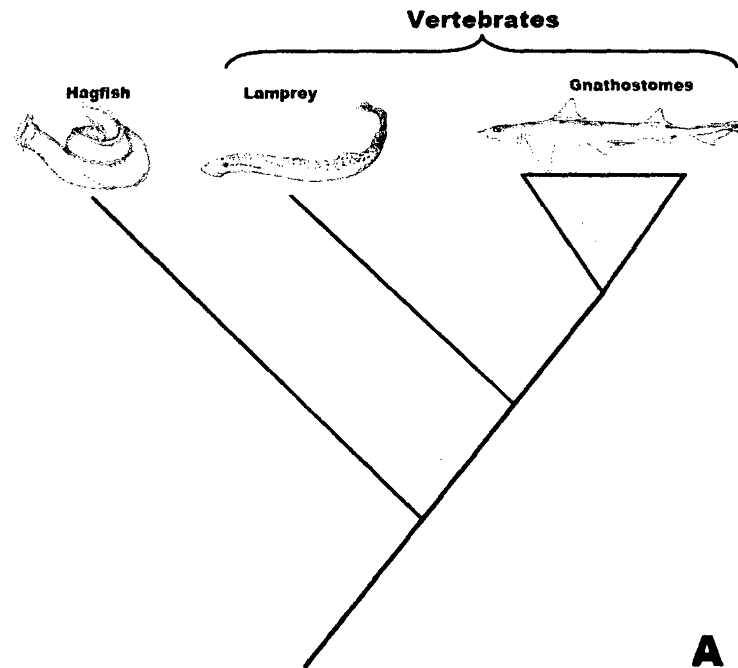
Despite the ease with which one can distinguish Cyclostomes from Gnathostomes using obvious macroscopic morphological features such as the absence of jaws or paired appendages in the former, there is serious ambiguity concerning the relationship between hagfishes and lampreys. Two major theories explaining the relationship of hagfish, lamprey, and Gnathostomes exist. The so-called "Vertebrate theory" (Figure 1A) places the hagfishes more basal on the chordate phylogeny, while the lamprey and the Gnathostomes form a monophyletic group, and are considered "Vertebrates". This theory is largely supported by morphological data including the presence of minuscule chondrified vertebral elements present in the lamprey which clearly classifies it as a vertebrate by the definition of the term, while no such elements have been observed in the

hagfish (Janvier 1991, 1996). The "Cyclostome Theory" (Figure 1B) posits that hagfishes and lamprey form a monophyletic group, separate from Gnathostomes. Multiple phylogenies constructed first using 18S and 28S rDNA genes, and subsequently a large number of other nuclear genes, some even on a massive scale incorporating whole-genome information from different clades, and using multiple methods from simple parsimony to Baysian analysis, show strong support for the Cyclostome Theory (Yu et al 2008, Delarbre et al. 2002, Furlong and Holland 2002, Delsuc et al 2006). A recently published phylogeny of the Deuterostomes (Figure 2; Delsuc et al 2006) depicts the most widely accepted position of hagfishes and lamprey according to molecular data.

Resolving this controversy is necessary in order to fully understand the earliest origins of the Craniate body plan. The difficulties in resolving this node lie both in the huge amount of time separating Gnathostomes from Cyclostomes, but also in the relatively short amount of time between the proposed divergence of Cyclostomes from Gnathostomes, and the divergence of hagfish from lamprey. Some of the most recent estimates derived from molecular data place the Cyclostome-Gnathostome divergence at approximately 500Mya, while it appears that hagfish and lamprey diverged from each other approximately 450Mya (Figure 3, Kuraku and Kuratani 2006). The relatively short (50Mya) period of time between the divergence of Cyclostomes from Gnathostomes, before the divergence of lamprey from hagfish complicates both morphological, and molecular methods of reconstructing phylogenies as both the phenotype and genetic sequence has had a significant amount of time to change.

For the purposes of this thesis the term Cyclostomes will be used in order to ensure consistency and to describe both hagfish and lamprey, regardless of whether or not

Cyclostomes are monophyletic or paraphyletic. One goal of the type of genomic analyses conducted within this thesis is to contribute to a more clear phylogenetic reconstruction of the relationship between Cyclostomes and Gnathostomes and therefore the question is still considered open.



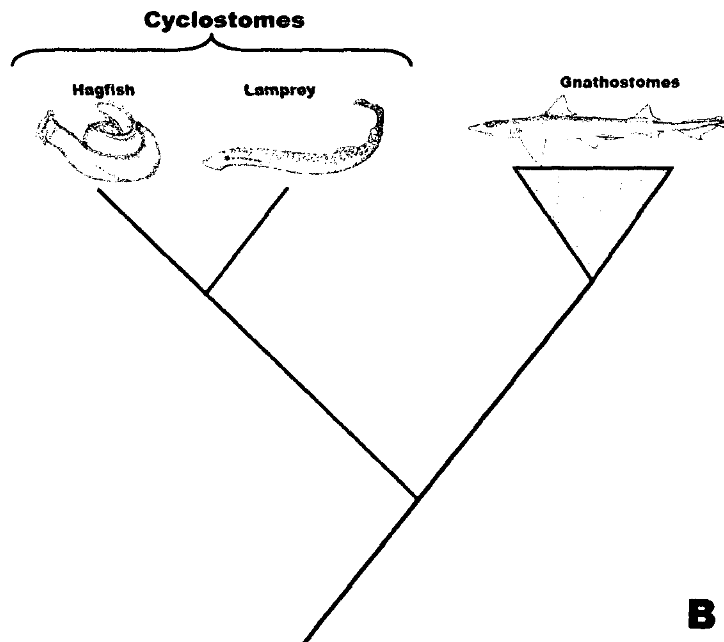


Figure 1: Theories on the phylogenetic relationship between hagfishes, lamprey, and Gnathostomes. The Vertebrate Theory (A) suggests that lamprey are more closely related to Gnathostomes than they are to hagfishes, while the Cyclostome Theory (B) suggests that hagfishes and lamprey are monophyletic and are more closely related to each other than to Gnathostomes.

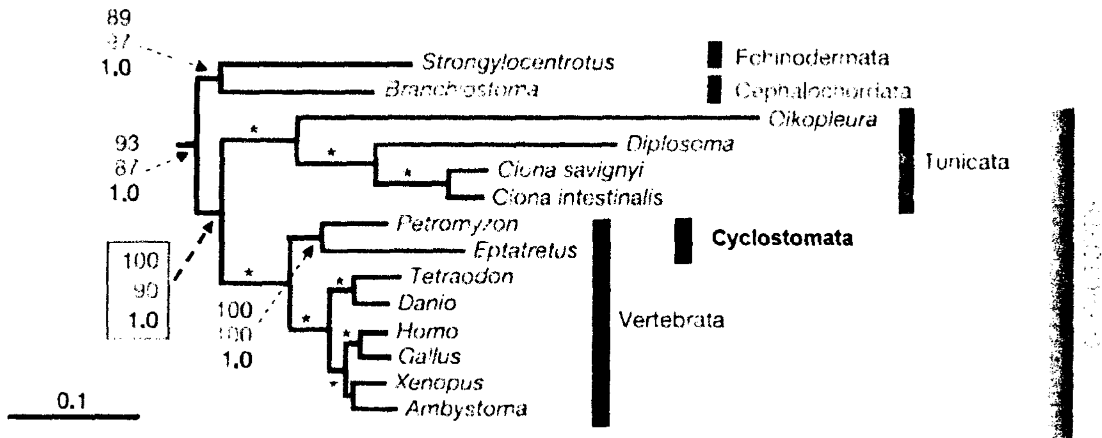


Figure 2: Molecular Phylogeny of the Deuterostomes. A recent phylogeny shows high support for the monophyletic grouping of lamprey (*Petromyzon*) and hagfishes (*Eptatretus*) into the Cyclostomata. (Figure modified from Delsuc et al 2006)

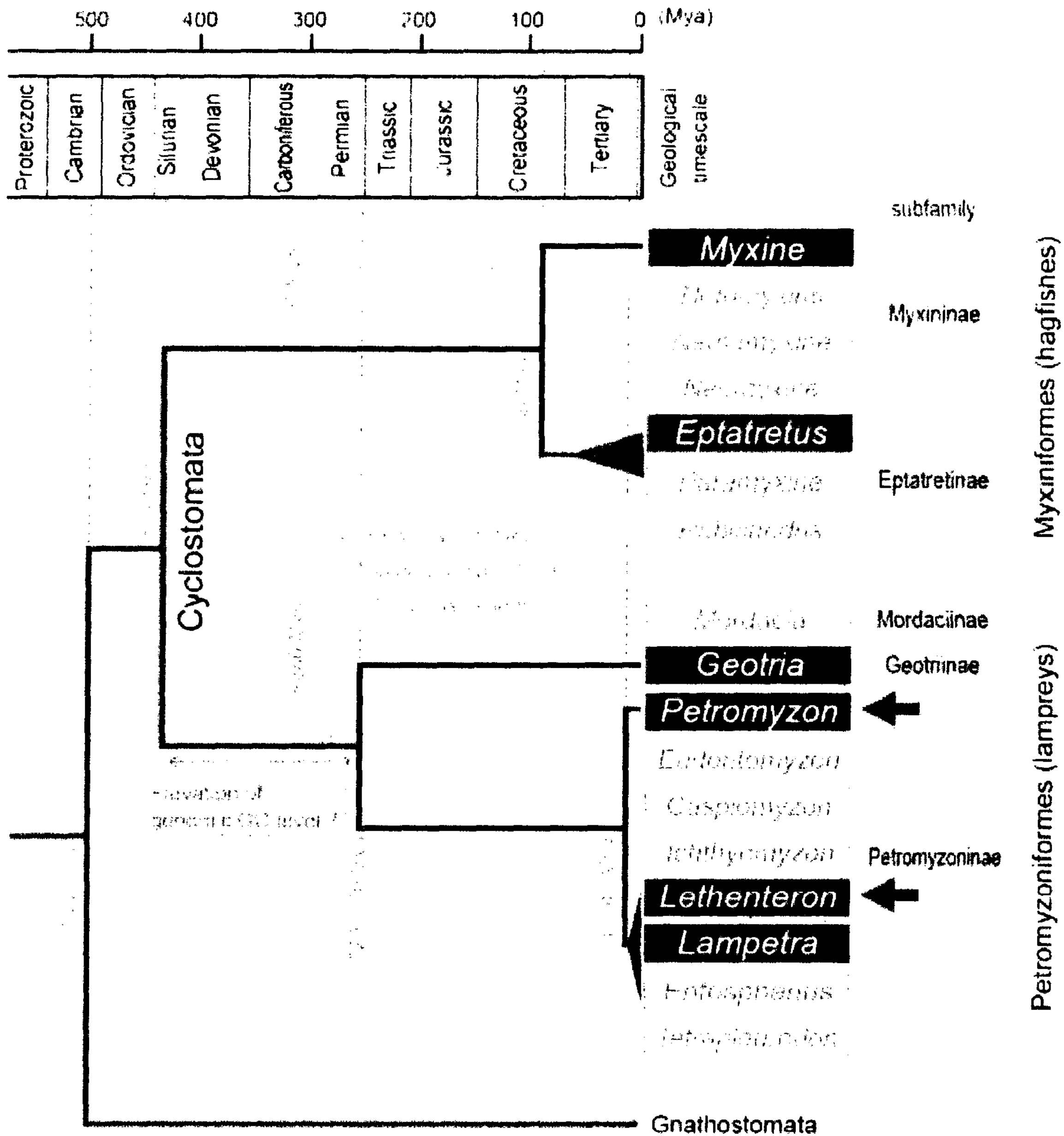


Figure 3: Time Scale for the Divergence of Cyclostomes and Gnathostomes. The divergence time estimated between hagfishes and lampreys is estimated at between 430-480 Mya while it is proposed that Gnathostomes and Cyclostomes diverged 500Mya. The short period of time between these two divergence points may complicate phylogenetic reconstruction and account for the massive morphological differences between lampreys and hagfishes even if they are monophyletic. (Figure from Kuraku and Kuratani 2006)

1.2.2 Cyclostomes: An Outgroup to Gnathostome Vertebrates

Craniates consist of two extant phyla, the Cyclostomes and the Gnathostomes. The Cyclostomes consist of only two living orders, the Petromyzontiformes (lampreys) and the Hyperotreti (hagfishes), while Gnathostomes are comprised of all other vertebrates. Cyclostomes lie in an invaluable position in animal phylogeny because they serve as the most closely related outgroup to Gnathostome vertebrates, and therefore serve as a reference organism with which the genomes, development, and morphology of Gnathostomes can be compared in order to determine the origins and functions of various Gnathostome specific features. Several major morphological differences separate Cyclostomes from Gnathostomes, including the presence of jaws, paired appendages, bone, and true teeth (Neidert et al 2001, Mallat 1996, Botella et al 2007, Kuratani et al. 2002). Interestingly, as described in Figure 4, the precise spatio-temporal expression of a particular family of developmental genes, the *Dlx* family of homeodomain containing transcription factors, is required for the normal development of each of these structures making them prime candidates for further investigation into their role in the evolution of these features (Neidert et al. 2001).

These evolutionary innovations are thought to have had a significant impact on the success of Gnathostomes and their massive speciation and radiation across the globe. The evolution of jaws, for example, may have allowed Gnathostomes to pass from simple filter feeding or scavenging organisms to active predation, and opened up new ecological niches into which Gnathostomes readily radiated. In order to determine how these Gnathostome specific features arose, a viable strategy is to compare the molecular basis for the development of these structures in Gnathostomes, with the basal condition in

Cyclostomes. Developmental comparisons between Cyclostomes and Gnathostomes have been traditionally confined to using the lamprey due to the inability to obtain hagfish embryos (Reviewed in Ota and Kuratani 2006). Despite this limitation, developmental studies using the lamprey have yielded insight into the origin of several Gnathostome features including brain organization (Murakami et al. 2005, Osorio et al. 2005), neural crest cells (Sauka-Spengler et al. 2007, McCauley and Bronner-Fraser 2003), and jaw apparatus (Shigetani et al. 2002,). Due to the interest in the lamprey as an evolutionary outgroup to the Gnathostomes and its availability as an experimental model of development, a whole genome sequencing project has been undertaken to sequence the genome of the sea lamprey, *Petromyzon marinus* (http://genome.wustl.edu/genome.cgi?GENOME=Petromyzon%20marinus). While the sequencing of the *Petromyzon* genome has been completed to 5.9X coverage, the process of assembling large amounts of raw sequence data has been complicated by the presence of large numbers of previously uncharacterized repetitive elements in the lamprey genome, resulting in a relatively poor quality assembly of the lamprey genome (Chris Amemiya, personal communication and Ensembl Genome Browser http://pre.ensembl.org/Petromyzon_marinus/index.html). Nevertheless, even this partial information is of tremendous use to comparative genomics projects searching for the most conserved features of chordate the prediction of novel functional elements in the genome of Gnathostomes, including humans, as well as to evolutionary studies searching to identifying key changes which have contributed to the diversification and radiation of Gnathostome vertebrates.

While large scale comparative genomics is a crucial tool in understanding and organizing the huge amount of information present in animal genomes and narrowing the field of view towards the most promising targets for further investigation, more precise information about the role of specific genomic elements necessitates a case-by-case assessment. Towards this end, I have selected to investigate a family of developmentally important transcription factors in the genome of Cyclostomes, the *Dlx* gene family. This family of transcription factors is known to be important in the normal embryogenesis of a wide array of Vertebrate morphologies, including many novelties which separate Cyclostomes from Gnathostomes (Neidert et al. 2001). The broad application of *Dlx* genes in the development of a several aspects of the Vertebrate body plan can be partially attributed to functional diversification and specialization that accompanied the expansion of this gene family from a single member in invertebrates to multiple members in vertebrates (Sumiyama et al. 2003, Stock et al. 1996, Zerucha and Ekker 2000). In fact, the expansion of a large number of gene families, including many developmental genes, accompanied the radiation of Vertebrates and is believed to have had significant influence on the change in complexity of the body plan along this lineage (Kasahara 2007, Panopoulou and Poustka 2005, Wagner et al. 2003, Furlong and Holland 2002).

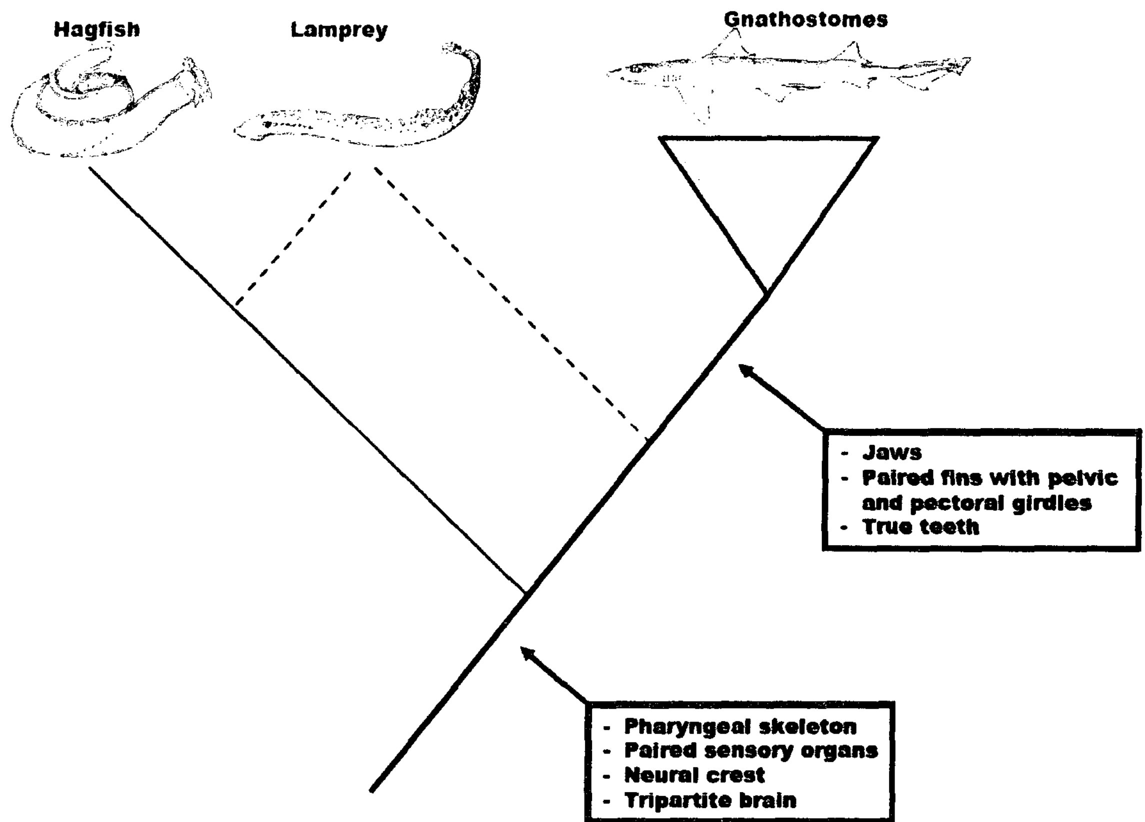


Figure 4: Morphological Novelties in the Chordate Lineage are Associated with the Expression of *Dlx* Genes. Many new morphological features which are considered key components of the Vertebrate body plan evolved just prior to, or subsequent to, the divergence of Cyclostomes. Each of these novelties listed here is associated with the expression of members of the *Dlx* family of homeodomain containing transcription factors.

1.3 Gene and Genome Duplication

1.3.1 The Fates of Duplicated Genes

Gene duplication is the major mechanism through which new genes are generated and contribute significantly to the diversification of the genetic network of an organism. Following the duplication of any given gene several fates are possible (reviewed in Li et al. 2005, Prince and Pickett 2002). The first and most likely fate is non-functionalization, or gene loss due to redundancy. Mutation accumulation will occur and one of the copies will be pseudogenized or deleted from genome through the action of purifying selection. A second possible fate of duplicate genes is sub-functionalization whereby each duplicate takes on partial responsibility of the original gene. In this case while the genes share some redundant functions, by splitting the role of the parent gene which is often highly pleiotropic, some constraints are alleviated and each gene is allowed to change in sequence more readily than when their functions were combined in the single ancestral gene. Finally, neo-functionalization occurs when one of the gene duplicates accumulates mutations which cause it to take on a new role in the genome. The mechanisms through which sub-functionalization and neo-functionalization take place may occur through changes in either their protein coding domains, or their cis-regulatory elements which control the spatiotemporal domain of their expression. One popular model of sub-functionalization permitting for differences in deployment of duplicated gene family members is the so called Duplication-Degeneration-Complementation, or DDC model (Reviewed in Force et al. 1999). This model stipulates that following gene duplication complementary degenerate changes in the regulatory elements which control the expression of each duplicate allow for duplicated sister genes to each assume a portion of

the expression domain of their original parent gene. Changes in the regulation of duplicated genes is an attractive hypothesis for explaining the ability of duplicates to undergo significant changes in their developmental roles since the risk of deleterious or dominant negative effects of protein coding changes is high and will necessarily affect all of the tissues and times during development during which that protein is expressed.

1.3.2 Genome Duplication

It was first proposed by Ohno that two successive large scale events of gene duplication occurred in the chordate lineage, possibly by whole genome duplication (Ohno 1970). While his evidence was based only on genome size and isozyme complexity, methods which by today's standards are weak, Ohno was remarkably accurate in his prediction and modern evidence largely supports his hypothesis. Today, while the relative timing of the events differs greatly from Ohno's prediction, it is widely accepted that two rounds of whole genome duplication took place early in the chordate lineage. The so-called "2R-Hypothesis" has been modified as more information is incorporated from a number of sources including whole genome data only recently made available from a number of diverse chordates (see Kasahara 2007 or Panopoulou and Poustka 2005 for a useful review). Several authors also correlate these major genome duplication events with opportunity for the evolution of Vertebrate morphological novelties (Shimeld and Holland 2000, Wagner et al. 2003), and while direct cause and effect is extremely difficult to demonstrate, theoretical evidence predicts that with the simple increase in the number of genes, the number of possible genetic interactions increases, opening up the possibility to develop new and complex anatomical features. In addition to the 2 rounds of whole genome duplication which are accepted to have occurred somewhere near the stem of the chordate lineage, there is clear evidence that a 3rd round of whole genome duplication took place at the base of the teleost lineage, leading to the so called "3R-hypothesis". In this case, diversification and speciation of the teleosts has been partially attributed to this 3rd genome duplication event, sometimes

referred to as the FSGD (Fish Specific Genome Duplication) (Reviewed in Meyer and Van de Peer 2005).

The evidence for whole genome duplication can be broken down into four main schools (Reviewed in Furlong and Holland 2002). The first and perhaps weakest evidence is derived from the analysis of gene number which reveals that while most invertebrates have no more than 20,000 genes vertebrates have considerably more, for example the most recent estimates, depending on the prediction tools used, count the number of human genes at approximately 35,000. Secondly by examining specific gene families it is evident that single genes from the Cephalochordate Amphioxus or other invertebrate genomes have multiple family members in vertebrates and often show a 1:2, 1:3, or 1:4 relationship (invertebrates:vertebrates). The third school of evidence is derived from the phylogenetic analysis of these gene families which often show a clear phylogenetic signal indicating how multiple vertebrate genes are related to one another and are derived from an ancestral invertebrate gene. Finally, large scale analysis of syntenic regions in various vertebrate genomes and the analysis closely maintained associations between multiple family members which form gene clusters add a physical dimension to the evidence for whole genome duplications in the vertebrate lineage. This final piece of evidence will be elaborated upon further as it is the basis upon which this thesis was constructed. Furthermore, because the so called Fish Specific Genome Duplication (FSGD) or "3R" event took place approximately 350 Mya, it has been possible to study what happens to gene duplicates on a more recent scale, leading to the development of multiple theories on the precise patterns and processes of gene

duplication which can be applied to the more ancient duplications which took place over 500 Mya and which are more difficult to resolve.

While the comparisons between Gnathostome and invertebrate genomes has made it evident that 2 rounds of whole genome duplication took place in the lineage leading to vertebrates, the precise timing and of the occurrence of these massive genomic events, remains unresolved. Significantly, while it is clear that basal chordates such as Cephalochordates and Ascidians have not experienced these events, and that Chondrichthyans, the most basal extant Gnathostomes have, this places the evolutionary occurrence of these large-scale events near the time of divergence of Cyclostomes from the Vertebrate lineage. To date, using only the limited data available to them authors have variously claimed that Cyclostome have undergone both, only one, or no shared rounds of whole genome duplication with the Gnathostomes (Reviewed in Escriva et al. 2002).

1.3.3 The Timing of Genome Duplication Events

Despite the general consensus that whole genome duplication events occurred and have clearly played an important role in the diversification of the vertebrate genetic landscape, and likely were involved in morphological diversification as well, the precise timing and mechanisms of these events are unknown. While Ohno first proposed that the first round of duplication occurred before the divergence of Cephalochordates from the Vertebrate lineage we know from several studies, including the recent whole genome sequencing of the model Cephalochordate *Branchiostoma floridae*, that Cephalochordates did not experience any of these whole genome duplication events (Putnam et al. 2008). It

has been quite evident for some time, from the first whole genome analyses of *Homo sapiens* and *Mus musculus* that tetrapods have experienced two rounds of whole genome duplication. Studies of specific gene families in a variety of Chondrichthyans, in combination with survey sequencing of a Chimaerid, the Elephant Shark *Callorhinchus milii,* strongly supports the hypothesis that both of these events occurred prior to the divergence of Chondrichthyans from other Vertebrates (Venkatesh et al. 2005, Venkatesh et al. 2007). Therefore, the window into which the occurrence of these major genomic restructuring events occurred has been narrowed is to a very specific group of enigmatic organisms: the Cyclostomes. A schematic representing the phylogeny of chordates and the proposed timing of whole genome duplication events is depicted in Figure 5.

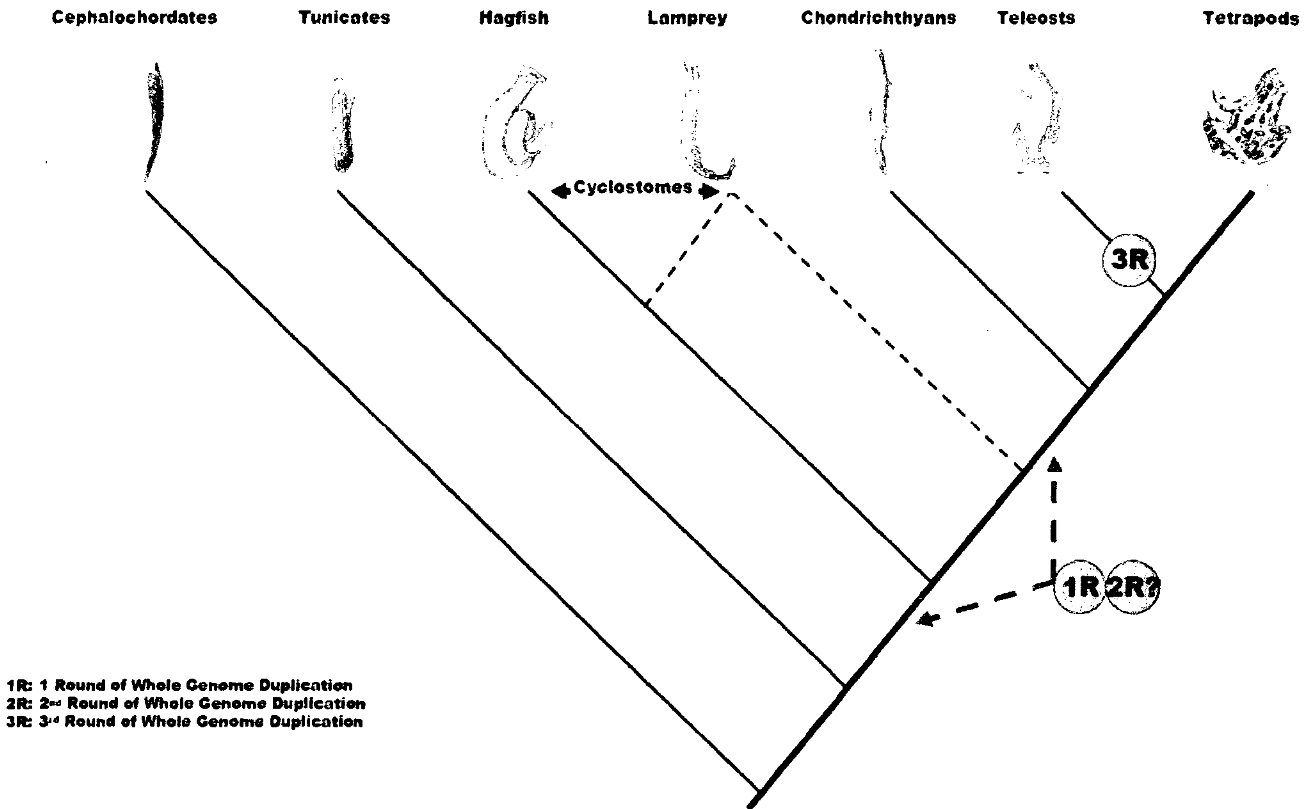


Figure 5: The Timing of Whole Genome Duplication Events in the Chordate Lineage. While abundant data exists to demonstrate that neither Cephalochordates nor Tunicates have undergone Whole Genome Duplication, while Chondrichthyans and all other Gnathostomes have experienced at least 2 rounds of whole genome duplication, it is still unknown whether these events occurred before or after the divergence of Cyclostomes.

1.3.4 Gene Clusters

Evidence for whole genome duplication can perhaps be most easily seen by examining families of physically clustered genes. These cases yield the most potent evidence because of the ability to combine phylogenetic analysis with physical position in the genome. This is important especially when examining ancient duplicated genes which have accumulated many mutations, complicating even the most advanced methods of phylogenetic reconstruction. Perhaps the most famous and most fruitful analysis of gene duplication in the animal lineage has been the examination of the ANTP superfamily of homeobox containing genes, including *Hox*, *Nkx*, and *Dlx* gene clusters (see Garcia-Fernàndez 2005 for a comprehensive review). *Hox* genes encode transcription factors important for embryonic development and have a role in the early patterning of the whole anterior-posterior axis of both invertebrate and vertebrate embryos (reviewed in Wellik 2007), including the lamprey (Takio et al. 2007). To date there have been 14 paralogy group members of *Hox* genes identified and orthologs to most, but not all, members of each paralogy group genes have been found across Animalia. Importantly, for reasons which are still debated but are likely due to the necessity for tight spatio-temporal regulation of gene expression during development, the multiple paralogous *Hox* genes in vertebrates have maintained a close linkage with one another following their initial series of tandem duplications early in animal evolution, pre-dating the protostome-deuterostome split (Lee et al. 2006). *Hox* genes therefore tend to form a coherent cluster in Vertebrate genomes, and their physical position with respect to one another within this cluster has also been maintained. While invertebrates and Cephalochordates have a single *Hox* cluster, all non-teleost Vertebrates that have been

examined have 4. This type of signal derived from physical linkage data is the strongest evidence for large scale genome duplication events as the maintenance of the physical and phylogenetic relationship of the genes eliminates the possibility that the overall increase in gene number has occurred solely by independent gene duplications. Despite the usefulness of *Hox* clusters to the resolution of whole genome duplications the relatively large number of genes and large cluster size makes this a daunting task. There is evidence based on gene number and phylogeny for between 3 and 4 *Hox* clusters in lamprey (Force et al. 2002, Irvine et al 2002), and in hagfish a PCR based survey has indicated that as many as 8 members of one *Hox* paralogy group exist opening the door to the possibility of multiple independent tandem duplications in this lineage, complicating the analysis (Stadler et al 2004). The largest *Hox* cluster fragment assembled from overlapping cosmids from the lamprey *P. marinus* consists of only 5 genes (*Hox4w*, *Hox5x*, *Hox6w*, *Hox8Q*, and *Hox9y*) and was named the *Pethox* (*Pet*romyzon *hox*) cluster, but other isolated *Hox* genes could not be assembled into larger clusters (Force et al. 2002). Primarily due to increased intron size, and space between *Hox* genes, this lamprey *Hox* complex is slightly larger than the *Hox* complement of mammals, previously the largest examined in vertebrates. The inability to isolate larger *Hox* complements from lamprey may stem from the highly repetitive nature of the genome and difficulty in assembling larger contiguous sequences. Alternatively, it may be possible that the *Hox* complexes of the lamprey have expanded, and/or become fragmented and scattered throughout the genome. This is the case of the *Hox* complement of Ascidians such as *Ciona intestinalis* (Ikuta et al. 2004), and in protostomes such as Drosophilid flies where the *Hox* complement is broken in at least 7

separate ways in different species (Negre and Ruiz 2007). It seems evident that lampreys have multiple copies of each *Hox* paralogy group and therefore if this second hypothesis is true, lamprey would be the first organism with multiple *Hox* complements that have been secondarily broken in the genome. By examining the *Hox* complements of a variety of deuterostomes in addition to vertebrates (Duboule 2007) has recently suggested that the ancestral *Hox* complex was large and relatively disorganized, albeit coherent unit, and that the tight linkage is a secondarily derived condition of the vertebrates. By examining the more primitive chordates such as *Ciona* and Amphioxus we see that the *Hox* complements of these organisms are either atomized or broad and disorganized, respectively (Ikuta et al. 2004, Amemiya et al. 2008, Garcia-Fernandez and Holland 1994). If the theory of Duboule is correct and the highly organized, tight nature of vertebrate *Hox* complexes is secondarily derived, it may be possible that whatever changes occurred in the vertebrate lineage to secondarily constrain *Hox* complexes did not occur in the Cyclostome lineage despite their possessing multiple complexes and resulted in their expansion and fragmentation. Further analysis of Cyclostome *Hox* complements will be necessary to make this assertion. Unfortunately, no evidence for the linkage of any *Hox* genes has been found in the hagfish to support the idea that at least micro-clusters of *Hox* genes may exist.

