

Jeffrey Borden McCallum. DNA Sequence Analysis and Characterization of a *Morone saxatilis* (Striped Bass) Genomic Clone Containing the *HoxB2* and *HoxB3* Genes. (Under the direction of Dr. Edmund J. Stellwag) Department of Biology, July 1997.

A single λ EMBL-3 genomic clone 15 kb in length from the teleost *Morone saxatilis* (striped bass) was characterized by restriction endonuclease mapping, Southern hybridization, and DNA sequence analysis. Restriction mapping and Southern hybridization provided evidence for two linked homeobox-containing regions separated by 8.5 kb. Comparison of the DNA sequences of the homeoboxes from this clone to those of other vertebrates revealed extensive DNA sequence homology and distinctive nucleotide signatures indicative of *HoxB2* and *HoxB3*. DNA sequence analysis of the *HoxB2* - *HoxB3* intergenic region provided evidence for the likely existence of two distinct promoters, located 200 bp and 1,356 bp upstream of the *HoxB2* first exon. The promoter located 200 bp upstream of the first exon of *M. saxatilis HoxB2* bears striking similarity to the human *HOXB2* promoter in organization and DNA sequence elements. Despite the similarities between the *M. saxatilis* and human *HoxB2* promoter regions, there are notable differences, including the absence of a *GATA-1* site in striped bass, the presence of a second *AP2* site in striped bass, and a significant shift in the location of both *Krox-20* and Box1 motifs between the two promoters. Limited information available for the *HoxB2/HoxB3* intergenic region of *Fugu rubripes* (pufferfish) revealed a *Krox-20/Box1* complex very similar in sequence and location to that in *M. saxatilis*. Although its significance is as yet uncertain, I have identified an alternate promoter located 1356 bp upstream of the promoter nearest the first exon of *HoxB2* in striped

bass. The overall conservation of the basal promoter elements between human and striped bass, with interposition of transcription elements unique to each promoter, suggests that *HoxB2* expression in these evolutionarily divergent species is likely quite different.

DNA Sequence Analysis and Characterization of a *Morone saxatilis* (Striped Bass)

Genomic Clone Containing the *HoxB2* and *HoxB3* Genes.

A Thesis

Presented to

the Faculty of the Department of Biology

East Carolina University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Molecular Biology and Biotechnology

by

Jeffrey Borden McCallum

July 1997

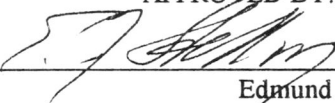
DNA Sequence Analysis and Characterization of a *Morone saxatilis* (Striped Bass)

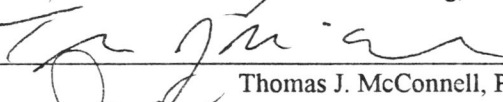
Genomic Clone Containing the *HoxB2* and *HoxB3* Genes.

by

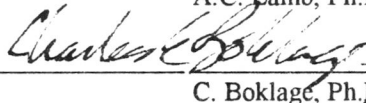
Jeffrey Borden McCallum

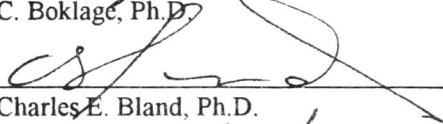
APPROVED BY:

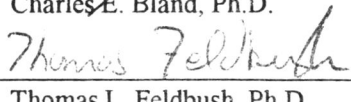
DIRECTOR OF THESIS  _____
Edmund J. Stellwag, Ph.D.

COMMITTEE MEMBER  _____
Thomas J. McConnell, Ph.D.

COMMITTEE MEMBER  _____
A.C. Lamb, Ph.D.

COMMITTEE MEMBER  _____
C. Boklage, Ph.D.

CHAIR OF THE DEPARTMENT OF BIOLOGY  _____
Charles E. Bland, Ph.D.

DEAN OF THE GRADUATE SCHOOL  _____
Thomas L. Feldbush, Ph.D.

Dedication

This thesis is dedicated to the memory of my grandfather William Andrew Weeks who was a dearly loved friend. His presence is greatly missed, and the memories of the times we spent together are treasures.

Acknowledgments

I owe a debt of gratitude to many people for their encouragement and support in completing this thesis. First, I would like to acknowledge the Department of Biology and my committee members for all of their support, without which this thesis would have been impossible. My thesis director Dr. Stellwag deserves a heartfelt thank you for taking me under his tutelage.

Special thanks go to Jean-Luc for teaching me that French guys are really Rednecks at heart.

My fellow graduate students conjure up many entertaining stories and have provided me with friendships that are cherished. I thank Tony Pavell for his willingness to include me in his research when I was an undergraduate and for all of his advice. To the members of TEAM HOX (Libby, Scott and Tom), I extend my gratitude for making life interesting, and the monotony enjoyable. Libby, the mini-prep Queen, I thank for serving as the lab D.J.

In our lab, the wall of quotes provides guidance and wisdom for all ambitious students. I will always remember Scott's words: "There are 24 hours in a day, and I need only 8 to sleep" and Eric's "What is there to writing a scientific paper?". Thus, I have endeavored to **not** achieve recognition on the wall of quotes.

My parents merit my gratefulness for their unwavering support in all that I do. My grandmother is a good friend who has helped me achieve my dreams.

I will never be able to say or do enough to thank my best friend for all of her encouragement and especially her friendship. Samara you are truly a blessing, thanks for everything!

Table of Contents

Figures and Tables	v
Symbols and Abbreviations	vii
Introduction	1
Materials and Methods	18
Library Screening and Clone Identification	18
Phage DNA Isolation	20
Restriction Analysis	24
Southern Blot Analysis	25
Subcloning	26
DNA Sequencing	29
Data Analysis	30
Results	36
Discussion	49
References	58
Figures	73
Tables	97
APPENDIX A: Synthetic Oligonucleotide Primers	100
APPENDIX B: Complete DNA Sequence of clone 2H	102

Figures and Tables

Figures

1. Diagram of hoxons from the cephalochordate amphioxus and selected vertebrates. 74
2. Schematic diagram depicting the regulation of the murine *HoxB1* and *HoxB2* genes. 76
3. Restriction Endonuclease Cleavage Map of the *Morone saxatilis*-derived insert of λ clone 2H 78
4. Schematic diagram depicting the overall strategy for subcloning λ clone 2H. 80
5. Schematic diagram of PCR reactions used to analyze the location and orientation of the individual subclones (2.2, 2.2.1, 2.2.2 and 2.3) generated from λ clone 2H. 82
6. Nucleotide sequence alignment of paralogue group 2 homeoboxes. 84
7. Amino acid sequence alignment of the region corresponding to the 1st exon of the *Hox* genes representative of paralog group 2. 86
8. Amino acid sequence alignment of homeodomains representative of paralog group 3 genes. 88
9. Amino acid sequence alignment of the region corresponding to the 2nd exon of the *Hox* genes representative of paralog group 2. 90

10. Amino acid sequence alignment of human <i>HOXB2</i> protein with <i>Morone saxatilis HoxB2</i> protein.	92
11. Schematic diagram of <i>Morone saxatilis HoxB2</i> putative alternate promoter located at positions 8665 - 8897 on λ clone 2H.	94
12. Comparison of <i>HoxB2</i> promoters from the human, <i>M. saxatilis</i> and <i>Fugu rubripes</i> *.	96

Tables

1. Primers used for PCR confirmation of correct subclone assemblage with predicted and observed product lengths.	97
2. A list of DNA sequence elements that were used to identify similar elements in <i>M. saxatilis</i> λ clone 2H	98
3. Notable DNA sequence-based landmarks identified from λ clone 2H	99

Symbols and Abbreviations

The following symbols and abbreviations are used throughout this thesis.

a.a. - amino acid

Abd-A - *Drosophila* abdominal A homeotic gene

Abd-B - *Drosophila* abdominal B homeotic gene

Ant-C - antennapedia complex

Antp - *D. melanogaster* antennapedia gene

bp - base pair

BX-C - bithorax complex

C - degrees Celsius

cDNA - complementary DNA

Dfd - *Drosophila* deformed homeotic gene

D. melanogaster - *Drosophila melanogaster*

DNA - deoxyribonucleic acid

dpm - disintegration per minute

F - degrees Fahrenheit

ftz - *Drosophila* fushi tarazu segmentation gene

g - force of gravity

kb - kilobase, one thousand bases of DNA

Lab - *Drosophila* labial homeotic gene

M - molar moles per liter

<i>MAT</i>	- yeast mating type proteins
ml	- 10^{-3} liter
mm	- 10^{-3} meter
MOI	- multiplicity of infection
ORF	- open reading frame
<i>Pb</i>	- <i>Drosophila</i> probiscipedia homeotic gene
p.f.u.	- plaque forming units
r#	- rhombomere number of the vertebrate hindbrain
<i>Scr</i>	- <i>Drosophila</i> sex combs reduced homeotic gene
T2	- second thoracic segment of <i>Drosophila melanogaster</i>
T3	- third thoracic segment of <i>Drosophila melanogaster</i>
<i>Ubx</i>	- <i>Drosophila</i> ultrabithorax homeotic gene
μ l	- 10^{-6} liter

Introduction

While working with Thomas Hunt Morgan in 1915, Calvin Bridges described mutations in *Drosophila melanogaster* that transform the halteres (flight balancers) of the third thoracic segment (T3) into a second thoracic segment (T2). Remarkably this homeotic transformation results in the appearance of a second set of wings on the homeotically transformed third thoracic segment. These mutations were referred to as *bithorax* mutations, and Bridges termed them homeotic mutations after Bateson's terminology describing the translocation of a body structure to a site distinct from its normal location (Gehring 1994; Lawrence 1992). Progress in studying homeotic mutations accelerated in 1946 when Ed Lewis mapped a group of homeotic mutations to a defined region he called the bithorax complex (*BX-C*). Lewis discovered that deletion of the *BX-C* resulted in a homeotic transformation of the third thoracic and all abdominal segments into segments resembling the second thoracic segment (T2) (Lawrence 1992). This observation led Lewis to the conclusion that T2 determines the default or "ground state" value for all segments posterior to T2 and that the *BX-C* genes function to direct the developmental pathways of segments posterior to T2 (Bender *et al.* 1983a).

Detailed molecular genetic analysis of the *Drosophila BX-C* led to the discovery of the homeobox, a 180 bp sequence that encodes a helix-turn-helix DNA binding motif that acts to regulate the expression of genes involved in development. Welcome Bender, who was collaborating with David Hogness, reported the recombinant DNA cloning of the *BX-C* in 1983 using the techniques of chromosome walking and chromosome jumping

(Bender, Spierer, and Hogness 1983b). The cloning of the *BX-C* proved to be a milestone in *Drosophila* research because it was the first complex of genes cloned from *Drosophila* that provided a high resolution physical map of the *BX-C*. Further molecular genetic analysis of the *Drosophila BX-C* by Scott *et al.* (1983) revealed that specific DNA fragments cloned from the *BX-C* could hybridize to the genes within and outside the *BX-C*. Scott *et al.* discovered that portions of the *Drosophila* homeotic gene *Antennapedia* (*Antp*) hybridized to portions of the *fushi tarazu* (*ftz*) and *ultrabithorax* (*Ubx*) genes (Kuroiwa, Hafen, and Gehring 1984). Simultaneously, McGinnis determined that portions of the *D. melanogaster Antp* cDNA could hybridize to other genomic fragments in *Drosophila* (McGinnis *et al.* 1984b). DNA sequence analysis of the regions of cross hybridization among these three genes uncovered a highly conserved region 180 base pairs (bp) in length that encodes a 60 amino acid (a.a.) stretch that is conserved between the three genes (Scott and Weiner 1984). The 180 bp region was termed the “homeobox” because of the high similarity of the DNA sequence (*Antp* was shown to be 77% homologous to *ftz*, and 79% homologous to *Ubx*, with *Ubx* and *ftz* sharing a 75% homology), and the 60 a.a. translation of the homeobox was termed the “homeodomain” (McGinnis *et al.* 1984a). The homology of the amino acids within the so called homeodomain is even more striking, with *Antp* and *ftz* sharing 50/60 amino acid positions (83% homology), *ftz* and *Ubx* sharing 45/60 amino acids (75% homology) and *Antp* and *Ubx* sharing 52/60 amino acids (87% homology) (McGinnis *et al.* 1984a). Using the *Antp* homeobox as a probe in Southern hybridization analysis, McGinnis discovered

strong hybridization to DNA prepared from three other invertebrates (*Drosophila hydei*, the beetle *Tenebrio molitor*, and the earthworm *Lumbricus sp.*) as well as to DNA prepared from chicken, mouse, and human (McGinnis *et al.* 1984a). Based on these results, McGinnis concluded that the homeobox region represents an evolutionary conserved DNA sequence in diverse metazoan species. The evolutionary conservation of the homeobox-containing genes was underscored by subsequent discovery of homeodomain encoding genes in eukaryotes from *Saccharomyces cerevisiae* to *Xenopus*.

The remarkable evolutionary conservation of homeobox-encoding genes was underscored by the discovery of other interesting relationships. For example, the *Xenopus* gene (*AC1* now designated *HoxC6*) shows a much higher homology to the *Drosophila Antp* gene than *Antp* shares with either the *ftz* or *Ubx* genes having 55 out of 60 amino acids conserved. Furthermore, it was shown that the *Xenopus* gene is expressed during embryogenesis, much like the *Drosophila* genes. The high degree of conservation of the amino acid sequence in the homeodomain and the apparent conservation in function sparked speculation that homeobox-containing genes represent a class of genes that act as "universal control mechanisms of development at the molecular level" (Carrasco *et al.* 1984; McGinnis *et al.* 1984a). This speculation was further fueled upon discovery that the homeodomains of both *Xenopus* and *Drosophila* genes share regions of homology with *Saccharomyces cerevisiae* $\alpha 1$ and $\alpha 2$ yeast mating type (*MAT*) proteins (Shepherd *et al.* 1984). The *MAT* genes are responsible for controlling cell differentiation by regulating other genes through a helix-turn-helix DNA binding motif

found in the *MAT* protein product. The DNA-binding domains of prokaryotic DNA-binding proteins have been shown to possess homology with the homeodomain (Laughon and Scott 1984). Thus, it was postulated that homeobox-containing genes function to control the program of development by acting as DNA binding proteins, and are evolutionarily related through a common precursor with multiple copies arising through duplication events (Laughon and Scott 1984; Lewis 1978; McGinnis *et al.* 1984a).

Hox genes are a subset of homeobox-containing genes and are characterized by several features. *Hox* genes possess a 180 bp DNA sequence that encodes a highly conserved 60 amino acid DNA binding motif. The DNA sequence of the homeoboxes found in *Hox* genes is closely related to the *Antennapedia* homeobox DNA sequence (Boncinelli *et al.* 1988; Duboule and Dolle 1989; Gaunt *et al.* 1988). Mutations and deletions of *Hox* genes result in homeotic transformations and alterations in the development of the embryo (Alexandre *et al.* 1996; Condie and Capecchi 1994; Johnson and O'Higgins 1996; Pollock *et al.* 1995; Vieille-Grosjean *et al.* 1997). *Hox* genes are grouped in clusters on the chromosome and are expressed in a colinear fashion, which means that their expression reflects their location in the cluster with respect to the other genes and to the region of the embryo in which they are expressed. For example, the genes found at the 3' end of the *Hox* cluster are expressed the most anteriorly, with the 5' genes of each cluster expressed most posteriorly in the developing embryo. This spatial colinearity of expression is also reflected in a temporal fashion as the 3' genes are activated first, and the 5' genes expressed last (De Robertis 1994; Duboule and Morata

1994; Graham, Papalopulu, and Krumlauf 1989). The genes are also arranged in the cluster so that the direction of transcription is conserved. The preservation of this organization is unprecedented in other gene complexes and suggests that there is strong selection involved in its maintenance. (Duboule and Morata 1994; Lewis 1978; Scott *et al.* 1983).

Hox proteins are found in the nucleus and possess a tri-helix DNA binding motif used to bind specific DNA sequences. The homeodomain consists of three alpha helices that act together to form the DNA binding domain of the Hox protein. The third helix binds to specific bases in the major groove of the target DNA whereas the first and second helices interact with the phosphodiester backbone of the target DNA and also make contacts with bases in the minor groove of the DNA molecule (Boulikas 1992; Corsetti *et al.* 1995; Laughon 1991; Shen *et al.* 1996). Hox proteins act as transcription factors early in development to specify the anterior-posterior axis of the developing embryo.

I propose the term "hoxon" - defined as a complex of homeobox- containing genes in which genes composing the complex have (1) the same transcriptional orientation, (2) order of transcriptional activation, and (3) colinear expression along the body axis of developing embryos- to describe the clustered arrangement of *Hox* genes in the chordates. The conserved nature of the hoxon stimulates the investigation of the mechanisms that control the development and evolution of species. *Hox* genes have been identified in all metazoans examined to date (Bürglin *et al.* 1991; *et al.* 1993; Degnan *et al.* 1995; Di Gregorio *et al.* 1995; Garcia-Fernández and Holland 1994; Kuhn, *et al.*

1996; Pendleton *et al.* 1993; Pfeffer and Holt 1991; Schummer *et al.* 1992; Seimiya *et al.* 1994; Snow and Buss 1994; Stein *et al.* 1996).

Hox gene organization

In the following sections, the organization (and function, if known) of the homeobox-containing genes in the organisms whose hoxon has been characterized is described and briefly discussed. The standardized nomenclature proposed by Scott (1992) for naming *Hox* genes is used throughout this thesis.

Amphioxus (*Branchiostoma floridae*), is a cephalochordate with a dorsal nerve cord, has recently been shown to possess a hoxon composed of a single cluster of 10 *Hox* genes (spanning 270 kb). All are linked on a single chromosome, are expressed in the embryo in a manner colinear to their occurrence in the cluster, and are activated temporally in accordance with their location within the cluster. Given the taxonomic position of amphioxus, and its relatively simple hoxon, the amphioxus hoxon is presumed to represent the basal form of *Hox* gene evolution in vertebrates (Garcia-Fernàndez and Holland 1994). Amplification of an amphioxus-like hoxon, presumably by repeated cluster duplications or chromosomal duplications is proposed to have generated a hoxon composed of multiple clusters as evidenced by the four-cluster arrangement observed among extant vertebrates. Each of the ten amphioxus *Hox* genes is assigned to a specific *Hox* paralog group based upon homology of the homeodomain with mammalian homeodomains (Figure 1). Members within a paralogous group are more related to each other based on the DNA sequence than they are to genes in other paralogous groups. The

expression of the amphioxus genes appear to maintain colinearity and resemble the expression of *Hox* genes in mice and humans, as demonstrated by the retinoic acid-induced expression of *Amphihox-1* in the anterior end of the nerve cord (Holland and Holland 1996).

Molecular genetic analysis of the mammalian hoxon, as represented by mice and humans, revealed 39 *Hox* genes located on four chromosomes and arranged in four clusters, with nine to eleven linked genes per cluster (Figure 1). DNA sequence analysis and comparison of the hoxon-associated genes revealed that the mammalian genes could be arranged into 13 distinct paralogous groups, in which the members of each paralogous group could be distinguished from other groups by virtue of DNA sequence differences in the homeobox-encoding genes both within and outside of the homeobox.

Individual *Hox* genes associated with the hoxon are assigned a specific alphanumeric designation based on their cluster association as well as their location in the cluster relative to other paralogous genes. Each of the four mammalian clusters have been given single letter designations A, B, C, and D. Further, individual genes representative of the 13 paralogous groups are numbered 1-13 depending on their position in each cluster as well as their sequence relatedness to genes in other clusters. For example, the *HoxA1* gene designation would be assigned to a gene on cluster A located at the most 3' region of the cluster and expressed in the most anterior region of the developing embryo.

Interestingly, the complement of *Hox* genes on the four clusters is different (Figure 1). The mammalian *HoxA* cluster is missing cognate genes from paralogous groups 8 and 12, whereas cognate genes from paralogous groups 1-9 and 13 are present on the *HoxB* cluster (Zelster *et al.* 1996). Similarly the *HoxC* cluster is missing cognate genes from paralogous groups 1-3 and 7, whereas the *HoxD* cluster is missing cognate genes from paralogous groups 2, and 5-7. If, as hypothesized, the hoxon of vertebrates is simply the product of two successive genomic duplication events beginning with a single ancestral cluster reminiscent of amphioxus (Holland and Garcia-Fernández 1994; Holland and Garcia-Fernández 1996; Ruddle *et al.* 1994; Schughart *et al.* 1989; Zhang and Nei 1996), then the organization of *Hox* genes on each of the four clusters should be identical. However, this is clearly not the case in mammals, nor is it the case in the recently characterized hoxon of the pufferfish.

Recently, the complete cluster arrangement has been reported for the pufferfish *Fugu rubripes* (Aparicio *et al.* 1997). *Fugu* has four clusters of *Hox* genes but has a different complement of *Hox* genes than human and mouse (Figure 1). *Fugu* is missing paralog 6 from the *HoxA* cluster, paralog 7 from both the *HoxA* and *HoxB* cluster, paralog 13 from the *HoxB* cluster, has 2 pseudogenes representing *HoxC1* and *HoxC3*, has a new representative of *HoxD2*, and is missing paralogs 1,3,4,8 and 12 from the *HoxD* cluster. Interestingly, data generated in our laboratory has shown that *Morone saxatilis* (striped bass) is missing paralogs 6 and 8 from the *HoxA* cluster. Overall, the vertebrate hoxon appears to be conserved as far as the number of clusters and function, but the constitution of each cluster varies between species. The regulation and expression

of the vertebrate *Hox* genes is another feature of the vertebrate hoxon that is conserved (Nonchev *et al.* 1996a; Prince and Lumsden 1994; Vesque *et al.* 1996).

Expression, Regulation & Function of *Hox* genes in vertebrates

The transcriptional expression domains of the *Hox* genes in mice and chickens are the most completely characterized among vertebrates. As previously discussed, *Hox* genes are expressed in a colinear fashion with their order of occurrence on DNA, meaning that the genes are turned on in a sequential order based on their location in the cluster and along the anteroposterior axis of the developing embryo (Duboule and Dolle 1989; Gaunt *et al.* 1988; Graham *et al.* 1989). The genes that are expressed in the region of the embryo fated to form the anterior structures are located at the 3' end of the cluster, with the most posteriorly expressed genes at the 5' end of the *Hox* cluster. For example, *HoxB1* is initially expressed in rhombomere 2 and is thus the most anteriorly expressed *Hox* gene. The anterior limit of *HoxB2* occurs at the boundary between rhombomeres 2 and 3 of the developing mouse, and extends through rhombomeres 4 and 5. *HoxB3* is expressed at the junction of rhombomeres 4 and 5, with the highest expression found in rhombomere 5. *HoxB4* is expressed posterior to *HoxB3* at the rhombomere 6 and 7 junction, and anterior to the *HoxB5* anterior expression limit. Clearly, the expression domains of these genes overlap, and provide a code of positional information. A recent paper reports that the colinear expression of the *HoxB* genes extends throughout the developing foregut and lung of the mouse embryo (Bogue *et al.* 1996). Transcriptional expression of the *Hox* genes in clusters other than the *HoxB* cluster also follow a colinear

expression pattern and also have domains of expression that overlap with the *HoxB* genes. Recently, a survey of the branchial *Hox* gene expression was conducted in human embryos. It showed that the genes of paralogous groups 1 through 4 are expressed in very defined domains and are tightly regulated in their expression. The expression domains were shown to be homologous with those in the mouse and other vertebrates (Vieille-Grosjean *et al.* 1997).