Dlx genes are in a unique position to help resolve the history of gene duplications at the stem of the chordate lineage. *Dlx* genes show a simple bigenic clustered arrangement in the genome of all Gnathostomes examined: three pairs of genes present on different chromosomes and, importantly, each pair is linked to an individual *Hox* complex and shared the same events of whole genome duplication (See Figure 6 for a

schematic of the joint duplication history of *Dlx* clusters and *Hox* complexes). Because this gene family is relatively small (1 gene in invertebrates and 6 genes in Gnathostomes), shows a characteristic clustered arrangement, and because they are ancestrally linked to *Hox* complexes, *Dlx* genes represent a key gene family for preliminary investigation into the organization of the Cyclostome genome.

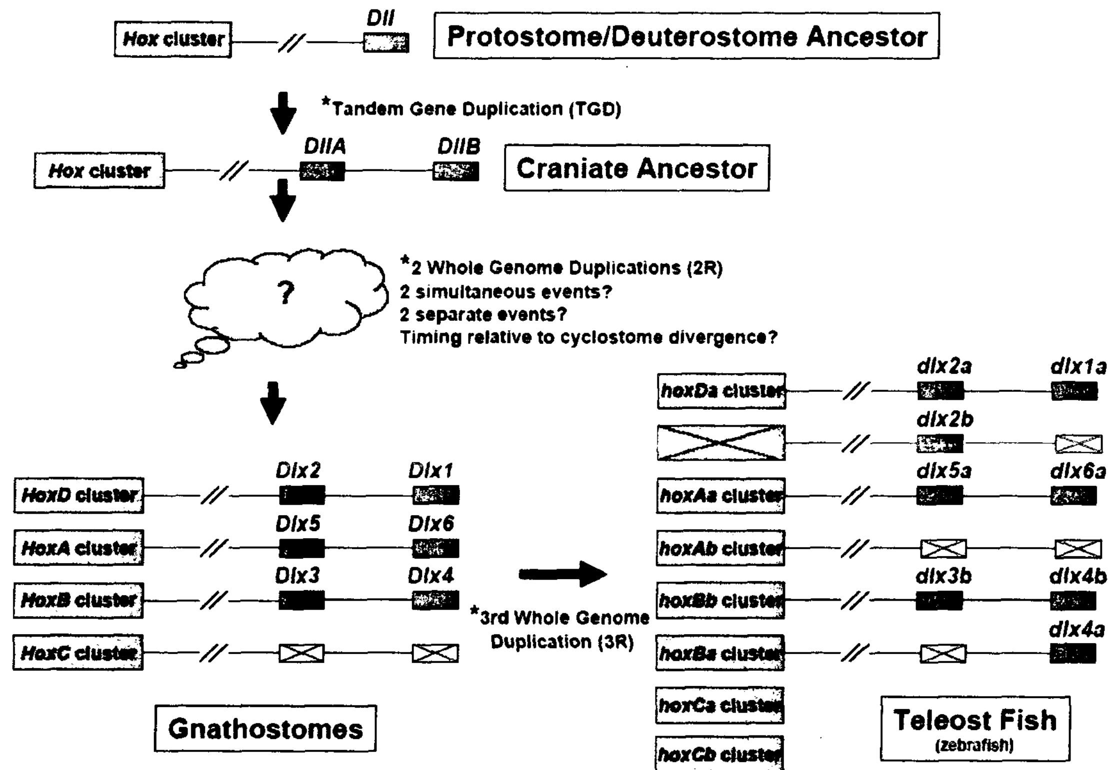


Figure 6: Duplication of the *Dll/Dlx* Gene Clusters and the *Hox* Complexes Can Be Used to Resolve the Timing of Ancient Whole Genome Duplications. An ancestral *Dll* gene duplicated in tandem, and following several whole genome duplications diversified to form the multigene *Dlx* family of transcription factors. The gap in our understanding of the duplication history of *Hox* Complexes and linked *Dlx* bigene clusters can be filled by examining the condition of these gene families in Cyclostomes.

1.4 The *Dlx* Gene Family

1.4.1 Introduction to the *Dll/Dlx* Gene Family

The *distal-less* gene, *Dll*, is a homeobox containing transcription factor gene and was originally cloned in *Drosophila* (Cohen, Bronner et al 1989). Homeobox genes encode transcription factors containing a highly conserved DNA binding domain, known as the homeodomain. Recently, evolutionary changes in the number, cis-regulation, function, and expression pattern of various homeobox genes have been shown to have a significant impact on the evolution of the animal body plan (reviewed in Gellon and McGinnis 1998, Wray 2003, and many more). In *Drosophila*, *Dll* has an indispensable role in the development of distal appendages including the limbs, antennae, and mouthparts (Reviewed in Panganiban and Rubenstein 2002). Furthermore, *Dll* is expressed in various components of the peripheral and central nervous systems of the fly (Panganiban 2002). As a result of a complex history of gene and genome duplication events vertebrates, including humans, have multiple co-orthologs of the invertebrate *Dll* gene, collectively named the *Distal-less-related homeobox* gene family or *Dlx*. All vertebrates examined have been shown to possess at least 6 *Dlx* genes, termed *Dlx*1 through *Dlx*6, organized in 3 convergently transcribed bigene clusters, each ancestrally linked to the *Hox* clusters in the genome. The *Dlx*1/*Dlx*2 cluster is linked to the *Hox*D cluster, the *Dlx*3/*Dlx*4 cluster is linked to the *HoxB* cluster, and the *Dlx* 5/*Dlx*6 cluster is linked to the *Hox*A cluster (Banerjee-Basu and Baxevanis 2001), also shown in Figure 6. It is hypothesized that following a tandem gene duplication of an ancestral *Dll* gene, forming the original bigene cluster, two whole rounds of genome duplication took place which resulted in the present complement of 4 *Hox* gene clusters in vertebrates and also

resulted in the duplication of the ancestral *Dlx* cluster (Sumiyama et al. 2003, Zerucha and Ekker 2000). While it is presumed that originally there was a 4th *Dlx* gene cluster linked to the *Hox*C complex this cluster seems to have been universally lost in all vertebrates examined, as is often the case of duplicated genes. As discussed above, the timing of these whole genome duplication events which resulted in the expansion of the *Dlx* gene family is nearly coincident with the divergence of Cyclostomes. The cephalochordate *Branchiostoma floridae* (also known as Amphioxus), recently demonstrated to be the most primitive chordate model discovered to date (confirmed by recent analyses of Delsuc et al. 2006 and Putnam et al. 2008), has been shown to possess a single *Dll* gene (Holland et al. 1996), while in the ascidian *Ciona intestinalis* there are 3 *Dll* genes in the genome, including a single tail to tail bigene cluster comprised of *DllA* and *DllB*, within less than 3Mb of a remainder of the atomized C. *intestinalis Hox* complex comprised of *Hox13* and *Hox12* orthologs, in addition to an isolated gene, *DllC* which is not linked to any *Hox* genes (Irvine et al. 2007, Caracciolo et al. 2000). Therefore, it appears that prior to the divergence of Ascidians that the ancestral *Dll* gene which was linked to the single *Hox* complex duplicated tandemly to form the first gene cluster. The presence of the third *Dll* gene, *DllC*, is most easily explained by implicating an additional independent tandem duplication since it is linked neither to any other *Dll* nor any *Hox* like genes in the C. *intestinalis* genome (Irvine et al. 2007).

The only studies of *Dll/Dlx* family members in the Cyclostomes have focused on the lamprey. The sea lamprey, *Petromyzon marinus*, have been shown to have at least 4 *Dlx* genes, *DlxA*, *DlxB*, *DlxC*, and *DlxD* (Neidert et al 2000), and unpublished results (Shigehiro Kuraku, personal communication) have found two additional *Dlx* paralogs,

termed *DlxE*, and *DlxF*, in addition to orthologs of the 4 previous genes, in the Japanese lamprey *Lethenteron japonicum*. While both lamprey and Gnathostomes have been shown to have 6 *Dlx* genes, it is still unclear whether they are orthologous or individual duplicates, an issue which must be resolved before any inferences about the possibility of shared genome duplications can be made. Unfortunately, lamprey *Dlx* genes do not form strict 1:1 orthologies with Gnathostome genes when phylogenetic analyses are undertaken (Neidert et al. 2001, Myojin et al. 2001). However, they can be grouped into the two major *Dlx* gene families. The *DlxA/B/C* genes group with the *Dlx2/3/5* family, while *DlxD/E/F* genes group with the *Dlx1/4/6* family (Neidert et al 2000, Shigehiro Kuraku personal communication, this study). The difficulty in resolving the phylogeny of lamprey *Dlx* genes complicates the effort to determine whether they represent true orthologs to the archetypal 6 Gnathostome *Dlx* genes or that they are independent duplicates as seems to be the case for *DllC* in C. *intestinalis*. In order to make such an assertion it will be necessary to examine whether or not lamprey *Dlx* genes are in fact linked in tail to tail bigene clusters in a manner homologous to vertebrates. No evidence of lamprey *Dlx* gene clusters has previously been found, but work in our lab has demonstrated that *DlxC* and *DlxD* are linked in a tail to tail fashion, as with other vertebrate *Dlx* genes (Maurya 2006). No studies have reported the identification of *Dlx* genes in the hagfish, a striking omission when considering its key position in animal phylogeny as the only other living members of the Cyclostomata. Investigation into the *Dlx* gene complement and organization in the hagfish will be necessary to determine whether the principles derived from investigation of the lamprey genome hold true for all

Cyclostomes or whether hagfish and lamprey are as different on a genomic level as they appear morphologically.

The presence of multiple members of the *Dlx* gene family in the vertebrate genome has significant functional implications for vertebrate development. It has been demonstrated that each gene forming a bigene cluster shares many expression domains and functional properties. For example the *Dlx*1 and *Dlx*2 genes, which form a cluster linked to the *Hox*D complex, share a nearly completely overlapping expression domain in the developing branchial arches of the mouse embryo at E10.5 (Depew et al. 2002). As is the case with many multi-gene families, certain roles of the genes seem to be shared allowing for a limited degree of functional redundancy between paralogs. However, clear differences in their expression patterns during development, regulation by upstream factors, and their downstream target genes are evident. The following section will highlight some of the major aspects of *Dll/Dlx* gene expression, regulation, and function across the chordates.

1.4.2 Cephalochordate and Ascidian *Dll* Expression

A single distal-less gene (Amphi*Dll*) has been isolated from the cephalochordate *Branchiostoma floridae*, also known as Amphioxus (Holland et al. 1996). This gene appears to have roles in the global regionalization of ectoderm, a part in the establishment of the dorsoventral axis, the specification of migrating epidermal cells, and the development of the forebrain. Because the role in nervous system development is shared with invertebrate *Dll* it has been proposed that the ancestral function of *Dll* was in nervous system development (Holland et al. 1996). While initially expressed throughout

the ectoderm during blastula stage, *AmphiDll* becomes down regulated in the presumptive neural plate during neurulation. Interestingly, *AmphiDll* is not expressed in mesodermal tissues, while vertebrate *Dlx* genes are, indicating that this role was acquired later in evolution lending support to the hypothesis that the original function of this gene was in specifying ectodermal derivatives, especially the nervous system. Arthropods also have only a single *Dll* gene and this is believed to be the ancestral condition of both the protostomes and deuterostomes. However, besides the neural tissue the ectodermal expression arthropod *Dll* is limited to the developing limbs, indicating that the role of this gene in the global specification of the ectoderm may be specific to the chordate lineage. Ascidians have been shown to be the closest living non-craniate chordate relative of modern vertebrates (Delsuc et al. 2006). The ascidian, *Ciona intestinalis* has been found to have 3 *Dll* genes termed *DllA*, *DllB*, and *DllC*. Interestingly, *DllA* and *DllB* form a tail-to-tail bigene cluster with an intergenic region spanning approximately 1.5kb, and linked over approximately 2.75Mb to C. *intestinalis Hox13* and *Hox12* orthologs, remnants of the decomposed *Hox* cluster found in this organism, while *DllC* is not flanked by any other *Dll* or *Hox* genes (Irvine et al. 2007, Caracciolo et al. 2000). This lends support to the theory that an ancestral *Dll* gene first duplicated in tandem to produce a single *Dll* gene cluster before undergoing a series of cluster or whole genome duplications during the evolution of vertebrates. By the time that *Ciona intestinalis* had diverged from the rest of the lineage of chordates, the *Dll* cluster had formed, but the whole genome duplications had not yet taken place. As in the case of *AmphiDll*, expression of *Ciona DllB* can be detected nearly globally in ectodermal lineages throughout gastrulation and becomes down regulated at the time of neurulation.

However, the expression patterns of *CiDllA* and *CiDllC* differ dramatically from this pattern early pan-ectodermal expression pattern. Only during tailbud stages does *CiDllA* begin to be expressed in the anterior ectoderm in the primordial atrial siphons, putative homologs of vertebrate ectodermal placodes, embryonic structures which express also express *Dlx* genes (Irvine et al 2007, Solomon and Fritz 2002). Expression of *CiDllC* is even more limited and has been shown to be expressed only during late hatched tailbud stage larvae in the adhesive palps at the anterior of the organism which are used to attach to solid substrates during metamorphosis. Therefore it appears that a single *Dll* ortholog in *Ciona intestinalis*, *CiDllB* has maintained the majority of the roles in global specification of the ectoderm of the single *AmphiDll* gene while the tandemly duplicated *CiDll*A and *CiDll*C have acquired specific roles in the specification of anterior ectodermal derivatives in later stage embryos. The sharing of developmental roles by duplicated *Dll/Dlx* family members and the acquisition of independent roles for duplicated genes is a theme that is also maintained in the more heavily duplicated *Dlx* genes of vertebrates.

1.4.3 Vertebrate *Dlx* Gene Expression

All vertebrates examined have been shown to possess at least 6 *Dlx* genes organized in 3 convergently transcribed bigene clusters each linked to a *Hox* cluster. During Gnathostome embryogenesis, *Dlx* genes are predominantly expressed in ectodermal derivatives such as the forebrain, where *Dlx*1, *Dlx*2, *Dlx*5, and *Dlx*6 are expressed in both a telencephalic and diencephalic domain (Liu et al. 1997), the otic placode, retina, apical ectodermal ridge of limb buds (Robledo et al. 2002) in which all

Dlx genes are expressed at some stage during development (Reviewed in Panganiban and Rubeinstein 2002), and ectomesenchymal cells of the branchial arches, where a complex spatiotemporal expression pattern of *Dlx* genes has been proposed to form a "*Dlx*-code" involved in the specification of the proxomodistal axis of the arches and results in the normal patterning of the Gnathostome jaw apparatus (Depew 2002, Depew 2005). Later in development, *Dlx* expression is found in developing skeletal tissues, teeth, and hematopoeitic cells. As a result of the expansion of the *Dlx* gene family, individual paralogs have both unique and overlapping functions depending on the time and tissue of expression. Interestingly, the paralogy relationships between *Dlx* family members seems to play an important role in this respect. In striking contrast to the case of Ascidians where the linked *DllA* and *DllB* do not share any expression domains, in vertebrates linked *Dlx* genes show highly overlapping expression patterns. While trans-paralogs (for example *Dlx*2 and *Dlx*3) are more closely related to each other in sequence than cis-paralogs (such as *Dlx1* and *Dlx2*), it is the cis paralogs which show the most significant overlap in expression pattern and likely also developmental role. For example, the expression of the physically linked cis paralogs, *Dlx3a* and *Dlx4a* are almost indistinguishable throughout zebrafish development (Ellies, Stock et al. 1997). There are several other examples of shared expression domains of linked *Dlx* genes, including nearly complete overlap and coordinate rise and fall in expression levels of *Dlx5* and *Dlx6* during murine osteoblast differentiation (Samee et al. 2007, Li et al. 2008). The relationship between physical linkage of vertebrate *Dlx* genes and their expression pattern is perhaps most evident in the developing branchial arches, the embryonic precursors of jaw, among other organs. The involvement of the *Dlx* genes in the development of the

branchial arches and patterning of the vertebrate jaw apparatus is of significant interest considering that the origin of this structure is considered to be a key novelty in the radiation of Gnathostomes, permitting the transition from passive filter feeding to active predatory lifestyle. Branchial arch 1 is the embryonic precursor of the vertebrate jaw and is divided into two components: a dorsal maxillary component and a ventral mandibular component (see Figure 7). Migrating cranial neural crest cells (CNCs) populate the branchial arches and contribute to the development of a significant portion of the craniofacial skeleton (reviewed in Cobourne and Sharpe 2003, Knecht and Bronner Fraser 2002). As mentioned above, the involvement of *Dlx* genes in the development of the craniofacial skeleton and their precise expression pattern within the branchial arches has led to the hypothesis that a "*Dlx*-code" exists which, depending upon the specific combination and/or the dose of different *Dlx* genes expressed in different compartments of the branchial arches, different cartilaginous and later ossified skeletal elements will form, patterning the complex structure of the vertebrate jaw (Depew 2002, Depew 2005). Within the ectomesenchymal cells of the branchial arches, the 6 vertebrate *Dlx* genes are expressed in a partially overlapping nested pattern along the dorsoventral axis. In the mouse E10.5, or embryonic day 10.5, is the time when the "*Dlx*-code" is most evident and this expression pattern appears to be transitory (Depew 2002). Figure 8 shows the expression of several murine *Dlx* genes in the branchial arches of the E10.5 stage mouse embryo. The physically linked genes *Dlx1* and *Dlx2* are expressed throughout the entire dorsoventral axis of the arches, and expands from the ventral most mandibular component to the dorsal most maxillary component of arch 1. *Dlx5* and *Dlx6* expression is confined to only the ventral mandibular component of arch 1 where it overlaps with

with *Dlx1* and *Dlx2* expression in its entire domain. *Dlx3* and *Dlx4* expression in the branchial arches is confined to an even more limited aspect of the ventral mandibular arch where it overlaps with both *Dlx1/2* and *Dlx5/6* over its entire expression domain (Depew et al 2005). The "*Dlx*-Code" proposed to pattern the dorso-ventral axis of the branchial arches is thought to function in a manner similar to the "*Hox*-Code" which patterns the anterior-posterior axis. See Figure 9 for a schematic representation of the *Hox* and *Dlx* codes in Gnathostome embryos. A similar expression pattern of orthologous genes are also found in zebrafish branchial arches (reviewed in Walker et al 2006). Evidence for the role of the *Dlx* code in specifying the branchial arches is derived not only from expression patterns but also from gene knockout or knockdown studies. For example, *Dlx1* and *Dlx2* double mutant mice showed profound dysmorphologies in the craniofacial bones and complete lack of maxillary molars (Depew et al 2005). In an even more extreme demonstration of the essential nature of *Dlx* genes in patterning the craniofacial region, while *Dlx5* inactivation in mouse results in only minor craniofacial deformities, combined inactivation of *Dlx5* and *Dlx6* (shown in Figure 10), disrupting the entire medial component of the *Dlx* code, results in a loss of maxillary-mandibular asymmetry and a homeotic transformation of the mandibular component of branchial arch 1 into a maxillary component in mouse, as demonstrated by specific bone development patterns (Beverdam et al. 2002, Depew et al. 2002). This not only demonstrates an essential role for the *Dlx5*/*Dlx*6 gene cluster in the development of the craniofacial skeleton, but also highlights an important feature of *Dlx* genes; partial functional redundancy. Interestingly, the oral region of the lamprey, while highly specialized in other ways, is also nearly symmetric and there is very little difference between the

structure of its "maxillary" and "mandibular" components. It has therefore been suggested by some authors that the inactivation of *Dlx5* and *Dlx6* resulted in an atavistic condition causing the oral region of Gnathostomes to revert to a more amorphous, symmetric oral region reminiscent of that of their Cyclostome ancestors (Graham 2002). Strikingly, when the expression of 4 *Dlx* genes were examined in the sea lamprey, *Petromyzon marinus*, it was found that they did not form a nested pattern of expression in the branchial arches as in Gnathostomes, and were instead ubiquitously expressed throughout the proximodistal axis of each arch (Neidert et al. 2001), leading to the hypothesis that the acquisition of the *Dlx* code in Gnathostomes may have facilitated the diversification of structures along the proximodistal axis of the arches and contributed to the evolution of the complex jaw apparatus from a less specialized symmetrical state.

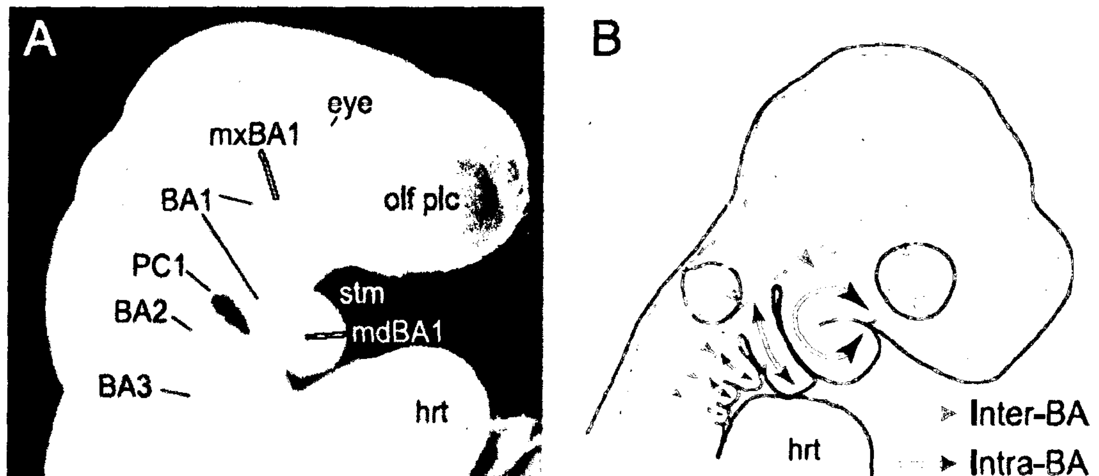


Figure 7: Branchial arches in the mouse embryo. (A) shows a scanning electron micrograph of an E10.5 day mouse embryo. BA1 = branchial arch 1, PC1 = pharyngeal cavity 1, m*Xba*1 = maxillary component of branchial arch 1, mdBA1 = mandibular component of branchial arch 1, hrt = heart, olf plc = olfactory placode. (B) shows the serially homologous nature of the branchial arches and illustrates that pattern formation in the mouse embryo requires specification both between different arches (Inter-BA), and within individual arches (Intra-BA). (Figure from Depew et al. 2005)

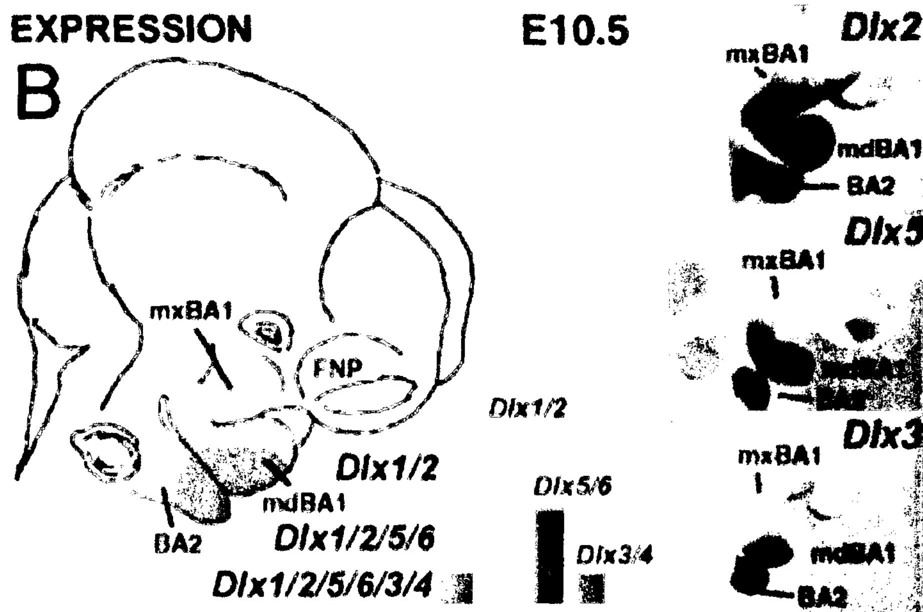


Figure 8: The *Dlx* code in the branchial arches of the mouse embryo. On the left, a schematic diagram depicting the nested nature *Dlx* gene expression in the branchial arches of an E10.5 mouse embryo is shown. On the right, in situ hybridization for representatives of each *Dlx* cluster clearly show the presence of this nested expression pattern. (Figure from Depew et al. 2002)

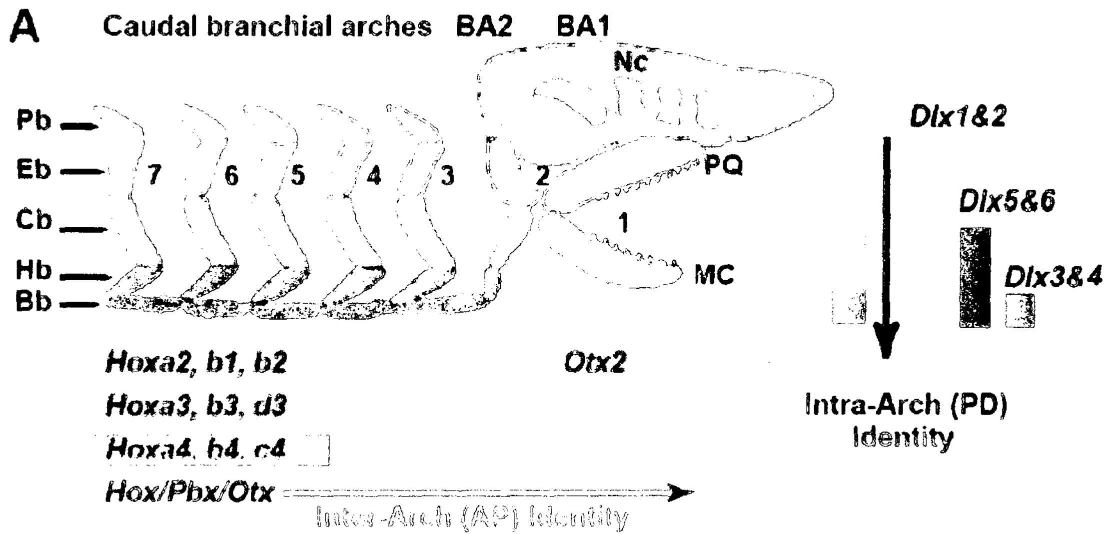


Figure 9: A *Dlx* Code patterns the Dorso-ventral axis of the branchial arches in Gnathostomes.
The expression pattern of 6 *Dlx* genes in the branchial arches helps determine the dorso-ventral identity of the gnathostome jaw apparatus. The expression of these genes must be precise or embryonic malformation may occur. (Figure from Depew et al 2002)

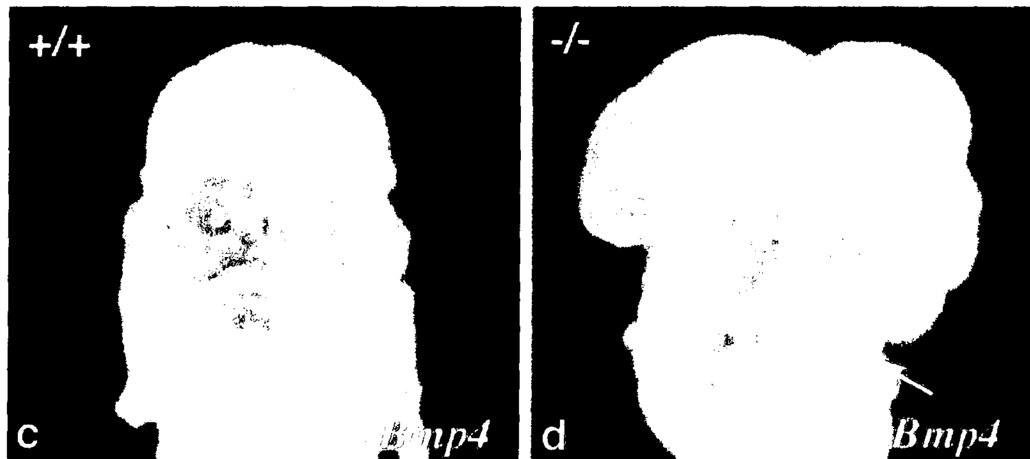


Figure 10: Loss of maxillary mandibular asymmetry in $Dlx5^{-/-}$ $Dlx6^{-/-}$ double mutant mice. *In situ* staining for *Bmp4* provides a superficial marker of the transformation, as maxillary papillae (dots stained for *Bmp4*) normally only present in the upper jaw in wild type mice (left) are present also on the lower jaw (right). (Figure from Beverdam et al 2002).

1.4.4 Cyclostome *Dlx* Gene Expression

Cyclostome *Dlx* gene expression has been examined only in the context of the lamprey. The most thorough study published describing lamprey *Dlx* gene expression used the sea lamprey, *Petromyzon marinus.* Like Gnathostomes, and invertebrate chordates including Amphioxus, lamprey *Dlx* genes tend to be expressed in ectodermal derivatives including the central nervous system, ectomesenchymal cells of the branchial arches, as well as dorsal and anal fin folds, despite the absence of paired appendages in these organisms. Expression of *Dlx* genes first becomes pronounced at 9 days following fertilization when *DlxA*, *DlxC*, and *DlxD* are expressed variously in the neural tube, premigratory and migratory neural crest cells, and ectoderm surrounding the mouth (Neidert et al. 2001). In later stages expression of *DlxA*, *DlxB*, *DlxC* and *DlxD* can be detected in the ectomesenchyme of the branchial arches in a broad uniform pattern which does not indicate any kind of nested code like system as has been found in Gnathostomes (Neidert et al. 2001). *DlxA*, *DlxC*, and *DlxD* are all expressed in the forebrain in domains consistent with the telencephalic and diencephalic expression domains of *Dlx* in Gnathostomes, although expression of *DlxD* is not detected in the telencephalic compartment of the forebrain (Neidert et al. 2001). Interestingly, *DlxB* expression is not detected in the forebrain of the lamprey embryo, as is found with Gnathostome *Dlx3* and *Dlx4* genes. In general, it was observed that *DlxA* and *DlxD* shared the most highly overlapping expression patterns, specifically in both the premigratory neural crest cells, and both telencephalic and diencephalic components of the forebrain, leading the authors to hypothesize that these genes may be physically linked in the lamprey genome. Recently, however, results from our lab conflict with this claim and show that in fact

lamprey *DlxC* and *DlxD* form a physically linked cluster (Maurya 2006). One of the most significant observations of lamprey *Dlx* gene expression is that, unlike the condition of Gnathostomes, within the branchial arches of the Stage 26 embryo *Dlx* genes do not appear to form a proximodistally nested expression pattern (seen in Figure 11). This is most easily emphasized seen by the expression of all 4 lamprey *Dlx* showing expression in the upper-lip region of the embryo, the proposed homolog of the maxillary component of Gnathostome branchial arch 1, while in Gnathostomes only the linked genes *Dlx1* and *Dlx2* are expressed in this tissue.

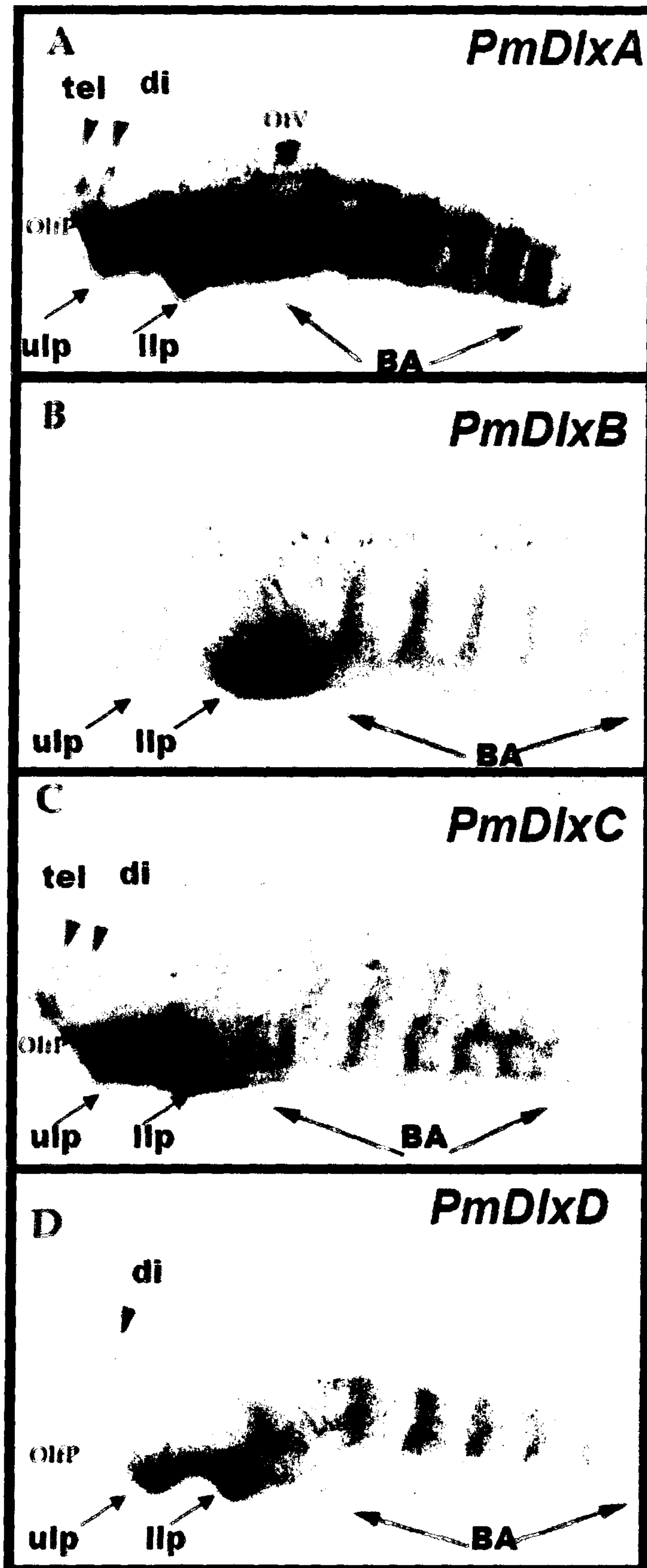


Figure 11: *In situ* Analysis of *Petromyzon marinus Dlx* expression. No *Dlx* code is evident in the branchial arches of *Petromyzon marinus* embryos at stage 26. Some differences in the expression domain of *PmDlx* genes are evident, but they also lack the archetypal *Dlx* code evident in Gnathostomes. Tel: telencephalon, di: diencephalon, ulp: upper lip, llp: lower lip, BA: branchial arches, OlfP: olfactory placode, Otv: otic vesicle (Figure modififed from Neidert et al 2001).

1.5 *Dlx* Gene Regulation

One scenario which accounts for the phenomenon of shared expression patterns of linked *Dlx* genes is that cis-paralogs share transcriptional regulatory elements, coordinately regulating their expression in space and time. This is also a favored hypothesis for why these clusters have remained intact over millions of years, resisting the insertion of repetitive elements, or chromosomal translocations which would disrupt the regulation of one or both genes contained within a cluster. Several such candidate regulatory elements have recently been discovered surrounding *Dlx* loci by phylogenetic footprinting, a method of searching for conserved blocks of sequence across a spectrum of distantly related species (See Figure 12). Across vertebrates, several regions of remarkable sequence conservation, or conserved noncoding elements (CNEs), have been found in each of the three intergenic regions separating *Dlx* genes, as well as upstream of the transcriptional start sites, and have been shown to act as tissue specific enhancers when assayed in transgenic mice and zebrafish (See Figures 13 and 14). Among these enhancers, those located within the intergenic region which is shared by each gene of a *Dlx* cluster are considered prime candidates for shared regulatory elements, coordinately regulating their expression and causing strong selective pressure to maintain the cluster intact. Within the *Dlx*1/*Dlx*2 cluster two elements of remarkable sequence conservation were originally detected by performing alignments between tetrapod and teleost genomes and were termed I12a and I12b (Ghanem et al. 2003). I12a consists of a sequence approximately 550bp in length showing approximately 92% identity between humans and zebrafish. When tested in reporter constructs the mouse version of this sequence was able to target expression of a transgene to only a subset of the mesenchymal cells of the

first and second branchial arches which normally express both *Dlx*1 and *Dlx*2 across the entire proximodistal axis in the developing mouse embryo (Park et al. 2004). I12b shows approximately 78% identity across 400bp between humans and zebrafish. When tested in transgenic organisms it is able to recapitulate expression of *Dlx*1 and *Dlx*2 in both the endogenous telencephalic and diencephalic domains of the embryonic forebrain of mice. Additional elements have been located upstream of the *Dlx*1 gene and have been named URE1 and URE2 and while characterization of these elements is still incomplete URE2 has been shown to drive expression of reporter genes in the embryonic telencephalon and diencephalon, apical ectodermal ridge of the forelimbs and hindlimbs, and ectomesenchymal cells of the branchial arches in the mouse embryo, while URE1 expression appears limited to a subset of neuronal precursors in the developing retina (Ghanem et al. 2007, Marc Ekker unpublished results). Within the intergenic region of the *Dlx5*/*Dlx*6 cluster two highly conserved regulatory elements have been located in paralogous positions to the two intergenic enhancers found in the *Dlx*1/*Dlx*2 cluster and have been named i56i and i56ii (Ghanem et al. 2003). The element i56i spans approximately 440bp and retains 81% sequence identity between humans and zebrafish. When tested in transgenic mice, this element is able to recapitulate the entire mesenchymal expression domain of *Dlx*5 and *Dlx*6 in the branchial arches of mouse embryos, and shows the appropriate limitation to the mandibular component of branchial arch 1 (Park et al. 2004). However, this element is not able to recapitulate the ectodermal expression of *Dlx*5 and *Dlx*6 within the branchial arches illustrating the requirement for the cooperation between several different enhancers to appropriately coordinate the complex spatiotemporal expression of *Dlx* within the developing embryo. The element

i56ii shows a similar expression domain to I12b in the telencephalic and diencephalic compartments of the embryonic forebrain. Interestingly, it has been demonstrated that while I12b and i56ii both target expression of transgenes to the same domain of the forebrain, with observable overlap at the cellular level, there are cells which respond only to I12b or only to I56ii (Zerucha et al. 2000). This may be why both members of the *Dlx*1/*Dlx*2 and the *Dlx*5/*Dlx*6 cluster are required during forebrain development despite the appearance of overlapping expression domains. It is currently unknown whether a similar partitioning of expression domains between different cell types within a contiguous tissue occurs where *Dlx* expression overlaps in the developing branchial arches. Within the *Dlx*3/*Dlx*4 cluster's intergenic region 5 elements have been identified which show very high sequence similarity between mouse and humans, but only one sequence I37-2 shows any detectable level of conservation between tetrapods and teleosts (Sumiyama et al. 2002). When a transgene incorporating this 79kb of this *Dlx* cluster including the entire intergenic region is tested in mice it was able to drive expression in many of the endogenous *Dlx*3/*Dlx*4 expression domains including the full mesenchymal component, but not the epithelial component, of the normal expression domain of these genes in the branchial arches (Sumiyama and Ruddle 2003). When individual elements were tested it was found that I37-2 alone was necessary and sufficient to drive branchial arch mesenchymal, but not epithelial, expression of the transgene in mouse embryos in a manner similar to i56i (Sumiyama and Ruddle 2003). While elements contained within orthologous clusters show high conservation across a number of different vertebrates, elements contained within different *Dlx* clusters do not appear to share high sequence similarity with one another, despite the fact that they are present in paralogous positions,

and target transgene expression to similar embryonic structures (Ghanem et al. 2003). It is hypothesized that ancestral elements were present in the intergenic region of the original bigene cluster, and following the whole genome duplications which resulted in *Dlx* cluster duplication, these elements were also duplicated but subsequently underwent rapid sequence evolution causing them to loose sequence similarity with each other while maintaining their ability to target *Dlx* expression to similar cell types during development. Genome wide identification of tissue-specific enhancers has also been carried out in the Ascidian, *Ciona intestinalis,* but curiously, while an Ascidian conserved element upstream of *CiDllA* was found to display enhancer activity (Harafuji et al. 2002), there were no elements found in the single *Dll* bigene cluster and even the known *CiDllA* enhancer shows no sequence similarity with the *Dlx* intergenic or upstream enhancers found in vertebrates. This may be one of the reasons for the lack of overlap between the expression patterns of *CiDll*A and *CiDll*B. The origin of the intergenic enhancers and the coordinate control of *Dlx* bigene clusters is therefore an open question and of significant interest given the importance of coordinate control of *Dlx* family member expression in a number of important vertebrate evolutionary novelties including the jaw apparatus.

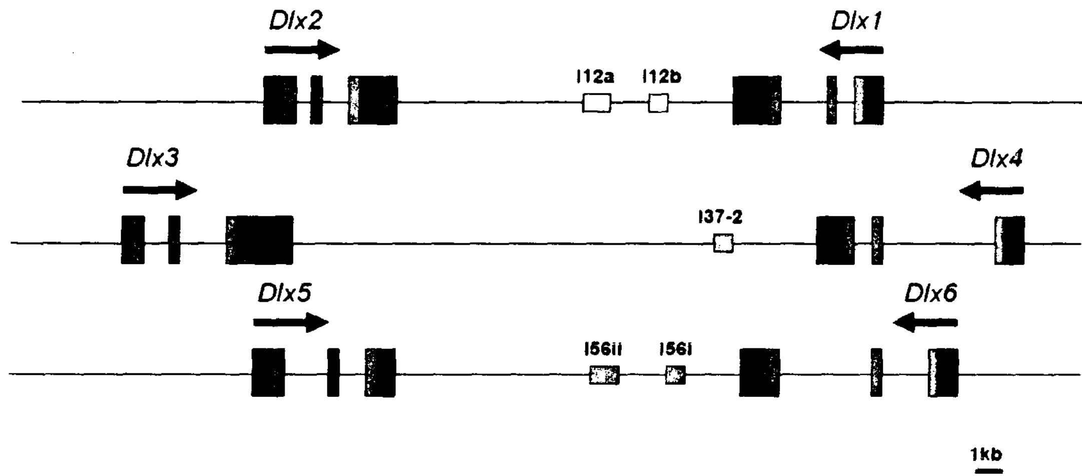


Figure 12: Conserved noncoding elements (CNEs) are found in the intergenic region of *Dlx* gene pairs.
Schematic of *Dlx* gene organization in the mouse. Phylogenetic footprinting identifies several highly conserved DNA elements in the regions flanking *Dlx* genes: I12a, I12b, I37-2, I56ii, I56i, (URE1, URE2 not shown). When these elements are tested for enhancer function, they display activity in domains consistent with *Dlx* expression. Only elements which show high similarity between tetrapods and teleosts are shown. Schematic is drawn to scale (scalebar shown). Colourcoding of the exons corresponds to which *Dlx* subfamily each gene belongs to (blue for the *Dlx2/3/5* subfamily; purple for the *Dlx1/4/6* subfamily). Black boxes represent UTRs. Arrows denote transcriptional orientations.

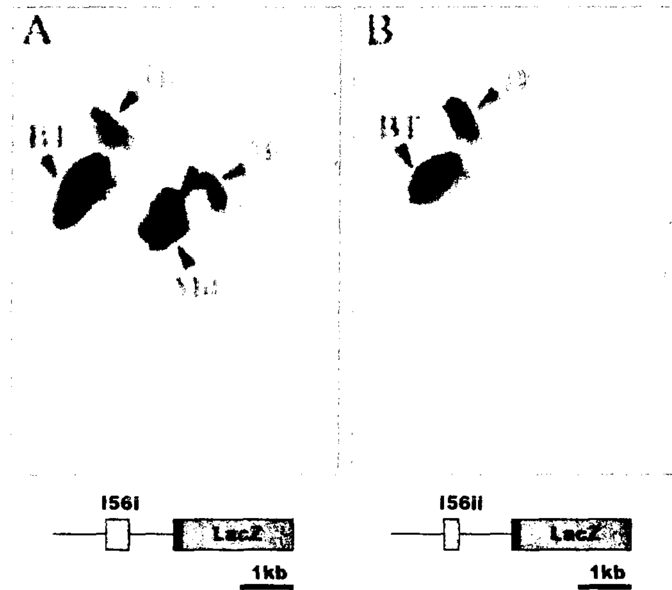


Figure 13: Activity of intergenic enhancers associated with the *Dlx5/Dlx6* bigene clusters in the mouse embryo. (A) The i56i-*lacZ* transgene shows expression in two domains in the forebrain, one telencephalic and one diencephalic, as well as expression in the mandibular component of the first branchial arch (Md: mandibular), and the second arch (Hy = hyoid). (B) The *i56ii*-lacZ construct shows both expression domains in the brain, but no expression in the branchial arches. Note that while i56i and i56ii show no sequence similarity, they overlap significantly in their macroscopic expression patterns in the forebrain. (Figure modified from Ghanem et al 2003)

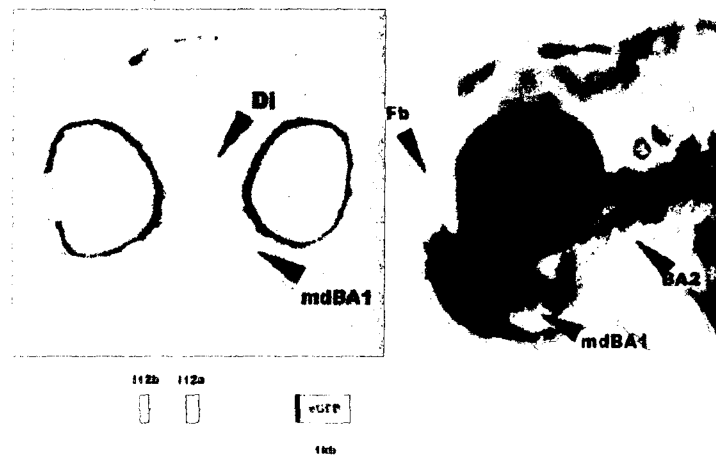


Figure 14: Activity of intergenic enhancers associated with the zebrafish *dlx1a/dlx2a* cluster. A 6kb transgene including zebrafish I12a and I12b and shows activity in the forebrain (fb) and branchial arches (mdBA1: mandibular branchial arch 1, BA2: branchial arch 2) of the developing zebrafish embryo. (Figure courtesy of Ryan MacDonald)

1.6 Major Questions

The major issues addressed by this thesis can be summarized by the following two questions. The first question is whether or not Cyclostomes have undergone the same two rounds of whole genome duplication known to be shared by all Gnathostome Vertebrates. The second question specifically relates to the origin of a defining morphological feature of the Gnathostomes; the jaw. That second question is whether changes in cis-regulatory elements of the *Dlx* family of homeodomain containing transcription factors have contributed to the evolution of the Gnathostome jaw apparatus, through the acquisition of a tight spatio-temporally regulated "*Dlx*-Code". Each of these questions have been addressed using slightly different combinations of molecular and developmental biology approaches.

Section 2: Materials and Methods

2.1 Genomic DNA extraction from hagfish tissue

Genomic DNA used in the isolation of novel hagfish *Dlx* genes was extracted from *Myxine glutinosa* lateral muscle, liver, and heart tissue from samples preserved in RNA later provided to us by Dr. Tom Moon. Tissue was digested with Proteinase K overnight at 55 °C in digestion buffer (50 mM Tris (pH 8.0), 100 mM EDTA (pH 8.0), 1% SDS, 100 mM NaCl, 350mg Proteinase K (fresh). Following digestion, extraction of genomic DNA was performed using 3 treatments of a 1:1 volumetric ratio of phenol:chloroform, followed by 2 treatments of chloroform to remove excess phenol. DNA was precipitated by adding 1/10 volume of 3M sodium acetate to the resultant aqueous phase followed by the addition of 3 volumes of ice cold 100% ethanol and incubation at -20 °C for 1hr.

Following a wash in 70% ethanol, the precipitated DNA was resuspended in 100-200ul of nuclease-free water, allowed to incubate overnight at room temperature, and concentration was measured spectrophotometrically. Genomic DNA was stored at 4 °C indefinitely for use in PCR reactions.

2.2 Isolation of novel hagfish *Dlx* gene fragments

Genomic DNA from *Myxine glutinosa* was used as a template in the PCR amplification of *Dlx* gene fragments. Degenerate primer pairs (Sequences: *Dlx*H*o*m1F GCCGGGATCCAARCCNMGNACNATHTAYTC, *Dlx*Hom*1R* TTYTGRAACCADATYTTNAC, *Dlx*H*o*m2F: CCGAATTCAARCCNMGNACNATHTA, *Dlx*H*o*m2R: CCGGATCCRTTYTGRAACCADATYTT) were used to amplify a fragment of the homeobox and the entire second intron of *Dlx* genes using a low-stringency PCR reaction. The standard PCR reaction mixture consisted of a 50ul reaction composed of 5ul of 10X PCR reaction buffer, 5ul 2uM dNTPs, 1ul 20ng/ul forward primer, 1ul 20ng/ul reverse primer, 1ul TAQ polymerase, 1ul of *M. glutinosa* genomic DNA, and 37ul autoclaved double distilled water. PCR thermocycler conditions used 30 cycles of 1min at 95°C for Denaturation followed by 1 min at temperatures between 45-55°C for primer annealing and 1-10min at 72°C for extension. PCR products were run on a 0.8% agarose gel in 0.5X TBE buffer at 100V for 30-60 minutes in order to isolate independent amplicons for gel extraction. Bands were cut out of the gel using a scalpel and gel extracted using the Qiagen QiaquickII Gel Extraction Kit. Individually isolated fragments were cloned into the pDrive T-A system PCR cloning vector by combining 2ul of the gel-extracted PCR products, 5ul of the 2X ligase/reaction buffer mixture, and 1ul

of the pDrive vector provided and ligated overnight at 16°C. The following day 5-10ul of the ligation reaction was used to transform 50ul of chemically competent XL-10 strain E. coli cells using the heat shock method. The ligation mixture was mixed with the thawed E. coli cells and incubated on ice for 40 minutes. Heat shock was then induced by incubating the cells at 42°C for 30 seconds, followed by an ice bath lasting 2 minutes. 500ul of fresh LB medium was added to each tube and the cells were incubated at 37°C for 30 minutes to allow recovery. Cells were then spun at 5000rpm for 30s and plated on LB-agar plates containing 50ng/ul ampicillin and previously smeared with 50ul of 100ng/ul X-gal and 10ul of 100ng/ul IPTG to permit selection of bacterial colonies containing an ampicillin resistant plasmid whose *lacZ* coding sequence had been interrupted by the incorporation of the cloned DNA fragment, as it had been designed to do, causing the colonies to appear white on the plate. White colonies were selected and picked from the plates using a plastic pipette tip and used to inoculate 5ml LB-Ampicillin (50ng/ul) cultures. These cultures were incubated overnight at 37°C, and DNA extraction was performed the following day using either a home made DNA mini-kit containing resuspension (P1), lysis (P2), and neutralization (P3), solutions or the Promega Wizard SV plus pure mini-prep kit. Insert size was verified by restriction digestion by using 3ul of prepared plasmid DNA, 3ul of Invitrogen React2 digestion buffer, 0.25ul of *EcoRI* enzyme in a 30ul total volume using water. The insert was flanked by *EcoRI* cut sites designed in the pDrive vector and therefore the entire fragment should be cut out of the 3.5kb cloning vector and be visible by gel electrophoresis and Ethidium bromide staining. Unique fragment sizes representing possibly different *Dlx* genes with larger or smaller

inserts were then sent for sequencing using the T7 primer designed to bind to the internal T7 promoter site present on the pDrive cloning vector.

2.3 Isolation of lamprey *Dlx* gene fragments

Previously published sequences of *Petromyzon marinus DlxA, DlxB, DlxC,* and *DlxD* (Neidert et al. 2001) and sequences of *Lethenteron japnomicum DlxE* and *DlxF* kindly provided by Dr. Shigehiro Kuraku were used to search the assembly of the *Petromyzon marinus* genome using the Ensembl pre! Species genome browser available at http://pre.ensembl.org/Petromyzon_marinus/index.html. Contigs containing each of these 6 *Dlx* genes were found in the *Petromyzon marinus* genome and used as templates for the design of primers to specifically amplify fragments of individual *Dlx* genes. Various primer pairs were used to amplify different fragments of each gene. The sequence of primers used to amplify fragments of *Petromyzon Dlx* genes were: *Dlx*A (PmarDlxA5'1: GCGCTCCAAGTTCAAGAAGC, PmarDlxA3'1: TGCTGGTGCTGCAGGTACCA, PmarDlxA3'UTRd1: CATAGGCCAGACCCAGATGT, PmarDlxA3'UTRu1: CGGCACTTAAAAGGGAATCA) *Dlx*B (PmarDlxB5'-3: ATTCCATGACCTGCGACTCC, PmarDlxB3'-1: GCCGGTACCACGGCTGGTAC), DlxC (PmarDlxC5'-1: ATCTGGTTTCAGAACCGGCG, PmarDlxC3'1: CGGGTACTGAGCGTGCAGCT), *Dlx*D (DlxDfor2: TTAGTGAACACTGGCTGCCT, DlxDrev2: TTGGCCATGCTCTCTTGATG), *Dlx*E (Lj*Dlx*Efor1: GTACAAGAAGATATGAAGCAGG, Lj*Dlx*Erev1: CCACGTTTGACTGCAAGTCCC, Lj*Dlx*Efor2: GGTTACGATGAGAAAGAGGAGG, Lj*Dlx*Erev2: CTCGTTTAGGTCGACTCGACC), *Dlx*F (Lj*Dlx*Ffor1:

AGAAGCTCCTCAAGCAGTCGG, Lj*Dlx*Frev1: GTGGTGGTTGTGATGATGGG, Lj*Dlx*Ffor2: AGATGACGACCTTCCTATTGGG, Lj*Dlx*Frev2: CGGTGGCTAGGTGTTAACTCC). These fragments were PCR amplified using a standard reaction with annealing temperatures ranging from 55-60°C from genomic DNA of *Petromyzon marinus* kindly provided to us by Dr. Sylvie Rétaux, cloned into the pDrive cloning vector, and sequenced, as described above.

2.4 Isolation of BACs containing hagfish and lamprey *Dlx* genes

A spotted array format *Petromyzon marinus* BAC library was purchased from CHORI BAC/PAC Ressources and an *Eptatretus stoutii* BAC library was kindly loaned to us by Dr. Chris Amemiya for the purpose of screening for clones containing cyclostome *Dlx* genes. Radiolabelled probes were prepared from either *Petromyzon* or *Myxine Dlx* gene fragments previously isolated by using the random priming method with αP^{32} ATP. For making double stranded DNA radioactive probes, 50-100 ng of DNA template was mixed with 2 µg of pdn6 (random hexanucleotide), 1 X react2 buffer (Invitrogen), 1.5 mM dNTP mix lacking dATP, 50 µCi of 32-P-alpha dATP and 10 units of Klenow polymerase (Invitrogen) and adjusted to a final volume of 25 µl with water. This mix was incubated for 1hr at 37°C. The reaction was stopped by adding 1 µl of 0.5M EDTA. Radioactive probes were isolated using the Qiagen PCR cleanup kit

These probes were used to perform Southern hybridization on the dot blot BAC libraries. The arrayed sheets were placed in 100ml tubes and prehybridization was done in Church medium (0.36 M Na_2HPO_4, 0.14 M NaH_2PO_4) containing 0.5 mg/ml of sheared, denatured salmon sperm DNA for 2 to 6 hours at 65°C. The medium was then removed, and replaced with fresh church medium. 20 to 50ng of the radioactive probe

was incubated with the membrane, and the hybridization was allowed to proceed overnight at temperatures ranging from 60 to 65°C, depending on the length of the probe. In general, lower hybridization temperatures were used for the isolation of *Eptatretus stoutii* BAC clones because the probes were amplified from a different species and were not 100% identical to the target sequences. The following day, blots were washed three times with 1X SSC, 0.1% SDS. They were then removed from the tubes and washed on a shaker for 15 minutes in 1X SSC, 0.1% SDS at room temperature. A second wash was then performed for another 15 minutes with 0.1X SSC, 0.1% SDS. A Geiger-Mueller counter was used to determine whether the blots were ready for imaging. Blots were then exposed to BIO RAD Imaging Screen – K for 2-4 hours and scanned with BIO RAD Molecular Imager® FX. If the image was considered good quality, the blots were then exposed to X-ray film for between 12 to 72 hours, depending on the strength of the signal, to capture a permanent, semi-transparent image of the blot which was used to determine the coordinates of positive clones. The blots were then stripped of radioactivity by washing in boiling hot 0.1% SDS solution and left to cool at room temperature. The blots were then washed with 1X SSC solution and stored at 4°C until further use.

2.5 PCR Screening BAC clones

BAC clones identified by southern hybridization were ordered and the stab cultures which were delivered were streaked on LB-Agar-Chloramphenicol (20ng/ul) plates, grown in 5ml LB-Chloramphenicol cultures, and amplified in large 200ml cultures in order to isolate large amounts of the low-copy BAC vectors carried in the cells using the Qiagen modified midi protocol with the Qiagen midi kit. The BAC DNA was then used

as a template for a standard PCR reaction using primers specific to various hagfish or lamprey *Dlx* genes in order to confirm whether or not the clones contained one or more *Dlx* genes. Specific primers used to confirm the identity of BAC clones containing hagfish *Dlx* genes were used.

2.6 Southern Hybridization to *Dlx* BAC Clones

Clones which tested positive by PCR were subsequently screened by Southern Hybridization in order to both confirm the identity of the clone, to determine the size of the *DlxW/DlxX* or *DlxC/DlxD* gene clusters, and to identify suitable sized fragments for subcloning. Approximately 10ug of BAC DNA was digested with a series of restriction enzymes including *Eco*RI, *Hind*III, *Bam*HI, *Xho*I, and *Xba*I and run on a 0.7% agarose gel overnight at 40V in order to achive accurate separation of the large restriction fragments which resulted from the digestion of the large insert clone. Standard southern blots were performed. The gel was equilibrated by soaking in a denaturation/transfer solution (0.5M NaOH) for 30 minutes. The DNA from the gel was then transferred to a HyBond-N membrane using the capillarity transfer method. A wick was setup 0.5M NaOH on which 3 sheets of 3MM Whatman filter paper. The gel was placed upside down and reverse orientation onto the Whatman paper followed by nylon membrane on the gel. On top of this setup, 10 cm of dry paper towels were stacked and a glass plate with approximately and weighted with textbooks. The gel was surrounded with parafilm to prevent the paper towels from coming in contact with the transfer solution. This setup was left overnight for transfer. The membrane was then baked for two hours at 80°C to fix the DNA onto the membrane. Following transfer, radiolabelled *Dlx* gene fragments

prepared as desribed above were used as probes to identify restriction fragments which contained *Dlx* genes.

2.7 Cloning BAC fragments in pBluescript

Restriction fragments containing cyclostome *Dlx* genes identified by Southern blotting were gel extracted and were subcloned into the pBluescript cloning vector. A 9kb *Hind*III fragment containing *Dlx*X and a 5kb *Xba*I fragment containing *Dlx*X were subcloned in this way. The fragments were subcloned using a 3:1 insert:vector molecular ratio with T4 DNA ligase in overnight ligation reactions at 16°C.

2.8 Fluorescent *in situ* hybridization to *E. burgeri* chromosomes

2.7.1 Chromosome Spread Preparation

Male specimens of *E. burgeri* were dissected and the testes were removed and minced in 2X Marine BSS (Balanced Salt Solution). The cell suspension was pipetted to induce shear force to break up cell clusters. The cells were briefly spun in a hand rotor in 15ml falcon tubes to remove large chunks of tissue and the supernatant was removed to continue the process. The cells in the supernantant were separated into several Eppendorf tubes and spun at 4°C for 30minutes at 4000 x g to pellet the cells. The cells were then resuspended, dropwise, in hypotonic solution (0.075M KCl) to induce swelling and treated in this solution for 45minutes. Fixation was carried out by first adding 200ul of Carnoy's Fixative (3:1 methanol:acetic acid) to 1000ul of cells in hypotonic solution for 30 minutes. The cells were then pelleted, and resuspended dropwise in Carnoy's fixative. This was repeated 3 times in order to ensure all aqueous solution had been removed. The final suspension of cells in Carnoy's was then dropped on glass slides from a height of approximately 10cm to spread the cells on the slide. Quality of spreading of metaphase

cells was assessed by Giemsa staining. Unused slides were stored at -80°C until needed for FISH experiments.

2.7.2 FISH Probe Preparation

1.5ug of *E. burgeri* BAC DNA was digested using a 1:2500 dilution of DNAseI according to the protocol described in the Invitrogen ULYSIS nucleic acid labeling kit protocols. 1-2ul of digest was checked on a 0.8% agarose gel in order to ensure adequate digestion. The entire quantity of digested BAC DNA was run on a gel and fragments between the size of 100bp – 600bp were gel extracted. For the 18S rDNA gene this step was skipped and as long as the DNAseI digestion produced fragments between mostly between 100-1000bp in size hybridization was still good. The digested and gel extracted DNA fragments were ethanol precipitated and air dried. Subsequently they were resuspended in labelling buffer (20ul – 24 μl depending upon the fluorophore used in the labelling reagent). The sample was vortexed and incubated at 37°C for 30min to ensure resuspension. The sample was denatured at 95°C for 5min, and flash cooled on ice. The ULYSIS labelling reagent 1-5ul depending upon the fluorophore (1ul for AlexaFluor 488 (Green) and 1.5ul AlexaFluor 534 (Red)) was then added and the mixture was incubated for 80C for 15 minutes. The reaction was stopped by incubation in an ice bath and the probe was purified using the Bio-Rad Micro Bio-Spin chromoatography column kit.

2.7.3 Slide Pretreatment

Metaphase spreads were removed from -80° and equilibrated in 2X SSC at room temperature for 30min. Slides were pretreated with 0.01N HCl (pH = 2.0) prewarmed to 37°C for 5 mins. Slides were next treated with 1mg/ml pepsin in 0.01N HCl to partially digest metaphase spreads and remove cytoplasm and incubated for 10min at 37°C. The

reaction was stopped by dipping slides in distilled water followed by 2 washes in 2X SSC for 5mins, brief fixation in 1% formaldehyde/2XSSC 5mins at 4°C, 2 washes in 2X SSC at room temperature 5 mins, and dehydrated through a graded ethanol series (70%, 95%, 100%, 100%). DNA on the slides was denatured by incubating in 50% deionized formamide 2XSSC at 76° for 3-4 minutes and flash cooled in 70% ethanol at -20°C followed by dehydration in a cold ethanol series, then air dried.

2.7.4 Hybridization

The hybridization mixture was prepared as follows: between 50 – 250ng of labeled probe, 2ul of sheared E. burgeri genomic DNA (7ug/ul), 1ul sheared salmon sperm DNA (10ug/ul), 2ul T/A primer (100ng/ul) was combined and brought to 50ul with distilled water. The probe/blocking DNA mixture was ethanol precipitated and resuspended in 5.5ul distilled water and 15ul deionized formamide, and resuspended by vortexing and incubation at 37°C for 30 minutes. To this mixture, 6ul of 50% dextran sulfate, 0.5ul of 10%SDS, and 3ul of 20X SSC was added and mixed. The hybridization mixture was denatured at 80°C for 15minutes and flash cooled on ice. The probe and blocking DNA was then allowed to prehybridize at 37°C for 60mins before the full 30ul volume was added to the slide, covered with a trimmed plastic coverslip and sealed with PaperBond adhesive. The hybridization was carried out in a humidified chamber at 37°C overnight, protected from light.

2.7.5 Wash Conditions

The coverslip was removed carefully and the wash conditions were: 2X SSC 0.1% SDS 50°C 5mins, 20% formamide 0.1X SSC 50°C 5mins three times, 2X SSC 0.01% SDS 50°C 5mins two times, 0.5X SSC 0.01% SDS 50°C 5 mins once, followed by a rinse

in room temperature distilled water. The slides were then allowed to airdry and were counterstained and mounted with DAPI mixed with the antifade agent DABCO in 50% glycerol, 1X Mcllevain's buffer.

2.9 Microinjection of DNA into zebrafish embryos

Zebrafish were bred and embryos were collected and arranged on an agarose plate for microinjection as described by (Westerfield 1995). Plasmid and cosmid DNA for microinjection was prepared using the promega wizard SV plus mini prep kit, resuspended in nuclease-free water pH = 7.0 and adjusted to concentrations ranging from 50-100ng/ul. A small amount of 1% phenol red was used to give the injection solution colour in order to visualize the injection process. Microinjection needles (ID:0.5mm, OD:1mm, Borosilicate with filament, Sutter Instruments) were pulled using a P90 micropipette puller. One cell stage embryos were microinjected using a Narashingi IM300 microinjector. The amount of DNA solution injected was adjusted manually from the microinjector such that the diameter of the injected solution was approximately one-fifth of the cell's diameter. Embryos were incubated in 1X E3 embryo medium (Westerfield 1995) for 24, 48, or 72 hours at 28°C before visualization of transgene expression was conducted using a UV equipped microscope and camera.

2.9 Screening zebrafish for transgene transmission

Primary injected embryonic zebrafish which were positive for transgene expression were raised to adulthood according to (Westerfield 1995) and crossed with each other in order to produce stable transgenic lines. Male and female zebrafish were sexed individually and placed in isolated containers to mate them pair-wise. Embryos were

collected from mated pairs and screened using the UV microscope for transgene expression.

2.10 Microinjection of DNA into *L. fluviatilis* embryos

Lamprey were obtained from fishermen in France and eggs and sperm were stripped from adults during mating season and used in artificial fertilization. Successfully fertilized eggs were transported to the lab where they were manually microinjected in a manner similar to (Westerfield 1995). Embryos were incubated at 15C in recirculating stream water by submerging petri dishes with holes bored into them into large tanks. Transgene expression was visualized with a UV microscope.