Since the transcriptional regulation of *Hox* gene expression in developing vertebrate embryos is baroque, I shall focus the discussion to those *Hox* genes most clearly related to the genes characterized in my thesis research, namely *HoxB1*, *HoxB2* and *HoxB3*. Based on previously published information, these three *Hox* genes are expressed within the anterior or branchial region of the developing embryo. Although their expression domain is limited to the branchial regions, it appears that these three genes, in concert with other members of paralogous groups 1, 2, and 3, are responsible for branchial specification of tissues derived from embryonic ectoderm, mesoderm and endoderm. At the molecular level, there appear to be several effector molecules that regulate *Hox* gene expression.

Retinoic acid stimulates the expression of the *Hox* genes in the developing vertebrate embryo (Dekker *et al.* 1992; Marshall *et al.* 1996). Sensitivity to retinoic acid has the greatest effect on the 3' genes and least effect on the expression of the 5' genes of the *Hox* clusters. Among the three signaling pathways (transcriptional activators) known to exist for mouse *HoxB1*, two pathways are modulated by retinoic acid (Figure 2). The first of these is located 3' of the gene and consists of a retinoic acid response element

(RARE) constructed of two functional elements that are conserved between chicken, mouse, human and pufferfish; DR-2B and DRE (Marshall *et al.* 1994; Ogura and Evans 1995a; Ogura and Evans 1995b). The DR-2B element is bound directly by the retinoic acid receptor/retinoid X receptor (RAR-RXR) heterodimer, whereas the DRE is directly bound by a coactivating protein referred to as retinoid-activating protein (RAP). Deletion or mutation of this RARE disrupts the expression of *HoxB1* observed in rhombomere 2 in the chicken, mouse and pufferfish. *HoxB1* is initially expressed in rhombomere 2, then the expression becomes restricted to rhombomere 4. (Dupe *et al.* 1997; Frasch *et al.* 1995; Marshall *et al.* 1994).

In the second pathway regulated by retinoic acid, a 5' RARE restricts the expression of *HoxB1* to rhombomere 4 and represses expression of *HoxB1* in rhombomeres 3 and 5 (Marshall *et al.* 1996). The 5' located RARE like the 3' located RARE consists of two elements, the first element DR-2A is bound by RAR-RXR and the second element URE is bound by the retinoid inducible protein (RIP). Deletion of the 5' RARE (including both DR-2A and URE) eliminates transcriptional regulation by retinoic acid, as does deletion of either of the two RARE components. Only when both DR-2A and URE elements are present is full inducibility of *HoxB1* exhibited in the presence of retinoic acid (Ogura and Evans 1995b).

The third signaling pathway that activates transcription of *HoxB1* is an auto-regulatory loop, which involves modulation of *HoxB1* transcription expression by the HoxB1 protein. Three sites are located upstream of the *HoxB1* gene that are bound by the HoxB1 protein and its cofactor Pbx. Members of the Pbx protein family are known

to modulate the DNA-binding activity of Hox-proteins (Chan and Mann 1996; Knoepfler and Kamps 1995; Neuteboom *et al.* 1995; Peltenburg and Murre 1997; Phelan *et al.* 1995). Deletion of the HoxB1 binding sites results in transcriptional inactivation of *HoxB1* in rhombomere 4 of the mouse. Comparison of the DNA sequence of the *HoxB1* intergenic region between the mouse and pufferfish revealed that the HoxB1 binding sites are conserved, which suggests that the HoxB1 binding sites may be of general importance in vertebrate development (Pöpperl *et al.* 1995).

Among vertebrates, the segmented development of the hindbrain is conserved with respect to the number of rhombomeres, the neuroanatomy and patterns of gene expression (Nonchev *et al.* 1996a). *HoxB2* is expressed at the rhombomere 2 and 3 boundary and is most intensely expressed in rhombomeres 3 and 5 (Nonchev *et al.* 1996a; Prince and Lumsden 1994; Vesque *et al.* 1996). *Krox-20* is a zinc-finger motif protein known to maintain the expression of *HoxB2* (and *HoxA2*) in rhombomeres 3 and 5 (r3/r5) of the mouse and chicken (Nonchev *et al.* 1996b; Sham *et al.* 1993). In *Krox-20*^{-/-} mice, mutants fail to express *HoxB2* in this region and the rhombomeres are lost shortly after forming, suggesting that *HoxB2* is important in the specification and maintenance of these structures (Schneider-Maunoury *et al.* 1997).

In mouse, *Krox-20* acts upon *HoxB2* through three conserved *cis*-binding sites located in the upstream enhancer region of *HoxB2*. Gel shift mobility assays demonstrate that one of these three *Krox-20* binding sites is a high affinity binding site, and the other two sites are low affinity *Krox-20* binding sites. The mouse *Krox-20* high affinity site is

identical to one of two *Krox-20* binding sites found in the mouse *HoxA2* enhancer (Nonchev *et al.* 1996a; Nonchev *et al.* 1996b). The high affinity site is absolutely required for activity of the *HoxB2* enhancer, as indicated by using point mutations to inactivate the *Krox-20* sites in transgenic mice.

Point mutations in all three *Krox-20* binding sites deactivate the r3/r5 enhancer. Only the high affinity site is necessary for directing the r3/r5 expression of the transgene, whereas point mutations in the two low affinity *Krox-20* binding sites do not result in a marked decrease of transgene expression. Constructs containing only the high affinity *Krox-20* site in multiple copies failed to induce expression of the transgene in r3/5 of the mouse embryo, suggesting that other *cis*-acting elements are needed to direct the r3/r5 expression of *HoxB2* in mouse (Vesque *et al.* 1996).

In addition to the *Krox-20* binding sites, another *cis*-acting element (Box1) has been found in the *HoxB2* enhancer and is conserved in sequence between human, mouse, pufferfish and chicken and in position relative to the high affinity *Krox-20* binding site (Nonchev *et al.* 1996a; Vesque *et al.* 1996). Replacing the Box1 motif with an unrelated sequence in a transgenic mouse demonstrates that Box1 is essential for correct expression of *HoxB2* in rhombomeres 3 and 5 by showing failure of the correct spatial expression pattern in r3 and r5 of the reporter gene. Box1 and the high affinity *Krox-20* binding site are both necessary for the correct function of the *HoxB2* r3/r5 enhancer in mouse, chicken and pufferfish. A transcription factor that binds the 8 bp palindromic Box1 has not yet been identified. The unknown Box1 factor is suspected to bind to the Box1 element and to cooperate with *Krox-20* in regulating the rhombomere 3 and 5 expression

of *HoxB2* (Nonchev *et al.* 1996a; Nonchev *et al.* 1996b; Vesque *et al.* 1996). Aside from the *Krox-20* and *Box1* elements controlling the r3/r5 restricted expression of *HoxB2*, other sequence elements have been identified that may play a role in *HoxB2* gene expression in other regions of the embryo.

HoxB2 is expressed in the erythroid cell line K562, in the erythro-megakaryocytic line HEL, megakaryocytic line Meg01, in two myeloid cell lines KG1a and KG1, in human bone marrow and progenitor cells (Sengupta, Lavelle, and DeSimone 1994; Vieille-Grosjean and Huber 1995). In the erythroid cell line K562, *HoxB2* is expressed under the control of a *GATA-1* binding site. Mutation of the *GATA-1* site located 48 bp upstream of the transcription start site resulted in an 85% decrease in reporter gene activity. Six *Ets* (a class of oncoproteins) recognition sites were also identified upstream of the human *HOXB2* gene but the importance of these sites in directing the expression of *HoxB2* has not been determined (Vieille-Grosjean and Huber 1995). Several factors have been identified that function upstream of *HoxB2*, whereas, only a few of the downstream targets of *HoxB2* have been identified.

Targeted disruption of the *HoxB2* locus in mouse results in altered expression patterns of *HoxB1* and *HoxB4*, suggesting that these genes are downstream targets of *HoxB2* (Barrow and Capecchi 1996). Seventy-five percent of the *HoxB2* targeted mutants died within 24 hours of birth, the surviving embryos showed serious defects. The *HoxB1*-like mutant mice, resulting from the targeted disruption of *HoxB2*, showed facial paralysis that resulted from the failure to form the VIIth facial nerve. The *HoxB4*-like mutant mice, generated by disrupting the *HoxB2* gene, exhibited sternal defects. *HoxB2*

and *HoxB4* are thought to work in concert to mediate the correct closure of the ventral thoracic body wall. In the vertebrate embryo, *HoxB2* is expressed most intensely in r3 and r5 with a lower level of expression in r6, and a basal level of *HoxB2* expressed along the remaining portion of the neural tube. In *HoxB4* mutant embryos, the expression of *HoxB2* was decreased in r6, suggesting that *HoxB4* acts as a mediator of *HoxB2* expression posterior of r5 (Barrow and Capecchi 1996). The *HoxB2* regulatory mechanisms are the most completely described to date, although many unknown factors probably act on this locus.

The rhombomere 5 expression of *HoxB3* is regulated by the *kreisler* protein *Kmr11*. Two domains are present upstream of the *HoxB3* gene (Kr1 and Kr2) and are essential for correct expression of *HoxB3* in rhombomere 5 (Manzanares *et al.* 1997). An *Ets-1* binding site is associated with Kr2 and forms the *Ets*-related activation site (ERAS). *Ets-1* interacts with the chicken *Krml* homologue *MafB* during avian haematopoiesis. *Ets-1* cannot be a required factor for expression in rhombomere 5 because it is only expressed in even numbered rhombomeres. As mentioned above, *Ets* recognition sites have been identified upstream of *HoxB2*, but their function remains unknown (Vieille-Grosjean and Huber 1995).

Aside from the individual regulatory mechanisms discussed above, experimental evidence points to regulation of the entire hoxon. Methylated *Hox* regions are expressed at a much higher rate than non-methylated regions (Flagiello *et al.* 1996). Also, the clustering of the *Hox* genes may serve as a form of regulation and thus may explain why these genes are clustered throughout evolution. First, a cascade of controls may be

initiated by the 3' most genes. These early genes probably act to turn on the successive *Hox* genes. Second, the *Hox* genes may respond in a differential manner to a common gradient molecule (much like the response seen to retinoic acid). Third, the genes may only become available to the transcriptional apparatus as the chromatin is transformed from an inactive state to an active state in a sequential manner (van der Hoeven, Zákány, and Duboule 1996). The regulatory mechanisms are likely to be conserved between species, much like the clusters and genes of the *Hox* complex. Differences in the *Hox* complement and how the *Hox* genes are regulated has helped generate the diversity of life.

Morone saxatilis was chosen for this project because it is a higher teleost, and this research will complement ongoing research in the pufferfish *Fugu rubripes* (a Tetraodontiform) and research previously conducted with the zebrafish and goldfish, both members of the Order Cyprinodontiformes, and in the killifish which is an atheriniform. DNA sequence comparisons of the *Hox* genes and the regulatory regions of *Hox* genes among other vertebrates will provide insight into the evolution of these genes and into the evolution of the developmental programs in vertebrates.

Morone saxatilis (Walbaum), also known as striped bass, stripers, or rockfish, is a perciform in the sea bass family Moronidae. The family Moronidae contains both freshwater and anadromous species of fishes and is represented in North America by four members of the genus *Morone*. Members of the sea bass family are called “true basses” to distinguish them from fishes like largemouth, smallmouth and spotted bass, which are members of the sunfish family Centrarchidae.

Striped bass (and other true basses) are characterized by a dorsal fin that is distinctly separated into spiny and soft-rayed portions. Striped bass are silvery, shading to olive-green on the back and white on the belly, with seven or eight uninterrupted horizontal stripes on each side of the body. Young stripers are distinguished from white bass (*Morone chrysops*) by 2 tooth patches on the back of the tongue and two sharp points on each gill cover (white bass have one tooth patch and one point on each gill cover). The second spine on the anal fin of stripers is about half the length of the third spine, while two-thirds the length of the third spine in white bass.

The purpose of this thesis is to identify a *Morone saxatilis* genomic clone that contains at least two homeoboxes and to determine the DNA sequence of the genomic clone. The gene structure of striped bass *Hox* genes including the exon and intron structure as well as the promoter regions of the genes encoded on the genomic clone will be determined. This research will help in determining the physical arrangement of *Hox* genes in striped bass by showing whether these genes are clustered in this species and whether colinearity is preserved in the clusters. The differences and similarities of the mechanisms controlling the expression of the *Hox* genes in striped bass and other vertebrates will be discussed.

Materials and Methods

Library Screening and Clone Identification

A λ EMBL-3 (GenBank accession numbers U02453 and U02425 right and left arms respectively) *Morone saxatilis* genomic library was screened using a polymerase chain reaction (PCR)-generated probe. The probe is specific for *Antennapedia* class *Hox* genes and was generated from striped bass genomic DNA using the following primers:

Hox48-71: 5'-GAGAAAGAGTTYCATTTCAACCG-3'

Hox141-163: 5'-CATGCGICGRTTCTGGAACCAG-3'

(Y = C or T, R = A or G, I = inosine). These primers are directed towards highly conserved regions of the homeobox and were designed by Anthony Pavell (Pavell 1994). The screening process involved titering the library by conducting serial ten fold dilutions in SM buffer (0.1 M NaCl, 0.008 M $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$, 50 mM Tris•cl pH 7.5, 0.01% gelatin), then using 100 μl of each dilution to infect 4.0×10^6 cells of *E. coli* strain P2-392 cultivated on Luria Broth (Luria, Adams, and Ting 1960; Luria and Burrows 1957) supplemented with 0.2% maltose and 10.0 mM MgSO_4 . After infection the cultures were incubated for 30 minutes at 37° C in an incubator. The cultures were plated on Luria agar using 3.0 ml of top agar, then incubated overnight at 37° C. The titer was determined by counting phage plaques on the plates. Once the titer was determined for the library, initial screening proceeded by plating 5.0×10^5 plaque forming units (p.f.u.) per 150 x 15 mm petri plates followed by overnight incubation at 37° C. A total of 20 petri plates at a plaque density of approximately 5.0×10^5 p.f.u./petri plate were plated for

use in the primary library screening step. After plaque formation, the plates were placed in the refrigerator at 4° C to cool for at least two hours. Phage were adsorbed from plaques onto 137mm circular Duralon-UV™ membranes (Stratagene, 11011 North Torrey Pines Road, La Jolla, CA 92037) then processed according to the Benton-Davis procedure (Benton and Davis 1977). The phage DNA was permanently bound to the Duralon-UV™ membranes using a Stratagene UV Stratalinker™ (model 1800). Hybridization using a α -dCTP ^{32}P labeled probe with a specific radioactivity of $\sim 5.0 \times 10^8$ dpm/ μg was carried out using relaxed hybridization conditions (55° C, 2X SSC) for 16 hours. After washing for two hours with three changes of 2X SSC and 0.1% SDS at 55° C, the filters were dried, mounted and exposed to X-ray film (Kodak X-OmatAR, Eastman Kodak Company, Rochester, N.Y. 14650) for 96 hours at -70° C. Once the film was developed, plaques that hybridized strongly were identified and subjected to secondary and tertiary screening steps. After the tertiary screening step, all plaques hybridized to the probe, indicating that each clone was purified and free of contaminating clones. A single clone designated λ Clone 2H was picked for further study because it was one of 14 clones that exhibited a strong hybridization signal during the primary library screen. It was hypothesized that λ clone 2H might contain two homeobox encoding regions due to the strong signal.

Phage DNA Isolation

During the course of this research two separate methods were used to produce λ phage DNA; both will be described here. Both methods involved generating broth lysates of *E. coli* using a multiplicity of infection (MOI) of 0.01. Lysates were generated by infecting 2.0×10^8 *E. coli* cells/ml (p2392 or XL-1 blue), which had been cultured in fresh Luria broth supplemented with 0.2% maltose and 10mM MgSO_4 with the appropriate amount of phage stock. The lysate progress was monitored by measuring the culture density using the Klett-Summerson photoelectric colorimeter (Model 800-3, Klett MFG. Co., Inc., New York) with a #66 red filter. When maximum lysis was achieved, as determined by a pronounced decrease in culture density, 1.0 ml chloroform/500 ml culture was added to the lysate and incubated for five minutes with shaking. Sodium chloride was added to a final concentration of 1.0 M and swirled until dissolved to facilitate release of the mature phage particles from cell debris. Cell debris and unlysed cells were pelleted by centrifugation for 10 minutes at 4°C at 11,000 G (Sorvall GSA rotor @ 7,000 rpm). This is the point where the two protocols diverge.

The first protocol used to isolate phage DNA is adapted from Yamamoto, *et al.* (1970). Briefly, after pelleting cells and cell debris, polyethylene glycol (PEG) 8000 was added to the supernatant to a final concentration of 10% w/v and mixed until it completely dissolved. The solution was then placed on ice for at least one hour or stored overnight at 4°C to precipitate the phage particles. The sample was then centrifuged in 250 ml OakRidge bottles at 11,000 G for 10 minutes at 4°C (Sorvall GSA rotor @ 7,000

rpm) to pellet the phage particles. The supernatant was poured off and the pellet resuspended in 4.0 ml of SM buffer per 200 ml of supernatant then transferred to 25 ml Corex tubes. Residual PEG was removed by two chloroform extractions using equal volumes of chloroform, mixing well and centrifuging at 3,000x G for 15 minutes at 4° C (Sorvall HB-4 rotor @7,000 rpm). The aqueous phase was collected and placed in a sterile 50 ml Falcon tube. Phage particles were isolated by layering the phage prep onto a CsCl block gradient constructed using 1.0 ml 1.7 g/ml CsCl/SM and 2.0 ml 1.4 g/ml CsCl/SM in Beckman Ultra-Clear 14 x 89 mm centrifuge tubes. Samples were centrifuged for three hours at 24,000 rpm at 4° C using a Beckman SW-41 Ti rotor in a Beckman L8-80M Ultracentrifuge (Beckman Instruments, Palo Alto, Ca. 94304). The purified mature phage particles were then collected from the gradient using a 21 gauge needle and a 1.0 ml syringe. Phage DNA was isolated from the phage by adding 0.1 volume of 2.0 M Tris•cl pH 8.5, 0.2 M EDTA mixing by inversion, then one volume of formamide was added and mixed. The solution was incubated for 30 minutes at room temperature to allow for denaturation of the phage particles. DNA was precipitated by adding two volumes of absolute ethanol and mixing by inversion. The tube was stored at -20° C for at least two hours to maximize the yield of DNA. The DNA was pelleted by centrifugation at 5,000 rpm for five minutes using an Eppendorf Micro Centrifuge (model 5415 C, Brinkmann Instruments, Inc. Westbury, N.Y. 11590). The supernatant was removed by pipetting and the pellet was washed several times with 250 µl of 70% ethanol. The DNA pellet was then dried using a Jouan Centrifugal Vacuum Concentrator

model RC 10.10 (Winchester, Va. 22601) and suspended in Tris low EDTA (10 mM Tris•Cl pH 8.0, 0.1 mM EDTA pH 8.0).

The second protocol used to isolate phage DNA was developed by Jean-Luc Scemama (personal communication) and produced a substantial yield of highly purified DNA. After centrifugation of the lysates to remove the cell debris, DNase I (Sigma, St. Louis, MO 63178) and RNase A (Sigma) were added to the phage solution and incubated overnight at 4° C or for two hours in a 37° C water bath. Polyethylene glycol (PEG 8000) was then added to a final concentration of 10% w/v and the solution then incubated on ice for two hours. After PEG treatment, the phage were pelleted by centrifugation in 250 ml Oakridge bottles (Sorvall GSA rotor 7,000 rpm for 10 minutes at 4° C) then suspended in 12.0 ml SM buffer. Residual PEG was removed from the solution by two chloroform extractions with a volume of chloroform equal to the volume of the aqueous phase. After chloroform extraction, a volume of 540.0 µl of 0.5M EDTA pH 8.0 was added to the tube and mixed well, followed by 300.0 µl of 10 % SDS then mixed by inversion. The solution was then treated with 120.0 µl of 10 mg/ml proteinase K (Sigma) and incubated for two hours at 58° C in a water bath. The solution was then extracted twice with an equal volume of phenol, chloroform, and isoamyl alcohol (PCI) (25:24:1 v/v). After PCI extraction, the aqueous phase was extracted twice with an equal volume of chloroform. The DNA was precipitated by adding an equal volume of isopropanol and harvested by spooling it on the end of a flame sterilized glass pipet, then washed twice with 500.0 µl of 70% ethanol and allowed to dry overnight in a sterile 1.5

ml microcentrifuge tube. The DNA was resuspended using 200.0 μ l of Tris low EDTA and heated at 65° C for 15 minutes.

Following the preparation of phage DNA, a 1.0 μ l sample of the preparation was electrophoresed on a 0.7% agarose gel containing 0.05 mg/ml ethidium bromide including molecular length standards (1 Kb DNA ladder, Gibco-BRL, 8400 Helgerman Court, Gaithersburg, MD 20884-9980) and High Molecular Weight DNA Markers (Gibco-BRL) to determine the size of the phage DNA and the concentration as well as the purity of the DNA. If RNA was present, the sample was treated with RNase A, PCI extracted then ethanol precipitated to remove the RNA.

The concentration of the DNA was determined using a Pharmacia Gene Quant spectrophotometer (Biochrom Ltd., Cambridge, England). Briefly, three serial two fold dilutions of a 1.0 mg/ml solution of λ DNA using Tris low-EDTA were conducted, then the absorptions at 260 nm for the 1.0 mg/ml, 0.5 mg/ml, 0.25 mg/ml, 0.125 mg/ml dilutions were measured and plotted against concentration to produce a standard curve. Tris low-EDTA was used to zero the instrument. The concentration of the DNA in the preparation was then determined by measuring the absorption of a sample and plotting this value against the standard curve. Purity was determined by observing a 1.0 μ l sample that was electrophoresed on a 0.7% agarose gel containing 0.05 mg/ml ethidium bromide, and by measuring the 260/280 ratio of absorption. Yields of greater than 1.0 μ g/ μ l of highly purified λ DNA were consistently prepared using Scemama's method.

Restriction Analysis

A restriction endonuclease map of λ clone 2H was constructed by conducting single digests of λ clone 2H DNA using the restriction endonucleases *Bam*H I, *Eco*R I, *Hind* III and *Sal* I (Gibco-BRL, 8400 Helgerman Court, Gaithersburg, MD 20884-9980) in conjunction with double digests using the enzymes *Bam*H I x *Eco*R I, *Bam*H I x *Hind* III, *Bam*H I x *Sal* I, *Eco*R I x *Hind* III, *Eco*R I x *Sal* I, and *Hind* III x *Sal* I. The restriction digests were conducted using 1.0 μ g of λ clone 2H DNA, 10 units of each restriction enzyme per reaction, the appropriate buffer for each enzyme, and incubated for two hours in a 37° C water bath. The reactions were stopped by heating the reaction tubes for fifteen minutes in a 65° C water bath, followed by two extractions with an equal volume of PCI. The digested DNA was then precipitated by adding 5.0 M NaCl to a final concentration of 0.3 M followed by the addition of an equal volume of absolute ethanol. To increase the yield of DNA precipitated, the tubes were incubated for two hours at -20° C. Immediately following incubation, the DNA was pelleted by centrifugation for five minutes at 10,000 rpm using an Eppendorf microcentrifuge. The supernatant was removed and the DNA pellet washed twice with 250 μ l of 70% ethanol, then dried in the Jouan Vacuum Concentrator. The pellet was resuspended in 20.0 μ l of Tris low-EDTA and heated at 65° C for fifteen minutes. A 0.7% agarose gel containing 0.05 mg/ml ethidium bromide was used to electrophorese the restriction endonuclease digestion products. Molecular weight standards were included in the gel to provide a guide in

determining the size of the restriction fragments. The gel was photographed for use in determining the locations of the restriction endonuclease cleavage sites on λ clone 2H.