2.11 Total RNA isolation from adult hagfish brain

Adult specimens of *Eptatretus stoutii* were sent to us by Dr. Bruce Cameron and sacrificed by anesthetizing them in MS-222 and decapitating the animals. The whole brain was then dissected from the cranium, frozen in liquid nitrogen, and stored at -80°. The whole brain was then homogenized in TRIZOL reagent and RNA was extracted using the standard TRIZOL protocol. Quality of RNA was assessed by running small sample (1-2ul) on a 1% 1X TAE gel run at 80V and inspected for the presence of strong bands rather than smeared degraded RNA.

2.12 Reverse-Transcription of adult hagfish brain total RNA

Reverse transcription of whole brain RNA was conducted using the random priming method in reactions consisting of 250ng pdN6 random hexanucleotides, 2ug total RNA, 5ul of 2mM dNTPs and distilled water to 12ul. The reaction was heated to 65°C for 5 mins, chilled on ice, centrifuged briefly to collect the reaction. 4ul of 5X first strand synthesis buffer, 2ul of 0.1M DTT and 1ul of RNAseOUT was added, the reaction was

mixed, and the sample was incubated at 25°C for 2 minutes. 1ul (200U) of SuperscriptII reverse transcriptase was finally added to the reaction, mixed and incubated at 42°C for 50 minutes followed by a fina lstep at 70° for 15 minutes.

2.13 Amplification of Hagfish *Dlx* cDNA

Eptatretus stoutii Dlx cDNA was amplified from the cDNA pools produced as described above by standard PCR reactions. Primers were designed shortly upstream of the predicted ATG start codon and downstream of the predicted stop codon, but upstream of the predicted AATAAA polyadenylation signal in order to attempt to amplify the full coding sequence and partial UTRs of the *Dlx* genes expressed in adult hagfish brain. Primer sequences were: *Dlx*W (EsDlxWfor1: ctacgaccgtgtcctggtct, EsDlxWrev1: gggatcaggttcagctcaca, EsDlxWcDNArev2: ggttctgacacgccgatatt), *Dlx*V (EsDlxVcDNAfor: cgccaccccaaattaac, EsDlxVcDNArev: gtcataggcgcgtcacaatac) *Dlx*X (EsDlxXcDNAfor: agtgttgcggcgactgtt, EsDlxXcDNArev: gcatgatgatcgaagggact), and *Dlx*Z (EsDlxZcDNAfor: ccctgctcttatccaagctg, EsDlxZcDNArev: gccaagtgtgtcttgtgctc). Amplification with *DlxW* cDNA primers in any combination (above) did not produce a product. PCR products were cloned in pDrive and sequenced as described above to confirm their identity.

Section 3: **Results**

3.1 Identification of *Dlx* Genes in the Lamprey Genome

Previously, the sequence and partial developmental expression patterns of 4 *Dlx* genes from the sea lamprey, *Petromyzon marinus*, had been published (Neidert et al. 2000). In order to decipher the genomic organization of these 4 *Dlx* genes a previous student in our laboratory amplified fragments of orthologues of each of these genes from genomic DNA of the European river lamprey, *Lampetra fluviatilis* (Maurya 2006). The amplified fragments of the 4 *Dlx* genes, termed *LfDlxA, LfDlxB, LfDlxC*, and *LfDlxD*, were used to screen a cosmid library of *L. fluviatilis* purchased from RZPD in order to isolate genomic clones containing these genes. Cosmid clones containing fragments of *LfDlxA, LfDlxC*, and *LfDlxD* were isolated in this way and partially sequenced using the shotgun method, and partially mapped through restriction digestion and Southern hybridization. It was determined that *LfDlxC* and *LfDlxD* were organized in a bigene cluster as with other Gnathostome *Dlx* genes. Sequencing and restriction mapping revealed that the intergenic region between these two genes was approximately 11kb, as large as that of model mammals and considerably larger than in model teleost fish. In the time since this work was completed orthologs of 2 additional *Dlx* genes were isolated from a Japanese lamprey, *Lethenteron japonicum* and named *LjDlxE* and *LjDlxF*, in addition to orthologs of the 4 previously identified genes (Shigehiro Kuraku personal communication).

In order to determine the genomic arrangement of *Dlx* genes in the lamprey, I sought to isolate genomic clones from the sea lamprey, *Petromyzon marinus*, an organism whose whole genome is currently being sequenced, but has run into significant problems

during assembly therefore requiring extensive manual mapping and annotation (Chris Amemiya personal communication, Ensembl genome browser website). A preliminary assembly of the *Petromyzon marinus* genome is available to browse in Ensembl Pre! Species (http://pre.ensembl.org/Petromyzon_marinus/index.html). Using the published cDNA sequences of *Petromyzon marinus PmDlxA*, *PmDlxB*, *PmDlxC*, and *PmDlxD* and the cDNA sequences of *Lethenteron japonicum LjDlxE* and *LjDlxF*, the preliminary assembly was searched using a BLAST algorithm (Zhang et al. 2000). Contiguous assemblies containing partial sequences of each of the 6 lamprey *Dlx* genes were found in this way using the genome browser (*Dlx*A: Contig7656, Contig42733, *DlxB*: Contig31902, Contig77102, *DlxC*: Contig7549 , *DlxD*: Contig40555, Contig33064, *DlxE*: Contig98301, Contig71202, *DlxF*: Contig38365) however as expected none of the contigs contained complete *Dlx* coding sequences, likely due to the inability to adjoin smaller contigs with high fidelity because of the presence of repetitive elements. Furthermore, while work in our lab has already demonstrated that *Lampetra fluviatilis LjDlxC* and *LjDlxD* are linked in a tail to tail cluster with an intergenic region of approximately 11kb this association was not detected by browsing the Ensembl Pre! Species database. In fact while more than 18kb of sequence has been assembled downstream of the *PmDlxC* containing Contig7549, no sign of *PmDlxD* is present in this region indicating that either missasembly has occurred at this locus, or that *Petromyzon marinus* and *Lampetra fluviatilis Dlx* genes are organized differently, which would be surprising considering the short amount of time between the divergence of these lamprey species (Kuraku and Kuratani 2006). Importantly because of the high similarity between members of the *Dlx* gene family, especially within their homeobox region, BLAST

searches for specific *Dlx* orthologs often returned hits with lower similarity to other paralogous members of the *Dlx* gene family. For example BLAST searches for *PmDlxA* would return contigs containing *PmDlxA*, *PmDlxB*, *PmDlxC*, and to a lesser degree due to their divergent sequence, *PmDlxD*, *PmDlxE*, and *PmDlxF*. During the extensive searches of contigs contining *Dlx*-like sequences no more than the 6 *Dlx* genes previously isolated from different species were found in the *Petromyzon marinus* genome. The sequences with the next highest similarity to *Dlx* located in these screens included members of the *Nkx* and *Hox* gene families, other developmentally important transcription factors anciently related to, but clearly not members of, the *Dlx* gene family. Because an approximately 6X sequencing of the *Petromyzon* genome has been conducted, we therefore believe it is unlikely that there are other *Dlx* genes present in the *Petromyzon* genome, bringing the total number of *Dlx* genes in this organism to 6, the same number as are present in the genome of Gnathostomes.

Despite the availability of a draft 6X sequence, assembly is incomplete and may be inaccurate. Because of these limitations in the quality of the genome sequencing project I sought to manually isolate and assemble genomic information about the *Dlx* genes of *Petromyzon marinus*. Based on information derived from the BLAST searches of the *Petromyzon* genome, target fragments between of 200bp to 600bp of each of the 6 *Dlx* genes in this species were identified with the intention of using them as probes to screen a BAC (Bacterial Artificial Chromosome) library of *Petromyzon marinus* purchased from CHORI BAC/PAC Ressources. This BAC library has the advantage of containing large inserts (90kb average insert size) and therefore has a high probability of containing both *Dlx* genes in a cluster should they exist. Effort was made to choose

fragments with low similarity between *Dlx* paralogs so that *Dlx* genes could be isolated independently. Candidate probe fragments were reverse-Blasted to the *Petromyzon* genome in order to assure that they were specific to the target gene and any fragments which returned BLAST hits of greater than 70% identity to other loci in the genome were rejected. Primers were designed in order to amplify each target fragment. Using genomic DNA isolated from adult *Petromyzon marinus* kindly provided to us by Dr. Sylvie Rétaux (CNRS Institute Alfred Fessard), fragments of each of the 6 *Dlx* genes were amplified, gel extracted, and cloned into a the pDrive PCR cloning vector. Sequencing of each probe was conducted using the internal T7 primer of pDrive to confirm the identity of the *Dlx* fragment before proceeding.

In order to identify BAC clones containing each of the *Dlx* fragments identified, probes were PCR amplified using T7 and SP6 vector primers and radiolabelled using αP^{32} ATP with the random priming method. Southern hybridization was conducted using the radiolabelled probes against a series of arrayed filter sets containing the thousands of BAC clones contained in the *Petromyzon marinus* library. Candidate clones were identified and ordered from the clone stock center for confirmation. BAC DNA was propagated and isolated for confirmation of clone identity. By performing a series of PCR checks using primers specific to the *Dlx* genes, and sending BACs for direct sequencing using *Dlx* specific primers the identity of BACs containing *PmDlxC*, *PmDlxD, PmDlxB*, and *PmDlxE* were confirmed. As we expected we found that *PmDlxC* and *PmDlxD* were present on the same BAC; clone 308022. This is in agreement with previous results indicating that *LfDlxC* and *LfDlxD* are linked in a tail to tail arrangement. This also confirms that while the *Petromyzon* genome project is a

powerful tool for gene identification, there are several assembly errors necessitating the use of manual and thorough assembly of regions of interest. Despite repeated efforts to identify clones containing *PmDlxA* and *PmDlxF*, no BAC clones tested positive for these genes. It is possible that they are not contained in this BAC library, or that our screening methods were not optimized for their recovery. It should be noted that a very high level of false positive BAC clones were isolated based on hybridization signals from the southern blot. Only 3 out of 15 isolated BACs actually contained *Dlx* genes, and many more BACs showed mild to weak hybridization signals despite stringent wash conditions. Using identical methods used to screen a hagfish BAC library (described below) we find that the false positive rate and background signal for *Petromyzon marinus* is much higher than other species which may be a result of specific properties of the lamprey genome which hinder high quality BAC library preparation, rather than our inability to optimize the protocols. In support of this, the company from whom the BAC library was purchased warns of the difficulty associated with screening their BAC library and makes no guarantee that specific clones will be found in their library (http://bacpac.chori.org/library.php?id=199).

Despite our inability to isolate all *Dlx* genes from the *Petromyzon* BAC library we were able to confirm that *PmDlxC* and *PmDlxD* are linked in a bigene cluster, and is likely an ancestral cluster orthologous to one, or several, Gnathostome *Dlx* bigene clusters. Interestingly by using PCR screens and southern hybridization with probes designed against multiple *Petromyzon Dlx* genes with digested BACs we were unable to detect the presence of any second *Dlx* gene on BACs containing *PmDlxB* or *PmDlxE*. This seems to indicate that these genes are isolated from other *Dlx* genes in the lamprey

genome and do not form clusters, a condition unlike that seen in any other vertebrates. However, it is impossible to rule out the possibility that these genes form clusters with exceptionally large intergenic distances, or are simply present towards the end of a BAC insert and that their sister gene was left off of the specific clones we identified. Full sequencing of these BAC clones would be necessary to make these conclusions. In support of our hypothesis, however, complementary information from *Lampetra fluviatilis* cosmids also indicates that *LfDlxA* has no other *Dlx* gene up to 20kb downstream of its last exon (Maurya 2006). This would make the intergenic distances in lamprey nearly twice the size of the distances found in any other vertebrate if indeed another gene is linked to *LfDlxA*. We find it more likely, that apart from the *DlxC/DlxD* cluster found in both *L. fluviatilis* and *P. marinus*, there are no other clusters of *Dlx* genes in the lamprey genome. This represents a significant divergence in lamprey from the archetypal organization of *Dlx* genes in vertebrates which are normally present in 3 highly conserved tangentially transcribed bigene clusters. See Figure 15 for a schematic representation of the organization of *Dlx* genes in the lamprey as inferred from the combined analysis of *Lampetra fluviatilis* and *Petromyzon marinus* genomic clones.

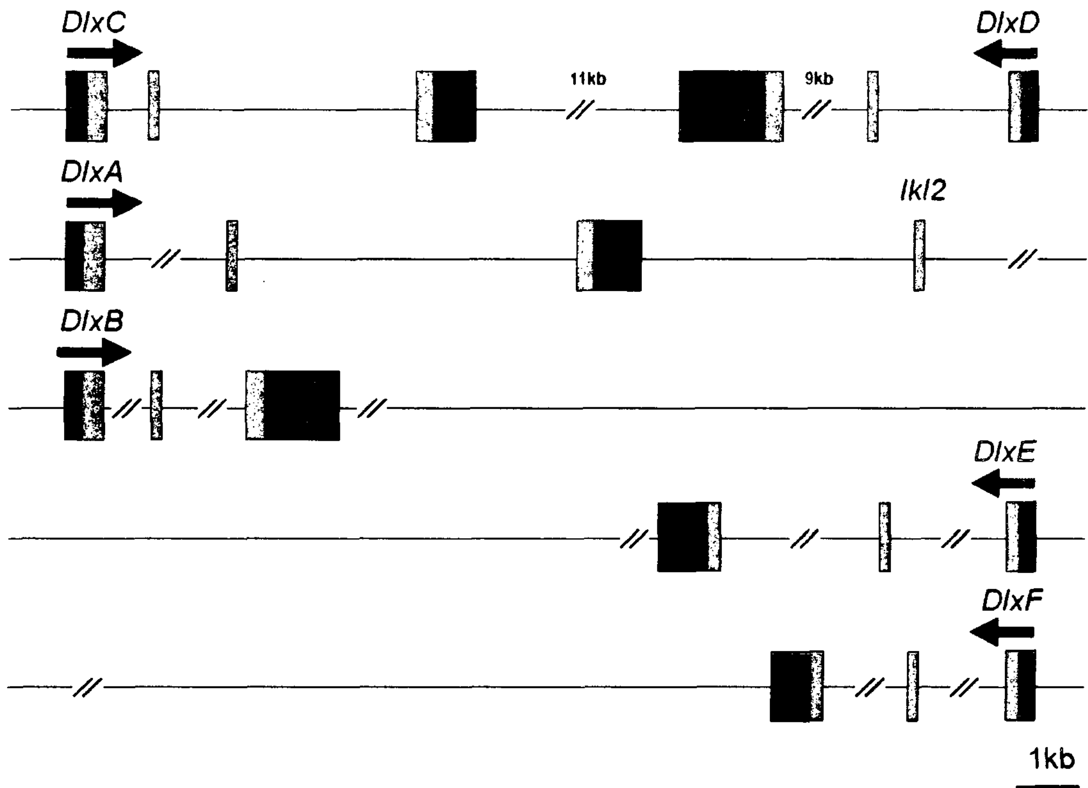


Figure 15: Schematic Representation of the Arrangement of *Dlx* Genes in *Petromyzon marinus*. Intron and intergenic distances are annotated with a number where known and separated by parallel lines where gaps in sequence assembly makes it ambiguous. Schematic is drawn to scale (scalebar shown). Colourcoding of the exons corresponds to which *Dlx* subfamily each gene belongs to (blue: *Dlx2/3/5* subfamily; purple: *Dlx1/4/6* subfamily). Black boxes represent UTRs. Arrows denote transcriptional orientations.

3.2 Identification of *Dlx* genes in the Hagfish Genome

In order to make any general statements about the condition of the ancestral Craniate genome it is of crucial importance to compare several species. Based on the information from the lamprey genome alone, it is impossible to state with any certainty whether the scattered nature of the *Dlx* genes in the lamprey represents an ancestral condition of the earliest craniates or whether the lamprey represents a derived condition and that the ancestral Craniate genomic organization was Gnathostome-like. The answer to this question has significant bearing on some of the central unanswered questions about early vertebrate evolution including whether or not Cyclostomes shared the same whole genome duplication events as Gnathostomes. The only other family of extant Cyclostomes are the hagfishes. Because of the limited availability of hagfish embryos (discussed above) this organism has received relatively little attention compared with its cousin the lamprey. Recently however, advances in the captive maintenance and breeding of Japanese lamprey *Eptatretus burgeri* has yielded important insights into the developmental biology and molecular genomics of the hagfish, including the demonstration that hagfish possess true neural crest cells as are found in lamprey and Gnathostomes but not in invertebrates (Ota et al. 2007).

In order to more finely resolve the state of the genome of primitive Craniate we sought to determine the number, sequence, and organization of *Dlx* genes in the hagfish. Previously, no work had been conducted on this family of genes in this organism and no sequences were available with which begin analysis. Unlike the lamprey, no genome sequencing project is currently underway, and no EST databases were available to search for *Dlx* homologs. We therefore designed an approach to isolate novel *Dlx* genes from

this species based on a degenerate PCR cloning strategy. The *Dlx* homeobox region is a highly conserved 180 nucleotide sequence. Based on a sequence comparison of this region across a large number of vertebrates, and invertebrate chordates including the Cephalochordate *Branchiostoma floridae* and the Ascidian *Ciona intestinalis*, degenerate primers were designed to capture *Dlx*-like homeobox sequences with similarity to the sequences of previously isolated genes. This approach has previously been used to isolate *Dlx* genes from the zebrafish (Stock et al. 1996) and to the isolation of lamprey *Dlx* genes from *Petromyzon marinus* (Neidert et al. 2001). Primers were designed within the *Dlx* homeobox, which in vertebrates is divided between exons 2 and 3, in order to amplify a fragment containing most of the homeobox and the entire 2nd intron of the *Dlx* gene. Lateral muscle, heart, and liver tissue from the Atlantic hagfish *Myxine glutinosa* was kindly provided to us by Dr. Tom Moon and was used to isolate genomic DNA. Genomic DNA was subsequently used as a template to amplify *Dlx* gene fragments using the degenerate primers. A wide variety of PCR conditions were tested in order to capture as many *Dlx*-like sequences as possible. PCR fragments were cloned in pDrive and DNA was isolated in order to determine the identity of the inserts. In total over 30 independent clones were chosen and sequenced. In this manner, 4 different *Dlx*-like sequences were identified in *Myxine glutinosa*, named *MgDlxW*, *MgDlxX*, *MgDlxV*, and *MgDlxZ*. The partial homeobox sequence of these genes was used to determine which *Dlx* family each gene belonged to. It was determined that *MgDlxW* and *MgDlxV* were more closely related to the *Dlx2/3/5* subfamily along with lamprey *DlxA/B/C*, while *MgDlxX* and *MgDlxZ* belong to the *Dlx1/4/6* subfamily with lamprey *DlxD/E/F*.

These sequences, which contain partial homeobox and the complete 2nd intron of the *Dlx* gene were used as radiolabelled probes to screen a BAC library of the pacific hagfish, *Eptatretus stoutii*, kindly loaned to us by Dr. Chris Amemiya (Benaroya Research Institute, CHORI) using a method identical to that employed in screening of the *Petromyzon marinus* BAC library. After amplification and isolation of BAC DNA, our PCR screening and southern hybridization methods confirmed that we had isolated BAC clones containing each of the 4 *Dlx* genes identified in the original degenerate PCR screen, with a much lower false positive discovery rate than found when screening the *Petromyzon marinus* library. In total, 3 clones containing *EsDlxW*, 2 clones containing *EsDlxX,* 2 clones containing *EsDlxV,* and a single clone containing *EsDlxZ* were identified. Interestingly 2 of 3 clones containing *EsDlxW* overlapped with the clones containing *EsDlxX*. We therefore conclude that *EsDlxW* and *EsDlxX* are linked. Using a series of restriction digestions and southern mapping we confirmed the presence of an *EsDlxW/EsDlxX* cluster with an intergenic region approximately 9kb in size. However, consistent with the results of the lamprey analysis we found no evidence that *EsDlxZ* or *EsDlxV* were present on complimentary BACs by either PCR or southern hybridization indicating that these genes are likely unlinked. Because of the nature of the screen we originally used to identify hagfish *Dlx* genes it was impossible for us to rule out the possibility of the existence of additional *Dlx* genes in the hagfish genome. However, thorough re-screening of the *EsDlxZ* and *EsDlxV* containing BACs using degenerate primers, and low-stringency screening of the BACs by southern blotting did not reveal the presence of any other *Dlx* gene on any of the 3 BACs. Given the relatively large size of these BACs (90kb average insert), we believe that these *Dlx* genes are unlinked.

Figure 16 shows a schematic representation of the organization of *Dlx* genes in the genome of *Eptatretus stoutii.* We conclude that the organization of the *Dlx* genes in the hagfish genome is similar to the lamprey; a single bigene cluster in addition to several unlinked *Dlx* genes each belonging to both major *Dlx* subfamilies. This represents a genomic synapomorphy of the Cyclostomes and a general divergence from the condition in Gnathostomes.

Subsequent to our initial screens, our collaborators have identified several members of the *Dlx* gene family in another species of hagfish, the Japanese inshore hagfish *Eptatretus burgeri* by screening embryonic cDNA pools (Kinya Ota personal communication). In addition to orthologs of each of the 4 genes identified here, at least 2 additional *Dlx* genes were identified, bringing the total number of known hagfish *Dlx* genes to 6 as in lamprey and Gnathostomes. Based on sequences kindly provided to me by Dr. Kinya Ota, PCR based screens were performed to confirm the presence of these additional *Dlx* genes (tentatively named *EbDlx1-3-3* and *EbDlx35B*) in the genome of *Eptatretus stoutii.* While we were able to amplify fragments of these genes from genomic DNA of *Eptatretus stoutii*, we were unable to amplify fragments of these genes from our BACs containing *EsDlxW/EsDlxX*, *EsDlxV,* or *EsDlxZ.* This further indicates that a single bigene cluster seems to be highly conserved from Ascidians, through Cyclostomes, and into Gnathostomes, and that other Cyclostome *Dlx* genes are unlinked.

We characterized clones containing *Eptatretus stoutii Dlx* genes by sequencing both subcloned fragments of the BACs and inserted into pBluescript, as well as by directly sequencing the BACs. Using this technique we were able to obtain the full sequence of the protein coding region of the *Eptatretus stoutii Dlx* genes. We sought to

thoroughly characterize the BAC 2M17 which contains both *EsDlxW* and *EsDlxX*. We subcloned a 9kb *Hind*III fragment containing the full coding sequence of *EsDlxX* as well as a 5kb *Xba*I fragment containing partial *EsDlxX* sequence. By fully sequencing these clones and primer walking on the 2M17 BAC we obtained the full sequence of the intergenic region separating these two genes (See Figure 17 for a schematic). This conclusively demonstrated that these genes are organized in a tail-to-tail fashion as other *Dlx* clusters are. Interestingly, we found a sequence with high similarity to a partial transposase coding sequence flanked with characteristic direct inverted repeats of a Tc1/mariner family transposon in the intergenic sequence separating *EsDlxW* and *EsDlxX*. Members of this family of transposable elements have previously been identified in a hagfish (Heierhorst et al. 1992). The size of the intergenic region between *EsDlxW* and *EsDlxX* is approximately 9kb in size. This is slightly shorter than the 11kb intergenic distance between *LfDlxC* and *LfDlxD* (Maurya 2006).

While comparison with other known *Dlx* genes allowed us to make guesses as to the intron-exon boundaries and approximate size of UTRs we sought to obtain more precise information about the structure of these genes by cloning cDNAs for hagfish *Dlx* genes. cDNA and predicted amino acid sequences of hagfish *Dlx* genes are presented in Figure 18. In vertebrates, 4 *Dlx* genes, *Dlx1, Dlx2, Dlx5* and *Dlx6* are expressed in the developing forebrain (Liu et al. 1997) including in developing cortical GABAergic neurons (Anderson et al. 1997). It has also been reported that several *Dlx* genes are expressed in adult brain tissue, presumably acting in adult neurogenesis and maintenance of neuron-specific gene expression throughout mature cell life (Saino-Saito et al. 2003). Adult specimens of *Eptatretus stoutii* were captured off the coast of British Columbia,

Canada and kindly sent to us by Dr. Bruce Cameron (Bamfield Marine Research Station). Whole brain tissue was used to extract RNA, which was then reverse transcribed to produce adult hagfish brain cDNA pools. Based on the genomic sequences obtained from *Eptatretus stoutii* BAC clones, primers were designed to amplify the full coding cDNA of hagfish *Dlx* genes. Primers were designed upstream and downstream of the predicted start and stop codons and used to amplify cDNA extracted from adult brain RNA tissue. Using this method, cDNA from *EsDlxX*, *EsDlxV*, and *EsDlxZ* was amplified, cloned, and sequenced to determine the intron exon boundaries of the genomic BAC clones. Despite several repeat efforts with alternate primer sets I was unable to amplify cDNA from *EsDlxW* from the hagfish brain cDNA pool I had produced. Without performing more exhaustive screens of cDNA derived from other tissues known to show *Dlx* expression, most desirably from embryonic materials, it is difficult to say whether *DlxW* is in fact not expressed in adult brain or whether I simply failed to detect its expression. Because only 4 of 6 Gnathostome *Dlx* genes are expressed in adult brain tissue it would not be surprising that not all hagfish genes are expressed in adult brain. However the Gnathostome genes which are expressed in the brain, *Dlx1/Dlx2* and *Dlx5/Dlx6* are linked and show highly overlapping patterns of expression in several tissues including the brain. This is significant because while I failed to detect *DlxW* expression in adult brain tissue cDNA I was able to isolate its clustered paralog *DlxX* in this way. Because these two genes are physically linked one might expect that they share expression domains through the use of shared enhancers as occurs in the Gnathostome case. Currently it is not clear whether the hagfish *Dlx* cluster shows highly overlapping expression as Gnathostome *Dlx* clusters do and therefore the significance of the lack of

DlxW transcripts in the adult brain cDNA pool is unclear. Using the cDNA sequences of the 3 genes I was able to clone, annotation of the BACs was performed. It was confirmed that, as in the case of lamprey and all other vertebrates, hagfish *Dlx* genes consist of 3 exons, with the homeobox split between exons 2 and 3 at the highly conserved splice point between the codons encoding the amino acids: (...)TQTQ/intron/VKIW(...), a characteristic splice site of the vertebrate *Dlx* gene family. Because of the strategy employed in construction of the cDNA pool I was unable to determine the length of the 3'UTR. However, the presence of the AATAAA polyadenylation signal is a useful indication of the approximate position of the cDNA transcript's termination point. In general, we found that the introns of hagfish *Dlx* genes were significantly shorter (200bp – 2kb range in the *EsDlxW/DlxX* cluster) than in lamprey (1kb – 9kb for the *LfDlxC/DlxD* cluster), but that the overall structure of individual genes was quite similar between hagfish, lamprey, and Gnathostomes.

Interestingly while *EsDlxX* and *EsDlxZ* produced a single strong band of approximately 1.2kb in length as expected based on the design of the primers, amplification of *EsDlxV* from adult brain cDNA consistently produced two bands. One band of approximately 1.1kb corresponded to the full cDNA sequence expected. A second band of approximately 900bp was cloned and sequenced and was shown to contain an alternatively spliced version of the *EsDlxV* gene which does not contain the second exon. This transcript, henceforth described as *δEsDlxV*, lacks the second exon which encodes the majority of the DNA binding homeodomain. Detailed inspection of the sequence of *δEsDlxV* further reveals that the splice position has shifted by 1 nucleotide, a guanine shown as *//g//* in Figure 18, at the exon1-exon3 boundary causing a

frameshift which prevents any of the homeodomain encoded in the 3rd exon from being translated due to an early stop codon. The result is a cDNA transcript containing the full exon1 and exon3 sequences but lacking the homeobox entirely. Several examples of alternatively spliced homeobox genes have been documented in model vertebrates including those lacking the homeobox itself (Lòpez 2005), and it has been suggested that these truncated transcripts may be involved in the regulation of gene expression via a dominant negative mechanism, for example by recruiting and titrating cofactors. Interestingly several alternatively spliced variants of murine *Dlx*5 have been identified, including a version lacking the 2nd exon and therefore most of the homeodomain (Lieu et al. 1997). While the precise nature of the splicing changes which result in the absence of the 2nd exon in the mature transcript are different in the case of the hagfish *δEsDlxV* gene and the truncated *Dlx5* gene, the result is similar: a truncated cDNA transcript which may function to regulate the full length *Dlx* expression Both *Dlx5* and *DlxV* belong to the same subfamily of *Dlx* genes but unfortunately orthologous relationships are difficult to decipher and it is not clear whether these genes are orthologues.

Recently, our collaborators have identified orthologs of the 4 *Eptatretus stouti Dlx* genes identified here in the Japanese inshore hagfish, *Eptatretus burgeri*. Furthermore, orthologs of at least 2 more *Dlx* genes, one belonging to each major subfamily, have been identified. Further investigation of the genomic arrangement of these additional *Dlx* genes will be necessary to lend support to our hypothesis that Cyclostome *Dlx* genes are not organized in the tightly linked bigene clusters seen in Gnathostomes. Preliminary PCR based screens of the *E. stoutii* BAC clones isolated here for the additional *Dlx* genes

found in *E. burgeri*, however, seem to indicate that these genes are also not linked, in concordance with our previous results.

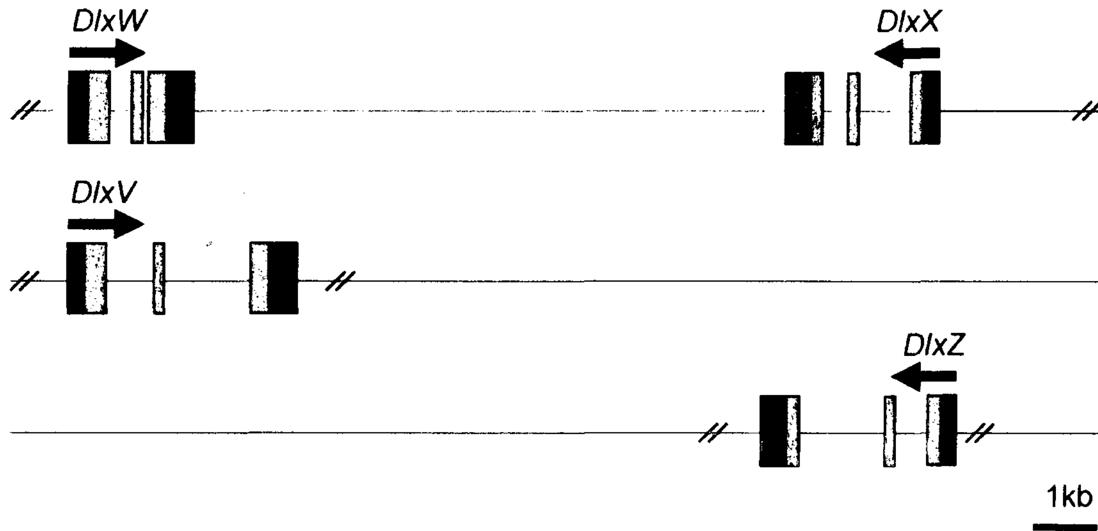


Figure 16: General Schematic of the Organization of *Dlx* Genes in the Hagfish *Eptatretus stoutii*. Schematic is drawn to scale (scalebar shown). Colourcoding of the exons corresponds to which *Dlx* subfamily each gene belongs to (blue: *Dlx2/3/5* subfamily; purple: *Dlx1/4/6* subfamily). Black boxes represent UTRs. Arrows denote transcriptional orientations.

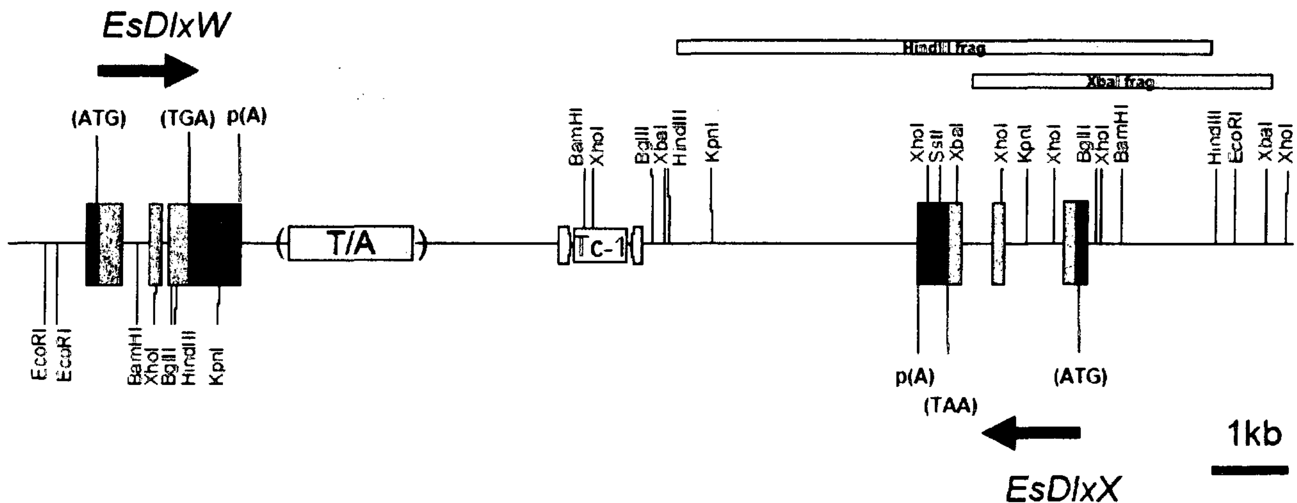


Figure 17: Detailed Structure of the *E. stoutii DlxW/DlxX* Gene Cluster. Resteriction sites are shown and the location of cloned 9kb *Hind*III and 5kb *Xba*I fragments used to resolve the cluster are shown. Note the presence of the Tc-1/Mariner repetitive element family member. The size of the T/A repeat region immediately downstream of *EsDlxW* is approximately 1.5kb but we were unable to sequence directly through this region.

>*EsDlxV* cDNA

cgccaaccccaaattaacccgtccggagcagctcgacactgagcgtttgtttctctctctctgtatttccattctcacccctcccttat
caccccctcacttttatacctgcggttacctgttatcgaacgagcacctggtcttagcctttatgtgcc**atg**acggacgtttatgactgc
ggagctccggacctacggccggaatttctcaaacgtgattcggacttcttcaacttgcagagcccccgttgctcggagagcctgcgcc
gggtgatggaggcttttattcacctcctccggccatgtaccagccgtacggatcgtactgcgcacctgcgacgtacagtcaggcggtcg
gacattactcgtacccgataccttcaacggggatgtgggcaattttgcgagcaagccgtatcagtacggctacggagctgctccgtac
ggcccgtacgaagtcggatactccagcttacccgcaccgacaaaaacgagcgaaGAA//ggaccatcgaagagcgaggagcggagacg
gccgtacgtatcgtcaacggcaagcccaagaagctccgtaaaccccgcaccatatactcgagcttccagctggcagccctgcagcggcg
ctttgagcgtactcAgtatctggcactgcccgagcgggccgagcttgccgccacactcggactcacacagacg//cagataaagatttg
gttccagaacaggcgctcaaagttcaagaagctcagcaagtcgggcgaggggattgcgccgccgcaggcgagtccgagttcgagcgacc
ccatggcgtgtaactcaccatcaccaccgactctatgggactcccccaatcccacgggcactgcagcgctgccaggtccagctttaatc
cagaactatgcgcaagctccttggtatcatagccagtacattcagcagcatcaacatcaactgcagcccttgcacagctcgcagacccc
cagtgaggtgaaccccacgccgtgcttc**tga**ccacggacccacttccgttccgtcgaacgagggactatcgtatcggacggagtcgtgc
tcttgctgcacccgcgtgccaaggagacccgtccgcacccatcgctcatgacgtgtcggtattgtgacgcgcctatgacg

>*EsDlxV* Amino Acid Sequence

MTDVYDCGAPDLRPEFLKRDSDFFNLQSPPLLGEPAPGDGGFYSPPPAMYQPYGSYCAPATYSQAVGHYSYPIPSTGDVGNFASKPYQY
GYGAAPYGPYEVGYSSLPAPTKTSEEGPSKSEERETAVRIVNGKPKK**LRKPRTIYSSFQLAALQRRFERTQYLALPERAELAATLGLTQ**
TQIKIWFQNRRSKFKKLSKSGEGIAPPQASPSSSDPMACNSPSPPTLWDSPNPTGTAALPGPALIQNYAQAPWYHSQYIQQHQHQLQPL
HSSQTPSEVNPTPCF

>*δEsDlxV* cDNA

cgccaccccaaattaacccgtccggagcagctcgacactgagcgtttgtttctctctctctgtatttccattctcacccctcccttatc
acccctcacttttatacctgcggttacctgttatcgaacgagcacctggtcttagcctttatgtgcc**atg**acggacgtttatgactgcg
gagctccggacctacggccggaatttctcaaacgtgattcggacttcttcaacttgcagagcccccgttgctcggagagcctgcgccg
ggtgatggaggcttttattcacctcctccggccatgtaccagccgtacggatcgtactgcgcacctgcgacgtacagtcaggcggtcgg
acattactcgtacccgataccttcaacggggatgtgggcaattttgcgagcaagccgtatcagtacggctacggagctgctccgtacg
gcccgtacgaagtcggatactccagcttacccgcaccgacaaaaacgagcgaagaa//**g**//ataaagatttggttccagaacaggcgct
caaagttcaagaagctcagcaagtcgggcgaggggattgcgccgccgcaggcgagtccgagttcgagcgaccccatggcgtg**taa**ctca
ccatcaccaccgactctatgggactcccccaatcccacgggtactgcagcgctgccaggtccagctttaatccagaactatgcgcaagc
tccttggtatcatagccagtacattcagcagcatcaacatcaactgcagcccttgcacagctcgcagacccccagtgaggtgaacccca
cgccgtgcctctgaccacggacccacttccgttccgtcgaacgagggactatcgtatcggacggagtcgtgctcttgctgcacccgcgt
gccaaggagacccgtccgcacccatcgctcatgacgtgtcggtattgtgacgcgcctatgac

> *δEsDlxV* Amino Acid Sequence

MTDVYDCGAPDLRPEFLKRDSDFFNLQSPPLLGEPAPGDGGFYSPPPAMYQPYGSYCAPATYSQAVGHYSYPIPSTGDVGNFASKPYQY
GYGAAPYGPYEVGYSSLPAPTKTSEEDKDLVPEQALKVQEAQQVGRGDCAAAGESEFERPHGV

>*EsDlxX* cDNA

agtgttgcggcgactgttccatatgcatcgcgcgcgagcgctcctcgcatgaacttttacccgcgcatccccgagcgcatgctcacgcc
gcaccagcgcggcccataaggtggtagcatctgagaggtcacctttgatctgatcatctatcagagccgtgcacggttgttgttatgta
ggatgccatcggtgcctggcgtgtcaatcgacc**atg**accatgagtgccgtggctgaaggtttggacgcgcccgagtgtggcaagtcggc
gttcctggagttcggccaaggttcccatcacccagggcaccatcagatgtctgcctcctccatggcgcaacatgcgcactaccagtcgc
actgcctgcccgggccaggtcccgatgcctcgtacccaacgccgacgcccccgtacggtggccgcgcacttgcctatccatacatgggt
ggtgcaggagggaacgcggcgagtgcgtacaacctgaaccctctcccatcgtaccaggcgagctccgcactcaaccagagcaggctcga
tgagcat//gatcccgagaaaacgccagtgattcagaatggggaacttcgcttcaacggaaagggaagaagatccgcaaacctcggac
catatactcgagccttcaacttcaggcgctgaaccggcgcttccagcaaacccagtacttggcgctaccggagcgggctgaactcgccg
cctcccttgggctcacccagacccag//gtgaagatttggtttcagaacaaacgctctaaatacaagaagctgatgaagcaaggaggta
gcggtctagagaacaaccctctcatggtgaacccgagctgtctgccatccagccctgccacgcagggtccagtcgccacgtgggacatc
catacgatgaacaagagcgccagcgtgcctgcagcggggtacatggcaccatacgcgcactggtactcagcggcagagacaatgcaaca
tgcgtcgctcatg**taa**tcgagagtaatggggagagggagagggcgagagaacgagaaagagagagagagagagagagaggaggcactgc
atacaaggccacgtaacgggcaccatgttacgcggacatcccgtgcaaagggtgcagagctcgaactgcttcccaccatgaatacccgt
cgttagtacctgtggcatcctgtttgtacagattaatgtgctattacgttaaagccttcctcggaagttaataatatcgagacagcagt
tgagtgatattatccgatggagcgcgcagtgaaggaatgatgcgatcacgtggtctaagtcccttcgatcatcatgc

>*EsDlxX* Amino Acid Sequence

MTMSAVAEGLDAPECGKSAFLEFGQGSHHPGHHQMSASSMAQHAHYQSHCLPGPGPDASYPTPTPPYGGRALAYPYMGGAGGNAASAYN
LNPLPSYQASSALNQSRLDEHDPEKTPVIQNGELRFNGKGKK**IRKPRTIYSSLQLQALNRRFQQTQYLALPERAELAASLGLTQTQVKI**
WFQNKRSKYKKLMKQGGSGLENNPLMVNPSCLPSSPATQGPVATWDIHTMNKSASVPAAGYMAPYAHWYSAAETMQHASLM

```
>EsDlxZ cDNA
ccctgctcttatccaagctggaatattgcaaacccctcccttcctcttcgactcctgcctcgttccccctgtgcctcctgggcgagctc
gcggcgaccatctcggtgtgcgggggcatgtatctcgatcttggtgcacgggggcatgtatttggctaatcgcggcgctagatcatgac
catgagagcaatggctgacggcttggatgtgcccgactgcggaaagtcagctttctcgagttcggacagagcggccaccctgcgcacc
cgcagatgtccgccgcgatggcgcatcatcatcatgcgcactatcctcctcagtgcctctcggcaccggggcctgatgccgcttactcg
gccccagcgtcgtcctactcccgcccgctgggctatccctacatcaaccccatcggtgggccgccgggaaacacggtcgcctcgtacgg
cctgaacccgctgcctgcttaccaggcgagcgccgcgattagccaaggcagactggaagataca//ggagacccagagaaaactgcggt
catccacaacggggagctgcgtttcaacgggaaaggcaagaagatacggaaaccccgaaccatctactcgagtctgcagctccaagcgc
ttaacaggcgcttccagcagacgcagtacctggcactacctgaacgtgccgagcttgccgcgtcactcggactcacccagacacag//g
tgaaaatttggttccagaacaagcgatccaaatacaagaagctgatgaagcaaggtggaggaggcctggaaaacaacccgctggtcaac
acgagctgtctgtcccccagcccatcagcacacccctcgattactacctgggatgtaaactgtgtgaacaagagcgcgagcatgccagc
aggtagctacatgtctgcttattcgcactggtattcgcccgctcaccaagaagccatacagcaccctgcaatcatg[illegible]agcaaacccc
ggactcagtgaatcccgtgttgttggtccaaagctgtgcagcgtgggtgaaatgcgttttcaaataaataggttaataaaaaaaatgaa
attatgagcacaagacacacttggc

>EsDlxZ Amino Acid Sequence
MTMRAMADGLDVPDCGKSAFLEFGQSGHPAHPQMSAAMAHHHHAHYPPQCLSAPGPDAAYSAPASSYSRPLGYPYINPIGGPPGNTVAS
YGLNPLPAYQASAAISQGRLEDTGDPEKTAVIHNGELRFNGKGKKIRKPRTIYSSLQLQALNRRFQQTQYLALPERAELAASLGLTQTQ
VKIWFQNKRSKYKKLMKQGGGGLENNPLVNTSCLSPSPSAHPSITTWDVNCVNKSASMPAGSYMSAYSHWYSPAHQEAIQHPAIM

>EsDlxW predicted cDNA coding sequence
atgacggacgtgtacgaaccgtgcggcagtgcggactcgcgtgtggactttctcaagcgtgaggccgagtactataatctgcacacaac
agcccctcacgccttctcgttaccggacagcgtgtctggagatgcgaccttctacccgccgcctcctccgaatgcgtactccgcgtacg
acccgtatgctccgtctccatacgggcaaagcatgggcccataccagtatccgttctcacagagctccgactccggtggatactgcgga
aagcagtaccagcaatattgcttcagtggag(NNNNNNNcggtccgtatgacacgtaccactcaagccaagcaccctcacccccaagg
aggccgatgaa)//ttgcagacgtcgcatgtcccgacagctccggccaaggaagaggaccgtgaggggtctgtgcgaatggtgggcggt
aaacccaagaagctgcggaaaccaagaaccatatactcgagcttccagctggctgccctgcagcggcgctttgagcgtactcagtacct
ggcgcttcccgagcgggctgaactcgccgcatcgttgggactcactcagacccag//gtgaagatctggttccaaaaccgtcgctcaaa
gttcaagaagcttggaaaatccggagacacaccagcaccttcccaggccagtccgagctcaagcgacccaatggcctgtaactctcccc
catccccaccgcattgggacgcacccagtccggtcggtactccaaacttgggacctatgggactgcagggctacccgcccagtgcgtgg
tactccaacctcccatacacgcagcagcagttgcagcaccaagcgggcttgccatatctccaccacacgggcagtgacgtaagtccggc
gctcgggtac[illegible]

>EsDlxW predicted partial amino acid sequence
MTDVYEPCGSADSRVDFLKREAEYYNLHTTAPHAFSLPDSVSGDATFYPPPPPNAYSAYDPYAPSPYGQSMGPYQYPFSQSSDSGGYCG
KQYQQYCFSGNNN (...) VRMVGGKPKKLRKPRTIYSSFQLAALQRRFERTQYLALPERAELAASLGLTQTQVKIWFQNRRSKFKKLGKSG
DTPAPSQASPSSSDPMACNSPPSPPHWDAPSPVGTPNLGPMGLQGYPPSAWYSNLPYTQQQLQHQAGLPYLHHTGSDVSPALGY
```

Figure 18: cDNA and Amino Acid Sequence of *E. stoutii Dlx* Genes. cDNA sequences were derived from adult brain except for *EsDlxW* which is only a partial predicted sequence based on genomic BAC clone sequencing. Start codons and stop codons are coloured green and red respectively. Intron/exon boundaries in nucleotide sequences are denoted by (//). The homeodomain of amino acid sequences is underlined.

3.3 Physical Mapping of Hagfish *Dlx* Genes using FISH

In order to determine the physical arrangement of hagfish *Dlx* genes that are not present on the same BAC clones physical mapping using Fluorescent *in situ* Hybridization to hagfish metaphase chromosomes was performed. Together with collaborators Dr. Kinya Ota and Dr. Shigeru Kuratani in Japan I developed a method for the preparation of meiotic metaphase cells from the gonads of adult male specimens of the Japanese inshore hagfish, *Eptatretus burgeri*, and performed FISH analysis to localize *Dlx* containing BAC clones to specific chromosomes. While it seems apparent that the orphan *Dlx* genes *DlxZ* and *DlxV* are not linked to each other over the average distance of a normal *Dlx* gene cluster, it still remains possible that expansion of the intergenic region has resulted in these genes being separated by a greater distance, but retain their association with the same chromosome. This would be an indication that an ancient gene cluster did exist and was dissociated over time. However, even if the genes are present on different chromosomes it does not rule out the possibility that they were once linked, as chromosomal translocation may have occurred. This type of FISH analysis may also yield insight into whether or not *DlxW* and *DlxV* are recent duplicates as indicated by phylogenetic analysis (discussed below). Recent duplicates are often present on the same chromosome and if the *DlxW/DlxX* cluster colocalizes to the same chromosome with *DlxV* this would be further evidence for the recent duplicate hypothesis.

E. *burgeri* has n=26 chromosomes in its germ cells, and n=18 chromosomes in adult somatic cells (Nakai et al. 1995). The difference in chromosome number is due to a chromosome diminution process which occurs during the first few rounds of cell division in the newly fertilized hagfish embryo and the full chromosome content is only retained

in the germ line. By examining the metaphase spreads (shown in Figure 19) we were usually able to identify 26 distinct ball-like chromosome condensations with similar appearance to Amphioxus chromosomes (Castro and Holland 2002), in agreement with previously published observations (Nakai et al. 1995). In the first experiments, we cloned a 2kb fragment of the E. *burgeri* 18S rDNA gene in order to use it as a positive control. We were able to specifically localize *E. burgeri* 18S in both interphase and metaphase nuclei (shown in Figure 20). While we did not develop a method to distinguish each individual chromosome we found that the 18S rDNA probe consistently hybridized to one of the largest *E. burgeri* chromosomes. We therefore conclude that *E. burgeri* has only a single 18S rDNA locus, like amphioxus (Castro and Holland 2002), and that this represents the ancestral Chordate condition and that the multiple 18S rDNA loci in other vertebrates are secondarily derived.

We performed FISH using BAC clones isolated from an *Eptatretus burgeri* BAC library prepared by Dr. Kinya Ota. Using the ULYSIS labeling kit (Invitrogen) probes were constructed from BAC clones containing *E. burgeri EbDlx35B* whose ortholog has not yet been cloned from *E. stouti,* and *EbDlx16A*, the ortholog of *E. stouti* gene *EsDlxZ.* We were able to localize these BACs to specific *E. burgeri* chromosomes in meiotic metaphase cells. Both *EbDlx35B* (Figure 21) and *EbDlx16A* (Figure 22, shown colocalized with *Eb18S*) localized to "medium" sized chromosomes. Colocalization of *EbDlx16A* and *Eb18S* was performed by labeling the *EbDlx16A* BAC with AlexaFluor-532 (Red) and *Eb18S* with AlexaFluor-488 (Green). We were able to demonstrate that these genes are present on separate chromosomes in *E. burgeri*. Therefore, using this method of FISH hybridization, it should be possible to colocalize multiple BACs labeled

with different fluorophores, resulting in a rough physical map of the arrangement of hagfish *Dlx* genes, and bringing to light any long range relationships which we were previously unable to detect.

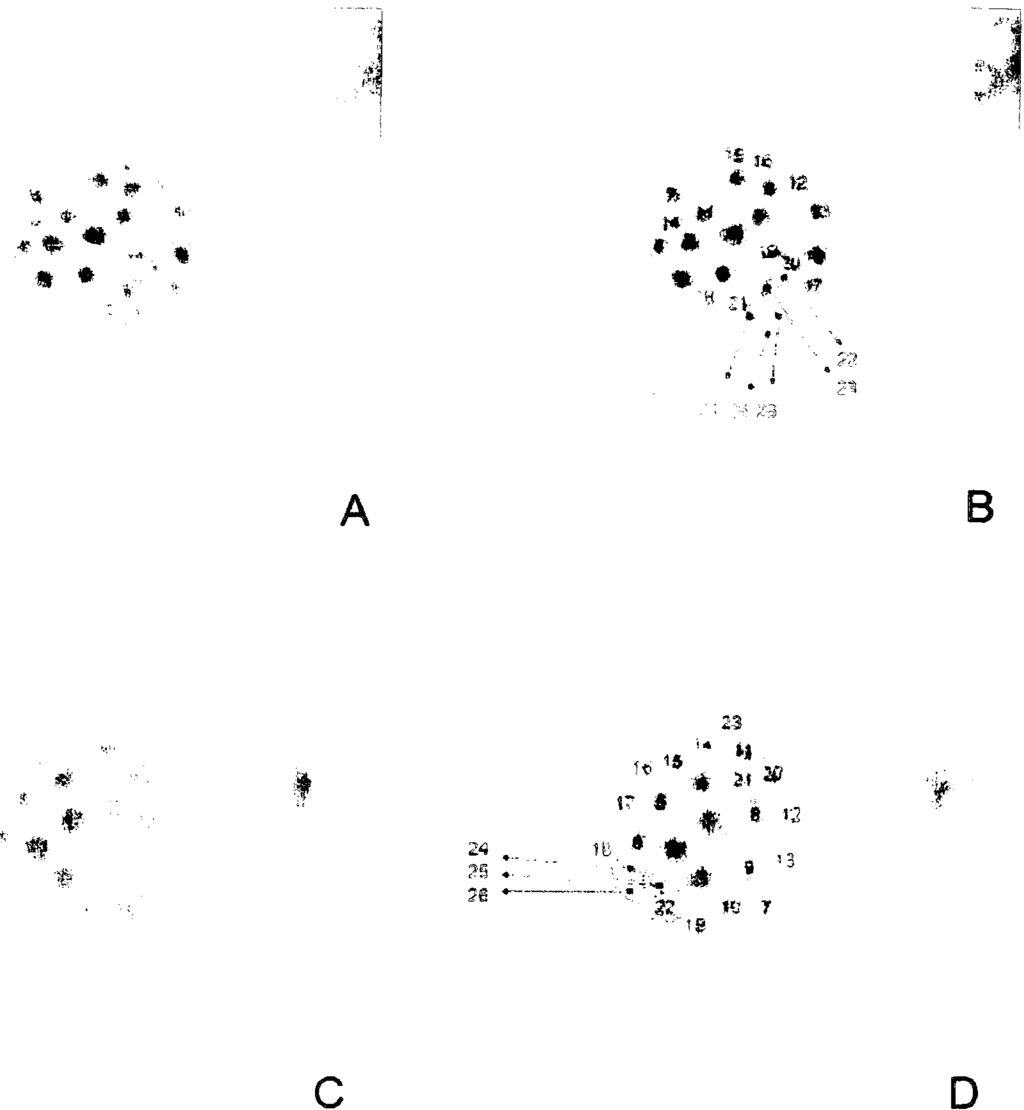


Figure 19: *E. burgeri* Metaphase Chromosome Preparations. Giemsa staining of *E. burgeri* metaphase chromosomes prepared from adult male gonads. Ball like chromosomal condensations containing both sister chromatids seen in (A, C) are numbered in (B, D) to illustrate the presence of the expected n = 26 chromosomes. These numbers do not represent specific chromosomes. Interphase cells are present on the same preparations adjacent to numerated metaphase cells.

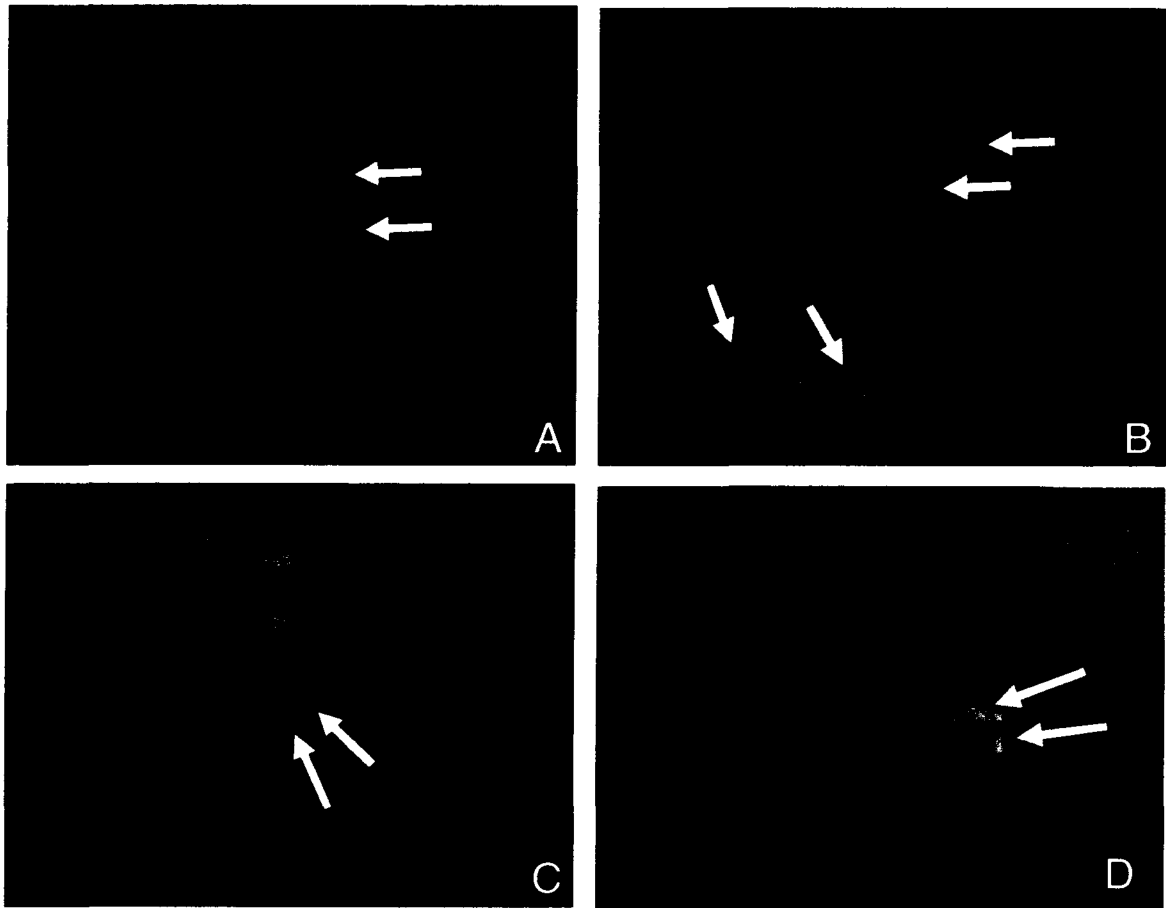


Figure 20: FISH Localization of *E. burgeri* 18S rDNA Cluster. A 2kb fragment of *Eb18S* was cloned and labeled with AlexaFluor532 (Red). Hybridization signals corresponding to the physical location of the *Eb18S* cluster can clearly be seen in both interphase (A, B) and metaphase (C, D) cells (white arrows). Note the characteristic twinspot on the large chromosome in the metaphase cells (C, D) which corresponds to separate signals from both sister chromatids. DAPI (blue) was used to counterstain chromosomal DNA.

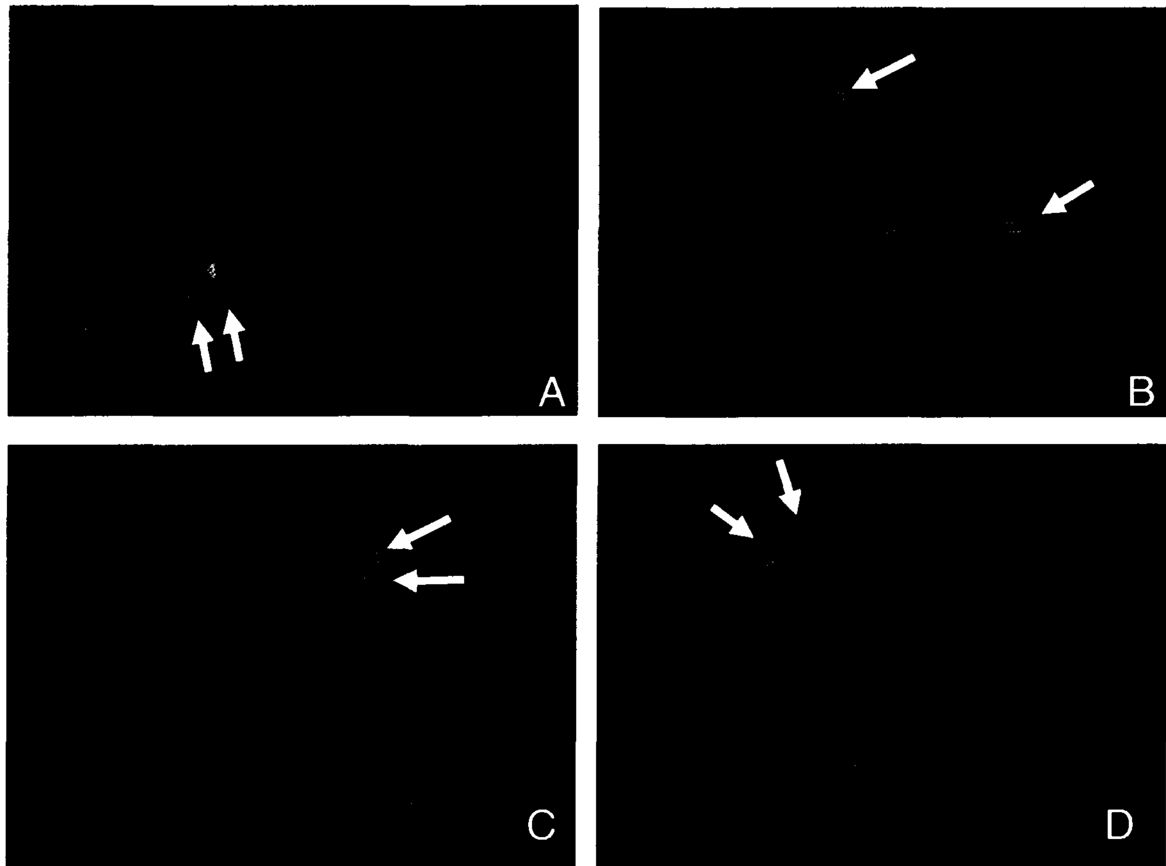


Figure 21: FISH Localization of *E. burgeri EbDlx35B*. Hybridization signals corresponding to the physical location of the *E. burgeri* BAC clone containing *EbDlx35B* can clearly be seen in both interphase (A, B) and metaphase (C, D) cells (white arrows). Note the characteristic twinspot on the large chromosome in the metaphase cells (C, D) which corresponds to separate signals from both sister chromatids. DAPI (blue) was used to counterstain chromosomal DNA.

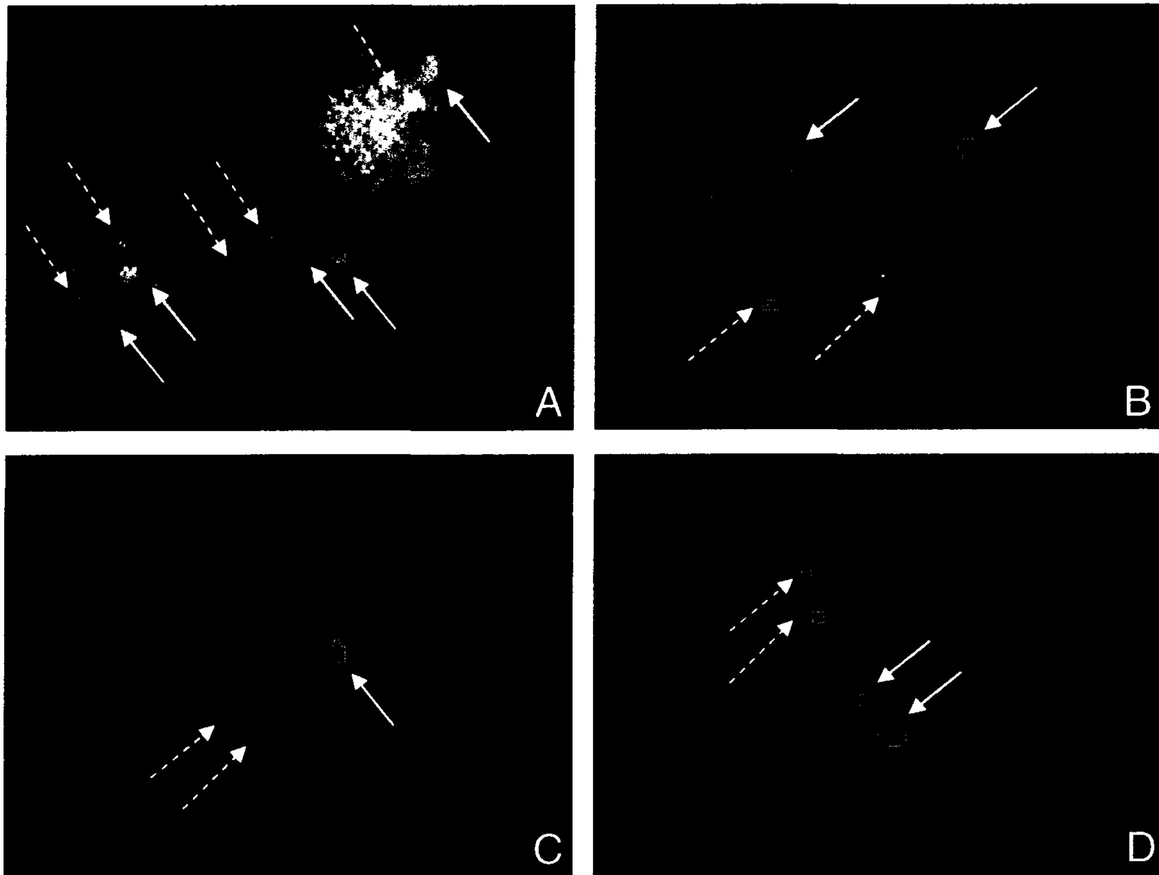


Figure 22: FISH Colocalization of *E. burgeri EbDlx16A* and *Eb18S*. Hybridization signals corresponding to the physical location of the *E. burgeri* BAC clone containing *EbDlx16A* labeled with AlexaFluor 532 (red) can clearly be seen in both interphase (A, B) and metaphase (C, D) cells (broken white arrows). Eb18S is colocalized with Eb*Dlx*16A by labelling with AlexaFluor 488 (green) and signals are correspondingly seen in both interphase (A, B) and metaphase (C, D) cells (white arrows). *EbDlx16A* and *Eb18S* do not colocalize. DAPI (blue) was used to counterstain chromosomal DNA.

3.4 Phylogenetic Analysis of Cyclostome *Dlx* Genes

In order to assign orthologous relationships between Cyclostome and Gnathostome *Dlx* genes, robust phylogenetic analysis must be performed. Currently we are unable to conclude whether or not the "orphan" Cyclostome genes were linked at some point during chordate evolution and subsequently became rearranged in or whether, apart from the ancestral bigene cluster, the additional *Dlx* genes represent independent tandem gene duplicates which have been scattered across the genome. The answer to this question has important implications for the attempt to resolve the timing of the whole genome duplication events which occurred early in vertebrate evolution. If it can be demonstrated that the additional genes are recent duplicates it would lend support to the hypothesis that Cyclostomes did not share the same rounds of genome duplication as Gnathostomes, while if it can be shown that cluster disintegration occurred and that each gene is orthologous to a specific Gnathostome gene then it would be evidence that the Cyclostomes had already experienced both rounds of whole genome duplication prior to their divergence from Gnathostomes.