Southern Blot Analysis

A Southern blot was conducted on λ clone 2H by blotting a 0.7% agarose gel in which the single and double digestions described above had been electrophoresed. The probe used for the Southern Blot was the same probe used in the library screening. A Schleicher & Schuell Turboblotter™ Rapid Downward Transfer System (P.O. Box 2012, Keene, NH 03431) was used to transfer the DNA from the agarose gel to a Maximum Strength Nytran® Nylon membrane following the instructions included with the kit. Briefly, the DNA was acid depurinated by soaking the gel in 0.25 M HCl for thirty minutes, changing the acid once at fifteen minutes. Following the acid depurination, the gel was rinsed with deionized water, then treated with the denaturation buffer (0.5 M NaOH, 1.5 M NaCl) for thirty minutes, rinsed with deionized water, neutralized by soaking the gel in neutralizing buffer (0.5 M Tris•Cl pH 7.0, 1.5 M NaCl) for thirty minutes, changing the solution once at fifteen minutes, then the gel was soaked in the 20X SSC transfer buffer (3.0 M NaCl, 0.3 M Na citrate, pH 7.0) for thirty minutes. The blotting apparatus was constructed as shown in the protocol booklet and the DNA transferred to the filter using 20X SSC. Following the transfer, the DNA was irreversibly bound to the filter using the Stratagene UV crosslinker on the autocrosslink setting.

Pre-hybridization was carried out using prehybridization buffer at 65° C in a Stratagene Hybridizer 600 Hybridization oven for twenty-four hours. Hybridization was

carried out using 0.25X prehybridization solution containing the ^{32}P -labeled DNA probe under high stringency conditions at 65°C. Hybridization was allowed to proceed for twenty-four hours followed by three washes with prehybridization solution over a period of two hours, then the filter was dried and placed on Kodak Xray film for two days at -70° C in a film cassette. The film was developed using Kodak developer and fixer. Results from the Southern blot were used to identify the restriction fragments containing a homeobox.

Subcloning

In order to conduct DNA sequence analysis of λ clone 2H, it was useful to subclone the insert of λ clone 2H into the plasmid vector pUC-19. Regions corresponding to the 5200 bp and 2300 bp *Sal* I restriction fragments of λ clone 2H were subcloned into the plasmid vector pUC-19 (GenBank accession # M77789) using a shotgun approach. Both the plasmid vector and λ clone 2H DNA were digested to completion using the restriction enzyme *Sal* I. After *Sal* I digestion of the pUC-19 plasmid vector and the clone 2H DNA, the DNAs were mixed using a molar ratio of 3:1 (insert to vector) and ligated using T4 DNA ligase (Gibco-BRL) by incubating for twelve hours at 14° C in a Lauda RM-6 recirculating refrigerated water bath (Brinkman Instruments, Westbury, N.Y. 11590). Following the ligation reactions, the DNA was extracted twice with PCI, ethanol precipitated and resuspended in 20.0 μl of Tris low-EDTA. A 5.0 μl aliquot of the ligated DNA was electrophoresed on a 0.7% agarose gel containing 0.05 mg/ml

ethidium bromide along with a negative control of unligated *Sal* I digested 2H λ DNA to ensure that ligation occurred. The ligated DNA was used to transform the *E. coli* ER 1458 (New England Biolabs) following the protocol of Hanahan (1983). The transformed cells were plated on Luria Agar plates containing 100 μ g/ml ampicillin, 50 μ g/ml X-Gal (5-bromo-4-chloro-3-indolyl- β -D-galactoside; Gibco-BRL), 0.0001 M IPTG (isopropylthio- β -galactoside; Gibco-BRL). Successful recombinant transformants were identified as white colonies, 20 of which were picked using a flame sterilized bacteria loop and streaked for isolation on fresh ampicillin, X-Gal, IPTG Luria Agar plates. These twenty colonies were characterized by conducting rapid plasmid isolations followed by restriction endonuclease digestions. Briefly, a single colony was picked from a petri plate using a flame sterilized bacteriological loop, which was used to inoculate 800 μ l of Terrific Broth (12.0 g tryptone, 24.0 g yeast extract, 4.0 ml glycerol, 2.31 g KH_2PO_4 , K_2HPO_4 per liter); (Tartof and Hobbs 1987) containing 100 μ g/ml ampicillin in a 1.5 ml microcentrifuge tube. The culture was incubated for eight hours in a shaking incubator at 37° C. The cells were harvested by centrifugation at 10,000 rpm for two minutes using an Eppendorf microcentrifuge, carefully pouring off the supernatant, and suspending the cell pellet using 300 μ l STE (300 mM NaCl, 10 mM Tris pH 8.0, 20 mM EDTA pH 8.0). The solution was extracted once with an equal volume of PCI, and the DNA precipitated by addition of two volumes absolute ethanol and incubation for thirty minutes at -20° C. The DNA was pelleted by centrifugation for five minutes at 10,000 rpm using an Eppendorf microcentrifuge, the supernatant poured off and the pellet washed using 100 μ l

70% ethanol. The pellet was dried using the Jouan Centrifugal Vacuum Concentrator then suspended in 20.0 μ l Tris low-EDTA. RNA was removed from the samples by adding 1.0 μ l of 10 mg/ml RNase A (Gibco-BRL) and incubating at 37° C for thirty minutes. *Sal* I restriction digests were conducted using 5.0 μ l of the DNA preparation, 10 units of restriction enzyme *Sal* I, and 2.0 μ l of React10 10X reaction buffer in a total volume of 20.0 μ l, incubating at 37° C for two hours. The reactions were stopped by heating at 65° C for fifteen minutes, then the complete reaction solution was electrophoresed on a 0.7% agarose gel containing 0.05 mg/ml ethidium bromide. DNA preparations identified as containing inserts with sizes corresponding to regions 2.2 and 2.3 of λ clone 2H were further digested with restriction enzymes known to cleave in these regions. Once the identity of the DNA preparations was determined, the corresponding colonies were maintained in continuous culture using Luria agar supplemented with 100 μ g/ml ampicillin. The 2.1 region of λ clone 2H was not found in any of the transformants.

In order to facilitate DNA sequencing of λ clone 2H , targeted subcloning was also used to subclone a small *Pst* I fragment contained within region 2.1 of λ clone 2H. Region 2.1 is too large (8138 bp) to transform *E. coli* efficiently. The procedure for subcloning the *Pst* I fragment involved excising a DNA band of interest (*Sal* I fragment 2.1) from an agarose gel, then purifying using the GENECLAN® kit (Bio 101, Inc., La Jolla, CA 92038) according to the protocol , followed by ligation into *Pst* I digested pGem-3zf (GenBank accession # X65306). The ligated DNA was then used to transform

E. coli ER 1458 following the protocol of Hanahan (Hanahan 1983). Ampicillin resistant white colonies were screened for an insert 900 bp in length using rapid DNA minipreps and restriction endonuclease digests.

Subclone 2.2 of λ clone 2H was further subcloned into 2 smaller *Sal* I x *Hind* III fragments, in order to generate molecules that were easier to sequence, by excising the *Sal* I fragment from an 0.7% agarose gel containing 0.05 mg/ml ethidium bromide, followed by purification of the appropriate fragments using the GENECLAN[®] procedure. The fragment was digested with *Hind* III, then ligated, using T4 DNA ligase, into the vector pUC-19 which had previously been digested with the restriction enzymes *Sal* I and *Hind* III. Transformation and screening of the two expected subclones was as previously described.

DNA Sequencing

Plasmid template for DNA sequencing was prepared using a modified alkaline lysis protocol as described in the Prism[™] Ready Reaction DyeDeoxy[™] Terminator Cycle Sequencing Kit Protocol (Revision A). λ clone 2H phage DNA template for sequencing was prepared using the Scemama method described above for isolating phage DNA. DNA sequencing was carried out using the protocol described in the ABI PRISM[™] Dye Terminator Cycle Sequencing Ready Reaction Kit With AmpliTaq[®] DNA Polymerase, FS Protocol (1995) using an ABI 373A Stretch DNA sequencer (Applied Biosystems, Inc., Foster City, CA 94404). Briefly, 1.0 μ g of template was used in each reaction (whether the template was plasmid DNA or λ DNA), with 9.0 pmol of primer and 8.0 μ l

TaqFS ready reaction mix in a total volume of 20.0 μ l. The reactions were cycled using a MJ Research MiniCycler™ (Watertown, MA) according to the following steps: denaturation at 96° C for 30 seconds, annealing at 50° C for 15 seconds, extension at 60° C for 4 minutes; repeated for 25 cycles. Following the thermal cycling, the extension products were purified using Princeton Separations Centri-Sep Spin Columns (Adelphia, N.J. 07710), following the instructions supplied with the spin columns, and dried using the Jouan vacuum concentrator. The samples were suspended (4 μ l) in 5 parts formamide/1 part 50mM EDTA pH 8.0 loading buffer, heated for two minutes at 96° C and stored on ice until they were loaded on the sequencing gel. Samples not electrophoresed immediately were stored in the dried form at -20° C. A 4.75% polyacrylamide sequencing gel was constructed using 40% (19:1 acrylamide/bis-acrylamide) stock solution (Bio-Rad Laboratories, Hercules, CA 94547) and 10X TBE (108.0 g Tris, 55.0 g Boric acid, 8.3 g Na₂EDTA per liter). The subclones 2.2 and 2.3, and region 2.1 of λ clone 2H were completely sequenced using primer walking. A list of the primers used is provided in Appendix A.

Data Analysis

DNA sequence data was processed on a MacIntosh PowerPC 7200/90 computer using the ABI DNA Sequencing Analysis program version 2.0.1 (Applied Biosystems 1994) to create an electrophoresis data file and a DNA sequence text file. The electrophoresis data file for each sequencing reaction was then processed using the ABI Factura™ program (Applied Biosystems 1995b) to identify vector sequences and to

identify ambiguous bases with an IUB code. The analyzed data files were then assembled into contiguous DNA segments using the ABI AutoAssembler™ program (Applied Biosystems 1995a). Individual DNA sequence data files were compared to identify regions of DNA sequence overlap. Once the overlaps were identified, differences were resolved by comparing the electropherograms from individual sequences and judging the most reliable information based on peak height, signal strength and signal to noise ratio with each other. Regions with extensive DNA sequence ambiguity were resequenced using the same primer, or a new primer if the original primer proved to be inefficient at priming the sequencing reactions.

The organization and orientation of individual subclones used in DNA sequencing were confirmed by two methods. First, the junctions between each subclone were sequenced in both directions using λ clone 2H DNA as the template and junction specific primers. Second, PCR amplification of λ clone 2H was conducted using primers whose predicted extension products overlapped the subclone junctions. The 5 primer combinations (and predicted product length) used were λ left arm x 2.1 R400 (530 bp), 2.1 F7200 x 2.3 F2256 (822 bp), 2.3 R142 x 2.2.2 R1200 (1845 bp), 2.2 F1300 x 2.2 R2295 (1712 bp), and 2.2.1 R1490 x λ right arm (1520 bp). A typical reaction consisted of 1.0 μ g λ clone 2H DNA, 30.0 pmol of each primer, 2.5 units Taq DNA polymerase (Gibco-BRL), 100 nmol dNTP's, 10.0 μ l of 10X PCR buffer (Gibco-BRL), 3.0 μ l 50 mM $MgCl_2$ in a total reaction volume of 100 μ l. PCR was carried out for 30 cycles in which

each cycle was 96° C for 1 minute, 52° C for 1 minute, 74° C for one minute for a total of 30 cycles.

Once the data for all of the subclones was assembled into one contiguous DNA sequence, the Basic Local Alignment Search Tool (BLAST) program was used to preliminarily identify regions of interest. BLAST uses the method of Altschul *et al.* (1990) to search for similarities between a query sequence and all the sequences in a database. In-depth database comparison was conducted using the Wisconsin GCG FASTA search program (GCG 1996). FASTA is a program that conducts a Pearson and Lipman search for similarity between a query sequence and a group of sequences in a database. These searches identified the positions of two homeoboxes on λ clone 2H DNA sequence. The GCG Pileup and Pretty programs were used to compare the homeobox nucleotide and peptide sequences of the two homeoboxes contained in λ clone 2H to homeobox sequences reported to GenBank. The direct sequence comparisons of λ clone 2H DNA sequence with homeobox-containing sequences from the database allowed assignment of both homeoboxes to a cognate group.

The open reading frames (ORF) of λ clone 2H were identified by using the web-based program Splice Site Prediction by Neural Network (Reese and Eeckman 1996; Reese, Harris, and Eeckman 1996) to identify and locate possible exon/intron junctions. The identified splice sites were mapped to the DNA sequence and to the translation of the three forward reading frames. Once a possible open reading frame was discovered, the DNA sequence encompassing the ORF was translated in all six reading

frames and used to search the GenBank non-redundant database using the GCG FASTA program. The search results were used to identify the correct ORF for each exon encoded in λ clone 2H. Once the open reading frames for each coding region had been defined, a direct sequence comparison of the exons of the corresponding genes in other vertebrates was used to ensure that the DNA sequence of λ clone 2H did not have any frameshifts and to define the exon coding regions.

The poly-A signals and AU rich elements (ARE) downstream of both homeoboxes were identified by scanning the region downstream of the homeoboxes for the consensus poly-A signal "AATAAA" and the consensus ARE sequence "TTATTTAWW", where W is the IUB code for either an A or a T, using the ABI Factura program (Applied Biosystems 1995b).

The DNA sequence of the intergenic region between the two homeobox containing genes were compared to corresponding regions in the chicken, mouse and human using the GCG program BESTFIT to identify regions with at least 70% homology with the λ clone 2H sequence. The homologous regions were then compared using the GCG program FINDPATTERNS to identify any common transcription factors. The λ clone 2H DNA sequence was examined for transcription factor binding sites using the Transcription Element Search Software (TESS) (Prestige 1995) on the World Wide Web. The TESS program used version 3.1 of the TRANSFAC database to search for all known transcription factor binding sites allowing no mismatches. Furthermore, the intergenic enhancer regions of the chicken *HoxB2*, mouse *HoxA1*, mouse *HoxA2*, mouse *HoxB2*,

and human *HOXB2* were searched for transcription factor recognition sites using TESS. The results of the TESS searches were used to identify sequence elements common to the striped bass sequence and the other vertebrate sequences searched.

Additional transcription factor binding elements and possible *cis*-acting elements were identified in λ clone 2H using the GCG FINDPATTERNS program. The consensus binding sequences of factors known to regulate genes orthologous to those contained in λ clone 2H were used as the search patterns. These consensus sequences included *Krox-20* consensus sites, Box 1 consensus sites, *GATA-1* consensus sites, *Ets* sites, and Retinoic Acid Response Elements (RAREs) and are listed in Table 3. Sites identified in this fashion were then mapped to the DNA sequence of λ clone 2H for comparison with corresponding regions in other vertebrate *HoxB2* and *HoxA2* genes.

The locations of possible promoter regions on λ clone 2H DNA were identified using the web based Signal Scan (Prestridge 1991) and Promoter Prediction by Neural Network (Reese and Eeckman 1995) programs. Visual inspection of the software proposed promoters on λ clone 2H, and direct comparison of these regions with the corresponding promoter region from the human *HOXB2* gene using the TESS and FINDPATTERNS results allowed the identification and elucidation of the promoter structure.

To ensure that no other known genes, partial gene or pseudo-genes are present in the non-coding regions of λ clone 2H, database searches using FASTA were conducted on these regions using overlapping blocks of 500 nucleotides. The 500 nucleotide blocks

were also translated in all six reading frames using the GCG PEPDATA program, and the translation was used to search the databases using the FASTA program.

The predicted primary amino acid sequence of the *M. saxatilis HoxB2* gene was determined using the TRANSLATE program from GCG. Comparison of the predicted *M. saxatilis* primary amino acid sequence to that of the human *HOXB2* gene was conducted using several other GCG programs. The two amino acid sequences were aligned using PILEUP with the -INS (in situ) function. The -INS function allows the PILEUP program to maximize the homology of an alignment by comparing the regions downstream of each match separately. Comparison of the predicted secondary structure of the human and striped bass *HoxB2* protein were conducted by analyzing the output of the PLOTSTRUCTURE program using the functions that predict the hydrophobic, hydrophilic, flexible, surface and antigenic regions of the protein sequence. PLOTSTRUCTURE uses the Chou-Fasman algorithm for predicting the secondary structures of amino acid sequences. The molecular weights and isoelectric points of each protein were determined using the PEPTIDESORT and ISOELECTRIC programs.

Results

A bacteriophage λ EMBL-3 recombinant DNA library representing the complete genome of *M. saxatilis* was screened using a homeobox-specific α -dCTP ^{32}P labeled DNA probe. Of the 1.2×10^6 recombinant plaques screened, 14 exhibited strong hybridization; a single such clone henceforth referred to as λ clone 2H was selected for further characterization.

A restriction map of purified λ clone 2H DNA was constructed from the results of reciprocal double digests with the restriction enzymes *Bam*H I, *Eco*R I, *Hind* III and *Sal* I. The composite restriction map of clone 2H is shown in Figure 3. The restriction map of λ clone 2H revealed an insert of approximately 15.5 kb. This result was expected based on the size selection of *Sau*3a partial digestion products used in the construction of the *M. saxatilis* genomic DNA library. The 15.5 kb insert was cleaved twice by *Bam*H I to produce three fragments of 10,600 bp, 2900 bp, and 2000 bp in length. *Eco*R I digestion of the λ clone 2H produced four insert fragments with lengths of 500 bp, 950 bp, 1900 bp, and 7600 bp. *Hind* III produced a single insert fragment of 5400 bp in length, whereas *Sal* I digestion generated three insert fragments of 2200 bp, 5200 bp and 8100 bp in length.

A Southern blot was conducted on λ clone 2H DNA digested either singly or in combination with the restriction enzymes *Bam*H I, *Eco*R I, *Hind* III and *Sal* I using the same α -dCTP ^{32}P labeled DNA probe used to screen the library. Comparison of the Southern blot results with the restriction map provided evidence of two homeoboxes in λ

clone 2H. The homeobox located on the 5200 bp *Sal* I fragment produced an intense hybridization signal, whereas the homeobox located on the 8100 bp *Sal* I fragment, produced a weaker hybridization signal than the 5200 bp *Sal* I fragment. These results indicate that the homeobox located on the 5200 bp *Sal* I fragment has greater homology to the homeobox-specific probe than the homeobox located on the 8100 bp *Sal* I fragment. The existence of two linked homeobox-encoding fragments on λ clone 2H was noteworthy because it suggested that I could establish the genomic organization of a region of the *M. saxatilis* homoxon and investigate details concerning the regulatory *cis*-acting elements in the intergenic region between the two *Hox* genes.

A strategy was developed to sequence λ clone 2H based on the results of the restriction mapping and Southern analysis. The strategy involved subcloning the three *Sal* I generated fragments, into the plasmid pUC-19 to generate shorter fragments more amenable to DNA sequence analysis (Figure 4). The 5200 bp *Sal* I fragment designated subclone 2.2 and the 2200 bp fragment designated subclone 2.3 were readily subcloned into appropriately digested pUC-19. The 8100 bp homeobox-containing fragment generated by *Sal* I digestion of λ clone 2H was designated as region 2.1, but proved to be too large to transform into the host *E. coli*, ER 1458. Subclone 2.2 was further subcloned to yield subclone 2.2.2, a 2900 bp *Sal* I x *Hind* III fragment and subclone 2.2.1, a 2300 bp *Hind* III x *Sal* I fragment. Subclone 2.2.2 contains the homeobox that exhibits the intense hybridization signal. Each of the three subclones and region 2.1 of λ clone 2H were completely sequenced using a strategy known as primer extension based

sequencing. Region 2.1 was sequenced exclusively using λ clone 2H DNA as the template. DNA sequencing of the three subclones and region 2.1 generated a total of 15,574 bp of non-ambiguous double stranded DNA sequence.

In order to assure that the DNA sequence generated from the aforementioned plasmids was assembled correctly, two approaches were used. In the first, primers were designed to sequence those regions representing junctions between each pair of subclones. The junction-specific primers were used to obtain DNA sequence directly from the λ clone 2H DNA template. The DNA sequence data generated from these reactions were used to verify the DNA sequence of junctions created by joining neighboring subclones. In the second approach, opposing primer pairs spanning the junctions of neighboring subclones were chosen to amplify short fragments using the polymerase chain reaction (PCR). Fragment sizes were predicted using the assembled consensus sequence. The predicted and observed fragment sizes of the amplification products are given in Table 1, and a diagram of the chosen PCR reactions is shown in Figure 5. The results of both of these approaches confirmed that the DNA sequences generated from the three subclones and region 2.1 from λ clone 2H represented all the DNA originally present in λ clone 2H. The final assembled DNA sequence for λ clone 2H was 15,574 bp in length and is shown in Appendix B.

The locations of relevant features including the homeoboxes, exons, introns, promoter elements, putative transcription factor binding sites, and poly-A sites are presented in Table 2. The sequence is enumerated from 5' to 3' starting with the *Sal* I

restriction site located in the cloning site of the λ EMBL-3 “left arm” and ending with the *Sal* I site located in the cloning site of the λ EMBL-3 “right arm”.

The regions encoding the two homeoboxes on λ clone 2H were identified based on direct DNA sequence comparisons to other vertebrate homeobox-encoding DNA sequence deposited in the GenBank database maintained by the National Institute of Health (Benson *et al.* 1996). The homeobox located within subclone 2.2.2 was assigned as a *HoxB2* cognate based on DNA sequence homology and a defined nucleotide signature characteristic of orthologous *HoxB2* genes. The DNA sequence comparison of the 2.2.2 homeobox with other paralogous group 2 members is shown in Figure 6. Nucleotides at selected positions within the homeobox-encoding sequence proved to be critical in assignment of the λ clone 2H homeoboxes to a specific cognate group. Nucleotides at positions 40, 82, 84 and 135 of the *HoxB2* cognates of examined species possess a C, C, G and a C, whereas the *HoxA2* cognates are characterized by a T, A, A, and a G. Nucleotides located at positions 40, 82, 84 and 135 of the *M. saxatilis* λ clone 2H subclone 2.2.2 homeobox encoding gene are C, C, G, and C and are all characteristic of those representative of *HoxB2* but not of *HoxA2*. Additional support for the assignment of the subclone 2.2.2 homeobox as a *HoxB2* cognate derives from the comparison of the amino acid translation of the first exons of group 2 paralogs as shown in Figure 7. Note that the *HoxB2* proteins are characterized by “AVLE” at positions 27 - 30 and by “TSSIKES” at positions 34 - 40, whereas the *HoxA2* proteins are characterized by “PVAD” and “SSSIKTS” at these locations. Amino acids at positions 27 - 30 and 34 -

40 in the exon 1 of *M. saxatilis* clone 2.2.2 are AVLE and TSSIKES and are identical to those of human *HoxB2* exon 1 but not chicken, mouse or rat *HoxA2* exon 1. These results clearly support the assignment of this homeobox-containing gene as *HoxB2*.