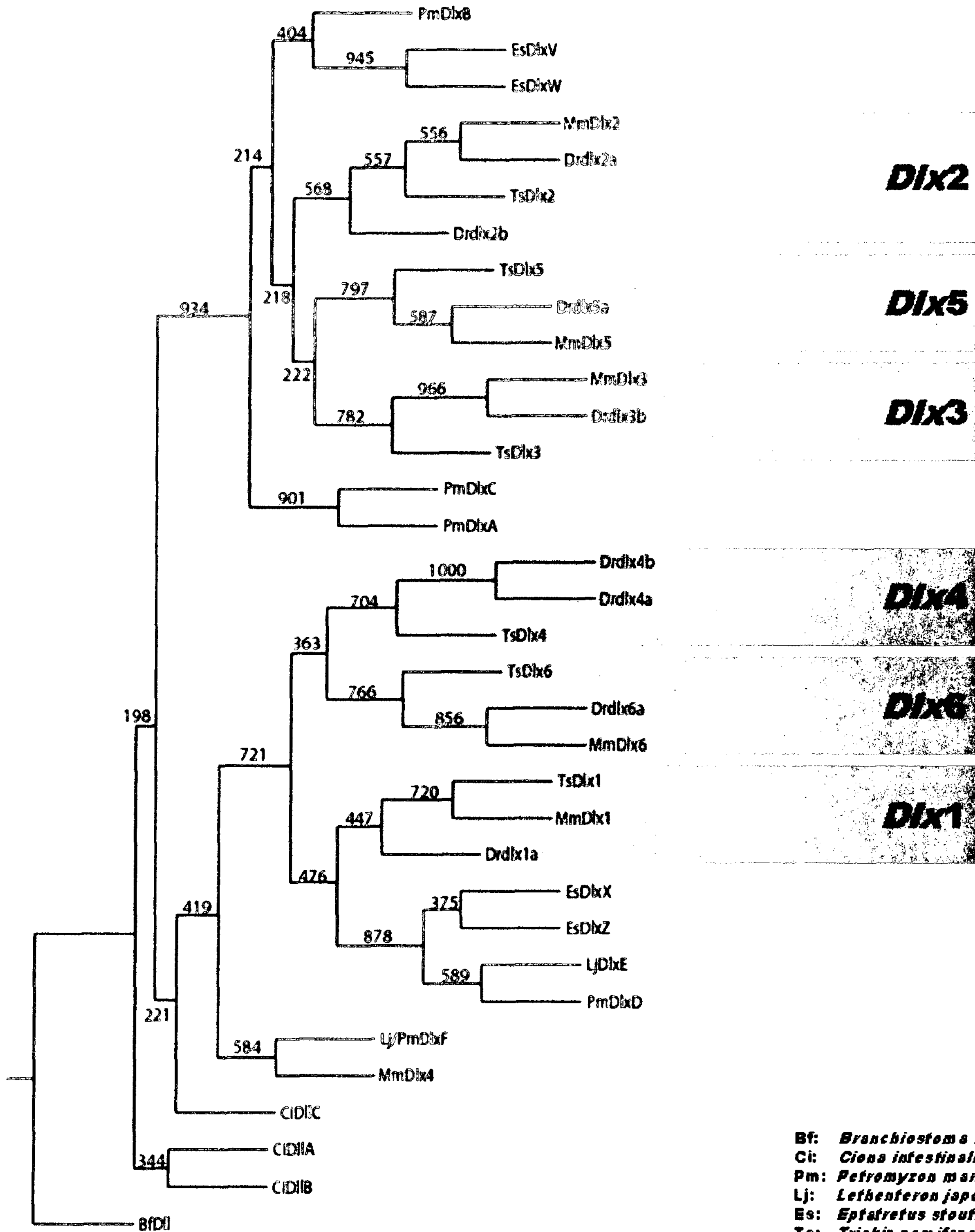
Alignments of nucleotide and amino acid sequences was performed with the ClustalX program available for download from (http://www.clustal.org/), and using the freely available PHYLIP phylogenetics package by Joseph Felsenstein available at (evolution.genetics.washington.edu/phylip.html). I conducted analyses of the relationships between several chordate *Dlx* gene family members. Due to the incomplete nature of several Cyclostome *Dlx* sequences available, and the ambiguity of splice positions in others including *EsDlxW*, all phylogenetic analyses were performed using the highly conserved 60 amino acid homeodomain region, in addition to 10 flanking amino

acids upstream and downstream, or the corresponding nucleotide sequence encoding this region. Even by examining the alignment of the homeodomain of chordate *Dlx* genes, it is evident that they can be divided into two major subfamilies, a family including Gnathostome *Dlx2/3/5* genes and one including Gnathostome *Dlx1/4/6* genes. See Figure 23 for an amino acid alignment of the homeodomain of Chordate *Dlx* genes.

Both maximum parsimony and neighbour joining methods were used in various phylogenetic reconstructions. Bootstrap support was included in order to judge the reliability of the inferred gene trees. I was able to determine the major subfamily of each of the Cyclostome *Dlx* genes but unfortunately clear 1:1 orthologies were impossible to assign due to high levels of sequence divergence between Cyclostome and Gnathostome genes. While our methods of phylogenetic reconstruction were not exhaustive, using both methods we employed we consistently found that the bootstrap support for any inferred gene tree was quite low and therefore the pairwise relationship between genes is unreliable. However, we found that lamprey *DlxA, DlxB, DlxC*, and hagfish *DlxW, DlxV* were shown to group with Gnathostome *Dlx2, Dlx3*, and *Dlx5* while lamprey *DlxD*, *Dlx*E, *DlxF* and hagfish *DlxX* and *DlxZ* were grouped with Gnathostome *Dlx1*, *Dlx4*, and *Dlx6* with high support. While the precise relationship between Cyclostome *Dlx* genes varied slightly depending on the method of phylogenetic reconstruction the general trend was clear: Cyclostome *Dlx* genes can be parsed into two major *Dlx* subfamilies but orthologous relationships cannot be determined. Because the methods used to construct the *Dlx* gene phylogenies were generally in high concordance, the Neighbour-joining method construted from the 80 amino acid homeodomain plus flanking amino acid methods is displayed in Figure 24.

Interestingly, while the bootstrap support for relationships between each Cyclostome *Dlx* gene is generally low, and therefore their precise relationships to one another unreliable, the close relationship between *EsDlxW* and *EsDlxV* consistently shows high support. This association was reconstructed 945 times out of 1000 trees produced by bootstrapping using the Neighbour-joining method with the amino acid sequence (shown in Figure 24). By examining an amino acid alignment of chordate *Dlx* genes we can see that *EsDlxV* and *EsDlxW* differ from one another by only 1 amino acid in the homeodomain (Figure 23). If *EsDlxV* and *EsDlxW* truly represent ancient duplicates derived from the same genomic event that led to the presence of 3 bigene clusters in all Gnathostomes this would indicate that incredibly strong selective pressures have acted to constrain the sequence evolution of these two genes. Pressures much stronger than between any other pair of *Dlx* genes in the genomes of any vertebrates examined. Alternatively, and more likely, these genes may be recent duplicates specific to the hagfish lineage. Likely, *EsDlxV* was derived from *EsDlxW* of the ancestral *EsDlxW/EsDlxX* gene cluster, and has subsequently been displaced elsewhere in the genome.

Figure 23: Amino Acid Level Alignment of Chordate *Dlx* Homeodomains. The 60 amino acid homeodomain of the Dll/*Dlx* gene family is highly conserved across Chordates but characteristic amino acid differences can clearly separate genes into two distinct subfamilies. Solid arrows show the position of *Eptatretus stoutii Dlx* homeodomain sequences, while broken arrows show the position of *Petromyzon marinus* or *Lethenteron japonicum Dlx* homeodomain sequences. Asterisks mark completely conserved amino acids. Dr: *Danio rerio*, Mm: *Mus musculus*, Ts: *Triakis semifasciata*, Lj: *Lethenteron japonicum*, Pm: *Petromyzon marinus*, Es: *Eptatretus stoutii*, Ci: *Ciona intestinalis*, Bf: *Branchiostoma floridae*.



Neighbor-joining, JTT, 80 sites, 1000 replicates
0.1 substitutions / site
PmDlx8
EsDlxV
EsDlxW
MmDlx2
Drdlx2a
TsDlx2
Drdlx2b
TsDlx5
Drdlx5a
MmDlx5
MmDlx3
Drdlx3b
TsDlx3
PmDlxC
PmDlxA
Drdlx4b
Drdlx4a
TsDlx4
TsDlx6
Drdlx6a
MmDlx6
TsDlx1
MmDlx1
Drdlx1a
EsDlxX
EsDlxZ
LjDlxE
PmDlxD
Lj/PmDlxF
MmDlx4
CiDllC
CiDllA
CiDllB
BfDll
Dlx2
Dlx5
Dlx3
Dlx4
Dlx6
Dlx1
Bf: Branchiostoma floridae
Ci: Ciona intestinalis
Pm: Petromyzon marinus
Lj: Lethenteron japonicum
Es: Eptatretus stouti
Ts: Triakis semifasciata
Dr: Danio rerio
Mm: Mus musculus

Figure 24: Protein Neighbour-Joining Gene Tree of Chordate *Dlx* Genes.
Gnathostome *Dlx* genes for which the two major *Dlx* subfamilies are named are shown in blue (*Dlx2/3/5* subfamily) and purple (*Dlx1/4/6* subfamily). It can be seen that both hagfish (Es: *Eptatretus stoutii*) and lamprey (Pm: *Petromyzon marinus*, Lj: *Lethenteron japonicum*) *Dlx* genes fall into one of these two major arms of the gene tree. Bootstrap support (out of 1000) is shown over each node.

3.5 Regulation of Cyclostome *Dlx* Genes

Vertebrate gene regulation occurs in part through the action of discreet genetic elements which act as either enhancers or repressors, fragments of DNA to which transcription factors and other proteins attach and either promote or suppress the transcription of the mRNA in specific tissues and at specific times during development processes and normal homeostasis (Blackwood and Kadonaga 1998). Comparative genomics has been highly successful in identifying many enhancers by searching for the most conserved region of animal genomes, looking for so-called conserved noncoding elements (CNEs), which do not encode a protein yet seem to have high selective pressures to maintain sequence similarity throughout evolution (Reviewed in Boffelli et al. 2004). Several conserved non-coding elements (CNEs) surrounding Gnathostome *Dlx* loci have been identified and functionally tested and shown to act as enhancers in tissues such as the embryonic forebrain, the apical ectodermal ridge (AER) in developing limbs, and the ectomesenchymal component of the branchial arches (Sumiyama et al. 2003, Ghanem et al. 2003, Ghanem et al. 2007). Several of these CNEs have been localized to the intergenic region between two *Dlx* genes within a cluster leading to the hypothesis that these enhancers are shared by each member of the cluster and likely contribute to the high degree of overlap between the expression domains of each gene in a cluster. These enhancers further act in coordinating a *Dlx* autoregulatory cascade in which earlier expressed *Dlx* genes subsequently activate other *Dlx* genes through the use of *Dlx* binding sites within the CNEs (Zerucha et al. 2000). These enhancers were first identified by comparing the genomes of zebrafish, fugu, mouse and humans and were shown to be highly conserved in sequence, indicating their potential functional roles in the genome.

Several whole genome wide searches for the most conserved regions of vertebrate genomes, including a search for so-called "ultraconserved elements" which are 100% identical between human and rodent genomes over more than 200bp have subsequently also found many of these same elements (Bejerano et al. 2004, Woolfe et al. 2005).

In order to study changes in the regulation of *Dlx* genes between Cyclostomes and Gnathostomes, we sought to identify orthologous CNEs in the lamprey and hagfish genomes in order to compare their sequence and ability to act as spatiotemporal enhancers of gene transcription. Because the intergenic region of Gnathostome *Dlx* clusters has been demonstrated to contain several enhancers active in the development of the brain, limbs, and branchial arches, we sought to examine the intergenic region of Cyclostome *Dlx* clusters, and the downstream region of apparently isolated *Dlx* genes in particular. Previously our lab obtained partial sequences of *Lampetra fluviatilis* cosmids containing *LfDlx*C, *LfDlx*D, and *LfDlx*A (Maurya 2006). Using PipMaker, a bioinformatics alignment program which searches through large amounts of sequence regions of high conservation, these sequences were compared with 100kb regions containing Gnathostome *Dlx* genes (Schwartz et al. 2000). However, while PipMaker successfully identified the similarity in the noncoding sequences of Gnathostomes which had previously been characterized as enhancers (Ghanem et al. 2003), no such elements were found in the *Lampetra fluviatilis* cosmids. In light of the recent availability of partial genome sequencing data from the sea lamprey, *Petromyzon marinus*, I sought to re-examine the issue of whether these CNEs were in fact present in the lamprey genome. By first searching the ensemble pre! Species web browser for all *Petromyzon Dlx* sequences, and then manually resequencing regions of ambiguity or poor sequence

quality by using BACs containing *Petromyzon Dlx* genes I had previously isolated I sought to obtain as much high quality sequence information from the *Petromyzon Dlx* loci as possible. Using the sequence of the contig found bys searching the genome browser, and by primer walking on *Petromyzon* BAC 308022 which contains *PmDlxC* and *PmDlxD* in order to fill in the many gaps or ambiguous "N" nucleotides, I was able to assemble approximately 18kb of noncoding sequence from the *PmDlx*C and *PmDlxD* cluster. Approximately 19kb of sequence downstream of the *PmDlx*A, and 7kb downstream of *PmDlxF* genes was derived entirely from the genome database. I performed Pipmaker alignments of *Petromyzon Dlx* loci with Gnathostome *Dlx* loci, and once again I found no evidence of the same CNEs present in Gnathostomes. Even manually searching the sequences for specific enhancers showed no regions of significant similarity to any of the enhancers. Our inability to locate CNEs at lamprey *Dlx* loci is surprising considering the high level of conservation of these elements across other vertebrates, which is higher than the conservation of portions of the protein coding region of the *Dlx* genes themselves. Because the *Petromyzon* genome is only partially assembled I hypothesized that these sequences may be present in the lamprey genome but incorrectly assembled by the *Petromyzon* genome project. In order to search for other contigs containing the Gnathostome *Dlx* CNEs I performed BLAST searches using mouse or zebrafish *Dlx* CNEs against the *Petromyzon* genome. Interestingly, while performing such searches with *Dlx* protein coding regions I was able to find contigs containing portions of all 6 lamprey *Dlx* genes, I was unable to find any sequences of greater than 50% similarity over 50bp to any of 6 *Dlx* CNEs searched for (I12a, I12b, I56i, I56ii, URE2, I37-2) shown to act as enhancers in previous studies.

In order to determine whether this apparent absence of *Dlx* CNEs in lamprey was specific to this family of organisms or is a shared feature of the Cyclostomes I sought to determine whether orthologs of these elements were present in the hagfish genome. Once again because of the demonstrated importance of the intergenic region in Gnathostome *Dlx* regulation and the documented reports of CNEs present in this region which act as enhancers, I focused my search for CNEs on the intergenic region between *Eptatretus stoutii DlxW* and *DlxX*. By performing sequence walking on BAC 2M17 which contains both of these genes I obtained the sequence of the entire intergenic region between the genes in this cluster. By performing PipMaker alignments with Gnathostome *Dlx* genes no sequences of greater than 50% similarity over 50bp was detected in the intergenic region of *EsDlxW/EsDlxX* and any Gnathostome *Dlx* cluster. This seems to indicate that while *Dlx* CNEs are a ubiquitous feature of the Gnathostomes and serve to coordinate tight spatiotemporal regulation of *Dlx* expression in these organisms, they are globally absent from the Cyclostomes indicating that the means of regulating *Dlx* expression in this clade may be substantially different. Because these CNEs have not been found flanking the single Amphioxus *Dll* gene, in the 1.5kb intergenic region of *Ciona Dll*A and *Dll*B nor flanking *Ciona Dll*C, and are ubiquitously present in Gnathostomes we suspect that these elements were acquired only after the divergence of Cyclostomes and represent Gnathostome genomic novelties. Alternatively, we cannot rule out the possibility that these elements were present in the common Craniate ancestor but was lost very early in the Cyclostome lineage. Interestingly the appearance of these genomic novelties surrounding Gnathostome *Dlx* genes is positively correlated with the appearance of the proposed "*Dlx* Code" in the branchial arches which serves to pattern

the proximo-distal axis of each arch, including branchial arch 1, the embryonic precursor of the jaw apparatus. It may be that only with the acquisition of these highly conserved regulatory elements that Gnathostomes were able to evolve the precise intra-*Dlx* regulatory mechanisms needed to control the spatially nested and temporally constrained expression patterns of individual *Dlx* paralogs that is necessary to pattern the complex structure of the jaw apparatus. Available evidence from the expression of *Petromyzon DlxA*, *DlxB*, *DlxC*, and *DlxD* which show that while each of the lamprey *Dlx* genes are expressed in the branchial arches of the developing organism they show no sign of a nested expression pattern and instead each show a wide expression pattern throughout both the proximal and distal components of each branchial arch, including branchial arch 1, supports this hypothesis (Neidert et al. 2001).

We sought to further refine the timing of the appearance of *Dlx* CNEs in the Gnathostome lineage in order to better correlate their appearance with the appearance of morphological novelties related to *Dlx* expression including the evolution of the jaw and paired appendages. Recently the chimaerid, *Calhorinchus milli* has been advanced as a model chondrichthyan genome (Venkatesh et al. 2005, Venkatesh et al. 2007). Chondrichthyans which include chimaerids, sharks, skates and rays are the most primitive living Gnathostomes and provide an important reference genome with which to compare the genomes of modern Gnathostomes and Cyclostomes. By searching the available 1.5X sequence of *Calhorinchus milli* for sequences orthologous to 6 of the Gnathostome *Dlx* CNEs which were previously identified (I12a, I12b, I56i, I56ii, I37-2, URE2) and for which no evidence was found in the genomes of Cyclostomes, I was able to find sequences with high similarity to 4 of the 6 sequences sought: I12a, I56i, I56ii, and

URE2. Chondrichthyans have been shown to have 6 *Dlx* genes also termed *Dlx1* through *Dlx6* to reflect the high support for their 1:1 orthologous relationships with other Gnathostome *Dlx* genes and previously published reports in the leopard shark have shown evidence that *Dlx1/Dlx2* and *Dlx5/Dlx6* form coherent gene clusters as in orther organisms (Stock 2005). Other studies in the European dogfish, *Squalus acanthurus*, have found that all 6 *Dlx* genes, including *Dlx*3/*Dlx*4 form coherent clusters (Mélanie Debiais-Thibeau and Didier Casane personal communication). While the status of the sequencing and assembly of the Elephant Shark genome is at an even more primitive state than the sea lamprey and we were unable to isolate contigs containing both *Dlx* genes and their flanking CNEs in a single contig, the high sequence similarity between Elephant Shark and Gnathostome sequences indicates they are true orthologs. Therefore, the appearance of *Dlx* CNEs occurred at least before the evolution of chondrichthyans, the most primitive Gnathostomes. This solidifies their status as Gnathostome synapomorphies.

Many other CNEs have been identified in vertebrate genomes and have been shown to act as enhancers, some over intergenic distances larger than 1Mb (Lettice et al. 2003). Interestingly, some studies have found that these CNEs cluster preferentially around developmental genes (Woolfe et al. 2005), possibly because of the need for very precise regulation of these genes during development. Whole genome comparisons with the Elephant Shark and other vertebrates have demonstrated that these elements appear to be globally present in the chondrichthyan genome, and even show more regions of similarity between sharks and mammals than between teleost fish and mammals, even though sharks are more distantly related to mammals than teleosts are (Venkatesh et al.

2005). It is of significant interest to determine whether these CNEs appeared globally throughout the Gnathostome genome prior to the divergence of chondrichthyans but after the divergence of Cyclostomes and if the lack of *Dlx* CNEs is an isolated case in Cyclostomes or whether it reflects a general property of the genome of Cyclostomata and they possess less/no vertebrate CNEs. Work in this field is ongoing but preliminary results using similar whole genome comparison methods to those employed in (Venkatesh et al. 2005) indicate that Cyclostomes share some CNEs with vertebrates, but may have considerably fewer (Byrappa Venkatesh, personal communication). In agreement with our study, these whole genome approaches in lamprey also find no evidence of CNEs similar in sequence to Gnahtostome *Dlx* CNEs (Byrappa Venkatesh, personal communication).

3.6 Transgenic Analysis of Noncoding DNA surrounding Lamprey *Dlx* Loci

The absence of CNEs surrounding Cyclostome *Dlx* loci is enigmatic considering that despite fine differences in the sub-domains of *Dlx* expression including the lack of a nested pattern in the ectomesenchyme of the branchial arches, lamprey *Dlx* genes appear to be expressed in many homologous tissues during equivalent developmental time points during embryogenesis, including the forebrain and branchial arches. In order to functionally assess the ability of lamprey *Dlx* noncoding DNA to act as enhancers and to localize the activity of lamprey enhancers which may exist but simply lack the same degree of sequence conservation observed in Gnathostomes a transgenic strategy was employed. Ashish Maurya built several transgenic constructs containing fragments of

*LfDlx*A, and *LfDlx*C/*Dlx*D tangentially linked to a fluorescent reporter construct containing *eGFP* (enhanced green fluorescent protein) in order to test the ability of lamprey noncoding regions to recapitulate domains of *Dlx* expression in the zebrafish embryo through microinjection (Maurya 2006). Because of the inability to localize specific CNEs by sequence comparison alone, we sought to test large regions of *Dlx* noncoding DNA simultaneously for its ability to recapitulate *Dlx* expression by inserting *eGFP* in frame with the first exon of *LfDlx*A or *LfDlx*C. Furthermore, because of special interest in the intergenic region of *Dlx* clusters as candidates for regulatory information, a 14kb fragment containing the whole *LfDlx*C/*Dlx*D intergenic region was subcloned and encorporated into a transgenic construct using a minimal human β-globin promoter linked to the *eGFP* reporter gene.

3.6.1 Zebrafish Transgenic Assay

I microinjected these constructs into single cell zebrafish embryos with the goal of producing stable transgenic lines in which a detailed expression analysis could be undertaken in order to compare the ability of lamprey *Dlx* noncoding DNA to recapitulate zebrafish *Dlx* expression with the ability of endogenous zebrafish *Dlx* noncoding DNA to perform this task. For the *LfDlx*A-*eGFP* construct I isolated approximately 300 embryos which showed expression of *eGFP* in regions consistent with normal *Dlx* expression. For the *LfDlx*C/*LfDlx*D-*eGFP* construct I also isolated approximately 300 embryos. Finally for the 14kb *LfDlx*C/*LfDlx*D intergenic-*eGFP*-β-globin construct I isolated approximately 200 embryos. Each of these constructs displayed expression of the *eGFP* transgene in regions consistent with endogenous *Dlx* expression in the zebrafish. Transgene

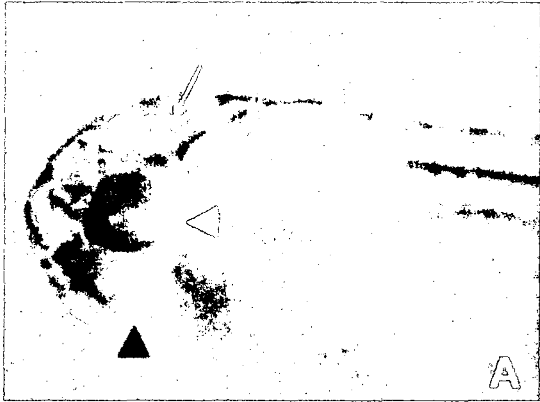
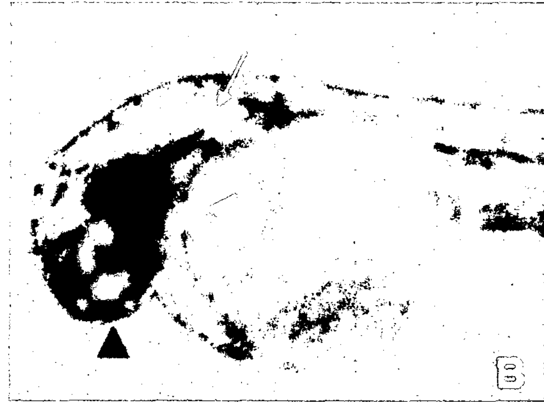
integration into the genome is not a ubiquitous event in all cells of the embryo, possibly because the first series of cleavage and division of the zebrafish is a rapid and transgene integration does not have time to integrate into the chromosome of the single cell embryo. In order to circumvent the problem of mosaicism which arises as a result, each of the strong positive primary transgenic zebrafish embryos were raised to adulthood and crossed in order to isolate animals which could stably transmit the transgene to their progeny, indicating that the transgene had integrated into the germline. Unfortunately, of over 400 animals which were screened, none were able to transmit the transgene to the next generation. Despite this setback, a careful inspection of the expression of the transgene in primary embryos can reveal some general properties of the ability of lamprey noncoding DNA to drive expression in a Gnathostome context.

3.6.1a *LfDlx*A-*eGFP* Activity in Zebrafish Embryos

The *Lampetra fluviatilis Dlx*A containing cosmid was modified in order to encorporate a fluorescent reporter gene, *eGFP*, in frame with the first exon thereby inducing the expression of *eGFP* anywhere where *Dlx*A would normally be expressed as a result of instruction from noncoding DNA present on this cosmid. This *Dlx*A-*eGFP* construct was manually microinjected into the single cell of newly fertilized zebrafish embryos and expression of the transgene was examined at 24hpf, 48hpf, and 72hpf (hpf: hours post fertilization). The *Dlx*A-*eGFP* construct drove very strong and consistent expression of the transgene in many of the same tissues in the zebrafish embryo which normally express *Dlx* genes (See Figure 25). These tissues appear primarily to be ectodermal in origin and include cells in a position consistent with the developing

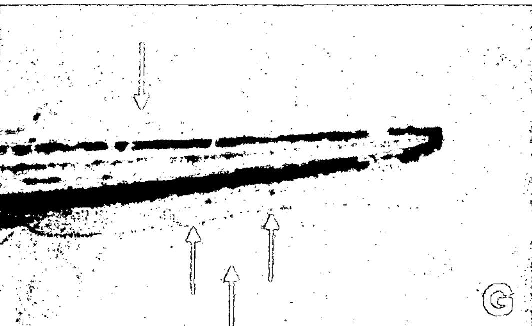
frontonasal placode (black arrowhead A), otic vesicle (white arrows A,B) in the 24hpf embryo, the branchial arches in both the 24hpf embryo and 48hpf embryo (white arrowheads A,B,C,D and ventral view E, F), and both the developing medial fin fold (white arrows G) and the pectoral fins (white arrow H). Each of these tissues is associated with expression of *Dlx* genes in the zebrafish embryo and we propose that the expression of *eGFP* driven in these tissues by the *Dlx*A-*eGFP* transgene reflects the conservation of the ability of the noncoding DNA surrounding lamprey *Dlx*A to act as tissue-specific enhancers in homologous tissues, despite the absence of overt sequence conservation between lamprey and zebrafish. Although *Petromyzon marinus Dlx*A is expressed in both the telencephalic and diencephalic domains of the embryonic forebrain, the *Dlx*A-*eGFP* construct did not drive expression of the transgene to the forebrain of the zebrafish embryo. Unfortunately, due to the highly mosaic nature of the expression pattern of the transgene in these organisms it is impossible to determine whether the *Dlx*A-*eGFP* transgene contains the necessary information to specifically restrict the expression of a specific *Dlx* ortholog to limited regions of the branchial arches in a manner consistent with the "*Dlx* code" of Gnathostomes, or whether the transgene is driven ubiquitously in the branchial arches in a more lamprey like manner. It is interesting that *DlxA-eGFP* transgene drives expression in what appears to be the otic vesicle of the zebrafish embryo, because *PmDlxA* is the only *Petromyzon Dlx* gene known to be expressed in this tissue (Neidert et al. 2001). The expression of the transgene detected in both the median fin folds and the developing pectoral fins of the zebrafish embryos is curious considering that lamprey diverged from the vertebrate lineage prior to the evolution of pectoral fins. Recently, it has been demonstrated that

many of the genetic mechanisms of both pectoral and pelvic fin development originally evolved in the median fins of primitive vertebrates (Freitas et al. 2006). We therefore believe that the ability of *LfDlx*A-*eGFP* to drive expression in both the median and lateral fins of the zebrafish reflects the conservation of regulatory mechanisms originally evolved to drive expression of *Dlx* in the median fins of the lamprey, and that the co-option of the genetic cascade responsible for fin development involved the recruitment of *Dlx* via these same regulatory elements.


A

B

D

E

F

G

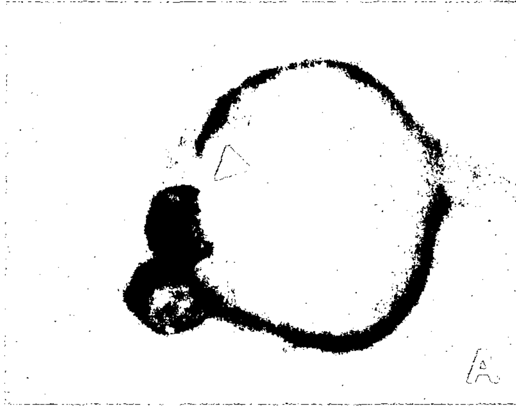
H

Figure 25: Activity of the *LfDlx*A-*eGFP* Cosmid Transgene in the Zebrafish Embryo. Zebrafish at 24 hours post fertilization (hpf) (A, B) 48 hpf (C, D, E, F, G) and 72 hpf (H) show expression of *eGFP* in many tissues associated with endogenous *Dlx* expression in the zebrafish. Expression is detected in the frontonasal placode (black arrowhead A, B), otic vesicle (white arrowhead A, B), branchial arches (white arrowheads A, B,C, D and ventral views of the same fish E, F), medial fin fold (white arrows G), and pectoral fin (white arrow H).

3.6.1b *LfDlx*C/D-*eGFP* Activity in Zebrafish Embryos

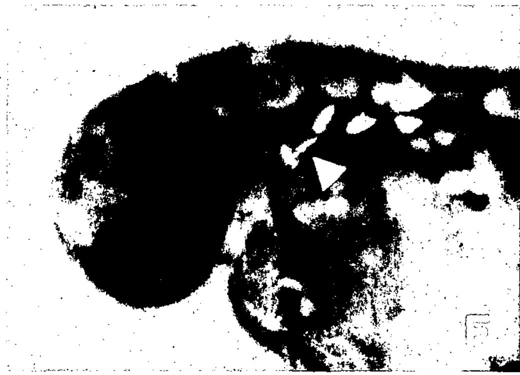
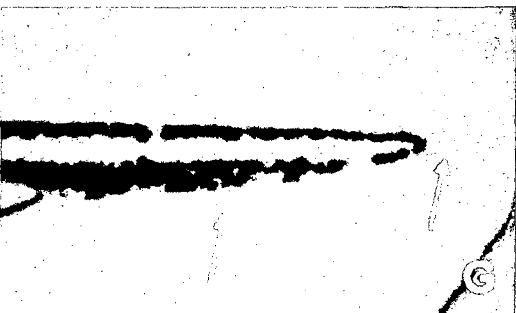
The *Lampetra fluviatilis* cosmid containing the entire coding sequence of *LfDlx*C and the second and third exons of *LfDlx*D modified by encorporating the *eGFP* reporter gene into the first exon of *LfDlx*C was microinjected into single cell zebrafish embryos and the expression of the transgene was examined at 24hpf, 48hpf, and 72hpf (shown in Figure 26). Because this construct contains the entire intergenic region of the lamprey *Dlx*C/*Dlx*D cluster it was hypothesized that it would show enhancer activity similar to the activity of the intergenic regions demonstrated from Gnathostome *Dlx* clusters. Expression of *eGFP* was detected in tissues consistent with endogenous expression of *Dlx* in the zebrafish embryo including the branchial arches (white arrowhead A, C, D, and F), the forebrain (white arrow E), the median fin fold (white arrows G), the pectoral fin (white arrow H), and the retina (white arrowhead H). These tissues are all associated with the endogenous expression of *Dlx* in the developing zebrafish embryo and we propose that the regulatory mechanisms responsible for driving *LfDlx*C/*LfDlx*D-*eGFP* in these tissues are conserved with zebrafish, despite the absence of high sequence similarity. Interestingly, there are several differences between the activity of the *LfDlx*C/*LfDlx*D-*eGFP* construct and the *LfDlx*A-*eGFP* construct. We are able to detect expression of *eGFP* in the forebrain of 48hpf zebrafish embryos whereas *LfDlx*A-*eGFP* did not drive expression in this domain. Both *PmDlx*A and *PmDlx*C transcripts are detectable in the telencephalon and diencephalons of the *Petromyzon marinus* embryo, while *PmDlx*D is detectable only in the diencephalic domain. It is unknown whether the absence of forebrain expression using *LfDlx*A-*eGFP* represents an absence of conservation in the regulatory mechanisms of this gene, or whether the region responsible

for driving forebrain expression was simply not included in the construct which was generated. Both *LfDlx*C/*LfDlx*D-*eGFP* and *LfDlx*A-*eGFP* were able to drive expression of the transgene in the developing branchial arches of the zebrafish embryo illustrating that at least basic mechanisms responsible for directing *Dlx* expression in this structure have been conserved. While quantitative measurements were not made, the overall expression levels of the *LfDlx*C/*LfDlx*D-*eGFP* construct appear to be weaker than the *LfDlx*A-*eGFP* construct. Furthermore, while the *LfDlx*A-*eGFP* construct was able to drive expression in as many as 4 discrete arches, the *LfDlx*C/*Dlx*D construct appearted to drive expression in only the first 2 arches. Furthermore, the expression of the transgene was not detected in any of the ectodermal placodes as it is in the case of the *LfDlx*A-*eGFP* transgene. Establishment of stable lines and a careful analysis of the cell types which express the transgene will be necessary in order to confirm these differences.


A

C

D



G

H

Figure 26: Activity of the *LfDlxC/DlxD-eGFP* Cosmid Transgene in the Zebrafish Embryo. 24hpf (A), 48hpf (B, C, D, E, F, G), and 72hpf (H) zebrafish embryos injected with *LfDlxC/DlxD-eGFP* show expression of the reporter gene in tissues consistent with domains of normal *Dlx* expression in the zebrafish embryo. Expression can be seen in the branchial arches (white arrowheads, 24hpf embryo A, 48hpf embryos C, D, F), the forebrain (white arrow, E, shown from ventral view), the medial fin fold (white arrows, G), the pectoral fin (white arrow, H) and the retina (white arrowhead, H).

3.6.1c *LfDlxC/D*-Intergenic-β-globin-*eGFP* Activity in Zebrafish Embryos

Because the intergenic region of Gnathostome *Dlx* genes has been shown to contain several highly conserved regulatory elements able to recapitulate *Dlx* expression in a tissue specific manner, a 14kb *Hind*III fragment containing the entire 11kb intergenic region from the single lamprey *Dlx* gene cluster, *Dlx*C/*Dlx*D was subcloned and flanked by a reporter construct, *eGFP* powered by a minimal human β-globin promoter which has minimal transcriptional activation activity on its own without the activity of flanking enhancers (Maurya 2006). Because this region is contained within the full *LfDlx*C/*Dlx*D-*eGFP* cosmid transgene it was hypothesized that the activity of the intergenic region would represent a portion of the activity of this construct. The construct was injected in single cell zebrafish embryos and the expression of the transgene was examined at 24hpf, 48hpf, and 72hpf. We were able to detect expression of *eGFP* in regions consistent with the expression of endogenous zebrafish *Dlx* in the 24hpf zebrafish embryo (shown in Figure 27) including the forebrain (arrows A, C) and the branchial arches (arrowheads A, B, C). Interestingly, expression of the transgene is greatly diminished at 48hpf compared with the expression at 24hpf, which is quite different from the case of the *LfDlx*C/*Dlx*D-*eGFP* or the *LfDlx*A-*eGFP* cosmid transgenes which maintain high activity well past 72hpf. It is possible that the intergenic region contains the necessary enhancers for driving early expression of *Dlx* in the brain and branchial arches, but not for maintaining high expression levels throughout later stages of development. A major difference in the design of the intergenic *Dlx*C/*Dlx*D-β-globin construct and the *LfDlx*C/*Dlx*D-*eGFP* full length construct is that the full length construct is driven by the endogenous lamprey *Dlx*C promoter. It may be that response elements contained within this region are

necessary for maintaining expression of *Dlx* past 24hpf, or that the cooperation between the specific *Dlx* promoter and various enhancers are necessary to faithfully recapitulate expression.

Although we were unable to locate specific quantifiable enhancer elements and test them individually due to the lack of high noncoding sequence conservation between lamprey and zebrafish, these transgenic assays demonstrate clearly that there is conservation of at least some level of regulation between Cyclostome and Gnathostome *Dlx* genes. Whether elements orthologous to the CNEs that drive Gnathostome *Dlx* genes were once present in the lamprey genome but have undergone significant sequence change making them unrecognizable, while maintaining many of the same transcription factor binding sites, or whether these elements evolved de novo in the Gnathostome lineage, and the role of driving Cylostome *Dlx* expression is fulfilled by other unrelated sequences is unknown.

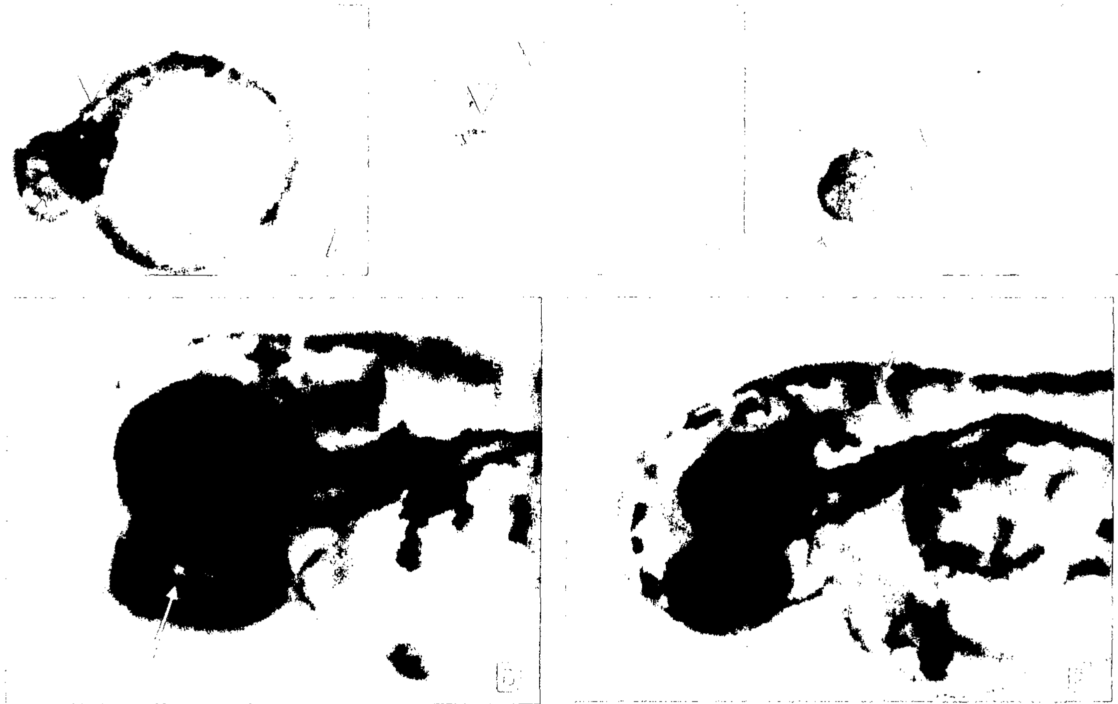


Figure 27: Activity of the 14kb *LfDlxC/DlxD*-intergenic-β-globin-*eGFP* Transgene in the Zebrafish Embryo. 24hpf (A, B, C) and 48hpf (D, E) zebrafish embryos injected with a 14kb *LfDlxC/DlxD*-intergenic-β-globin-*eGFP* transgene shows expression of the reporter gene in tissues normally associated with *Dlx* expression in the zebrafish embryo only at the 24hpf stage. Expression can be seen in the forebrain of 24hpf embryos (white arrows A, C), and the branchial arches (white arrowheads A, B, C). Expression of the *eGFP* is highly attenuated at 48hpf and only shows weak expression in the head region (white arrows D, E).

3.6.2 Lamprey Transgenic Assay

In order to draw any useful conclusions about the conservation or divergence in ability to recapitulate *Dlx* expression in the zebrafish, a baseline expression pattern in the original organism from which the DNA was derived was sought for comparison. In order to determine the endogenous ability of these lamprey constructs to drive *Dlx* expression in the host organism from which they were derived, I attempted a trial series of embryonic microinjections of *Lampetra fluviatilis* embryos in collaboration with Dr. Sylvie Rétaux (Institute Alfred Fessard CNRS, Gif-sur-Yvette France) and Dr. Sylvie Mazan (Institute de Transgenose CNRS, Orleans France). There has only been one published attempt to express transgenes in the developing lamprey embryo, and in this case a very highly conserved promoter of a teleost muscle actin gene was used to express *eGFP* in the muscle cells of the Japanese lamprey, *Lethenteron japonicum* (Kusakabe et al. 2006). Tests of lamprey noncoding DNA has been limited to analysis of lamprey promoter activity in cell culture (Takahashi et al. 2005). The embryonic microinjection of *Lampetra fluviatilis* described here therefore represents the first attempt to test endogenous lamprey noncoding sequences *in vivo* for their ability to drive expression of a transgene in a tissue specific manner. *Lampetra fluviatilis* only breeds once a year, from mid-March to mid-May, therefore any experimental manipulation of the early lamprey embryo must take place during this period, and the availability of embryos is subject to the success of collection and artificial fertilization of wild-captured adults. Furthermore, because the lamprey have a complex life cycle involving spending several years as an ammocoete larvae before undergoing metamorphosis and developing into a sexually

mature adult, it is impossible to produce stable transgenic lines using these organisms, and only mosaic primary injected embryos can be examined.

3.6.2a Activity of the CMV-*eGFP* Transgene in *L. fluviatilis* Embryos

In the first attempts to produce primary transgenic lamprey embryos, a construct containing the strong ubiquitous promoter CMV linked to the *eGFP* transgene was injected into either the single cell of the newly fertilized lamprey embryo or into one of two cells of the embryo at the two cell stage. The results of this pilot experiment are shown in Figure 28. While the first cleavage in the zebrafish embryo takes under 20 minutes, the first cleavage of the *Lampetra fluviatilis* embryo appeared to take several hours, widening the window available for microinjection. Early results of the preliminary trials were encouraging and expression of *eGFP* was observed in blastula stage embryos several hours after fertilization (A, B). Lamprey development is considerably slower than zebrafish development and by the time of head emergence (stage 21) between 4 to 5 days had passed. At this stage expression of the CMV-*eGFP* transgene was detected in a nonspecific mosaic fashion throughout the entire embryo. It was at this stage, however, that significant mortality of injected lamprey embryos was detected. As many as 90% of injected embryos died before they reached the stage of head emergence, most as a result of the microinjection process, but some also due to bad fertilizations which can be detected by monitoring the survival rate of uninjected control embryos from the same clutch of eggs. Despite several weeks of attempts to optimize the protocol, I was never able to reduce the mortality rate beyond 80 – 90 % at this stage of development. Furthermore, it was observed that the development of injected lamprey embryos was

significantly retarded compared with uninjected control embryos. Nevertheless, some CMV-*eGFP* injected embryos did survive to stage 26 (C: brightfield, D), the stage where (Neidert Ref) detected the expression of 4 *Petromyzon marinus* transgenes ubiquitously throughout the branchial arches, and when expression of *Dlx* transgenes would yield relevant information pertaining to their ability to faithfully recapitulate endogenous *Dlx* expression in the lamprey. Only 4 such CMV-*eGFP* injected embryos survived to this stage, one of which is depicted in (C, D) and shows expression of *eGFP* in a nonspecific manner throughout the yolk (white arrowheads), and in muscle cells (white arrows).

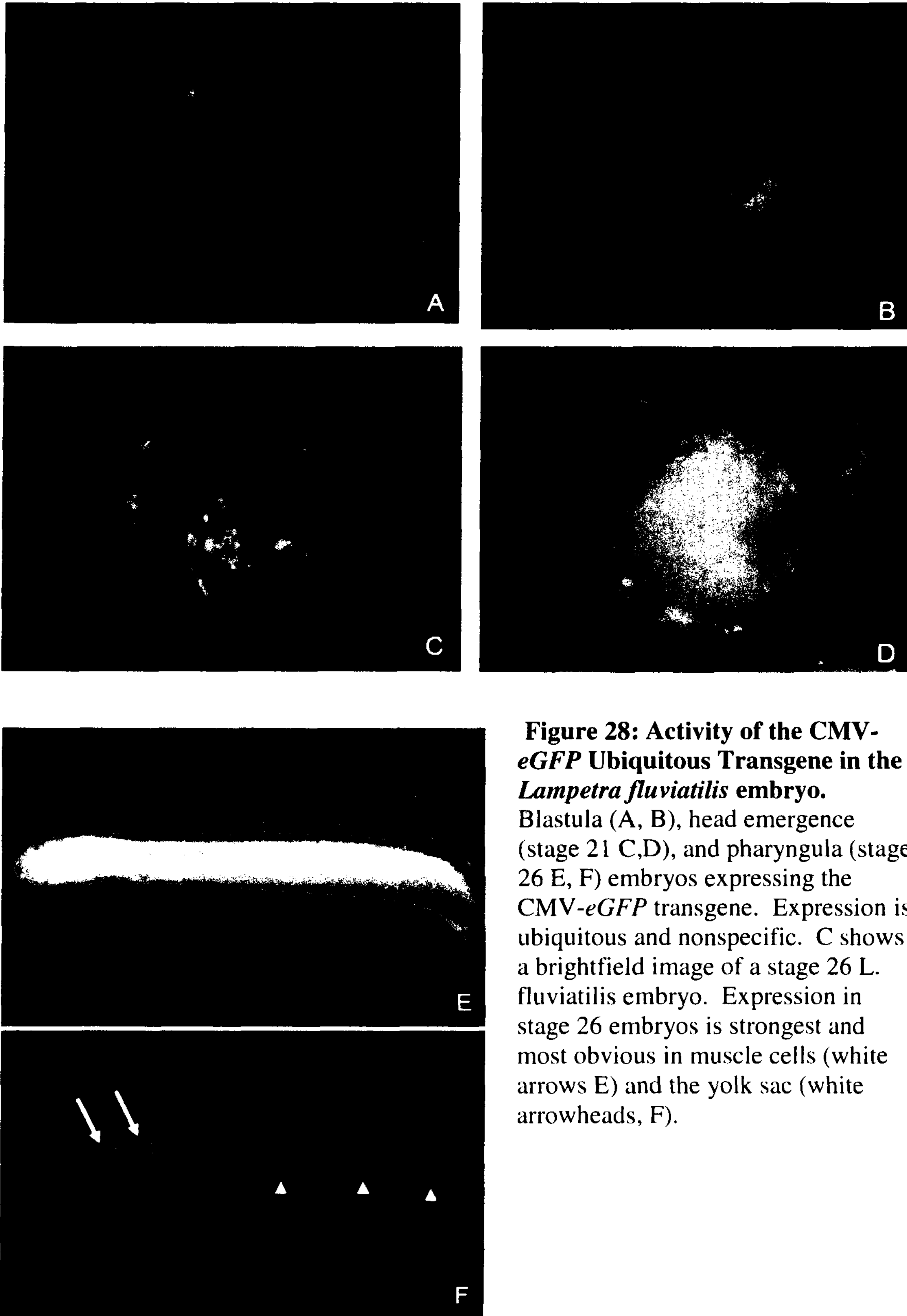


Figure 28: Activity of the CMV-*eGFP* Ubiquitous Transgene in the *Lampetra fluviatilis* embryo. Blastula (A, B), head emergence (stage 21 C,D), and pharyngula (stage 26 E, F) embryos expressing the CMV-*eGFP* transgene. Expression is ubiquitous and nonspecific. C shows a brightfield image of a stage 26 L. fluviatilis embryo. Expression in stage 26 embryos is strongest and most obvious in muscle cells (white arrows E) and the yolk sac (white arrowheads, F).

3.6.2b *LfDlx*A-*eGFP* Activity in Lamprey Embryos

The *Dlx*A-*eGFP* cosmid construct was microinjected into single or two cell stage embryos of *Lampetra fluviatilis*. This experiment suffered from the same problems of high mortality and embryonic deformity as the trial injections of the CMV-*eGFP* construt. Nevertheless, it was possible to detect the expression of the *eGFP* transgene in blastula and head emergence stage lamprey embryos indicating that the combination of the *LfDlx*A promoter and surrounding noncoding DNA was able to drive transcription of the transgene (shown in Figure 29). Expression was first detectable began at blastula stage (A, B) but no specific pattern was discernable. While detectable levels of *eGFP* were expressed in the head and branchial arch regions of the developing lamprey embryo at stage 21 (C, D) we also noticed high levels of what we believe is nonspecific expression of the transgene in other cells of the embryo especially the yolk, hindering our ability to make conclusions about the specificity of the expression of the transgene in the head region. Also, because no embryos injected with this construct survived past the intial stages of head emergence we were unable to compare the expression of the *LfDlx*A-*eGFP* transgene with the published expression patterns of *Petromyzon Dlx*A.

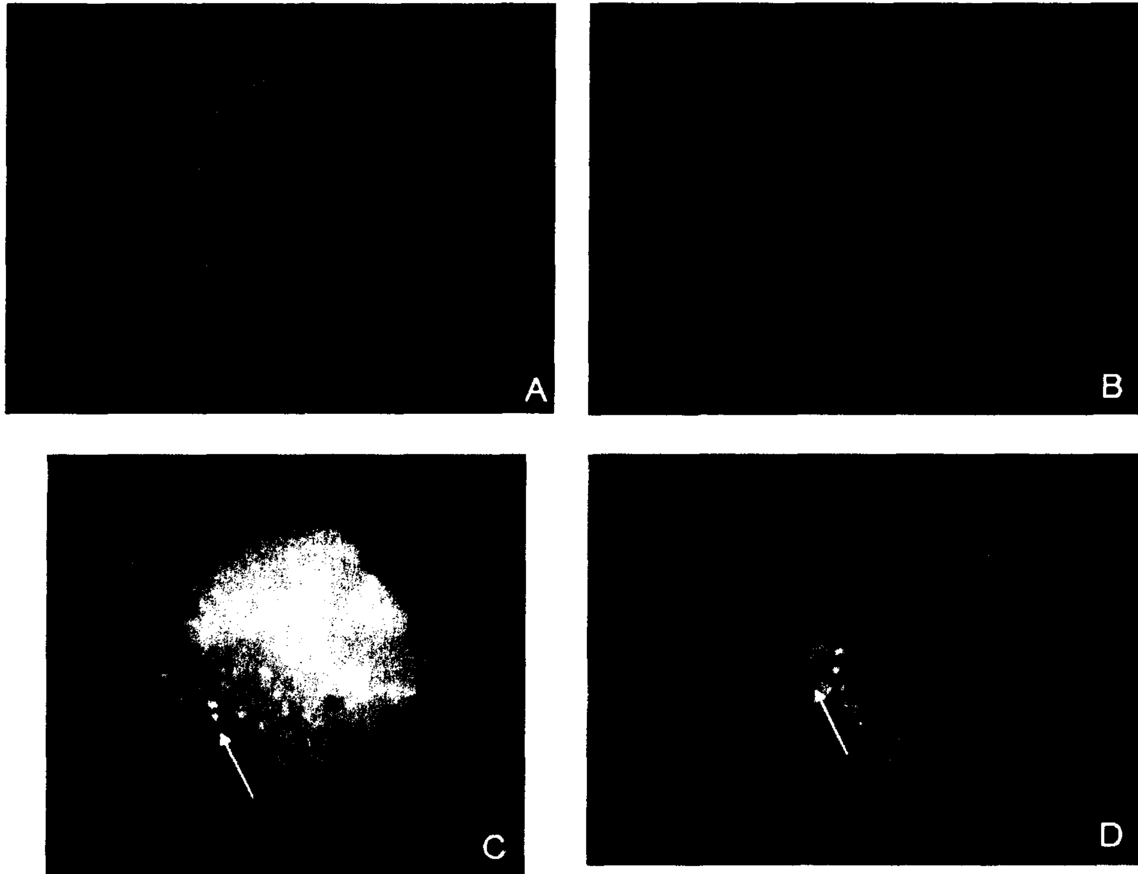


Figure 29: Activity of the *LfDlxA-eGFP* Cosmid Transgene in the *Lampetra fluviatilis* embryo. Blastula stage embryos (A, B), and head emergence stage embryos (C, D) showed expression of the *eGFP* transgene. While expression in head emergence stage embryos overlaps the area of the developing branchial arches (white arrows C, D) the high levels of nonspecific expression make this observation unreliable.

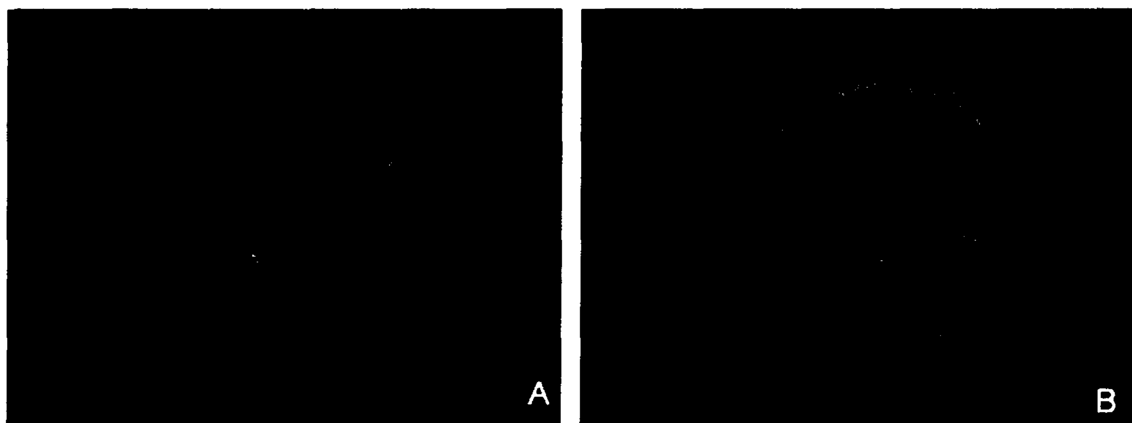


Figure 30: Activity of the *LfDlxC/DlxD-eGFP* Cosmid Transgene in the *Lampetra fluviatilis* embryo. Expression in blastula stage embryos (A, B) can be seen.

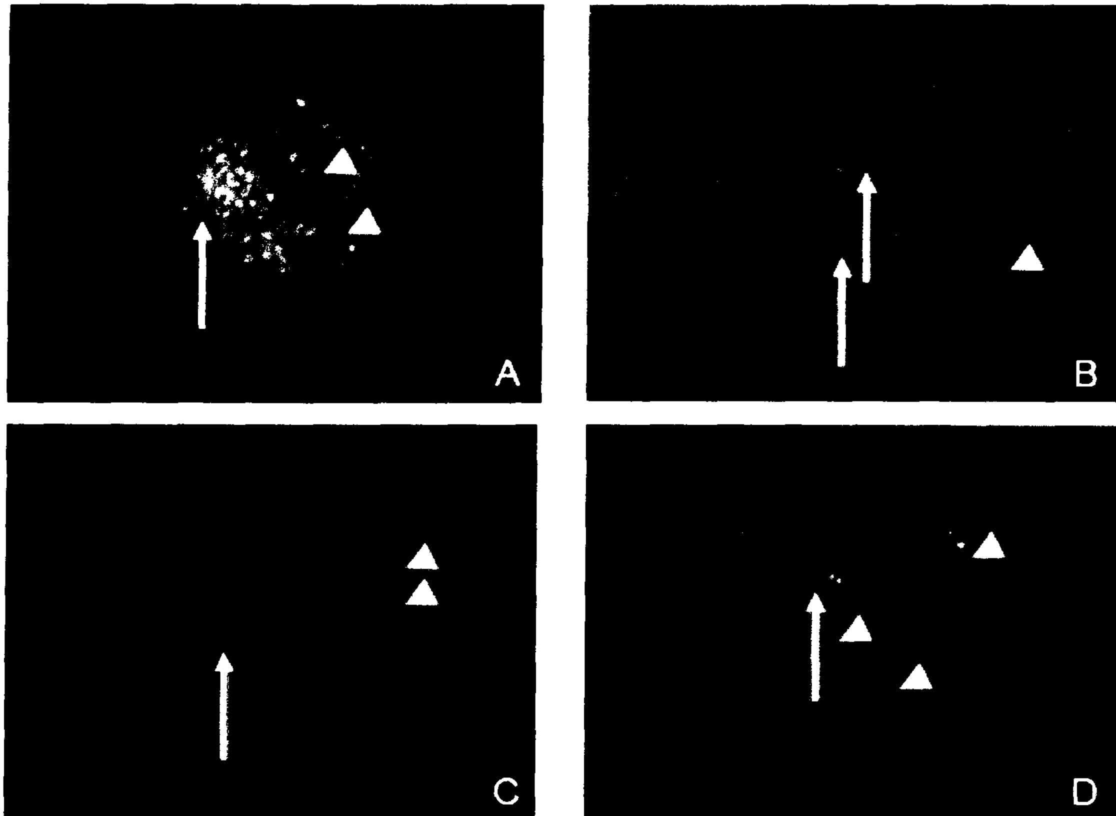


Figure 31: Activity of the zebrafish 6kb *Dlx1a/Dlx2a*-intergenic-β-globin-*eGFP* Transgene in the *Lampetra fluviatilis* Embryo. Expression of the transgene was not detectable in blastula stage embryos. Head emergence stage embryos show expression of the transgene in regions consistent with the developing branchial arches (white arrows A, B, C, D) but also show expression in surface ectodermal cells and yolk (white arrowheads A, B, C, D).

Section 4: Discussion

The novel information presented in this thesis pertaining to the number and genomic organization of *Dlx* genes in Cyclostomes fills a gap in our understanding of how the ancestral invertebrate *Dll* gene duplicated to form the multigene *Dlx* family in extant vertebrates. This work presents evidence for 6 *Dlx* genes in the lamprey *Petromyzon marinus* and 4 *Dlx* genes in the hagfish *Eptatretus stoutii*, while our collaborators provide evidence for at least 6 *Dlx* genes in the closely related *Eptatretus burgeri.* This thesis further demonstrates that only a single *Dlx* gene cluster exists in each of these animals and that the remaining 4 *Dlx* genes appear to be "orphans" in the genome. Figure 32 summarizes the known organization of the *Dlx* gene complement of Chordates and the predicted timing of major gene and genome duplication events. The genomic organization of these genes is in stark contrast to the situation in Gnathostomes where 3 *Dlx* bigene clusters are invariably present. It is striking, however, that the 6 *Petromyzon marinus Dlx* genes and 4 *Eptatretus stoutii Dlx* genes that I examined fit the phylogenetic pattern of Gnathostome *Dlx* genes by segregating in equal numbers into two distinct subfamilies. If Cyclostomes shared the same two rounds of whole genome duplication as Gnathostomes, one would expect this particular phylogenetic relationship between the different genes, but also that the physical arrangement of these genes would be Gnathostome-like. With the information available now it is only possible to speculate whether one of several scenarios truly accounts for the complement of *Dlx* genes found in the Cyclostome genome. It may be that Cyclostomes did not share the same rounds of whole genome duplication which resulted in the present complement of 6 *Dlx* genes in Gnathostomes and that the presence of multiple members of each *Dlx* subfamily may be

the result of independent tandem gene duplications. Evidence in support of this hypothesis is derived from the fact that only a single bigene cluster is present in Cyclostomes, while the other *Dlx* genes appear orphaned in the genome. This may reflect a condition similar to that of *Ciona intestinalis* whose genome is comparatively well characterized and has been shown to possess a single *Dlx* bigene cluster as well as an orphan *Dlx* gene derived from independent tandem duplication (Caracciolo et al. 2000, Irvine et al. 2007). Furthermore, the phylogenetic analysis presented here shows that *EsDlxW* and *EsDlxV* are very closely related to each other in sequence, much more so than would be expected for duplicates whose origin dates back between 450 and 500Mya which is the proposed time of the divergence between Cyclostomes and Gnathostomes and the earliest possible time that any shared duplication events could have occurred (Kuraku and Kuratani 2006). Therefore, unless some very high level of selection has maintained the sequences of the *EsDlxW* and *EsDlxV* genes very similar, it is likely that these genes are lineage specific duplicates which have arisen relatively recently in hagfish evolution. Interestingly, because our initial screen for hagfish *Dlx* genes was carried out in *Myxine glutinosa* and we identified orthologs of both of these genes in this species, this places the origin of these proposed recent duplicates at least prior to the divergence between the *Myxine* and *Eptatretus* genera estimated at approximately 40-90Mya (Kuraku and Kuratani 2006).

The second major scenario which can account for the present complement of *Dlx* genes in the Cyclostome genome is that Cyclostomes shared the same rounds of whole genome duplication which took place in the Gnathostome lineage, resulting in equal numbers of *Dlx* genes which fit a consistent phylogenetic pattern of segregation into two

distinct subfamilies. In order to believe this hypothesis, however, it is necessary to imply that following duplication of the ancestral *Dll* bigene cluster, two clusters were broken in the Cyclostome lineage following their divergence from the lineage leading to Gnathostomes. This would explain why an equal number of *Dlx* genes are present in both Cyclostomes and Gnathostomes, and why the phylogenetic relationship of Cyclostome *Dlx* genes fits the Gnathostome pattern. The tight association between pairs of *Dlx* genes has been conserved from Chondrichthyans to Tetrapods and if this is the case it would be the first documented evidence of any *Dlx* gene cluster being broken and scattered in the genome. It is not inconceivable that the association between linked *Dlx* genes have been broken in the Cyclostome lineage. While other highly conserved homeobox gene clusters such as the *Hox* genes have maintained extremely tight associations with one another throughout the vertebrates, possibly through the action of shared regulatory elements (Kikuta et al. 2007, Lee et al. 2006), there are several documented cases of cluster breakup. The *ParaHox* cluster, originally identified as the evolutionary sister of the *Hox* cluster in the Cephalochordate, Amphioxus (Brook et al. 1998), has been found to have broken following the 3R event specifically in the teleost lineage (Mulley et al. 2006). The *ParaHox* genes have also been investigated in the hagfish where it was found that *a ParaHox* cluster was present, but altered by the pseudogenization of the central *Xlox* gene which is universally preserved in the *ParaHox* clusters of other vertebrates (Furlong et al. 2007). There seems to be some lineage specific properties which constrain certain gene clusters in either vertebrates or invertebrates. For example, while the *Nkx* gene cluster is highly constrained in protostomes, it has been broken in the chordate lineage and exists as fragmented micro-

clusters in both Amphioxus and humans (Luke et al. 2003). It has even been known for some time that while it is widely misconceived that highly constrained *Hox* clusters are a universal trait of deuterostomes, it has been secondarily broken in several cases specifically in invertebrates, including being split into two subcomplexes in Drosophilid flies (Negre and Ruiz 2007), and atomized into small micro-clusters in the chordate *Ciona intestinalis* (Ikuta et al. 2004). Denis Duboule recently hypothesized that one reason for the preservation of *Hox* complexes in vertebrates may be the fact that they have been duplicated (Duboule 2007). According to this theory, the duplication of these developmentally important transcription factors required the genome to evolve complex and very precise mechanisms of regulating the multiple members at the level of the cluster itself in order to prevent their misexpression in the wrong tissues at the wrong times which could have disastrous consequences on development including homeosis. It has been believed for some time that one potential reason for the conservation of tightly linked genes within gene clusters is that genes share regulatory elements and the disruption of their linkage would lead to the disregulation of one or several genes of the complex (Lee et al. 2006, Kikuta et al. 2007, McEwen et al. 2006). Evidence for shared regulatory regions have been found in the *HoxD* complex where a so-called Global Control Region (GCR) is partially responsible for coordinating the nested expression patterns of 5' *HoxD* genes (Spitz et al. 2003). Several Conserved Noncoding Elements (CNEs) have been found in the intergenic region located between Gnathostome *Dlx* genes and are spaced nearly equidistant between both genes in a cluster. These elements have been shown to act as tissue specific enhancers, and it has been proposed that they are likely shared between both genes, leading to their coordinate expression in space and

time (Ghanem et al. 2003, Ghanem et al. 2005). Because several *Dlx* autoregulatory response elements have been located within these intergenic enhancers it is believed that they also serve to coordinate the sequential activation of *Dlx* genes in time and may well serve a role in establishing the proximodistally nested expression pattern of *Dlx* genes in tissues such as the branchial arches, leading to the establishment of the "*Dlx*-code". It is therefore significant that thorough searches of the intergenic region between Cyclostome *Dlx* gene clusters, as well as the downstream region of some of the orphan Cyclostome *Dlx* genes showed no evidence of the presence of these CNEs which are ubiquitously present in Gnathostomes including Chondrichthyans. The presence of these CNEs in the genome of Gnathostomes is correlated with properties of tightly coordinated expression of linked *Dlx* genes, and a nested expression pattern of these linked genes in the proximodistal axis of the developing branchial arches (See Figure 33). I believe it is likely that these CNEs which have been shown to act as enhancers, likely on both genes within a cluster, and which contain *Dlx* autoregulatory regions for the coordination of a *Dlx* expression cascade, are major contributors to these properties of coordinated and nested expression in Gnathostomes. The absence of these CNEs at Cyclostome loci may be a major reason for the differences in gene expression that are seen. In support of this idea, the total lack of overlap between the expression of linked *Ciona intestinalis Dll* genes is also correlated with a lack of vertebrate CNEs in its intergenic region.

It is possible that there is a relationship between the dispersed genomic organization of Cyclostome *Dlx* genes and the absence of *Dlx* CNEs. Assuming that multiple *Dlx* clusters were at one time present in Cyclostomes, one hypothesis is that the absence of these CNEs may have alleviated the strong purifying selection necessary to

maintain gene clusters intact over millions of years, resulting in the decay of the *Dlx* clusters. This theory, however, does not explain while one gene cluster has been maintained intact in both branches of the Cyclostome lineage (*Dlx*C/*Dlx*D in lamprey and *Dlx*W/*Dlx*X in hagfish). It may be simple chance that one bigene cluster remains in each cyclostome genome, or there may be other forces acting to constrain the cluster which have not yet been identified. It may be significant, however, that while the intergenic region of Gnathostome *Dlx* genes is free of recognizable transposon insertions, the *EsDlxW/EsDlxX* gene cluster is interrupted by the insertion of a Tc1/mariner family transposon. Transposon insertions are rare within gene clusters such as the *Hox* genes presumably because of their negative effects on regulation. If the *EsDlxW/EsDlxX* cluster was maintained intact over millions of years because of shared regulatory elements in the intergenic region, as proposed for Gnathostome *Dlx* genes, we would not expect to find repetitive element insertions in this region.