A comparison among the predicted amino acid translation of the region 2.1 homeobox with those of other vertebrate *Hox* paralog 3 homeodomains is shown in Figure 8. Note that the *HoxA3* homeodomain from mouse and *Fugu rubripes* differs from the *HoxB3* homeodomain by one amino acid at position 41. The *HoxA3* homeodomain is characterized by a threonine at this position and *HoxB3* is characterized by a serine at this position. The *HoxD3* homeodomain differs from the *HoxB3* homeodomain by two amino acids; in position 4 a valine is present in *HoxD3*, whereas an alanine at position 4 is a characteristic of the *HoxB3* homeodomain. The second difference between the *HoxD3* and *HoxB3* homeodomains occurs at position 41, where, like *HoxA3*, the *HoxD3* homeodomain is characterized by a threonine. The predicted amino acid translation for the region 2.1 homeobox contained one mismatch, with respect to other *HoxB3* homeodomains, at position 55 of the homeodomain possessing an asparagine in place of a lysine. Since the homeodomain encoded by the homeobox in region 2.1 does not contain any of the diagnostic residues characteristic of *HoxA3* and *HoxD3* homeoboxes, and an asparagine to lysine substitution is considered a conservative substitution, along with physical linkage to the *HoxB2* homeobox, the homeobox in this region is assigned as *HoxB3*.

In order to determine the structure of the *HoxB2* and *HoxB3* genes encoded on λ clone 2H, a search for the possible open reading frames was conducted by translating the entire sequence of λ clone 2H in all six reading frames using the GCG program PEPDATA (Genetics Computer Group, 1997). Open reading frames were identified at positions 1281 to 2330, 10205 to 10837, and 12147 to 13094. The translation was then used to search the GenBank non-redundant database for any related protein sequences using the GCG FASTA program. All of the protein sequences identified by this search were *Hox* protein sequences. The FASTA results were used to identify the open reading frames.

Once the open reading frames had been identified, additional DNA sequence analysis using the web-based program Splice Site Prediction by Neural Network (Reese, M.G. and F.H. Eeckman, 1996) provided information concerning putative exon-intron boundaries. A consensus donor site with a score of 0.92 was identified at position 10,675 by the web based program Splice Site Prediction by Neural Network and an acceptor site with a score of 0.95 was identified at position 12243. These two splice sites define the exon/intron boundaries of the *HoxB2* gene. A second acceptor site was identified from the λ clone 2H DNA sequence with a score of 0.99 just upstream of the *HoxB3* homeobox at position 1461; however, no donor site was identified upstream of this position. These results suggest that λ clone 2H contains the entire coding region for the *HoxB2* gene including 2 exons located at positions 10,223-10,675 (exon 1) and 12,243-13,094 (exon 2), but only a single complete exon, exon 2 of the *HoxB3* gene,

located at position 1,461 - 2,330. A search of the region immediately upstream of this exon using the FASTA program, failed to provide any evidence for the first exon of *HoxB3*. It is important to note that all of the regions outside of the exon regions defined above are interrupted by multiple stop codons in all three reading frames. This includes the intergenic region between *HoxB2* and *HoxB3*, as well as the introns of both genes. The occurrence of multiple stop codons in all three reading frames in regions outside the exons provides additional presumptive support for the genetic map of λ clone 2H.

Comparison of the amino acid sequence from the second exon of the striped bass *HoxB2* gene with the second exons of the chicken, mouse and rat *HoxA2*, and human *HOXB2* second exons reveals that the striped bass exon contains several insertions (Figure 9). Two large insertions of bases encoding 18 a.a. and 14 a.a. are found in the striped bass exon when compared to all of the other 4 second exons. A 7 a.a. and 8 a.a deletion is found in the human *HOXB2* exon when compared with the chicken, mouse and rat *HoxA2* second exons and the striped bass *HoxB2* second exon. Interestingly, the C-terminal end of these proteins show considerable amino acid sequence homology, thus possibly revealing a conserved function for this region of the protein.

Comparison of the predicted *M. saxatilis HoxB2* amino acid sequence with the human *HoxB2* protein amino acid sequence reveals that although the two proteins differ with respect to size (46,241.37 g/mol *M. saxatilis HoxB2* protein; 37,913.27 g/mol human *HOXB2* protein) and other physical properties, they share significant regions of amino acid similarity (Figure 10). There are four regions at which the two proteins share

significant amino acid homology. These four regions include the amino terminal 43 amino acid residues, the region corresponding to the homeodomain (residues 165-224 in striped bass and 143-202 in humans), a third region corresponding to the hexapeptide motif (residues 111-116 in striped bass and 94-99 in humans), and the region corresponding to the final 38 amino acids located at the carboxyl terminal of the proteins.

Several short amino acid segments are found upstream of the *M. saxatilis HoxB2* homeodomain, which are not found in the human *HOXB2* amino acid sequence. These are located at positions 65-66, 71-84, 108-110 of the *M. saxatilis HoxB2* amino acid sequence. The region downstream of the striped bass *HoxB2* homeodomain shows several amino acid segments that are not present in the human sequence. The extra amino acid residues are found in positions 288-289, 300-304, 320-357, 371-385 and 434-435 in the striped bass *HoxB2* amino acid sequence.

Secondary structure predictions reveal that the striped bass *HoxB2* protein is more flexible than the human protein downstream of the homeodomain. Based on the surface prediction and antigenic index plots, a short region downstream of the homeobox in the striped bass *HoxB2* protein is predicted to have a surface domain which is not predicted in the human protein. The striped bass protein is predicted to have a molecular weight of 46,241.37 g/mol with an isoelectric point of 6.30, whereas the human *HoxB2* protein is predicted to have a molecular weight of 37,913.27 g/mol with an isoelectric point of 4.87.

The locations of poly-adenylation signals and AU rich elements (ARE) for both genes were identified by searching for the consensus sequences: AATAAA and TATTATT. Two putative poly-A sites were identified 3' of both genes. The *HoxB2*

poly-A signals are located at positions 13,781 and 14,159 or 687 bp and 1,065 bp from the end of exon 2. The *HoxB3* poly-A signals are located at positions 2656 and 3011 or 326 bp and 681 bp from the end of the *HoxB3* exon 2. Four consensus ARE's were identified downstream of the *HoxB2* gene, and one possible ARE was identified downstream of the *HoxB3* gene. The locations of the *HoxB2* ARE's are positions 13452, 13707, 13742, and 14036 with the *HoxB3* ARE located at position 2,591.

The probable location of the *HoxB2* promoter was discerned using several independent DNA sequence analysis algorithms. Promoter Prediction by Neural Network (Reese, M.G. and F.H. Eeckman. 1995) predicted with a score of 0.95 that the transcription start site was located at base 10045. The Signal Scan program (Prestige, Dan S. 1995) and the Transcription Element Search Software (TESS) (Schug, J. and G.C. Overton. 1997) identified the location of transcription factor binding elements within the λ clone 2H DNA sequence. Both programs revealed that the core transcription factors TBP and TFIID binding sites are located at position 10015 suggesting that the basal transcriptional apparatus may assemble at a site within 30 bp of the start site from transcription initiation. This arrangement is common for eukaryotic promoters and provides confidence that the region extending from bp 10015 to 10045 of λ clone 2H represents the authentic promoter. An alternate promoter with a score of 0.99 was identified at position 8894 by the Promoter Prediction by Neural Network program, and both the Signal Scan program and TESS confirm that the transcriptional apparatus can

assemble at this site. This result provides presumptive evidence for the existence of multiple *HoxB2* promoters in *M. saxatilis*.

A diagram of the alternate promoter found at position 8,894 is shown in Figure 11. The TATA box is located 30 bp upstream of the putative transcription start site. An *AP2* recognition site is located 70 bp upstream of +1. Interestingly, a *HoxA5* consensus binding site is located 221 bp upstream of +1. Alternate promoters are known to exist for other *Hox* genes such as the *HoxB3* gene in the chicken, mouse, and human. The presence of an alternate promoter for *HoxB2* is significant because only one promoter has been reported for *HoxB2* in the chicken, mouse, human and pufferfish.

Additional transcription factor binding elements and possible *cis*-acting elements were identified in λ clone 2H using the GCG FINDPATTERNS program. The patterns searched for were chosen because they are identified as DNA sequence elements present in corresponding intergenic regions of the chicken, mouse, human and pufferfish. These sites included *Krox-20* consensus sites, Box 1 consensus sites, *GATA-1* consensus sites, *Ets* sites, and Retinoic Acid Response Elements (RAREs) and are listed in Table 3. Sites identified in this fashion were then mapped to the λ clone 2H DNA sequence for comparison with corresponding regions in other vertebrate *HoxB2* and *HoxA2* genes.

Three possible *Krox-20* binding sites are located in the λ clone 2H intergenic region at positions 3386, 2239, and 9717. One *Krox-20* binding site was identified in the intron of *HoxB2*. Only the *Krox-20* site located at position 9717 is closely associated with a possible Box 1 motif. Box1 is a conserved DNA sequence element that is closely

associated with the Krox-20 binding sites of the HoxB2 enhancer in chicken, mouse, human and pufferfish.

Ten *GATA-1* consensus sites were identified in the intergenic region between the striped bass *HoxB2* and *HoxB3* genes. The location of these sites are at positions 2677, 3123, 3497, 6305, 6503, 3266, 4403, 5672, 6152, and 6631. Six *Ets* sites were identified within 1200 bp of the promoter located at position 10015. The positions of these *Ets* sites are 8968, 9001, 9249, 9443, 9480 and 9643. Although it is possible to speculate that these *GATA-1* and *Ets* sites are significant their spatial organization in striped bass does not resemble the organization in the human *HOXB2* intergenic region.

Two partial RAREs were identified at positions 15,436 and 15,477 possibly marking the beginning of the enhancer region controlling the *HoxB1* gene. The RARE located at base 15477 is identical to the *Fugu rubripes HoxB1* RARE, and the RARE located at 15436 is identical to the RAR- α 2 consensus. Further DNA sequence data is needed downstream of λ clone 2H to resolve the structure of the *HoxB1* promoter and gene.

Among the various intergenic regions compared to that of the striped bass *HoxB2* - *HoxB3* intergenic region, a 155 bp region of the human *HOXB2* promoter region exhibited the most extensive homology. This region of homology includes the predicted promoter site extending from position 9,879 to 10,032 in λ clone 2H. A comparison of the *cis*-acting promoter elements common to the intergenic regions of striped bass and humans revealed several striking similarities, which are summarized in Figure 12. Both

the striped bass sequence and human sequence had a TATA box located approximately 30 bp upstream of the putative transcription start (+1) at position 10015, with an *AP2* binding site located around 67 bp upstream of +1. Further, a CCAATT-binding factor recognition site is located 72 bp upstream of the striped bass *AP2* site and 69 bp upstream of the human *AP2* site. Moreover, both promoters possess possible binding sites for *NF*-related transcription factors. The striped bass promoter has a *NF-I* recognition site at -113, whereas the human promoter has a *NF-kappaB* recognition site at -72. In addition to the promoter elements that are conserved both in terms of location and DNA sequence, there are additional elements that differ in location relative to +1 but have conserved sequences.

The striped bass *HoxB2* gene and the human *HOXB2* gene each have *Krox-20* binding sites and a Box 1 motif located upstream of the promoter. The human *Krox-20* site is located 1014 bp upstream of +1, and is located 19 bp downstream of the Box 1 motif. The overall conservation of the DNA sequence elements identified in the human *HOXB2* and striped bass *HoxB2* promoter regions suggests that the developmental regulation of these genes (and their role in the branchial region of the embryo) is under considerable selective pressure.

Comparison of the striped bass *HoxB2* - *HoxB3* intergenic region DNA sequence to information in a recently published manuscript describing the DNA sequence of the pufferfish *HoxB2* - *HoxB3* intergenic region revealed remarkable similarities between certain *cis*-acting DNA sequence elements (Nonchev *et al.* 1996a). The striped bass *Krox-20* site is identical to the pufferfish *Krox-20* recognition site and is located 492 bp

upstream of the ATG . The pufferfish *Krox-20* binding site is located approximately 424 bp upstream of the ATG (Vesque *et al.* 1996). The striped bass Box 1 motif differs from the pufferfish Box 1 motif by 1 bp and is located 44 bp upstream of the *Krox-20* recognition site. The pufferfish Box 1 motif is located 46 bp upstream of the *Krox-20* site.

To ensure that no other known genes, partial gene or pseudo-genes are present in the non-coding regions of λ clone 2H, database searches using FASTA were conducted on these regions using blocks of 500 nucleotides. The 500 nucleotide blocks were also translated in all six reading frames using the PEPDATA program, and the translation was used to search the databases using the FASTA program. None of theses searches revealed the presence of a gene, or portion of a gene with significant homology to any intergenic, or intronic region of λ clone 2H. These results support the conclusion that λ clone 2H does not contain any other genes, and does not contain any partial genes or pseudo-genes.

Discussion

The hoxon, defined herein, is a highly organized genomic structure in which the arrangement of individual, linked homeobox-containing genes shows considerable evolutionary conservation. Based on the results of restriction endonuclease mapping, Southern hybridization analysis and particularly DNA sequence analysis, I conclude that a single λ EMBL-3 genomic clone isolated from the *M. saxatilis* genome contained two *Hox* genes that are orthologs of the mammalian *HoxB2* and *HoxB3* genes. The identification of *HoxB2* and *HoxB3* genes on a single 15.5 kb DNA fragment strongly suggests that these two genes are linked and arranged in this fashion on the genome of *M. saxatilis*. The physical linkage of the *HoxB2* and *HoxB3* genes in *M. saxatilis* is consistent with the genetic linkage of these two genes in other vertebrates and reflects the evolutionary conservation of hoxon genomic structure.

In addition to the general conservation of hoxon structure, certain specific details of sub-genomic hoxon architecture also appear to be evolutionarily conserved. Exon/intron composition, as well as the length of the intergenic region between individual linked *Hox* genes representative of each cluster, are remarkably similar. Based upon DNA sequence analysis, I determined that the *HoxB2* and *HoxB3* genes of striped bass consist of two exons and one intron. The homeobox is located near the 5' end of the second exon in both genes. This mirrors the organization found in other chordates, specifically among the genes which have been completely characterized, i.e., the amphioxus *Amphihox-3* gene, the mouse *HoxB3* gene and the human *HOXB2* gene. As in

other vertebrates, the direction of transcription of the striped bass *HoxB2* and *HoxB3* genes is from the *HoxB3* side of the cluster (5' posteriorly expressed genes) towards the *HoxB2* side of the cluster (3' anteriorly expressed genes). The size of the intergenic region between *HoxB2* and *HoxB3* in striped bass (8.0 kb) is similar to that observed in the cephalochordate amphioxus between the *HoxB2* and *HoxB3* orthologs *Amphihox-2* and *Amphihox-3* (8.5 kb), and in the mammalian *HoxB2* - *HoxB3* intergenic region (9.0 kb) (Garcia-Fernández and Holland 1994; Holland and Garcia-Fernández 1996; Zhang and Nei 1996). The orthologous region in the pufferfish, a tetraodontiform, is approximately 8.0 kb, which is interesting because the pufferfish genome is compressed relative to other vertebrate genomes, and is proposed as a model vertebrate genome by a few researchers (Aparicio *et al.* 1995; Gilley, *et al.* 1997). The conservation of the exon/intron composition and of the size of the intergenic region between *HoxB2* and *HoxB3* among vertebrates suggests that the organization, both spatial and compositional, of this region is vital for correct development of the embryo.

The anterior *Hox* genes, which pattern the hindbrain, are conserved both in the complement of *Hox* genes, as well as the spatio-temporal patterns of expression in the branchial region of vertebrates. DNA sequence comparisons between the striped bass *HoxB2* promoter region and the orthologous region of the human *HOXB2* gene reveal a conservation of basal promoter elements. The conservation of factors regulating the *Hox* genes of the branchial region is expected due to the conservation of the expression patterns among the chicken, mouse and human embryos (Hunt *et al.* 1991a; Hunt *et al.* 1991b; Marshall *et al.* 1996; Vieille-Grosjean *et al.* 1997). The basal promoter elements

that are conserved between humans and striped bass are the TATA box, an *AP2* binding site, and a CAAT box. The location of the CAAT box in striped bass and human differs by 3 bp with respect to the transcription start site, whereas the location of the other conserved basal promoter elements is identical in both species.

The striped bass and human promoter regions both contain a *Krox-20/Box1* motif that is known to direct the expression of *HoxB2* in rhombomeres 3 and 5 of the embryonic vertebrate hindbrain (Aparicio *et al.* 1995; Nonchev *et al.* 1996a; Nonchev *et al.* 1996b; Sham *et al.* 1993; Vesque *et al.* 1996). The human *Krox-20/Box1* complex is found much farther upstream of the *HoxB2* gene than that of the same complex in striped bass. A *Krox-20/Box1* complex has been identified in the upstream promoter region of the Japanese pufferfish, *Fugu rubripes* (Aparicio *et al.* 1995). The *Krox-20* binding site found in the pufferfish is identical to the striped bass *Krox-20* binding site, and the *Box1* motifs differ by one nucleotide (Vesque *et al.* 1996). The *Krox-20/Box1* complex is found in the same relative position in the striped bass *HoxB2* promoter region as it is in the pufferfish *HoxB2* promoter region. Experimental data show that the pufferfish *Krox-20/Box1* complex is capable of directing r3/r5 restricted expression of a reporter gene in a transgenic mouse (Vesque *et al.* 1996). The conservation of elements controlling the expression of *HoxB2* in the human, striped bass and pufferfish suggests that *HoxB2* in striped bass is expressed in r3 and r5 of the hindbrain during development. Thus, I predict that the *HoxB2* and *HoxB3* genes in striped bass will be expressed in the branchial rhombomeres 3 and 5 of the embryo between 19 - 24 hours after fertilization.

This is the period during the embryogenesis of the striped bass when the neural tube is being established, and the somites first start to appear.

The core elements promoting the transcription of *HoxB2* are conserved between the human and the striped bass promoter regions. However, analysis of the intergenic region also reveals a number of noteworthy differences between the two species. The human *HOXB2* gene promoter has a *GATA-1* site located 48 nucleotides upstream of the transcription start site that is absent in the striped bass *HoxB2* - *HoxB3* intergenic region. The *GATA-1* transcription factor is known to promote the expression of *HOXB2* in erythroid cell lines (Vieille-Grosjean and Huber 1995). The *NF-kappaB* recognition site found in the human promoter (-72 with respect to the transcriptional start site) but not in the striped bass *HoxB2* promoter is another transcription factor that may promote *HoxB2* expression in erythroid cell lines. *NF-kappaB* was identified as a factor promoting the expression of the *kappa-b* light chain in immune cell lines, but is now known to be a ubiquitous transcription factor (Sen and Baltimore 1986). The presence of a factor that directs *HoxB2* expression in human erythroid cell lines suggests that in addition to the "normal" function attributed to *HOXB2*, the human ortholog has been co-opted for directing a cell lineage-specific developmental program. This hypothesis is supported by the fact that haematopoiesis in striped bass is likely to be much different from that found in humans because striped bass have nucleated red blood cells, whereas humans do not. Another difference observed between the two *HoxB2* promoter regions in striped bass and human is the presence of a second *AP2* binding site in the striped bass promoter. This *AP2* site is located 280 bp upstream of the proposed transcription start site. A

corresponding *AP2* site is not present in the human sequence. It is likely that as new data becomes available, more homologies and differences will be found among the factors controlling the expression of these genes, and additional functions will be assigned to *HoxB2* beyond its role in patterning the hindbrain during embryogenesis.

A particularly unusual feature associated with the striped bass *HoxB2* - *HoxB3* intergenic region is the presence of a second promoter upstream of *HoxB2*. Only one promoter has been characterized for this gene in the human. The upstream promoter was identified based on the assemblage of transcription factor recognition sites 1356 bp upstream of the ATG in exon 1. The promoter has a TATA box, an *AP2*, and a *HoxA5* binding site located 30 bp, 70 bp and 221 bp, respectively, upstream of the proposed transcription start site. Multiple promoters are known to exist for several other *Hox* genes, including the mouse *HoxB3* and human, mouse and *Xenopus* *HoxC6* genes. The murine *HoxB3* gene is transcribed under the direction of at least two promoters, yielding transcripts of differing lengths but producing the same protein (Sham *et al.* 1992). The human, mouse and *Xenopus* *HoxC6* gene is transcribed from two separate promoters, resulting in two distinct transcripts and different proteins products (Cho *et al.* 1988; Coletta *et al.* 1991; Simeone *et al.* 1988). The second promoter of striped bass *HoxB2* may function to provide an alternate transcript of *HoxB2* as in the case of the *HoxC6* promoters. It is also possible that the upstream promoter functions as a long-range promoter to produce the same protein as the first promoter, as is the case of the murine *HoxB3* promoter (Coletta *et al.* 1991). Multiple promoters add flexibility to gene regulation and allow for the gene expression to be attenuated in a precise manner. It is

likely that the second promoter of *HoxB2* in striped bass functions during later stages of development and may be involved in the specification of discrete tissues, because the first promoter has a *Krox-20/Box1* complex that is located in a similar position in the pufferfish, and is known in the pufferfish to direct the early expression of *HoxB2* in r3 and r5 of the hindbrain.

A *HoxA5* recognition site was identified upstream of the promoter based on the consensus binding site published in the Transfac database. The *HoxA5* recognition site identified in the alternate promoter may be of functional significance. It appears that all central class *Hox* proteins recognize similar binding sites containing the ATTA core sequence (Boulikas 1992; Gehring and Qian 1994; Laughon 1991). Cross-talk between *Hox* genes is presumably an important factor regulating their expression. Products of neighboring *Hox* genes may act to repress expression of a gene, or to enhance the expression of a *Hox* gene, depending upon the spatio-temporal context of the cell at a particular moment in time. Targeted disruption of the *HoxB2* locus in the mouse results in an alteration in the expression of both *HoxB1* and *HoxB4* (Barrow and Capecchi 1996). These results suggest that the *HoxB2* protein is interacting with elements upstream of both *HoxB1* and *HoxB4*. Cross-talk between genes located on other chromosomes is also suspected to regulate the expression of the *Hox* genes.

Direct comparison of the pufferfish and zebrafish DNA sequence for the *HoxB2* - *HoxB3* intergenic region with that of the striped bass DNA sequence should reveal common elements that could then be tested for regulatory function, but such a comparison is not possible, as the pufferfish data is not yet available. The conservation

found among the *Hox* genes of the vertebrates suggests that a common method exists for building a vertebrate organism. The differences observed in the regulatory regions are likely to be one of the factors directly responsible for the evolutionary diversity observed among vertebrates. Moreover, these regulatory regions are also responsible for the maintenance of the *Hox* clusters as they may promote (or repress) the expression of both neighboring genes. In fact, Krumlauf hypothesizes that the *Hox* genes are maintained in clusters because the genes within a cluster share essential regulatory elements located within the non-coding regions (Krumlauf, personal communication). Non-coding regions within and surrounding the *Hox* genes are likely to reveal a wealth of regulatory elements that strictly govern the expression of *Hox* genes as more DNA sequence data becomes available. The evolution of the vertebrate hoxon will be further elucidated by examination of these regulatory regions and by direct comparison of the DNA sequence elements between homologous regulatory regions in the vertebrates.

Given the considerable homology of function of *HoxB2* in vertebrates, it is expected that the proteins would be quite similar. However, the proteins encoded by the human and striped bass *HoxB2* proteins are different biochemically. The striped bass *HoxB2* protein is much larger (46,241.37 g/mol) than the human ortholog (37,913.27 g/mol). The additional residues in the striped bass protein are predicted to allow the protein more flexibility than the human protein. A surface domain is predicted to exist on the striped bass *HoxB2* protein that does not exist on the human protein, and may serve as a binding sight for a cofactor. The isoelectric point of the proteins is different, with the human protein isoelectric point being 4.87 and the that of striped bass protein being 6.30.

The conservation of the N-terminal, hexapeptide, homeodomain, and the C-terminal domains suggests that these regions are fundamental to the action of *HoxB2*. The homeodomain is known to directly bind to DNA targets, while the hexapeptide is bound by members of the *Pbx* protein family. The association of a *Pbx* protein with a *Hox* protein stabilizes the protein-DNA complex (Knoepfler and Kamps 1995; Neuteboom *et al.* 1995; Peltenburg and Murre 1997; Phelan *et al.* 1995; Shen *et al.* 1996). The N-terminal domains of both the human and striped bass *HoxB2* proteins have hydrophilic residues, and the C-terminal domains are largely hydrophobic. The function of the N-terminal and C-terminal domains is unknown, but they are likely responsible for conferring the proper conformation of the protein during folding.

At this point it is impossible to discern whether the extra amino acids found in the striped bass *HoxB2* protein represent the ancestral state, or if they are new acquisitions. Comparison of the *HoxB2* amino acid sequences between striped bass, zebrafish, pufferfish and *Fundulus*, as well as the amphioxus *Amphiox-2* protein should reveal the ancestral state. Currently the only amino acid data available for the complete *HoxB2* protein is that of human *HOXB2*.

The conservation of the *Hox* genes in form and function appears to be central to their role in directing regional tissue specification during ontogeny. Differences in *Hox* gene expression are likely to be due to changes in regulatory elements found in the promoter and enhancer regions associated with these genes. The significant differences observed in the striped bass *HoxB2* promoter region when compared to the human *HOXB2* promoter suggest that the *HoxB2* gene in perciforms is regulated differently than

the orthologous genes in other vertebrates. Differences in the regulation of the expression of *Hox* genes during ontogeny may prove to be critical determinants in generating phylogenetic divergence.

References

- Alexandre, D., J.D. Clarke, E. Oxtoby, Y. L. Yan, T. Jowett, and N. Holder. 1996. Ectopic expression of *Hoxa-1* in the zebrafish alters the fate of the mandibular arch neural crest and phenocopies a retinoic acid-induced phenotype. *Development* 122 (3):735-746.
- Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic Local Alignment Search Tool. *Journal of Molecular Biology* 215:403-410.
- Aparicio, Sam, A. Morrison, A. Gould, J. Gilthorpe, C. Chaudhuri, Peter W. J. Rigby, Robb Krumlauf, and Sydney Brenner. 1995. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. *Proceedings of the National Academy of Science of the United States of America* 92 (5):1684-1688.
- Aparicio, Sam, Kelvin Hawker, and Amanda Cottage. 1997. Organization of the *Fugu rubripes Hox* clusters: evidence for continuing evolution of vertebrate *Hox* complexes. *Nature Genetics* 16:79-83.
- AutoAssembler 1.4.0. 1995. Perkin-Elmer Corporation, Foster City, CA.
- Barrow, Jeffery R., and Mario R. Capecchi. 1996. Targeted disruption of the *Hoxb-2* locus in mice interferes with the expression of *Hoxb-1* and *Hoxb-4*. *Development* 121:3817-3828.

- Bender, Welcome, Michael Akam, François Karch, Philip A. Beachy, Mark Peifer, Pierre Spierer, E. B. Lewis, and David S. Hogness. 1983a. Molecular genetics of the Bithorax complex in *Drosophila melanogaster*. *Science* 221:23-29.
- Bender, Welcome, Pierre Spierer, and David S. Hogness. 1983b. Chromosomal walking and jumping to isolate DNA from Ace and rosy loci and the *Bithorax* complex in *Drosophila melanogaster*. *Journal of Molecular Biology* 168:17-33.
- Benton, W.D., and R.W. Davis. 1977. Screening λ gt recombinant clones by hybridization to single plaques *in situ*. *Science* 196:180.
- Benson, D., M. Boguski, D.J. Lipman, and J. Ostell. 1996. GenBank. *Nucleic Acids Research* 24:1-5.
- Bogue, Clifford W., Lily J. Lou, Hema Vasavada, Christine M. Wilson, and Harris C. Jacobs. 1996. Expression of *Hoxb* genes in the developing mouse foregut and lung. *American Journal of Respiratory Cell and Molecular Biology* 15 (2):163-171.
- Boncinelli, Edoardo, Renato Somma, Dario Acampora, Maria Pannese, Maurizio D'Esposito, Antonio Faiella, and Antonio Simeone. 1988. Organization of human homeobox genes. *Human Reproduction* 3 (7):880-886.
- Boulikas, Teni. 1992. Homeotic protein binding sites, origins of replication and nuclear matrix anchorage sites share the ATTA and ATTTA motifs. *Journal of Cellular Biochemistry* 50:111-123.

- Bürglin, Thomas R., G. Ruvkun, A. Coulson, Nancy C. Hawkins, James D. McGhee, D. Schaller, C. Wittmann, F. Muller, and R. H. Waterston. 1991. Nematode homeobox cluster. *Nature* 351:703.
- Carrasco, Andrés E., William McGinnis, Walter J. Gehring, and Eddy M. De Robertis. 1984. Cloning of an *X. Laevis* gene expressed during early embryogenesis coding for a peptide region homologous to *Drosophila* homeotic genes. *Cell* 37:409-414.
- Cartwright, Paulyn, Matthew Dick, and Leo W. Buss. 1993. *HOM/Hox* type homeoboxes in the chelicerate *Limulus polyphemus*. *Molecular Phylogenetics and Evolution* 2 (3):185-192.
- Chan, S.K., and R.S. Mann. 1996. A structural model for a homeotic protein-extradenticle-DNA complex accounts for the choice of *HOX* protein in the heterodimer. *Proceedings of the National Academy of Science of the United States of America* 93 (11):5223-5228.
- Cho, Ken W. Y., J. Goetz, Christopher V. E. Wright, J. Hardwicke, Andreas F. Fritz, and Eddy M. De Robertis. 1988. Differential utilization of the same reading frame in a *Xenopus* homeobox gene encodes two related proteins sharing the same DNA-binding specificity. *The EMBO Journal* 7:2139-2149.
- Coletta, P.L., S.M. Shimfield, C. Chaudhuri, U. Muller, J.P. Clarke, and P.T. Sharpe. 1991. Characterization of the murine *Hox-3.3* gene and its promoter. *Mechanisms of Development* 35:129-142.

- Condie, Brian G., and Mario R. Capecchi. 1994. Mice with targeted disruptions in the paralogous genes *hoxa-3* and *hoxd-3* reveal synergistic interactions. *Nature* 370:304-307.
- Corsetti, Maria Teresa, Giovanni Levi, Flavia Lancia, Lorenza Sanseverino, Silvano Ferrini, Edoardo Boncinelli, and Giorgio Corte. 1995. Nucleolar localization of three *Hox* homeoproteins. *Journal of Cell Science* 108:187-193.
- De Robertis, Eddy M. 1994. The homeobox in cell differentiation and evolution. In *Guidebook to the Homeobox Genes*, edited by D. Duboule. New York: Oxford University Press.
- Degnan, Bernard M., Sandie M. Degnan, Andrew Giusti, and Daniel E. Morse. 1995. A *hox/hom* gene in sponges. *Gene* 155 (2):175-177.
- Dekker, Erik-Jan, Maria Pannese, Erwin Houtzager, Ans Timmermans, Edoardo Boncinelli, and Antony Durston. 1992. *Xenopus Hox-2* genes are expressed after the onset of gastrulation and are differentially inducible by retinoic acid. *Development Supplement*:195-202.
- Di Gregorio, Anna, A. Spagnuolo, F. Ristoratore, M. Pischetola, F. Aniello, M. Branno, L. Cariello, and R. Di Lauro. 1995. Cloning of ascidian homeobox genes provides evidence for a primordial chordate cluster. *Gene* 156 (2):253-257.
- Duboule, Denis, and Pascal Dolle. 1989. The structural and functional organization of the murine *HOX* gene family resembles that of *Drosophila* homeotic genes. :1497-1505.

- Duboule, Denis, and Gines Morata. 1994. Colinearity and functional hierarchy among genes of the homeotic complexes. *Trends in Genetics* 10 (10):358-364.
- Dupe, Valerie, Marc Davenne, Jacques Brocard, Pascal Dolle, Manuel Mark, Andree Dierich, Pierre Chambon, and Fillippo M. Rijli. 1997. In vivo functional analysis of the *Hoxa-1* 3' retinoic acid response element (3'RARE). *Development* 124:300-410.
- Factura 1.2.0r6. 1995. Perkin-Elmer Corporation, Foster City, CA 94404.
- Flagiello, Domenico, Marie-France Poupon, Clemente Cillo, Bernard Dutrillaux, and Bernard Malfoy. 1996. Relationship between DNA methylation and gene expression of the *HOXB* gene cluster in small cell lung cancers. *FEBS Letters* 380 (1-2):103-107.
- Frasch, Manfred, Xiaowei Chen, and Thomas Lufkin. 1995. Evolutionary-conserved enhancers direct region-specific expression of the murine *Hoxa-1* and *Hoxa-2* loci in both mice and *Drosophila*. *Development* 121:957-974.
- Garcia-Fernàndez, Jordi, and Peter W. H. Holland. 1994. Archetypal organization of the amphioxus *Hox* gene cluster. *Nature* 370:563-566.
- Gaunt, Stephen J., Paul T. Sharpe, and Denis Duboule. 1988. Spatially restricted domains of homeo-gene transcripts in mouse embryos: relation to a segmented body plan. *Development* 104 (Supplement):169-179.
- Gehring, Walter J., and Yan Qiu Qian. 1994. Homeodomain-DNA recognition. *Cell* 78:211-223.

- Gehring, Walter J. 1994. A history of the homeobox. In *Guidebook to the Homeobox Genes*, edited by D. Duboule. New York: Oxford University Press.
- Gilley, Jonathan, Niall Armes, and Mike Fried. 1997. *Fugu* genome is not a good mammalian model. *Nature* 385:305-306.
- Graham, Anthony, Nancy Papalopulu, and Robb Krumlauf. 1989. The murine and *Drosophila* homeobox gene complexes have common features of organization and expression. *Cell*:367-378.
- Hanahan, Douglas. 1983. Studies on transformation of *Escherichia coli* with plasmids. *Journal of Molecular Biology* 166:557-580.
- Holland, Linda Z., and Nicholas D. Holland. 1996. Expression of *AmphiHox-1* and *AmphiPax-1* in amphioxus embryos treated with retinoic acid: insights into evolution and patterning of the chordate nerve cord and pharynx. *Development* 122 (6):1829-1838.
- Holland, Peter W. H., and Jordi Garcia-Fernàndez. 1994. Gene duplications and the origins of vertebrate development. *Development Supplement*: 125-133.
- Holland, Peter W. H., and Jordi Garcia-Fernàndez. 1996. *Hox* genes and chordate evolution. *Developmental Biology* 173 (2):382-395.
- Hunt, Paul, Jenny Whiting, Ian Muchamore, Heather Marshall, and Robb Krumlauf. 1991a. Homeobox genes and models for patterning the hindbrain and branchial arches. *Development Supplement* 1:187-196.
- Hunt, Paul, Jenny Whiting, Stefan Nonchev, Mai-Har Sham, Heather Marshall, Anthony Graham, Martyn Cook, Rudolf Allemann, Peter W. J. Rigby, Massimo Gulisano,

- Antonio Faiella, Edoardo Boncinelli, and Robb Krumlauf. 1991b. The branchial *Hox* code and its implications for gene regulation, patterning of the nervous system and head evolution. *Development Supplement* 2:63-77.
- Johnson, D.R., and P. O'Higgins. 1996. Is there a link between changes in the vertebral "hox code" and the shape of vertebrae? A quantitative study of shape change in the cervical-vertebral column of mice. *Journal of Theoretical Biology* 183 (1):89-93.
- Knoepfler, Paul S., and Mark P. Kamps. 1995. The pentapeptide motif of *Hox* proteins is required for cooperative DNA binding with *Pbx1*, physically contacts *Pbx1*, and enhances DNA binding by *Pbx1*. *Molecular and Cellular Biology* 15 (10):5811-5819.
- Kuhn, Kerstin, Bruno Streit, and Bernd Schierwater. 1996. Homeobox genes in the Cnidarian *Eleutheria dichotoma*: Evolutionary implications for the origin of *Antennapedia*-class (*HOM/Hox*) genes. *Molecular Phylogenetics and Evolution* 6 (1):30-38.
- Kuroiwa, Atsushi, Ernst Hafen, and Walter J. Gehring. 1984. Cloning and transcriptional analysis of the segmentation gene *fushi tarazu* of *Drosophila*. *Cell* 37:825-831.
- Laughon, Allen, and Matthew P. Scott. 1984. Sequence of a *Drosophila* segmentation gene: protein structure homology with DNA-binding proteins. *Nature* 310:25-31.
- Laughon, Allen. 1991. DNA binding specificity of homeodomains. *Biochemistry* 30 (48):11357-11366.

- Lawrence, Peter A. 1992. *The Making of a Fly; The Genetics of Animal Design*. Boston: Blackwell Scientific Publications.
- Lewis, E.B. 1978. A gene complex controlling segmentation in *Drosophila*. *Nature* 276:567-570.
- Luria, S.E., and J.W. Burrous. 1957. *Journal of Bacteriology* 74:471.
- Luria, S.E., J.N. Adams, and R.C. Ting. 1960. *Virology* 12:348.
- Manzanares, Miguel, Sabine Cordes, Chung-Tin Kwan, Mai-Har Sham, Greg S. Barsh, and Robb Krumlauf. 1997. Segmental regulation of *Hoxb-3* by *kreisler*. *Nature* 387:191-195.
- Marshall, Heather, Michèle Studer, Helke Pöpperl, Sam Aparicio, Atsushi Kuroiwa, Sydney Brenner, and Robb Krumlauf. 1994. A conserved retinoic acid response element required for early expression of the homeobox gene *Hoxb-1*. *Nature* 370:567-571.
- Marshall, Heather, A. Morrison, Michèle Studer, Helke Pöpperl, and Robb Krumlauf. 1996. Retinoids and *Hox* genes. *The FASEB Journal* 10 (9):969-978.
- McGinnis, William, Richard L. Garber, Johannes Wirz, Atsushi Kuroiwa, and Walter J. Gehring. 1984a. A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. *Cell* 37:403-408.
- McGinnis, William, M.S. Levine, Ernst Hafen, Atsushi Kuroiwa, and Walter J. Gehring. 1984b. A conserved DNA sequence found in homeotic genes of the *Drosophila Antennapedia* and *Bithorax* complexes. *Nature* 308:428-433.

Neuteboom, Saskia T., Lucy T.C. Peltenburg, Marc A. van Dijk, and Cornelis Murre.

1995. The hexapeptide LFPWMR in *Hoxb-8* is required for cooperative DNA binding with *Pbx1* and *Pbx-2* proteins. *Proceedings of the National Academy of Science of the United States of America* 92 (20):9166-9170.

Nonchev, Stefan, Mark Maconochie, Christine Vesque, Sam Aparicio, Linda

Ariza-McNaughton, Miguel Manzanares, Kamala Maruthainar, Atsushi Kuroiwa, Sydney Brenner, Patrick Charnay, and Robb Krumlauf. 1996a. The conserved role of *Krox-20* in directing *Hox* gene expression during vertebrate hindbrain segmentation. *Proceedings of the National Academy of Science of the United States of America* 93:9339-9345.

Nonchev, Stefan, Christine Vesque, Mark Maconochie, Tania Seitanidou, Linda

Ariza-McNaughton, Monique Frain, Heather Marshall, Mai-Har Sham, Robb Krumlauf, and Patrick Charnay. 1996b. Segmental expression of *Hoxa-2* in the hindbrain is directly regulated by *Krox-20*. *Development* 122:543-554.

Ogura, Toshihiko, and Ronald M. Evans. 1995a. Evidence for two distinct retinoic acid response pathways for *HOXB1* gene regulation. *Proceedings of the National Academy of Science of the United States of America* 92 (2):392-396.

Ogura, Toshihiko, and Ronald M. Evans. 1995b. A retinoic acid-triggered cascade of *HOXB1* gene activation. *Proceedings of the National Academy of Science of the United States of America* 92 (2):387-391.

- Pavell, Anthony M. 1994. Survey of *Hox*-like genes in a neoteleost (striped bass): implications for evolutionary expansion of the *Hox* gene family in vertebrates, Master's Thesis, Department of Biology, East Carolina University, Greenville.
- Peltenburg, Lucy T. C. , and Cornelis Murre. 1997. Specific residues in the *Pbx* homeodomain differentially modulate the DNA-binding activity of *Hox* and *Engrailed* proteins. *Development* 124:1089-1098.
- Pendleton, John W., Brian K. Nagai, Michael T. Murtha, and Frank H. Ruddle. 1993. Expansion of the *Hox* gene family and the evolution of chordates. *Proceedings of the National Academy of Science of the United States of America* 90:6300-6304.
- Pfeffer, Peter L., and Claus von Holt. 1991. Stage- and adult tissue-specific expression of a homeobox gene in embryo and adult *Parechinus angulosus* sea urchin. *Gene* 108:219-226.
- Phelan, Michael L., Isabel Rambaldi, and Mark S Featherstone. 1995. Cooperative interactions between *HOX* and *PBX* proteins mediated by a conserved peptide motif. *Molecular and Cellular Biology* 15 (8):3989-3997.
- Pollock, Robert A., Taduru Sreenath, Lien Ngo, and Charles J. Bieberich. 1995. Gain of function mutations for paralogous *Hox* genes: implications for the evolution of *Hox* gene function. *Proceedings of the National Academy of Science of the United States of America* 92 (10):4492-4496.
- Pöpperl, Helke, M. Bienz, Michèle Studer, S.K. Chan, Sam Aparicio, Sydney Brenner, R.S. Mann, and Robb Krumlauf. 1995. Segmental expression of *Hoxb-1* is

controlled by a highly conserved autoregulatory loop dependent upon *exd/pbx*.

Cell 81 (7):1031-1042.

Prestridge, D.S. 1991. SIGNAL SCAN: A computer program that scans DNA sequences for eukaryotic transcriptional elements. *CABIOS* 7:203-206.

Prestige, Dan S. 1995. Prediction of Pol II promoter sequences using transcription factor binding sites. *Journal of Molecular Biology* 249:923-932.

Prince, Victoria, and Andrew Lumsden. 1994. *Hoxa-2* expression in normal and transposed rhombomeres: independent regulation in the neural tube and neural crest. *Development* 120:911-923.

Reese, M.G., and F.H. Eeckman. 1995. Novel Neural Network Algorithms for Improved Eukaryotic Promoter Site Recognition. Paper read at The Seventh International Genome Sequencing and Analysis Conference, at Hilton Head Island, South Carolina.

Reese, M.G., and F.H. Eeckman. 1996. Splice Sites: A detailed neural network study. Paper read at 1996 Genome Mapping & Sequencing Meeting, at Cold Spring Harbor Laboratory, N.Y.

Reese, M.G., N.L. Harris, and F.H. Eeckman. 1996. Large Scale Sequencing Specific Neural Networks for Promoter and Splice Site Recognition. Paper read at the Pacific Symposium on Biocomputing, January 2-7, 1996, at Kona, Hawaii.

Ruddle, Frank H., Kevin L. Bentley, Michael T. Murtha, and Neil Risch. 1994. Gene loss and gain in the evolution of the vertebrates. *Development Supplement*: 155-161.

- Schneider-Maunoury, Sylvie, Tania Seitanidou, Patrick Charnay, and Andrew Lumsden. 1997. Segmental and neuronal architecture of the hindbrain of *Krox-20* mouse mutants. *Development* 124:1215-1226.
- Schug, J., and G.C. Overton. 1997. TESS: Transcription Element Search Software on the WWW. In *Technical Report CBIL-TR-1997-1001-v0.0, of the Computational Biology and Informatics Laboratory*, : School of Medicine, University of Pennsylvania.
- Schughart, Klaus, Claudia Kappen, and Frank H. Ruddle. 1989. Duplication of large genomic regions during the evolution of vertebrate homeobox genes. *Proceedings of the National Academy of Science of the United States of America* 86:7067-7071.
- Schummer, Michel, Imke Scheurlen, Chica Schaller, and Brigitte Galliot. 1992. *HOM/HOX* homeobox genes are present in hydra (*Chlorohydra viridissima*) and are differentially expressed during regeneration. *The EMBO Journal* 11:1815-1823.
- Scott, Matthew, and Amy J. Weiner. 1984. Structural relationships among genes that control development: sequence homology between the *Antennapedia*, *Ultrabithorax* and *fushi tarazu* loci of *Drosophila*. *Proceedings of the National Academy of Science of the United States of America* 81:4115-4119.
- Scott, Matthew P., Amy J. Weiner, Tulle I. Hazelrigg, Barry A. Polisky, Vincenzo Pirrotta, Franco Scalenghe, and Thomas C. Kaufman. 1983. The molecular organization of the *Antennapedia* locus of *Drosophila*. *Cell* 35:763-776.