The genomic organization and absence of CNEs of Cyclostome *Dlx* genes is compatible with the theory of Duboule discussed above if Cyclostomes diverged from the lineage leading to Gnathostomes after the genome duplications but before the evolution of the CNEs which tightly regulate *Dlx* expression and may prevent the decay of the clusters. It will be interesting to expand upon the study of noncoding conservation in the genome of Cyclostomes and examine whether or not other highly conserved Gnathostome CNEs, of which several thousand have been identified, depending upon the method used (Woolfe and Elgar 2007, McEwen et al. 2006, Muller et al. 2002, Bejerano et al. 2004, Woolfe et al. 2005), are present or not in the genome of Cyclostomes. It has already been demonstrated that the conservation of *Dlx* CNEs in the genome of the

Elephant Shark, *Calhorinchus milli* is a reflection of a broader conservation of many of the same CNEs previously identified in the genomes of other Gnathostomes (Venkatesh et al. 2005, Venkatesh et al. 2007). It is possible that the global appearance of developmental gene associated CNEs is a Gnathostome specific feature and may very well have played a role in the evolution of a huge number of Gnathostome novelties in addition to those specifically associated with *Dlx* expression. A global survey of the genome of multiple Cyclostomes for the presence of CNEs, as well as more thorough understanding of how these CNEs act in the regulation of Gnathostome genes including the *Dlx* clusters will be necessary in order to begin thoroughly testing this hypothesis.

The absence of CNEs surrounding the Cyclostome *Dlx* genes is curious considering that these elements are ubiquitously present in the genomes of all other vertebrates, and their demonstrated ability to act as important *Dlx* enhancers in many embryonic structures also present in Cyclostomes.. This scenario is similar to the case of *Ciona intestinalis* whose single *Dll* cluster with an intergenic region of only 1.5kb does not show any high sequence similarity to the intergenic region of Gnathostome *Dlx* genes (Ikuta et al. 2004). Nevertheless the results presented herein support the idea that some level of regulatory information is conserved between Cyclostomes and Gnathostomes. The ability of *Lampetra fluviatilis* noncoding DNA surrounding *Dlx* loci to recapitulate several aspects of endogenous zebrafish *Dlx* expression in primary transgenic zebrafish embryos demonstrates that the information necessary to drive *Dlx* expression in homologous tissues including the branchial arches, forebrain, and fins is conserved, despite our inability to find any sequences conserved more than 50% over 50bp between the genomes of Cyclostomes and Gnathostomes at *Dlx* loci. The ability to identify

enhancers based on sequence conservation is a powerful tool and has led to the successful identification of important tissue specific enhancers responsible for coordinating expression of several developmentally important genes by testing them in transgenic organisms including teleost fish (Reviews in Muller et al. 2002, Boffelli et al. 2004). However there has been recent evidence demonstrating that high sequence conservation is not a necessary prerequisite for conservation of enhancer function, such as the recent example from the zebrafish and human RET loci (Fisher et al. 2006). Because enhancers are thought to act as a staging platform for the cooperative binding of a number of transcription factors and cofactors there may be much higher lability in the order of the transcription factor binding sites within an enhancer compared with the linear order of codons in an mRNA sequence, and changes in the order of binding sites within the enhancers, as well as the relative orientation and position of the enhancers in the genome with respect to the genes they regulate may be prevalent features chordate genomes (Sanges et al. 2006). It may therefore be inappropriate to talk about the appearance of *Dlx* CNEs in the Gnathostome lineage, or their loss in Cyclostomes. Sequence originally shared in the Craniate ancestor may simply have been shuffled so extensively as to eliminate its conservation with Gnathostome sequence. In this case the element is not so much lost as it is simply changed. Alternatively, multiple disorganized binding sites scattered over a few kb could have been congealed in into discreet continuous units in Gnathostomes but not in Cyclostomes. In either case it is still a provocative mystery what the evolutionary forces responsible for the molding of the *Dlx* genomic landscape were and why they differed so dramatically in the Cyclostome and Gnathostome lineages.

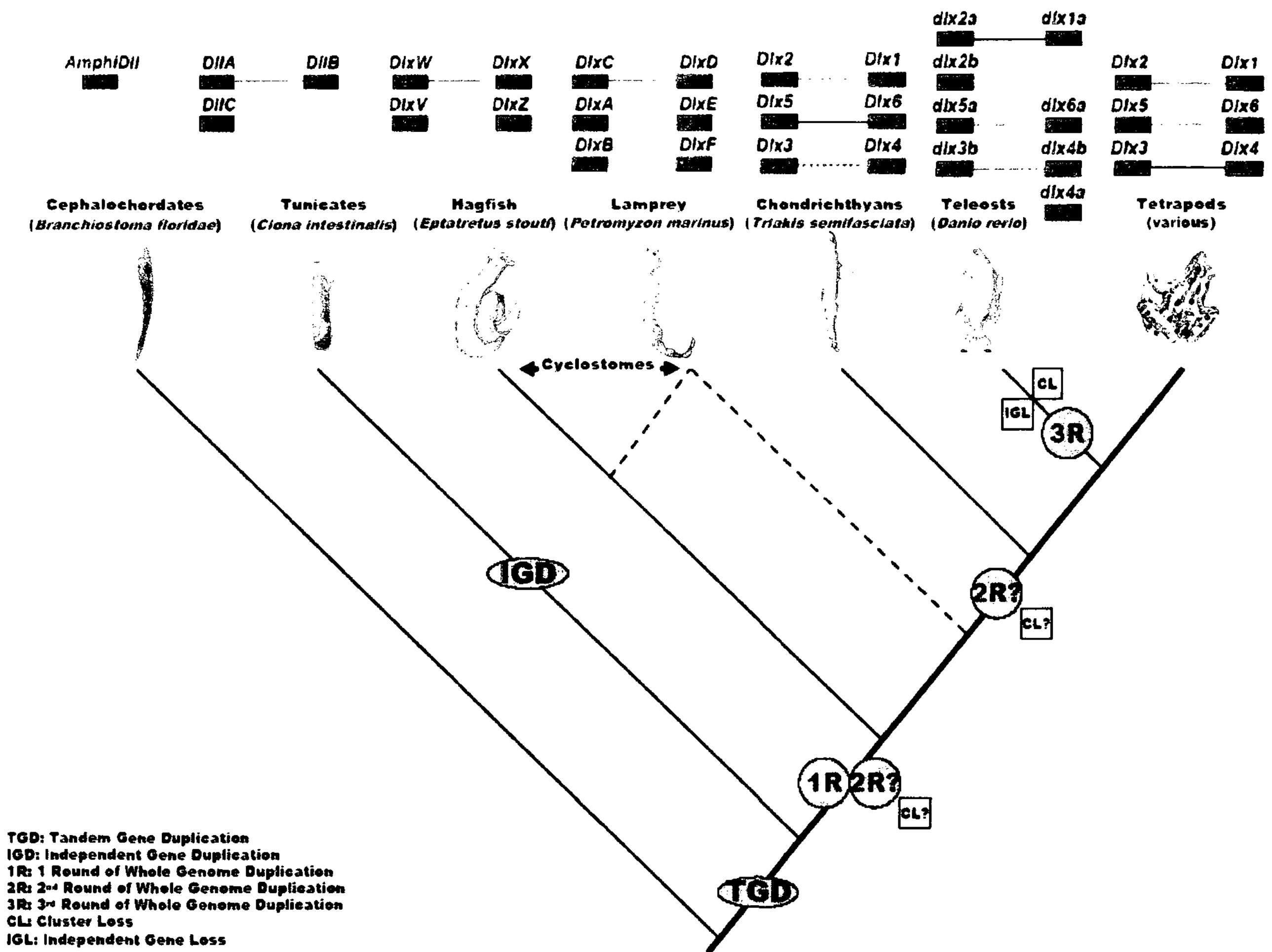


Figure 32: Summary of *Dlx* Gene Organization and Gene/Genome Duplication Events Across Chordate Phylogeny. The *Dlx* gene complement and its organization from representatives of 6 major branches of Chordate phylogeny are superimposed. The proposed timing of major genome duplication events are indicated (green circles). The expansion of the *Dll/Dlx* gene family can be used to trace duplication events. However, phylogenetic ambiguity of cyclostome *Dlx* genes, and the absence of more than 1 *Dlx* cluster makes it difficult to claim definitively whether duplication events occurred before or after Cyclostome divergence and information from multiple other gene families is necessary.

		Number of Confirmed *Dlx* Genes in the Genome	Number of *Dlx* Gene Clusters Present	Coordinate Expression of Clustered *Dlx* Genes	Nested Expression of *Dlx* Genes in BAs	CNEs Present in *Dlx* Gene Cluster Intergenic Regions
Cephalochordates		1	N/A	N/A	N/A	N/A
Tunicates		3	1	NO	NO	NO
Cyclostomes	Hagfish	4	1	?	?	NO
	Lamprey	6	1	Possible	NO	NO
Chondrichthyans		6	3	?	?	YES
Telosts		> 6 (8 in zebrafish)	3	YES	YES	YES
Tetrapods		6	3	YES	YES	YES

Figure 33: Summary of Major Properties of *Dlx* Gene Organization and Regulation in the Chordate Lineage. The origin of the properties of coordinate expression of clustered *Dlx* genes and the proximodistally nested expression pattern of clustered genes in the branchial arches (BAs) coincides with the origin of CNEs present in the intergenic region, and the presence of multiple *Dlx* clusters. Expression analyses of hagfish and chondrichthyan *Dlx* genes has not been published and would be necessary in order to resolve this correlation further.

Section 5: Conclusions

In order to reconstruct the condition of the ancestral Craniate genome and elucidate what genomic and gene regulatory changes have taken place during the morphological diversification of the Gnathostomes I sought to determine the gene number, genomic organization, and regulatory code of the *Dlx* family of homeobox containing transcription factors of both living orders of Cyclostomes, the hagfishes and the lampreys. Because Cyclostomes diverged from the chordate lineage prior to the evolution of a suite of morphological characteristics ubiquitious to the Gnathostomes and important for their evolutionary success including the jaw apparatus, true teeth, and paired appendages these organisms are an ideal outgroup for comparison with Gnathostomes in order to determine the origin of these novel morphologies. Gnathostomes possess at least 6 *Dlx* genes, different members of which are expressed in each of these morphological novelties making them prime candidates for involvement in the evolution of these structures. In agreement with unpublished results in a different species, I was able to confirm the presence of 6 members of the *Dlx* gene family in a lamprey, *Petromyzon marinus*. I was also able to clone 4 novel members of the *Dlx* gene family from a hagfish, *Eptatretus* stouti, while our collaborators were able to isolate 2 additional genes bringing the number of known hagfish *Dlx* genes to 6. By isolating and characterizing BAC clones containing Cyclostome *Dlx* genes I was able to demonstrate that only a single *Dlx* gene cluster exists in each organism; a *Dlx*C/*Dlx*D cluster in lamprey and a *Dlx*W/*Dlx*X cluster in hagfish. This is in stark contrast to the condition of Gnathostomes who invariably possess 3 bigene clusters. Furthermore, while all Gnathostomes have been shown to share several highly conserved noncoding elements

(CNEs) which act as *Dlx* enhancers in various embryonic tissues including the branchial arches, these elements are absent from the Cyclostome *Dlx* loci. Because these elements are ubiquitously absent from the genomes of all other invertebrates, and we have found orthologs in the genome of the basal Gnathostome, the Elephant Shark *Callhorinchus milli*, we have been able to resolve the point of origin of these CNEs as discreet functional units to the period of time following the evolution of Craniates, but prior to or coincident with the origin of Gnathostomes.

The possession of unique *Dlx* bigene clusters and the absence of CNEs surrounding *Dlx* loci is shared between both hagfish and lamprey and lends genomic support to the "Cyclostome theory" which postulates that hagfish and lamprey form a monophyletic group, and not that lamprey are more closely related to Gnathostomes than to hagfish. The absence of *Dlx* CNEs correlates with the absence of a highly temporally and spatially constrained "*Dlx*-Code" in the branchial arches of Gnathostomes which is known to be necessary for the appropriate patterning of the jaw apparatus, and may be why significantly less overlap is seen between the expression of linked lamprey *Dlx* genes. While the highly specific regulation necessary for the establishment of the "*Dlx*-code" may be absent from Cyclostome *Dlx* loci, by assaying the ability of lamprey or zebrafish noncoding DNA flanking the *Dlx* genes to drive expression of reporter constructs in both the zebrafish and in the lamprey embryo I demonstrated that at least the level of regulation responsible for targeting *Dlx* to homologous tissues between lamprey and zebrafish is conserved. While unreliable phylogenetic support makes orthology assignment between lamprey, hagfish, and Gnathostome *Dlx* genes difficult, it is clear that an equal number of members of each major *Dlx* subfamily exist in the Cyclostomes,

consistent with the hypothesis that they shared one or both major genome duplication events with Gnathostomes.

Section 6: References

Amemiya CT, Prohaska SJ, Hill-Force A, Cook A, Wasserscheid J, Ferrier DE, Pascual-Anaya J, Garcia-Fernàndez J, Dewar K, Stadler PF: The amphioxus Hox cluster: characterization, comparative genomics, and evolution. J Exp Zoolog B Mol Dev Evol. 2008, 310(5):465-477.

Anderson S, Eisenstat D, Shi L and Rubenstein J: Interneuron migration from basal forebrain to neocortex: dependence on *Dlx* genes. Science 1997, 278:474-476.

Banerjee-Basu S and Baxevanis AD: Molecular evolution of the homeodomain family of transcription factors. Nucl. Acids Res. 2001, 29(15): 3258-3269.

Banerjee-Basu S, Baxevanis AD: Molecular evolution of the homeodomain family of transcription factors. Nucleic Acids Research 2001, 29:3258-3269.

Bardack D: First Fossil Hagfish (Myxinoidea): A Record from the Pennsylvanian of Illinois. Science 1991, 254(5032):701-703.

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

Beverdam A, Merlo GR, Paleari L, Mantero S, Genova F, Barbieri O, Janvier P, Levi G: Jaw transformation with gain of symmetry after *Dlx5/Dlx6* inactivation: Mirror of the past? Genesis 2002, 34:221-227.

Blackwood EM, Kadonaga JT. Going the distance: a current view of enhancer action. *Science* 1998, 281:60-63.

Boffelli D, Nobrega MA, Rubin EM. Comparative genomics at the vertebrate extremes. *Nat Rev Genet* 2004, 5:456-465.

Botella H, Blom H, Dorka M, Ahlberg PE, Janvier P: Jaws and teeth of the earliest bony fishes. Nature 2007, 448(7153):583-586.

Brooke NM, Garcia-Fernàndez J, Holland PW: The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. Nature 1998, 392(6679):920-922.

Caracciolo A, Di Gregorio A, Aniello F, Di Lauro R, Branno M: Identification and developmental expression of three Distal-less homeobox containing genes in the ascidian Ciona intestinalis. Mechanisms of Development 2000, 99:173-176.

Castro LF, Holland PW: Fluorescent in situ hybridisation to amphioxus chromosomes.
Chang MM, Zhang J, Miao D: A lamprey from the Cretaceous Jehol biota of China. Nature 2006, 441:972–974.

Cobourne MT, Sharpe PT: Tooth and jaw: molecular mechanisms of the patterning in the first branchial arch. Archives of Oral Biology 2003, 48:1-14.

Darwin, C. On The Origin of Species by Means of Natural Selection. John Murray, London 1859.

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

Depew MJ, Lufkin T, Rubenstein JLR: Specification of jaw subdivisions by *Dlx* genes. Science 2002, 298;381-385.

Depew MJ, Simpson CA, Morasso M, and Rubenstein JLR: Reassessing the *Dlx* code: the genetic regulation of branchial arch skeletal pattern and development. J. Anat. 2005, 207:501-561.

Duboule D: The rise and fall of Hox gene clusters. Development 2007, 134:2549-2560.
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):1440-1450.

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

Force A, Amores A, Postlethwait JH: Hox cluster organization in the jawless vertebrate *Petromyzon marinus*. J. exp. Zool. (Mol Dev Evol) 2002, 294:30-46.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: Preservation of duplicate genes by complementary, degenerative mutations. Genetics 1999, 151(4):1531-45

Freitas R, Zhang G, Cohn MJ: Evidence that mechanisms of fin development evolved in the midline of early vertebrates. Nature 2006, 442(7106):1033-7.

Furlong RF, Holland PW: Were Vertebrates Octoploid? Philos Trans R Soc Lond B Biol Sci. 2002, 357(1420):531-44.

Furlong RF, Younger R, Kasahara M, Reinhardt R, Thorndyke M, Holland PW: A degenerate ParaHox gene cluster in a degenerate vertebrate. Mol Biol Evol. 2007, 12:2681-2686.

Garcia-Fernández J, Holland PW: Archetypal organization of the amphioxus Hox gene cluster. Nature 1994, 370(6490):563-6.

Garcia-Fernàndez J: The genesis and evolution of homeobox gene clusters. Nature Reviews Genetics 2005, 6:881-891.

Gess RW, Coates MI, Rubidge BS: A lamprey from the Devonian period of South Africa. Nature 2006, 443:981–984.

Ghanem N, Jarinova O, Amores A, Long Q, Hatch G, Park BK, Rubenstein JLR, and Ekker M: Regulatory roles of conserved intergenic domains in vertebrate *Dlx* bigene clusters. Genome Res 2003, 13:53-543.

Ghanem N, Yu M, Long J, Hatch G, Rubenstein JLR, Ekker M: Distinct cis-regulatory elements from the *Dlx*1/*Dlx*2 locus mark different progenitor cell populations in the ganglionic eminences and different subtypes of adult cortical interneurons. The Journal of Neuroscience 2007, 27(19):5012-5022.

Graham A: Jaw Development: Chinless Wonders. Current Biology 2002, 12:R810-R812.

Harafuji N, Keys DN, and Levine M: Genome-wide identification of tissue-specific enhancers in the Ciona tadpole. PNAS 2002, 99;6802-6805.

Heierhorst J, Lederis K, Richter D: Presence of a member of the Tc1-like transposon family from nematodes and Drosophila within the vasotocin gene of a primitive vertebrate, the Pacific hagfish Eptatretus stouti. PNAS 1992, 89(15):6798-802.

Holland ND, Panganiban G, Henyey EL, Holland LZ: Sequence and developmental expression of AmphiDll, an amphioxus Distal-less gene transcribed in the ectoderm, epidermis and nervous system: insights into evolution of craniate forebrain and neural crest. Development 1996, 122;2911-2920.

Ikuta T, Yoshida N, Satoh N, Saiga H: Ciona intestinalis Hox gene cluster: Its dispersed structure and residual colinear expression in development. PNAS 2004, 101(42):15118-23.

Irvine SQ, Cangiano MC, Millette BJ, Gutter ES: Non-overlapping expression patterns of the clustered Dll-A/B genes in the ascidian Ciona intestinalis. J. Ex Zool (Mol Dev Evol) 2007, 308B;1-14.

Irvine SQ, Carr JL, Bailey WJ, Kawasaki K, Shimizu N, Amemiya CT, Ruddle FH: Genomic analysis of Hox clusters in the sea lamprey Petromyzon marinus. Journal of Experimental Zoology (Mol. Dev. Evol.) 2002, 294:47-62.

Janvier P, Lund R: Hardistiella montanensis (Petromyzontida)from the Lower Carboniferous of Montana, with remarks on theaffinities of the lampreys. J Vertebr Paleontol. 1983, 2:407–413.

Janvier P: Palaeontology: modern look for ancient lamprey. Nature 2006, 443:921–924

Kasahara M: The 2R hypothesis: an update. Current Opinion in Immunology 2007, 19:547-552.

Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D: Human genome ultraconserved elements are ultraselected. *Science* 2007, 317:915.

Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engström PG, Fredman D, Akalin A, Caccomo M, Sealy I, Howe K, Ghislain J, Pezeron G, Mourrain P, Ellingsen S, Oates A, Thisse C, Thisse B, Foucher I, Adolf B, Geling A, Lenhard B, Becker TS: Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res 2007, 17:545-555.

Knecht AK and Bronner-Fraser M: Induction o the neural crest: a multigene process. Nature Reviews Genetics 2002, 3:453-461.

Kuraku S, Kuratani S: Time scale for cyclostome evolution inferred with a phylogenetic diagnosis of hagfish and lamprey cDNA sequences. Zoolog Sci 2006, 23(12): 1053-64.

Kuratani S, Kuraku S, Murakami Y: Lamprey as an Evo-Devo model: lessons from comparative embryology and molecular phylogenetics. Genesis 2002, 34:175-183.

Kusakabe R, Tochinai S, Kuratani S: Expression of foreign genes in lamprey embryos: an approach to study evolutionary changes in gene regulation. Journal of Experimental Zoology (Mol. Dev. Evol.) 2003, 296B:87-97.

Lee AP, Koh EGL, Tay A, Brenner S, and Venkatesh B: Highly conserved syntenic blocks at the vertebrate Hox loci and conserved regulatory elements within and outside Hox gene clusters. PNAS 2006, 18;6994-6999.

Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaf E: A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, 2003, 12:1725–1735.

Li H, Marijanovic I, Kronenberg MS, Erceg I, Stover ML, Velonis D, Mina M, Heinrich JG, Harris SE, Upholt WB, Kalajzic I, Lichtler AC: Expression and function of *Dlx* genes in the osteoblast lineage. Developmental Biology 2008, 316:458-470.

Li W-H, Yang J, Gu X: Expression divergence between duplicate genes. Trends in Genetics 2005, 21(11):602-607.

Liu JK, Ghattas I, Liu S, Chen S, Rubenstein JL: *Dlx* genes encode DNA-binding proteins that are expressed in an overlapping and sequential pattern during basal ganglia differentiation. Developmental Dynamics 1997, 210(4):498-512.

Lòpez AJ: Developmental role of transcription factor isoforms generated by alternative splicing. Developmental Biology 2005, 172:396-411.

Luke GN, Castro LF, McLay K, Bird C, Coulson A, Holland PW: Dispersal of NK homeobox gene clusters in amphioxus and humans. PNAS 2003, 100(9):5292-5295
Mallatt, J., 1996. Ventilation and the origin of jawed vertebrates: a new mouth. Zool. J. Linn. Soc. 117, 329– 404.

Maurya, AK. Regulatory mechanisms controlling Distal-less related gene expression in lamprey. University of Ottawa. Ottawa, 2006.

McCauley DW, Bronner-Fraser M: Neural crest contributions to the lamprey head. Development 2003, 130:2317-2327.

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 Res 2006, 16:451-465.

Meyer A, Van de Peer Y: From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). BioEssays 2005, 27:937-945.

Muller F, Blader P, Strahle U: Search for enhancers: teleost model sin comparative genomic and transgenic analysis of *cis* regulatory elements. *Bioessays* 2002, 24:564-572.

Mulley JF, Chiu CH, Holland PW: Breakup of a homeobox cluster after genome duplication in teleosts. PNAS 2006, 103(27):10369-10372.

Murakami Y, Uchida K, Rijli FM, Kuratani S: Evolution of the brain developmental plan: Insights from agnathans. Developmental Biology 2005, 280:249-259.

Myojin M, Ueki T, Sugahara F, Murakami Y, Shigetani Y, Aizawa S, Hirano S, Kuratani S: Isolation of *Dlx* and Emx gene cognates in an agnathan species, Lampetra japonica, and their expression patterns during embryonic and larval development: Conserved and diversified regulatory patterns of homeobox genes in vertebrate head evolution. Journal of Experimental Zoology (Mol Dev Evol) 2001, 291:68-84.

Nakai Y, Kubota S, Goto Y, Ishibashi T, Davison W, Kohno S: Chromosome elimination in three Baltic, south Pacific and north-east Pacific hagfish species. Chromosome Res. 1995, 5:321-330.

Negre B, Ruiz A: HOM-C evolution in Drosophila: is there a need for Hox gene clustering? Trends in Genetics 2007, 23(2):55-9.

Neidert AH, Virupannavar V, Hooker GW, and Langeland JA: Lamprey *Dlx* genes and early vertebrate evolution. PNAS 2001, 98;1665-1670.

Ohno S: Evolution by Gene Duplication. Springer 1970.

Osorio J, Mazan S, Rétaux S: Organisation of the lamprey (Lampetra fluviatilis) embryonic brain: Insights from LIM-homeodomain, Pax and hedgehog genes. Developmental Biology 2005, 288:100-112.

Ota KG, Kuraku S, Kuratani S: Hagfish embryology with reference to the evolution of the neural crest. Nature 2007, 446(7136):672-675.

Ota KG, Kuratani S: The history of scientific endeavors towards understanding hagfish embryology. Zoological Science 2006, 23: 403-418.

Panganiban G, Rubenstein JLR: Developmental functions of the Distal-less/*Dlx* homeobox genes. Development 2002, 129;4371-4386.

Panopoulou G and Poustka AJ: Timing and mechanism of ancient vertebrate genome duplications – the adventure of a hypothesis. Trends in Genetics 2005, 21(10):559-567.

Park BK, Sperber SM, Choudhury A, Ghanem N, Hatch GT, Sharpe PT, Thomas BL, and Ekker M: Intergenic enhancers with distinct activities regulate *Dlx* expression in the mesenchyme of the branchial arches. Developmental Biology 2004, 268;532-545.

Poitras L, Ghanem N, Hatch G, Ekker M: The proneural determinant MASH1 regulates forebrain *Dlx*1/2 expression through the I12b intergenic enhancer. Development 2007, 124;1755-1765.

Prince VE, Pickett FB: Splitting Pairs: the diverging fates of duplicated genes. Nature Reviews Genetics 2002, 3:827-837.

Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK, Benito-Gutiérrez EL, Dubchak I, Garcia-Fernàndez J, Gibson-Brown JJ, Grigoriev IV, Horton AC, de Jong PJ, Jurka J, Kapitonov VV, Kohara Y, Kuroki Y, Lindquist E, Lucas S, Osoegawa K, Pennacchio LA, Salamov AA, Satou Y, Sauka-Spengler T, Schmutz J, Shin-I T, Toyoda A, Bronner-Fraser M, Fujiyama A, Holland LZ, Holland PW, Satoh N, Rokhsar DS: The amphioxus genome and the evolution of the chordate karyotype. Nature 2008, 453(7198):1064-1071.

Robledo R, Rajan L, Li X, Lufkin T: The *Dlx*5 and *Dlx*6 homeobox genes are essential for craniofacial, axial, and appendicular skeletal development. Genes & Development 2002, 16:1089-1101.

Saino-Saito S, Berlin R, Baker H: *Dlx*-1 and *Dlx*-2 expression in the adult mouse brain: relationship to dopaminergic phenotypic regulation. J Comp Neurol. 2003, 461(1):18-30
Samee N, de Vernejoul M-C, Levi G: Role of *DLX* regulatory proteins in osteogenesis and chondrogenesis. Critical Reviews in Eukaryotic Gene Expression 2007, 17(3):173-186.

Sanges R, Kalmar E, Claudiani P, D'Amato M,Muller F and Stupka E: Shuffling of *cis*-regulatory elements is a pervasive feature of the vertebrate lineage *Genome Biology* 2006, 7:R56.

Santagati F, Rijli FM: Cranial neural crest and the building of the vertebrate head. Nat Rev Neurosci. 2003, 4(10):806-818.

Sauka-Spengler T, Meulemans D, Jones M, Bronner-Fraser M: Ancient evolutionary origin of the neural crest gene regulatory network. Developmental Cell 2007, 13:405-420.

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W. PipMaker – A web server for aligning two genomic DNA sequences. Genome Research 2000, 10:577-586.

Shigetani Y, Sugahara F, Kawakami Y, Murakami Y, Hirano S, Kuratani S: Heterotopic shift of epithelial-mesenchymal interactions in vertebrate jaw evolution. Science 2002, 296:1316-1319.

Shimeld SM, Holland PW: Vertebrate innovations. PNAS 2000, 97(9):4449-4452.
Solomon KS, Fritz A: Concerted action of two *Dlx* paralogs in sensory placode formation.

Sperber SM, Saxena V, Hatch G, Ekker M: Zebrafish *Dlx*2a contributes to *Hind*brain neural crest survival, is necessary for differentiation of sensory ganglia and functions with *Dlx*1a in maturation of the arch cartilage elements. Developmental Biology 2008, 314:59-70.

Spitz F, Gonzalez F, and Duboule D: A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* 2003, 113:405-417.

Stadler PF, Fried C, Prohaska SJ, Bailey WJ, Misof BY, Ruddle FH, Wagner GP: Evidence for independent Hox gene duplications in the hagfish lineage: a PCR-based gene inventory of Eptatretus stoutii. Molecular Phylogenetics and Evolution 2004, 32:686-694.

Stock DW, Ellies DL, Zhao Z, Ekker M, Ruddle FH, Weiss KM: The evolution of the vertebrate *Dlx* gene family. PNAS 1996, 93;10858-10863.

Stock DW: The *Dlx* gene complement of the Leopard Shark, Triakis semifasciata, resembles that of mammals: implications for genomic and morphological evolution of jawed vertebrates. Genetics 2005, 169:807-817.

Sumiyama K, Irvine SQ, Ruddle FH: The role of gene duplication in the evolution and function of the vertebrate *Dlx*/distal-less bigene clusters. Journal of structural and functional genomics 2003, 3:151-159.

Sumiyama K, Irvine SQ, Stock DW, Weiss KM, Kawasaki K, Shimizu N, Shashikant CS, Miller W, and Ruddle FH: Genomic structure and functional control of the *Dlx*3-7 bigene cluster. PNAS 2002, 780-785.

Sumiyama K, Ruddle FH: Regulation of *Dlx*3 gene expression in visceral arches by evolutionarily conserved enhancer elements. PNAS 2003, 100; 4030-4034.

Takahashi A, Nakata O, Kosugi T, Kako K, Sakai T, Ito M, Takamatsu N, Shiba T, Sower S, Kawauchi H: Promoter activity of sea lamprey proopiocortin and proopiomelanotropin genes in AtT-20/D16v cells. General and comparative endocrinology 2005, 144;182-187.

Takio Y, Kuraku S, Murakami Y, Pasqualetti M, Rijli FM, Narita Y, Kuratani S, Kusakabe R: Hox gene expression patterns in Lethenteron japonicum embryos – Insights into the evolution of the vertebrate Hox code. Developmental Biology 2007, 308(2):606-620.

Tautz D: Evolution of transcriptional regulation. *Curr Opin Genet Dev* 2000, 10:575-579.

Venkatesh B, Kirkness EF, Loh Y-H, Halpern AL, Lee AP, Johnson J, Dandona N, Viswanathan LD, Tay A, Venter JC, Strausberg RL, Brenner S: Survey sequencing and comparative analysis of the elephant shark (Callorhinchus milii) genome. PLOS biology 2007, 4;e101.

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

Venkatesh B, Tay A, Dandona N, Patil JG and Brenner S: A compact cartilaginous fish model genome. Current Biology 2005, 3;R82-R83.

Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio LA: Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nature Genetics 2008, 40(2):158-160.

Wagner GP, Amemiya C, Ruddle F: Hox cluster duplications and the opportunity for evolutionary novelties. PNAS 2003, 100(25):14603-14606.

Walker MB, Miller CT, Talbot JC, Stock DW, Kimmel CB: Zebrafish furin mutants reveal intricacies in regulating Endothelin1 signaling in craniofacial patterning. Developmental Biology 2006, 295:194-205.

Wellik DM: Hox patterning of the vertebrate axial skeleton. Developmental Dynamics 2007, 236(9):2454-2463.

Westerfield, M. The Zebrafish Book. University of Oregon Press. Eugene, 1995.

Woolfe A, Elgar G: Comparative genomics using Fugu reveals insights into regulatory subfunctionalization. Genome Biology 2007, 8:R53.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T,Smith SF, North P, Callaway H, Kelly K, *et al.*: Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* 2005, 3:e7.

Wray GA: Transcriptional regulation and the evolution of development. Int J Dev Biol. 2003, 47(7-8):675-84.

Yu S, Zhang W, Li L, Huang H, Ma F, Li Q: Phylogenetic analysis of 48 gene families revealing relationships between Hagfishes, Lampreys, and Gnathostomata. J. Genet. Genomics 2008, 35:285-290.

Zerucha T, Ekker M: Distal-less-related homeobox genes of vertebrates: Evolution, function and regulation. Biochem. Cell Biol. 2000, 78:593-601.

Zerucha T, Stühmer T, Hatch G, Park BK, Long Q, Yu G, Gambarotta A, Schultz JR, Rubenstein JLR, Ekker M: A highly conserved enhancer in the *Dlx5*/*Dlx6* integenic region is the site of cross-regulatory interactions between *Dlx* genes in the embryonic forebrain. The Journal of Neuroscience 2000, 20(2):709-721.

Zhang Z, Schwartz S, Wagner L, and Miller W: A greedy algorithm for aligning DNA sequences. JComput Biol 2000, 7(1-2):203-14.