- Scott, Matthew P. 1992. Vertebrate Homeobox Gene Nomenclature. *Cell* 71:551-553.
- Seimiya, Makiko, Hiroshi Ishiguro, Keiji Miura, Yoko Watanabe, and Yoshikazu Kurosawa. 1994. Homeobox-containing genes in the most primitive metazoa, the sponges. *European Journal of Biochemistry* 221:219-225.
- Sen, R., and D. Baltimore. 1986. Multiple nuclear factors interact with the immunoglobulin enhancer sequences. *Cell* 46:705-716.
- Sengupta, Pritam K., Donald E. Lavelle, and Joseph DeSimone. 1994. The 87-kD A γ -globin enhancer-binding protein is a product of the *HOXB2* (*HOX2H*) locus. *Blood* 83:1420-1427.
- Sequencing Analysis 2.0.1. 1994. Perkin-Elmer Corporation, Foster City, CA.
- Sham, M., P. Hunt, S. Nonchev, N. Papalopulu, E. Boncinelli, A. Graham, and R. Krumlauf. 1992. Analysis of the murine *HoxB3* gene: conserved alternative transcripts with differential distributions in the nervous system and the potential for shared regulatory regions. *The EMBO Journal* 11:1825-1836.
- Sham, Mai-Har, Christine Vesque, Stefan Nonchev, Heather Marshall, Monique Frain, Romita Das Gupta, Jenny Whiting, David Wilkinson, Patrick Charnay, and Robb Krumlauf. 1993. The zinc finger gene *Krox-20* regulates *HoxB2* (*Hox2.8*) during hindbrain segmentation. *Cell* 72:183-196.
- Shen, Wei-Fang, Ching-Pin Chang, Sofia Rozenfeld, Guy Sauvageau, R. Keith Humphries, Ming Lu, H. Jeffrey Lawrence, Michael L. Cleary, and Corey

- Largman. 1996. *Hox* homeodomain proteins exhibit selective complex stabilities with *Pbx* and DNA. *Nucleic Acids Research* 24 (5):898-906.
- Shepherd, John C.W., William McGinnis, Andrés E. Carrasco, Eddy M. De Robertis, and Walter J. Gehring. 1984. Fly and frog homoeo domains show homologies with yeast mating type regulatory proteins. *Nature* 310:70-71.
- Simeone, Antonio, Dario Acampora, Edoardo Boncinelli, Maurizio D'Esposito, and Maria Pannese. 1988. At least three human homeoboxes on chromosome 12 belong to the same transcription unit. *Nucleic Acids Research* 16:5379-5387.
- Snow, Peter, and Leo W. Buss. 1994. *HOM/Hox*-type homeoboxes from *Stylaria lacustris* (Annelida: Oligochaeta). *Molecular Phylogenetics and Evolution* 3 (4):360-364.
- Stein, Stefan, Rüdiger Fritsch, Lydia Lemaire, and Michael Kessel. 1996. Checklist: Vertebrate homeobox genes. *Mechanisms of Development* 55:91-108.
- Tartof, K.D., and C.A. Hobbs. 1987. Improved media for growing plasmid and cosmid clones. *Bethesda Research Laboratory Focus* 9:12.
- van der Hoeven, F., József Zákány, and Denis Duboule. 1996. Gene transpositions in the *HoxD* complex reveal a hierarchy of regulatory controls. *Cell* 85 (7):1025-1035.
- Vesque, Christine, Mark Maconochie, Stefan Nonchev, Linda Ariza-McNaughton, Atsushi Kuroiwa, Patrick Charnay, and Robb Krumlauf. 1996. *Hoxb-2* transcriptional activation in rhombomeres 3 and 5 requires an evolutionarily conserved cis-acting element in addition to the *Krox-20* binding site. *The EMBO Journal* 15 (19):5383-5396.

- Vieille-Grosjean, Isabelle, and Philippe Huber. 1995. Transcription factor *GATA-1* regulates human *HOXB2* gene expression in erythroid cells. *Journal of Biological Chemistry* 270 (9):4544-4550.
- Vieille-Grosjean, Isabelle, Paul Hunt, Massimo Gulisano, Edoardo Boncinelli, and Peter Thorogood. 1997. Branchial *HOX* gene expression and human craniofacial development. *Developmental Biology* 183:49-60.
- Wisconsin Package Version 9.0 1996. Genetics Computer Group. Madison, Wisc.
- Yamamoto, K. R., B.M. Alberts, R. Benzinger, L. Lawhorne, and G. Treiber. 1970. Rapid bacteriophage sedimentation in the presence of polyethylene glycol and its application to large-scale virus purification. *Virology* 40 (3):734-744.
- Zhang, J., and M. Nei. 1996. Evolution of *Antennapedia*-class homeobox genes. *Genetics* 142 (1):295-303.
- Zelster, Lori, Claude Desplan, and Nathaniel Heintz. 1996. *Hoxb-13*: a new Hox gene in a distant region of the *HOXB* cluster maintains colinearity. *Development* 122 (8):2475-2484.

Figure 1. Schematic diagram depicting the relationships between the amphioxus hoxon, mammalian hoxon and pufferfish hoxon. The diagram does not represent the actual distances between the *Hox* genes (boxes). Note that paralog group 9 is the only group to have a representative in all of the clusters in each hoxon.

Figure 1. Diagram of hoxons from the cephalochordate amphioxus and selected vertebrates.

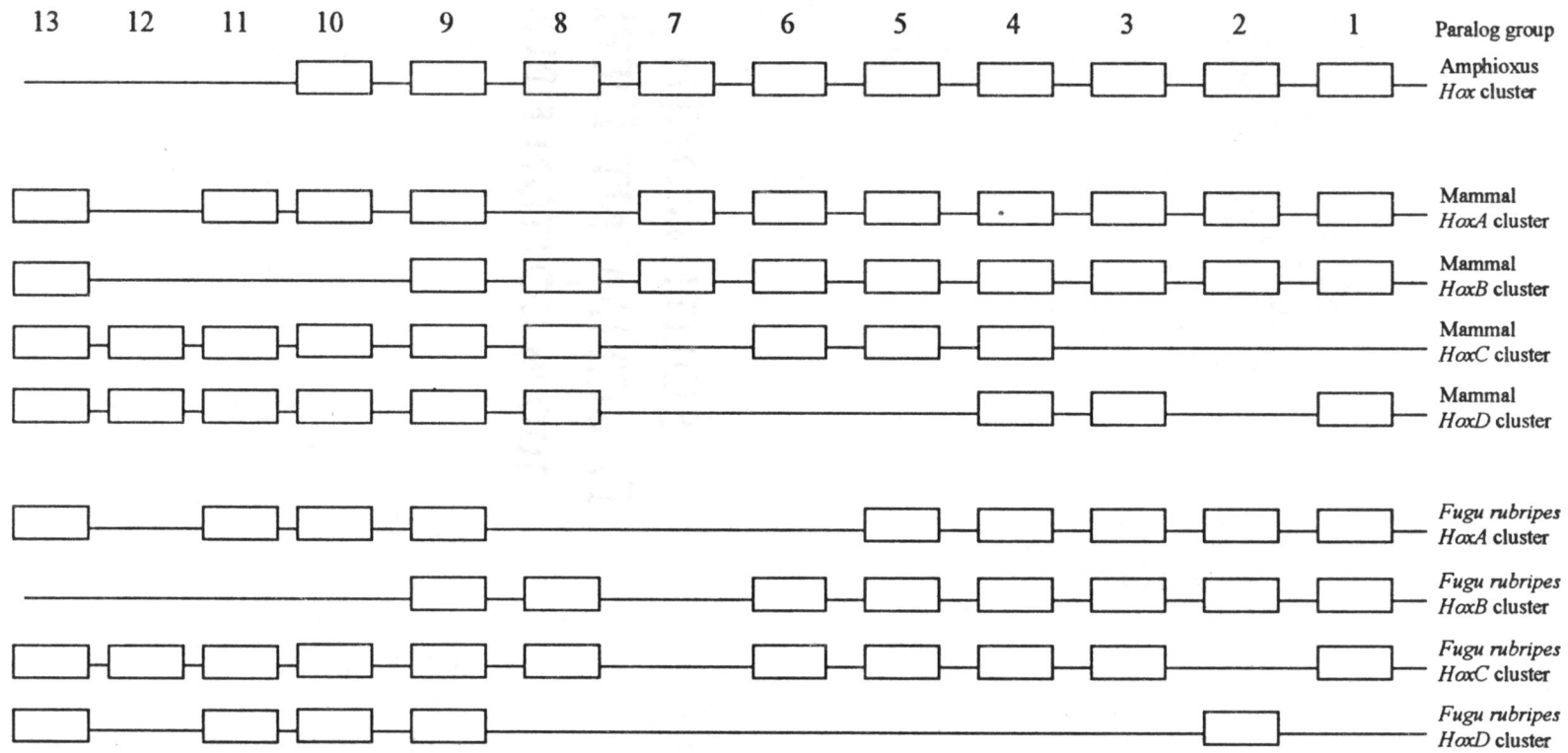


Figure 2. A schematic diagram depicting the regulation of the murine *HoxB1* and *HoxB2* genes. The 3' RARE (retinoic acid response element) directs the expression of *HoxB1* in rhombomeres 2 and 4. The autoregulatory loop of *HoxB1* is enhanced by the presence of a *Pbx* protein that binds to the *HoxB1* protein and increases the stability of the protein/DNA complex. The homeobox regions are designated as black boxes, the shaded boxes represent DNA sequence elements which are directly bound by regulatory proteins. The Box1 motif and *Krox-20* binding site direct the expression of *HoxB2* in rhombomeres 3 and 5. RA - retinoic acid; RAP - retinoid activating protein; RAR:RXR - retinoic acid receptor/retinoid X receptor heterodimer; RIP - retinoid inducible protein. (Ogura and Evans 1995b; Marshall *et al.* 1994; Nochev *et al.* 1996a; Nonchev *et al.* 1996b; Vesque *et al.* 1996)

Figure 2. Schematic diagram depicting the regulation of the murine *HoxB1* and *HoxB2* genes.

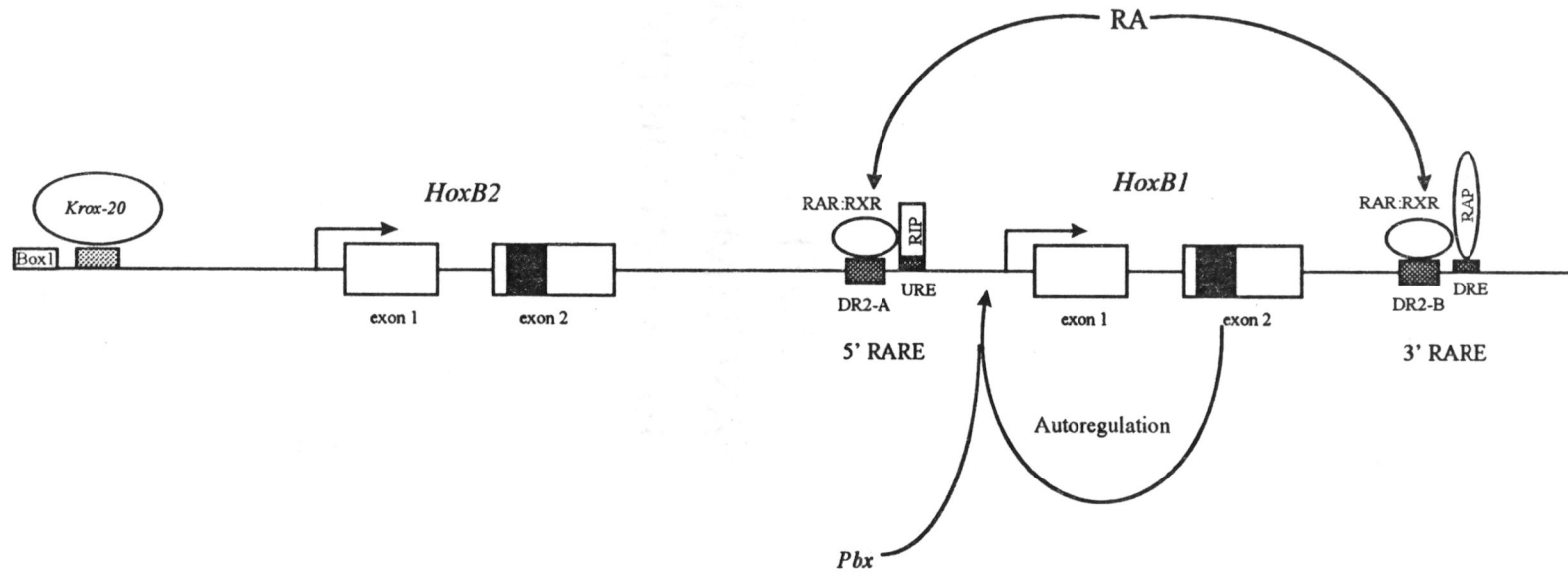


Figure 3. Restriction endonuclease cleavage map of λ clone 2H. The scale bar represents 1,000 base pairs.

Figure 3. Restriction Endonuclease Cleavage Map of the *Morone saxatilis*-derived insert of λ clone 2H

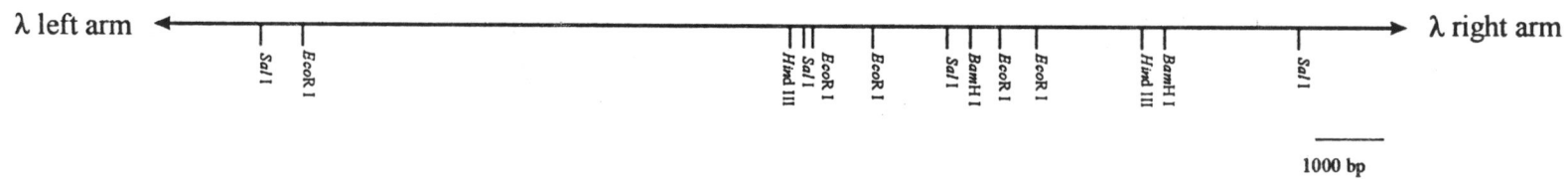


Figure 4. A schematic diagram of the strategy used to subclone λ clone 2H. Regions containing homeoboxes are depicted with shaded boxes.

Figure 4. Schematic diagram depicting the overall strategy for subcloning λ clone 2H.

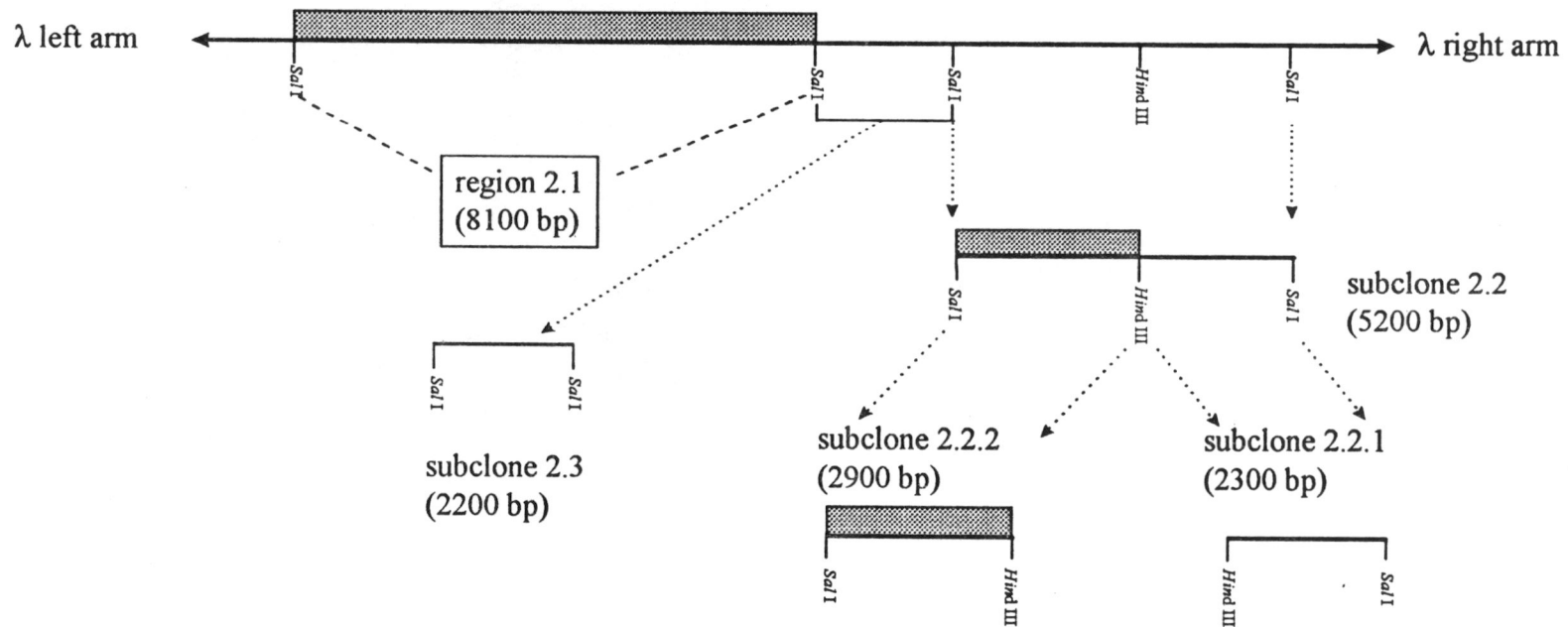


Figure 5. Schematic diagram of PCR reactions used to analyze the location and orientation of the individual subclones (2.2, 2.2.1, 2.2.2 and 2.3) generated from λ clone 2H. Predicted PCR products are represented by diamonds connected by solid lines. The diamonds indicate the location of the primer annealing sites. Numbers in circles correspond to the reaction number in the inset box, and designate the predicted PCR product.

Figure 5. Schematic diagram of PCR reactions used to analyze the location and orientation of the individual subclones (2.2, 2.2.1, 2.2.2 and 2.3) generated from λ clone 2H.

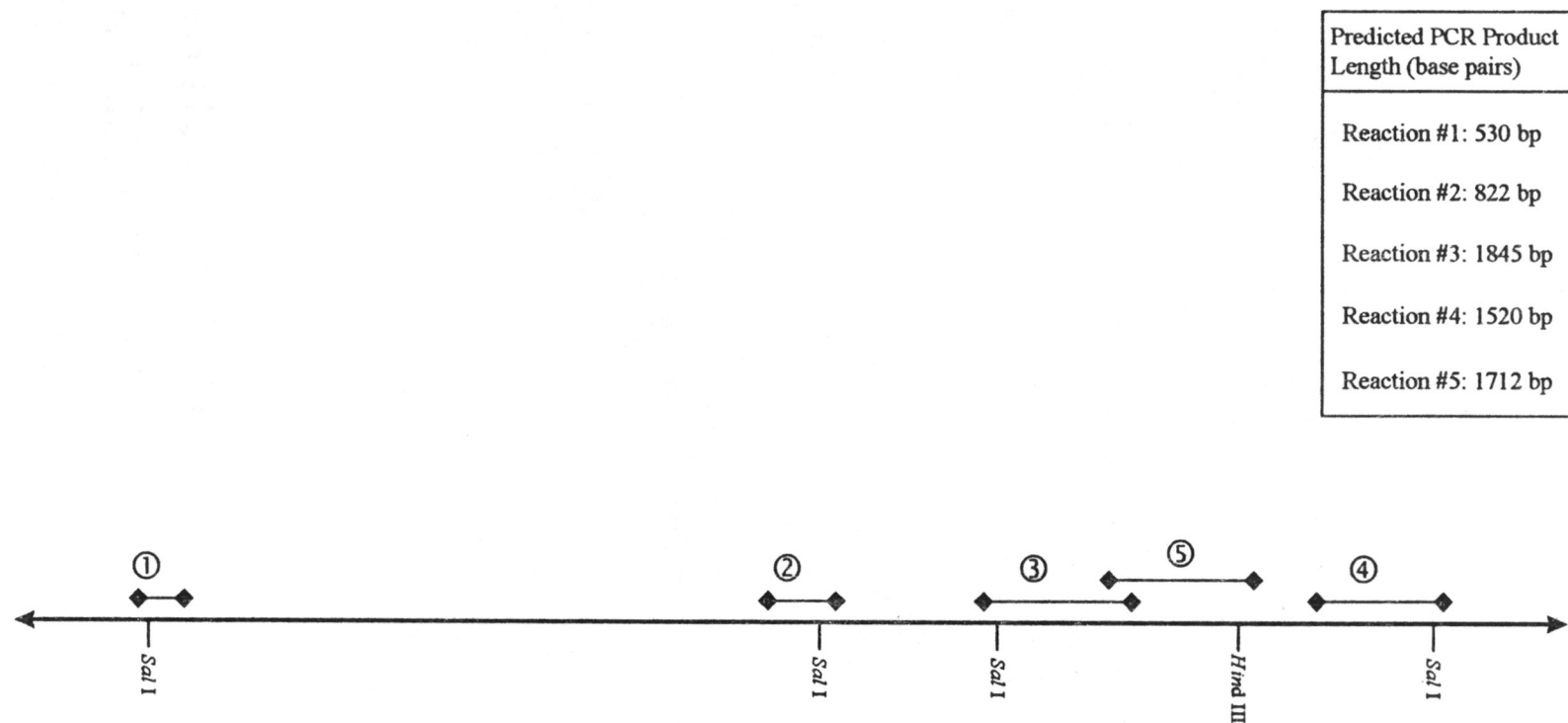


Figure 6. Multiple DNA sequence alignment of vertebrate *Hox* paralogous group 2 homeoboxes. The consensus is calculated from all input sequences, with differences from the consensus designated by a lower case character. In cases where the consensus could not be determined, a “-” is used. The shaded nucleotides are the signature bases used to assign the *M. saxatilis* homeobox as *HoxB2*. GenBank Accession numbers: Amphioxus (*Branchiostoma floridae*) *Hox2*, Z35143; Chicken (*Gallus gallus*) *HoxA2*, X74323; Mouse (*Mus musculus*) *HoxA2*, M93148; Rat (*Rattus norvegicus*) *HoxA2*, M91802; Salmon (*Salmo salar*) *HoxB2*, M18905; Human *HOXB2*, X16665; Mouse (*Mus musculus*) *HoxB2*, M34004; Newt (*Notophthalmus viridescens*) *HoxB2*, M84002; *Morone saxatilis* (Msax) *HoxB2* data from λ clone 2H bases 12282-12,461.

Figure 6. Nucleotide sequence alignment of paralogue group 2 homeoboxes

	1								80
Consensus	TCCCGGCGGC	TGCG-AC-GC	CTACACCAAC	ACGCAGCT-C	TGGAGCTGGA	GAAGGAGTTC	CACTTCAACA	AGTACCTCTG	
amphioxusHox2	TCCCGtCGat	TGCGtACaGt	CTtCACaAAC	ACtCAaCTtC	TGGAGtTGGa	GAAaGAGTTC	CACTaCAACA	AaTACgTgTG	
mouse HoxA2	TCCaGGCGtC	TGaGaACcGC	gTACACCAAC	ACtCAGCTtC	TGGAGCTGGA	aAAGGAaTTt	CAtTTCAACA	AGTACCTtTG	
rat HoxA2	TCCaGGCGtC	TGaGaACtGC	tTACACCAAC	ACtCAGCTtC	TGGAaCTGGA	aAAGGAaTTt	CAtTTCAACA	AGTACCTCTG	
chick HoxA2	TCtaGGCGGC	TGaGgACgGC	CTACACCAAC	ACGCAGCTtC	TGGAGCTGGA	GAAaGAaTTC	CACTTCAACA	AGTACCTCTG	
Msax HoxB2	TCCCGcCGGC	TGCGcACcGC	aTACACCAAC	ACaCAGCTgC	TGGAGCTGGA	GAAGGAGTTC	CACTTCAACA	AGTACCTaTG	
human HOXB2	gCgCGcaGGC	TGCGcACgGC	tTACACCAAC	ACGCAGCTgC	TGGAaCTGGA	GAAGGAaTTC	CACTTtAaTa	AGTACCTgTG	
mouse HoxB2	TCCCGcaGaC	TGCGcACgGC	CTACACCAAC	ACGCAaCTgC	TGGAGCTGGA	GAAGGAGTTC	CACTTCAaTa	AGTACCTgTG	
newt HoxB2	TCCCGGCGcC	TGCGgACcGC	CTACACCAAC	ACGCAGCTcC	TGGAGCTGGA	GAAGGAGTTC	CACTTCAACA	AGTACCTCTG	
salmon HoxB2	cCCgGaaGGC	TGaGgACtGC	CTACACCAAC	ACGCAGCTaC	TGGAGCTGGA	GAAGGAGTTC	CACTTCAACA	AaTAtCTCTG	
Consensus	TCCCGGCGGC	TGCG-AC-GC	CTACACCAAC	ACGCAGCT-C	TGGAGCTGGA	GAAGGAGTTC	CACTTCAACA	AGTACCTCTG	
	81								160
Consensus	CCGGCCCCGG	CGGGTGGAGA	TCGCCGCCCT	GCTGGATCTG	ACCGAGAGGC	AGGTGAAAGT	GTGGTTCCAG	AACCGGCG-A	
amphioxusHox2	CaaCCtCGt	CGGaaGGAGA	TCGCgtCtta	ctTGGActTG	AaCGAGAGaC	AGGTGAAAaT	tTGGTTCCAG	AACCGaCGcA	
mouse HoxA2	CaGaCCCCGc	aGGGTGGAAa	TCGCCGcGCT	GCTGGAtTtTG	ACCGAGAGaC	AaGTGAAAGT	GTGGTTtCAG	AACCGGaGaA	
rat HoxA2	CaGaCCCCGG	aGGGTGGAAa	TCGCCGcGCT	GCTGGAtTtTG	ACCGAGAGaC	AaGTGAAAGT	GTGGTTtCAG	AACCGGaGgA	
chick HoxA2	CaGaCCcGcG	CGGGTcGAGA	TCGCaGcGCT	GCTGGATCTG	ACCGAGAGGC	AaGTGAAgGT	cTGGTTCCAG	AACCGcaGgA	
Msax HoxB2	CCGGCCCaGg	CGGGTGGAGA	TCGCCGCCCT	cCTGGATCTc	ACCGAGcGGC	AGGTcAAAGT	GTGGTTCCAG	AACCGGCGgA	
human HOXB2	CCGGCCCaGc	CGcGTcGAGA	TCGCgGCtT	GCTGGAcCTc	ACCGAaAGGC	AGGTcAAAGT	cTGGTTtCAG	AACCGGCcGA	
mouse HoxB2	CCGGCCcGtG	CGcGTcGAGA	TCGCtGCctT	GCTGGAcCTc	ACCGAaAGGC	AGGTcAAAGT	cTGGTTCCAG	AACCGaCGcA	
newt HoxB2	CCGGCCgaGG	aGGGTGGAGA	TCGCCGCCCT	GCTGGAcCTG	ACCGAGcGGC	AGGTGAAgGT	tTGGTTCCAG	AACCGcCGcA	
salmon HoxB2	tCCGGCCCCG	CGaGTGGAGA	TaGctGCCCT	tCTGGATCTa	ACCGAGAGaC	AGGTcAAAcT	GTGGTTCCAG	AACCGGaGgA	
Consensus	CCGGCCCCGG	CGGGTGGAGA	TCGCCGCCCT	GCTGGATCTG	ACCGAGAGGC	AGGTGAAAGT	GTGGTTCCAG	AACCGGCG-A	
	161								180
Consensus	TGAAGCACAA	GCGGCAGACC							
amphioxusHox2	TGAgaCAgAA	GaGGagGgaC							
mouse HoxA2	TGAAGCAtAA	GaGGCAaACC							
rat HoxA2	TGAAGCAtAA	GaGGCAaACC							
chick HoxA2	TGAAGCACAA	GaGGCAGACC							
Msax HoxB2	TGAaACACAA	GCGcCAGAcT							
human HOXB2	TGAAGCACAA	GCGGCAGACg							
mouse HoxB2	TGAaACACAA	GCGGCAGACg							
newt HoxB2	TGAAGCACAA	aCGcCAGACC							
salmon HoxB2	TGAAGCACAA	GCGGCaaACg							
Consensus	TGAAGCACAA	GCGGCAGACC							

Figure 7. Amino acid sequence alignment generated from the translation of the DNA sequence corresponding to the preserved 1st exon of paralogous group 2 *Hox* genes from all vertebrates characterized to date. Each single letter designation corresponds to a single amino acid. Gaps which were inserted to maximize the alignment are represented by “.”. Genbank Acession numbers: Chicken (*Gallus gallus*) *HoxA2*, X74323; Mouse (*Mus musculus*) *HoxA2*, M93148; Rat (*Rattus norvgicus*) *HoxA2*, M91802; Human *HOXB2*, X16665. *Morone saxatilis* (*Msax*) *HoxB2* data from λ clone 2H bases 10,223-10,675.

Figure 7. Amino acid sequence alignment of the region corresponding to the 1st exon of the *Hox* genes representative of paralog group 2.

		1																	80
chick	<i>HoxA2</i>	MNFEFEREIG	FINSQPSLAE	CLTSFPPVGD	TFQSSSIKNS	TLSHSTLIPP	PFEQTIPSLN	P..GGHPRHS	G..GGRPKAS										
mouse	<i>HoxA2</i>	MNYEFEREIG	FINSQPSLAE	CLTSFPPVAD	TFQSSSIKTS	TLSHSTLIPP	PFEQTIPSLN	P..GSHPRHG	AGVGGRPKSS										
rat	<i>HoxA2</i>	MNYEFEREIG	FINSQPSLAE	CLTSFPPVAD	TFQSSSIKTS	TLSHSTLIPP	PFEQTIPSLN	P..GSHPRQS	AGAGGRPKSS										
human	<i>HOXB2</i>	MNFEFEREIG	FINSQPSLAE	CLTSFPAVLE	TFQTSSSIKES	TLIPP...PP	PFEQTFPSLQ	P..GASTLQR	PRSQKRAEDG										
Msax	<i>HoxB2</i>	MNFEFEREIG	FINSQPSLAE	CLTSFPAVLE	SFQTSSSIKES	TV.....IPP	PFEHTIPSLS	PCTASQPRPT	ARSQQNRRTS										
		81																	157
chick	<i>HoxA2</i>		PRARSG	SPGPAGAPPP	.P...EYPWM	KEKKASKRSS	LPPASASASA	AGPACLSHKD	PLEIPDS										
mouse	<i>HoxA2</i>		PAGSRG	SPVPAGALQP	.P...EYPWM	KEKKAACKTA	LPPAAAS...	TGPACLGHK.											
rat	<i>HoxA2</i>		PAGSRG	SPVPARALQP	.P...EYPWM	KEKKAACKTA	LPPAAAS...	TGPACLGHKE	SL.....										
human	<i>HOXB2</i>		PALPPP	PPPPLPAAPP	AP...EFPWM	KEKKSARK..	...PSQSATS	PSPAASAVPA	SGVGSPA										
Msax	<i>HoxB2</i>	ATNGLQTHHR	GAQQPCPPGQ	APATATAAPA	GPLAHEFPWM	KEKKSSKKCP	.KPGSSSSGG	GSIIPPASSS	PSPTASG										

Figure 8. Amino acid sequence alignment generated from the translation of the DNA sequence corresponding to the homeodomain of paralogous group 3 *Hox* genes from all vertebrates characterized to date. Each single letter designation corresponds to a single amino acid. GenBank Accession numbers: Amphihox3 (*Branchiostoma floridae*), X68045; Mouse (*Mus musculus*) *HoxA3*, U56399; Fugu (*Fugu rubripes*) *HoxA3*, *HoxB3* as published in Aparicio, *et al* 1997; Mouse (*Mus musculus*) *HoxB3*, U02278; Human *HOXB3*, X16667; Chick (*Gallus gallus*) *HoxB3*, X74506; Xenopus (*Xenopus laevis*) *HoxB3*, M91588; Mouse (*Mus musculus*) *HoxD3*, U56401; Human *HOXD3*, D11117. *Morone saxatilis* (Msax) *HoxB3* data from λ clone 2H bases 1531-1710.

Figure 8. Amino acid sequence alignment of homeodomains representative of paralog group 3 genes

		1						60
Consensus		SKRARTAYTS	AQLVELEKEF	HFNRYLCRPR	RVEMANLLNL	SERQIKIWFQ	NRRMKYKKDQ	
AmphiHox3		gKRARTAYTS	AQLVELEKEF	HFNRYLCRPR	RVEMAamLNL	tERQIKIWFQ	NRRMKYKKeQ	
mouse <i>HoxA3</i>		SKRARTAYTS	AQLVELEKEF	HFNRYLCRPR	RVEMANLLNL	tERQIKIWFQ	NRRMKYKKDQ	
Fugu <i>HoxA3</i>		SKRARTAYTS	AQLVELEKEF	HFNRYLCRPR	RVEMANLLNL	tERQIKIWFQ	NRRMKYKKDQ	
Msax <i>HoxB3</i>		SKRARTAYTS	AQLVELEKEF	HFNRYLCRPR	RVEMANLLNL	SERQIKIWFQ	NRRMnYKKDQ	
Fugu <i>HoxB3</i>		SKRARTAYTS	AQLVELEKEF	HFNRYLCRPR	RVEMANLLNL	SERQIKIWFQ	NRRMKYKKD	
mouse <i>HoxB3</i>		SKRARTAYTS	AQLVELEKEF	HFNRYLCRPR	RVEMANLLNL	SERQIKIWFQ	NRRMKYKKDQ	
human <i>HOXB3</i>		SKRARTAYTS	AQLVELEKEF	HFNRYLCRPR	RVEMANLLNL	SERQIKIWFQ	NRRMKYKKDQ	
chick <i>HoxB3</i>		SKRARTAYTS	AQLVELEKEF	HFNRYLCRPR	RVEMANLLNL	SERQIKIWFQ	NRRMKYKKDQ	
Xenopus <i>HoxB3</i>		SKRARTAYTS	AQLVELEKEF	HFNRYLCRPR	RVEMANLLNL	SERQIKIWFQ	NRRMKYKKDQ	
mouse <i>HoxD3</i>		SKRvRTAYTS	AQLVELEKEF	HFNRYLCRPR	RVEMANLLNL	tERQIKIWFQ	NRRMKYKKDQ	
human <i>HOXD3</i>		SKRvRTAYTS	AQLVELEKEF	HFNRYLCRPR	RVEMANLLNL	tERQIKIWFQ	NRRMKYKKDQ	
Consensus		SKRARTAYTS	AQLVELEKEF	HFNRYLCRPR	RVEMANLLNL	SERQIKIWFQ	NRRMKYKKDQ	

Figure 9. Amino acid sequence alignment generated from the translation of the DNA sequence corresponding to the 2nd exon of paralogous group 2 *Hox* genes from all vertebrates characterized to date. Each single letter designation corresponds to a single amino acid. Gaps which were inserted to maximize the alignment are represented by ".". Genbank Acession numbers: Chicken (*Gallus gallus*) *HoxA2*, X74323; Mouse (*Mus musculus*) *HoxA2*, M93148; Rat (*Rattus norvgicus*) *HoxA2*, M91802; Human *HOXB2*, X16665. *Morone saxatilis* (*Msax*) *HoxB2* data from λ clone 2H bases 10,223-10,675.

Figure 9. Amino acid sequence alignment of the region corresponding to the 2nd exon of the *Hox* genes representative of paralog group 2.

		1								80
chick	<i>HoxA2</i>	SRRLRTAYTN	TQLLELEKEF	HFNKYLCRPR	RVEIAALLDL	TERQVKVWFQ	NRRMKHKRQT	QCKENQNGEG	KFKGSEDPEK	
mouse	<i>HoxA2</i>	SRRLRTAYTN	TQLLELEKEF	HFNKYLCRPR	RVEIAALLDL	TERQVKVWFQ	NRRMKHKRQT	QCKENQNSEG	KFKNLESDSK	
rat	<i>HoxA2</i>	SRRLRTAYTN	TQLLELEKEF	HFNKYLCRPR	RVEIAALLDL	TERQVKVWFQ	NRRMKHKRQT	QCKENQNSEG	KFKNLESDSK	
human	<i>HOXB2</i>	ARRLRTAYTN	TQLLELEKEF	HFNKYLCRPR	RVEIAALLDL	TERQVKVWFQ	NRRMKHKRQT	QHREPPDGEP	ACPG.....	
Msax	<i>HoxB2</i>	SRRLRTAYTN	TQLLELEKEF	HFNKYLCRPR	RVEIAALLDL	TERQVKVWFQ	NRRMKHKRQT	TPHRDGGGGG	GHESSDPGGF	
		81								160
chick	<i>HoxA2</i>	AAEDDEEEKA	LFEQALGTVS	GALLEREGYA	FQQNALSQQQ	AQNAHNGES.		QSF	PVSPLTSNEK	
mouse	<i>HoxA2</i>	.VEEDEEEKS	LFEQAL.SVS	GALLEREGYT	FQQNALSQQQ	APNGHNGDS.		QTF	PVSPLTSNEK	
rat	<i>HoxA2</i>	.VEEDEEEKS	LFEQAL.SVS	GALLEREGYT	FQQNALSQQQ	APNGHNGDS.		QTF	PVSPLTSNEK	
human	<i>HOXB2</i>	.ALEDICDPA	..EPPAASPG	GPSASRA..A	WEACCHPPEV	VPGALSADP.		RPL	AV.....	
Msax	<i>HoxB2</i>	EPLEGADASS	PYSSQPLEAS	GTGEDTSESE	PGSGNLFSAV	APTAYANDSG	DNAOPTPEER	GCLSRPERPP	QPGPLGSSDV	
		161								240
chick	<i>HoxA2</i>	NLKHFFQ..HQ	SPTVQNCLST	MAQNCAAGL.		...NNDSPEA	LEVPSLQDFN	VFSTDSCQL	SDAVSPSLPG	
mouse	<i>HoxA2</i>	NLKHFFQ..HQ	SPTVPNCLST	MGQNCGAGL.		...NNDSPEA	IEVPSLQDFN	VFSTDSCQL	SDALSPSLPG	
rat	<i>HoxA2</i>	NLKHFFQ..HQ	SPTVPNCLST	MGQNCGAGL.		...NNDSPEA	LEVPSLQDFN	VFSTDSCQL	SDALSPSLPG	
human	<i>HOXB2</i>	RLEGAG..AS	SPGCA.LRGA	GGLEPGPLP.		...EDVFSGR	QDSPFLPDNL	FFAADSCLQL	SGGLSPSLQG	
Msax	<i>HoxB2</i>	VASHHSPLRS	APRTPQTTRV	LMQLSEAGPA	AAASDPSE	QRDSSITSTS	SSAPPAASS	ALPDLTFFAG	DACLSPSLQS	
		241			271					
chick	<i>HoxA2</i>	SLDSPVDISA	DSFDFFDTL	TTIDLQHLNY	*					
mouse	<i>HoxA2</i>	SLDSPVDISA	DSFDFFDTL	TTIDLQHLNY	*					
rat	<i>HoxA2</i>	SLGSPVDISA	DSFDFFDTL	TTIDLQHLNY	*					
human	<i>HOXB2</i>	SLDSPVPFSE	EELDFFTSTL	CAIDLQF..P	*					
Msax	<i>HoxB2</i>	SLDSPVDFSE	EDFDLFTSTL	CTIDLQHLNF	*					

Figure 10. Amino acid sequence alignment generated from the translation of the DNA sequence corresponding to the human and striped bass HoxB2 genes. Each single letter designation corresponds to a single amino acid. Gaps which were inserted to maximize the alignment are represented by ".". Genbank Accession number: Human *HOXB2*, X16665. *Morone saxatilis* (*Msax*) *HoxB2* data from λ clone 2H bases 10,223-10,675. Shaded box with bold letters corresponds to the homeodomain. The shaded box with regular letters corresponds to the hexapeptide motif.

Figure 10. Amino acid sequence alignment of human *HOXB2* protein with *Morone saxatilis* *HoxB2* protein.

	1		80
Human <i>HOXB2</i>	MNFEFEREIG FINSQPSLAE CLTSFPAVLE TFQTSSIKES TLIPPPPPFE QTTPSLQPGA STLQRP..RS QK.....		
<i>Msax HoxB2</i>	MNFEFEREIG FINSQPSLAE CLTSFPAVLE SFQTSSIKES TVI..PPPF HTIPSLSPCT ASQPRPTARS QQNRRTSATN		
	81		160
Human <i>HOXB2</i>	RAED GPALPPPPPP PLPAAPPAP. ..EFPWMKEK KSAKKPSQSA TSPSPAASAV PASGVGSPAD GLGLPEAGGG		
<i>Msax HoxB2</i>	GLQTHHRGAQ QPCPPGQAPA TATAAPAGPL AHEFPWMKEK KSSKKCPKPG SSSSGGGSII PPASSSPSPT ASGRPEPGQS		
	161		240
Human <i>HOXB2</i>	G.....ARRL RTAYTNTQLL ELEKEFHFNK YLCRPRRVEI AALLDLTERQ VKVWFQNRMM KHKRQTOHRE PPDGEPACPG		
<i>Msax HoxB2</i>	GCWRSRARRL RTAYTNTQLL ELEKEFHFNK YLCRPRRVEI AALLDLTERQ VKVWFQNRMM KHKRQTPHR DGGGGGGHES		
	241		320
Human <i>HOXB2</i>	ALEDICDPAE EPAASPGGPS ASRAAWEACC HPPEVVPGA. .LSADPRPLA VRLEGAGASS P.....GCAL RGAGGLEPGP		
<i>Msax HoxB2</i>	SDPGGFEPLE GADASSPYSS QPLEASGTGE DTSESEPGSG NLFSAVAPTA YANDSGDNAQ PTPEERGCLS RPERPPQPGP		
	321		400
Human <i>HOXB2</i>	L.....P EDVFSGRQDS PF.....LPD LNFFAADSCL		
<i>Msax HoxB2</i>	LGSSDVVASH HSPLRSAPRT PQTTRVLMQL SEAGPAAAAS DPSFSEQRDS SITSTSSSAP PAASSSALPD LTFFAGDAC.		
	401	443	
Human <i>HOXB2</i>	QLSGGLSPSL QGSLDSPVPF SEEELDFFTS TLCAIDLQFP ..*		
<i>Msax HoxB2</i>	LSPSL QSSLDSPVDF SEEDFDLFTS TLCTIDLQHL NF*		

Figure 11. A schematic diagram representing a presumptive alternate promoter for the *HoxB2* gene from *M. saxatilis*. The presumptive alternate promoter was identified by analysis of the intergenic region between *HoxB2* and *HoxB3*. The location of promoter elements common to eukaryotic class II transcription units are enclosed in boxes and the numbers in brackets represent the nucleotide position of the nucleotide base nearest to the presumed transcription initiation start site. Numbers bracketed by directionally opposing arrows represent the distances in nucleotides separating the 5' and 3' nucleotides of the nearest neighboring promoter elements. The shaded box represents the first exon of *HoxB2*.

Figure 11. Schematic diagram of *Morone saxatilis* *HoxB2* putative alternate promoter located at positions 8665 - 8897 on λ clone 2H.

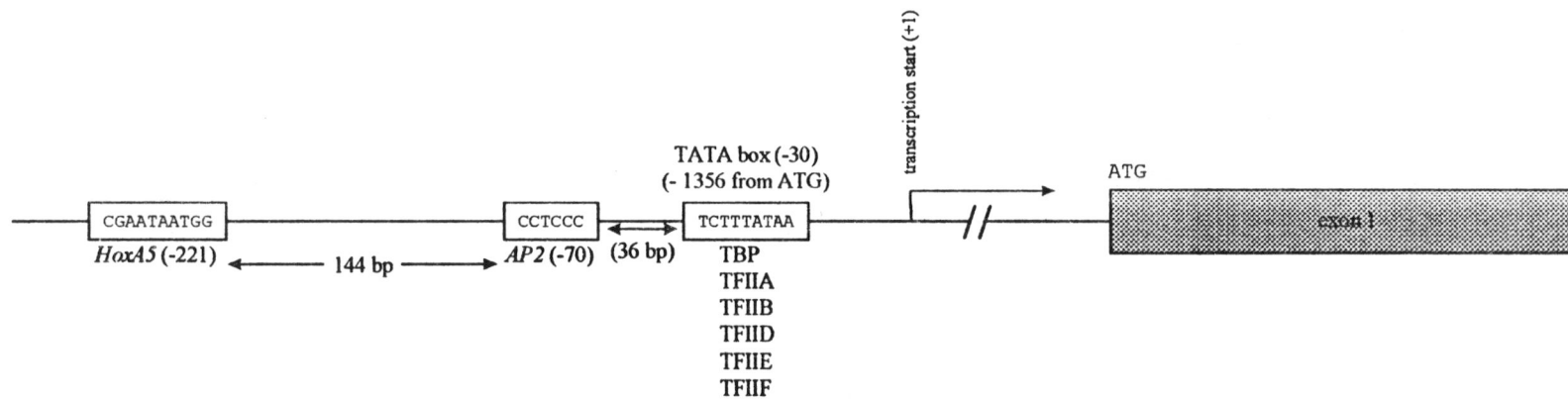
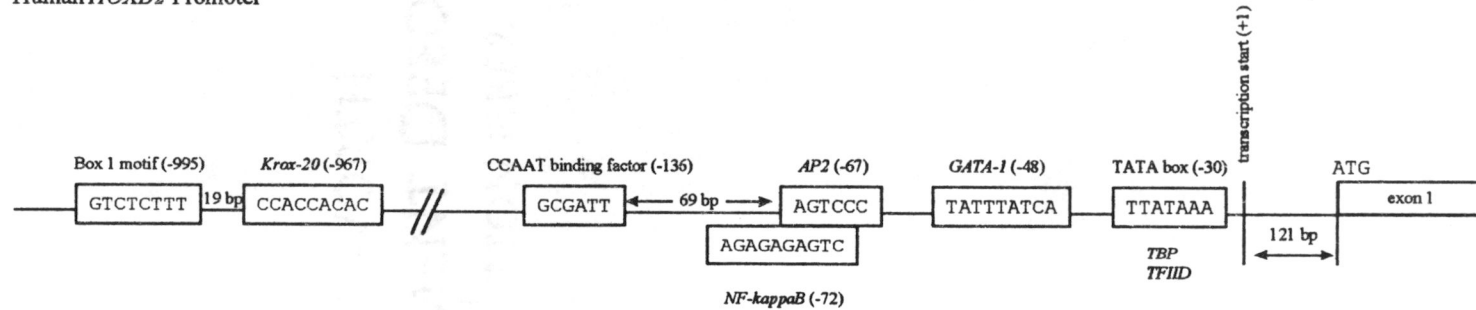


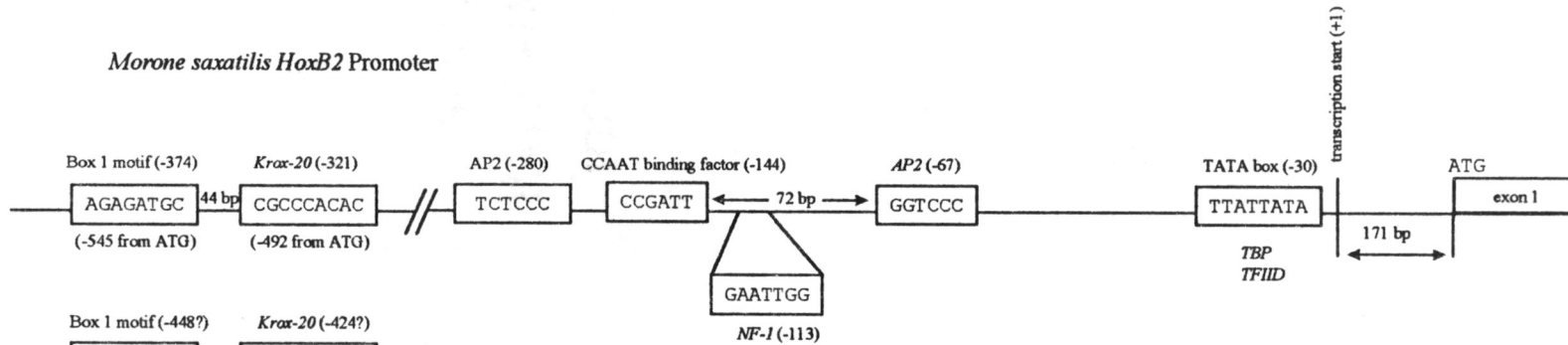
Figure 12. A schematic diagram representing a comparison among the conserved promoter elements shared by the three vertebrate species for which DNA sequence information is available. Numbers bracketed by directionally opposing arrows represent the distances in nucleotides separating the 5' and 3' nucleotides of the nearest neighboring promoter elements. Numbers in brackets represent the nucleotide position of the nucleotide base closest to the presumed transcription start site, which corresponds to the +1 nucleotide. *Limited data is available for the *Fugu rubripes HoxB2* promoter, although this region has been sequenced.

Figure 12. Comparison of *HoxB2* promoters from the human, *M. saxatilis* and *Fugu rubripes**.

Human *HOXB2* Promoter



Morone saxatilis *HoxB2* Promoter



Fugu rubripes Krox-20 enhancer (distances are estimated from the ATG)

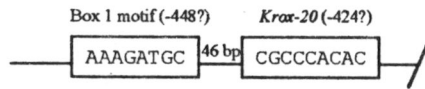


Table 1. Primers used for PCR confirmation of correct subclone assemblage with predicted and observed product lengths.

Primer #1	Primer #2	Predicted Length	Observed Length
λ left arm	2.1 R400	530 bp	530 bp
2.1 F7200	2.3 F2256	822 bp	822 bp
2.3 R142	2.2.2 R1200	1845 bp	1845 bp
2.2 F1300	2.2 R2295	1712 bp	1712 bp
2.2.1 R1490	λ right arm	1520 bp	1520 bp

Table 2. A list of DNA sequence elements that were used to identify similar elements in *M. saxatilis* λ clone 2H.

Pattern Name	Sequence
Repeat 1	TCAGATGGATGG
Repeat 2	TGATTGAAG
Block 1	CTTTGTCRTGCTAATGTC
Repeat 3	TGATGGATGAG
3' RARE HoxA1	GGTTCA
RAR-a2	AGTTCA
mouse HoxB1 RARE	AGGTCA
Fugu HoxB1 RARE	AAGTCA
mouse HoxB1 RARE	AGGTCA
Krox consensus	GCGNGGGCG
Krox consensus	GYGTGGGTG
mouse HoxA2 Krox-20 binding site	GCGTGGGTG
mouse HoxA2 Krox-20 binding site	CTGTGGGCA
chicken HoxA2 Krox-20 binding site	GCGTGGGTG
mouse HoxB2 Krox-20 binding site	GCGTGGGTG
chicken HoxB2 Krox-20 binding site	GTGTGGGTG
chicken HoxB2 Krox-20 binding site	CACCCACAC
Fugu HoxB2 Krox-20 binding site	GTGTGGGCG
human HoxB2 Krox-20 binding site	GTGTGGTGG
Box A	CTTTGT
Fugu Box 1	GCATCTTT
human Box 1	GTCTCTTT
mouse Box 1	GAAACTTT
chicken Box 1	GAAACTTT
GATA-1 consensus	WGATAR
Ets elements	AGGAAR
Ets elements	GAGAA
Ets elements	CTTCCT
Ets elements	TTTCCT
Ets elements	TTCCTC
Ets elements	RYWTCCKGY
Repeat consensus	WGATRGAWG

Table 3. Notable DNA sequence-based landmarks identified from λ clone 2H

Feature Name	Description	Position
exon	<i>HoxB2</i> exon #1	10223 - 10675
exon	<i>HoxB2</i> exon #2	12243 - 13094
exon	<i>HoxB3</i> exon #2	1461 - 2330
homeobox	<i>HoxB2</i> homeobox	12282 - 12461
homeobox	<i>HoxB3</i> homeobox	1531 - 1710
intron	<i>HoxB3</i> intron	20 - 1460
intron	<i>HoxB2</i> intron	10676 - 12242
poly-A signal	<i>HoxB2</i> poly-A signal	13781 - 13786
poly-A signal	<i>HoxB2</i> poly-A signal	14159 - 14164
poly-A signal	<i>HoxB3</i> poly-A signal	2656 - 2661
poly-A signal	<i>HoxB3</i> poly-A signal	3012 - 3017
ARE	<i>HoxB2</i> AU-rich element	13453 - 13460
ARE	<i>HoxB2</i> AU-rich element	13708 - 13715
ARE	<i>HoxB2</i> AU-rich element	13742 - 13750
ARE	<i>HoxB2</i> AU-rich element	14036 - 14044
ARE	<i>HoxB3</i> AU-rich element	2591 - 2600
TATA box	<i>HoxB2</i> promoter element	10010 - 10017
<i>Krox-20</i>	Intergenic element	3386 - 3394
<i>Krox-20</i>	Intergenic element	2239 - 2247
<i>Krox-20</i>	<i>HoxB2</i> promoter element	9717 - 9725
<i>GATA-1</i>	Intergenic element	2677 - 2682
<i>GATA-1</i>	Intergenic element	3123 - 3128
<i>GATA-1</i>	Intergenic element	3497 - 3502
<i>GATA-1</i>	Intergenic element	6305 - 6310
<i>GATA-1</i>	Intergenic element	6503 - 6508
<i>GATA-1</i>	Intergenic element	3266 - 3271
<i>GATA-1</i>	Intergenic element	4403 - 4408
<i>GATA-1</i>	Intergenic element	5672 - 5677
<i>GATA-1</i>	Intergenic element	6152 - 6157
<i>GATA-1</i>	Intergenic element	6631 - 6636
<i>Ets</i>	Intergenic element	8969 - 8974
<i>Ets</i>	Intergenic element	9002 - 9007
<i>Ets</i>	Intergenic element	9250 - 9255
<i>Ets</i>	Intergenic element	9444 - 9449
<i>Ets</i>	Intergenic element	9482 - 9487
<i>Ets</i>	Intergenic element	9644 - 9649
RARE	Possible <i>HoxB1</i> enhancer element	15438 - 15443
RARE	Possible <i>HoxB1</i> enhancer element	15478 - 15483
Box1	Box 1 element	9665 - 9672
<i>AP2</i>	<i>HoxB2</i> promoter element	9760 - 9765
<i>AP2</i>	<i>HoxB2</i> promoter element	9973 - 9978
<i>AP2</i>	Alternate <i>HoxB2</i> promoter element	8818 - 8823
CCAAT binding factor	<i>HoxB2</i> promoter element	9896 - 9901
<i>HoxA5</i>	Alternate <i>HoxB2</i> promoter element	8665 - 8674
TATA box	Alternate <i>HoxB2</i> promoter element	8859 - 8867

APPENDIX A: SYNTHETIC OLIGONUCEOTIDE PRIMERS

The primer name is the name given to the primer when designed and ordered from the manufacturer. The designation is based on the location relative to the first base of the primer with relation to the first base of the λ clone 2H DNA sequence from Appendix B. The sequence is the actual sequence of the primer.

Name	Designation	Position	Sequence 5'-3'
pUC-19 Forward primer	pUCF	n/a	CCC AGT CAC GAC GTT GTA AAA CG
pUC-19 Reverse primer	pUCR	n/a	AGC GGA TAA CAA TTT CAC ACA GG
EMBL3 Right Arm (-39)	λ RA	n/a	TAA CGA GTG GAT CTG GGT CG
EMBL3 Left Arm (-37)	λ LA	n/a	GCA GAC AAA CTG CGC AAC TCG TG
2.1 R170	1R	168	AAA CGG TGT CAG AGA GG
2.1 F355	2F	336	TCC CCA GAT CTC AGC AG
2.1 R400	3R	472	AGC CTC TGT CAC CGC CAT TC
2.1 F800	4F	809	ATA CAC ATT GCC ATT CTG
2.1 F1200	5F	1,222	ATT CTC TTC CTC TCT GG
2.1 F1526	6F	1,509	ACC GGT GGG TCG TCA GCG
2.1 F1684	7F	1,655	ACG CCA GAT CAA GAT CTG G
2.1 F2043	8F	2,044	AGG CCA GTC CAG TCT ACG
2.1 R5822	9R	2,300	AGC TTG GGT GCT TCC TGG
2.1 F2427	10F	2,410	ACT AAC TAA CAT GTT GGG G
2.1 F2618	11F	2,617	TGT TAC CAA TTA GCA GTG
2.1 F2941	12F	2,943	AGC GAA GCA TAT TGC AGG
2.1 F3353	13F	3,355	TAA TTC ATT ATT GGA CTG
2.1 R4117	14R	4,006	TAA GAG TGA TGG GAG GGG
2.1 F3808	15R	4,312	ATA AAT AAA TAG TTC TCC
2.1 F3347	16R	4,775	ACA ATG ATG CCC AGA TGT G
2.1 R4800	17R	5,082	TAA AAA ATG TAA CTT CAG
2.1 R2935	18R	5,187	ACC AAT CCC CAC CCT TGC C
2.1 R5000	19R	5,297	AAA AGG AGC TAC AAA GGG
2.1 R5600	20R	5,785	AAT AAT GGC GAT GTC TGG
2.1 R6000	21R	6,190	ACC ATC TAG TCT GTA AAG
2.1 R6400	22R	6,579	ATG GGA TTG GAG GGC TG
2.1 R6800	23R	6,982	AAA CAG TGG ATT AGC TAG
2.1 R7200	24R	7,341	TAC AGA GAG CGA GGA GAG
2.1 F7200	25F	7,371	ACT CGG AGG CCT GAG AGC AG
2.1 F 2.3 (2992)	26R	7,702	AGT CAA ATG GAT AGC TAC TG
2.3 F2256	27R	8,177	TGG TGG AGA GCA GAA GC
2.3F 1925	28R	8,821	AGC AGA GGA TAT TTT GGG G
2.3F 1597	29R	9,137	CTG CCC TCT CTC TCT CTC
2.3R 1490	30F	9,244	GCA CTG AGG AGA GAT CCA
2.3 F879	31R	9,485	ATC CCC AAC TTC CAC ACA GG
2.3R 1040	32F	9,692	ACC GGG CCG GGA TGC G

Name	Designation	Position	Sequence 5'-3'
2.3 Forward 490	33R	9,856	ATG GCT GTG AGA AGC AGC
2.3 R142	34F	10,189	CGC CGT TTG AGG GTG AA
2.2 into 2.3	35R	10,375	AGG GGC TGA GGC TGG GG
2.2 Forward 320	36F	10,658	TCG CCG ACC GCC TCC GGG
2.2 F650	37F	11,075	GCT TAA CAC TGA GTG CAG
2.2 F900	38F	11,261	ACT GGG TGG TGT GAT TTG
2.2 F1300	39F	11,628	TGC CTT TAT CGC ATC CAT G
2.2.2 R1200	40R	12,017	AGC TGC TTC AGC ATG GAG
2.2.2 R900	41R	12,311	AGC TCC AGC AGC TGT GTG
2.2.2 R580	42R	12,602	TCC CCG GCT CGC TCT CC
2.2.2 Reverse 240	43R	12,986	GCT CTG CAG GCT GGG TG
2.2.2 R2600	44R	13,014	TCG GAG AAA TCC ACT CGG
2.2 Forward 2500	45F	13,192	GAC GGC ACT CGG CGA GGA C
2.2 R2295	46R	13,324	AAA CTA CAT TAC AGC CG
2.2.1 Reverse 360	47F	13,585	CTA GGT GCA GAT AGC CGG
2.2.1 R 859	48F	14,072	CAA TAC ATT TAT CTA TTG C
2.2.1 R1490	49F	14,095	AGC CTA TAA TTT CCA GGT T
2.2.1 F1085	50R	14,473	ATT AGC TCC TAC TCC GGG
2.2 R 777	51R	14,825	TCA CAG TGA GCC AGT CAT C
2.2 Reverse 350	52R	15,237	TGC ATT GGG CAA TAG GCC
2.2 R114	53F	15,444	AAT CCA TAC AGA AGA GGC

APPENDIX B: Complete DNA Sequence of λ clone 2H

1	GTCGACCTGC	AGGTCAACGG	ATCATTTTGG	AGGTGTTTCC	ATTATTTACT	50
51	CTACCCTCAT	TGATATATCA	GGTTGGACAT	AACATTGCAA	ACAACCTCAG	100
101	TATGATGCAA	CATAATTTAA	CAGCACCAAA	CTGCAGCCTC	TAAAATAGCC	150
151	ATACAGCTGA	ATCAACACCT	CTCTGACACC	GTTTCAACAA	AAACTGAACA	200
201	TTATAACCTA	ATTTGAGCTA	GCTAATAAAC	TTGGTACACA	GTGTAGATGA	250
251	ACTGACTTGT	TTACAACCTG	TCTGAACACA	GAATCACACG	TATTTCTTCA	300
301	ACTGTCGGAC	GTATTTTTTA	CTGAGATTGC	AGTGTCCCC	AGATCTCAGC	350
351	AGTGACTTTA	GTCACCACAA	TGTTATTCTA	ATCAATAAAA	CTAATGACCA	400
401	GAGCTAAGAA	AATAAACTT	AAGTTGCAAT	AAACTTTTTT	GCCTCCTCAC	450
451	TATGGTTGCA	GCACTGCTGA	AGAATGGCGG	TGACATAGGC	TGATTTGAGG	500
501	TTTACCAGCT	TTTATGGCTT	GCAAACAAAT	GATGTGAAAC	ATTGTGCAAA	550
551	CATTTCTTTG	TAATGTTTTA	GAAGGAATTC	TTTATTGGAA	ACCGATTGAG	600
601	AGGATGGAGG	ATAATCATCG	GCTATGGGCT	AGTTAATTGT	GGCTAGTATA	650
651	GACTACTTGA	GCAATTTTGA	GTCAGATAAC	CAATTATCAC	CTAGAGCCTT	700
701	ATTCTCTTAA	ACGTGTCTTA	TAAAGCAGTG	ACAGCATCTG	TTGGATTTTA	750
751	AGGTTTAGCG	GGCTCTATGG	CCTTAAGACA	AAATGATTAG	CTTCTGCTG	800
801	AGAATAACAT	ACACATTGCC	ATTCTGAGCA	AATATTTACA	GTAGGTAGCA	850
851	AGTTTCCACT	GGCGTGAGC	TGAATGCCTT	TCAGCTGTTT	AAGAGTCAGT	900
901	ACAACATATCA	CATAAACGTG	GTGAGAAGTT	TGACTTTTCT	TGGCGAGCTG	950
951	ATTGAAGTGA	GTGCAGATCA	ACCTCGGTTT	CTGCTCGTAG	TCCTTGTGCT	1000
1001	GCAGTGCTG	TTGCTGTGGG	GGGTTTGTCT	TTCCAGTCGA	GATGCAATTT	1050
1051	AGCTGCCCTC	GGCCAGCGCT	GGGGCCTCCG	TCTCTTAATA	ACTGCCACTG	1100
1101	TTATGAAAGT	TTGACAACTA	ATATCAAGGG	ATAATTCAGC	CAGCATGCCC	1150
1151	GGACCCCTT	CTGCTTTATT	AAGAACCTGA	ATGCTTTCAA	TTTCCACCAG	1200
1201	CATGCAGCGT	GATTGTGTTT	TATTCTCTTC	CTCTCTGGTT	CTCTCAATGC	1250
1251	TCAATTCATC	ATCTTTTTAC	CCTTTTTTAA	TATCCACACT	TTTCTCCCTT	1300
1301	TTTGTACATT	TCTCCCACTT	TTGCTCTCTT	CATTTCCCAG	GGGTGTCCCT	1350
1351	TCCCTGTGTT	CTTTTGTGTT	TCTTTCTCTG	TGTCCTTGCC	TTTTTTAATT	1400
1401	TCATTCTCCA	TCCAACCTC	TCACTTGCCC	TTTCTCGGCG	TAATGTTTTT	1450
1451	TCTTCTCTCC	TCAGCATCAG	GAGGAGCTGA	GAGTGGTAGC	AGCGGGGAGA	1500
1501	AGAGCCCTAC	CGGTGGGTCG	TCAGCGTCAT	CTAAACGGGC	TCGTACGGCG	1550
1551	TACACCAGCG	CCCAGCTGGT	TGAGCTGGAG	AAGGAGTTCC	ACTTCAACCG	1600
1601	GTACCTGTGC	CGGCCGCGCC	GTGTGGAGAT	GGCCAACCTG	CTTAACCTGT	1650
1651	CCGAACGCCA	GATCAAGATC	TGGTTCAGAA	ACCGGCGCAT	GAAGTACAAG	1700
1701	AAGGACCAGA	AGTTGAAGGG	CCTGAGCTCC	AGCTCCGGAG	GGCCTTCACC	1750
1751	GACTGGCTCC	CCACCCCTGA	CGATGCAGTC	CTCCGCCACT	TTCTCTCACT	1800
1801	CAATGCACCC	CATGGGAGGG	GGAGGAGGAG	GGTCAGGCTA	CGACGTGCCA	1850
1851	TCTCCGCCAT	CCTTTGGCAA	GCCCCACCAG	GGAGTGCTTT	ACTCCATGTC	1900
1901	CACTGCCTAC	TCCAATGTCC	CCATGAAGGG	CTGCCCCCCC	CAGCAGAAGT	1950
1951	ATGGGGCCCC	AGACCCGGAC	TACGGCGACC	CTCATCCCCA	CCACGGCCTA	2000
2001	GTGCAGGCCA	ACAACGCTGG	CTATGGGACT	CCCACCAACA	TGCAGGCCAG	2050
2051	TCCAGTCTAC	GTGGGGGGTG	GCAGCTATGT	GGAACCCATG	CCTGGCTCGG	2100
2101	GGCCCTCTAT	GTATGGGCTA	AACCACCTCG	GCCCCACGCC	ACCCACCCAC	2150
2151	CAAACCATGG	ACTACAATGG	CGCGGGGGCC	ATGTCGGCCG	CCAACCAGCA	2200
2201	CCACGTGGGC	GGAGGGGGCC	CCCCGGGTCC	CTGTGACCCG	CCCACACACC	2250
2251	CTACGTACAC	CGAGCTGTCC	GCGCACCA	CCTCGACTCA	GGCAGAATC	2300
2301	CAGGAAGCAC	CCAAGCTCAC	TCACCTGTAG	CAGGTTTGA	ACAAAAATGA	2350
2351	CAACGTAGGA	AATAGTGAAG	ACGACACAGA	ACAAGTAACA	CGACGCGACT	2400
2401	CACTAACAGA	CTAACTAACA	TGTTGGGGAC	TTTACTGCTG	CAACGGGACG	2450
2451	GCAGGAGAGA	CAGAACGAGA	CAGACAGGAG	CCATCCAGTG	GTTTGTCCCC	2500
2501	GTGCTCTTAT	AGCCACAAAG	GGGAAACAAT	CAGTGTTTGG	TGTTGTCCAG	2550
2551	CACCACCCTG	TTTCCACTGC	CTTCATCAAT	TGTAAGTTCA	TTTAATTAAA	2600
2601	AAAACAAAGC	TGTGCCTGTT	ACCAATTAGC	AGTGATGGAC	CCAAACCATT	2650
2651	CATGTCAATA	AAAGTACTGA	AATCAATGAT	AGTTGCCAAA	AATTCCAAGA	2700

2701	ATTTTTTAAAG	CTGGGACCAA	ACTTACTAAT	TTACTCTCCA	CTACTGAAAT	2750
2751	CCTCATTGAG	GAGAACAGGA	AACCAAAACA	TTTAAACCTT	TTAACATTAA	2800
2801	TCAGTTTGTT	AACTTCCCAT	AAAAATGGTG	GCTTGCTTAA	TAATCAGAGA	2850
2851	TTTATTTTTC	ATTTTAGAAC	TCTTGGCTAC	TTTTACGCAT	GCTTATGAGT	2900
2901	AATTTTACACT	CTACTCTGAG	TTGTATTTCT	TGGTCCCTGG	GCAGCGAAGC	2950
2951	ATATTGCAGG	CGTGGGAGAC	GAGTGTTACA	GTCCAATTGT	CTCCCACGTC	3000
3001	CCCCTGGGGA	CAATAAATGT	CAGGACACTC	GTCTTGTGGC	ACTGGTGCAG	3050
3051	AGACATATTG	GCAGCTGTAT	GGCATAACAG	GGTCTATGGT	GAGCACCATT	3100
3101	GTTTGGTCAG	TATTCATTAG	ACTGATAATC	TACTGCACAG	TCAGTCTGAA	3150
3151	AACTGCAACA	TGGACTACTG	TATTCAAACA	AAGCAGATGT	CAGAAGGCAA	3200
3201	AATGTCAGAT	TTTGATTAAAG	AAAACCTTCTC	AAGAGACCAA	AAAAAACTA	3250
3251	AAATGAAAAA	AAAATCTATC	AGTATGATAT	GATATATATA	TTTAAATAAT	3300
3301	TTAAATAATT	GCAGCATTAT	TACACTGTCA	CAAAAGTCTT	CCTCTGTGGA	3350
3351	TTCCTAATTC	ATTATTGGAC	TGCTAAGCTA	TTTTTGC GTG	GGCGGAGCAA	3400
3401	ATTTGCAGCA	GTAGCTCAGT	TAGCAGTAGC	ATGTTGTAAA	AAAGAAAAGC	3450
3451	CTGTAATGGC	TCTGTTAGAG	CTAGTATGTG	CTTTAAAATT	GTTCTTTGAT	3500
3501	AATAAAATCT	CCCCCCTAAG	ACACACACAT	ATACAAACAG	CTCTCCC GCC	3550
3551	TGGTAAATGA	AATAGTTGAG	ATTTTAAATGG	GCAGTTAACG	AGGTGGTTAG	3600
3601	ATTTATGACG	CTGAGCTCTC	TCCCCCTCCA	GGCAACACAT	GTTGTTGTCA	3650
3651	GCCGAGCTGA	GCTATGTGGA	GGGTTCTGCC	TCTCCATCAG	CCGGAGAGAG	3700
3701	AGCGAGAGAG	AGAACAGAAA	TATATGGCTG	GAGCTGGCAG	CACCTCGTCTT	3750
3751	CCTCTGTTCC	CCTCGTTCCA	CGAAGCACAT	AAAAACCACT	GGACGGTCAT	3800
3801	CTCTTCACTG	TCTCTTTTTG	CCTCATGCCG	TACCCGTCTC	TGTCCGTCCA	3850
3851	GA CTGAATCA	CTACTAACCC	TCTGTCTCCT	GTCTGTGGCA	GCAGCACATG	3900
3901	GCTTCTGAAG	GAGGCAAAAT	TTCCCACAAT	CCGCCACAGA	CCACCCCCCT	3950
3951	CCCACCCAC	ACACACACAC	ACACACATCC	ATCACCACAC	ATGCACCACT	4000
4001	CAGAACCCCT	CCCATCACTC	TTAACAGGGT	TTAAGAAGCT	ACAAGCTCAA	4050
4051	AGGCTAAACT	AGCTCCCAT	GACACACATC	TGTGCTAAGC	TACATTAGCT	4100
4101	AATCACTCCT	TTACACACAC	ACACACACAC	ACACACACAC	ACACACACAC	4150
4151	ACACACAGCA	ATGCTAAGCT	AGCTTCCTTT	CGTTGACACA	CATCTACAGC	4200
4201	CAGGTACGC	CAGCTGTTGA	CATCTTCAAA	CAAACACATG	AGAGGCATG	4250
4251	GAAGAGAAAT	TTTTAATGTT	CTTACTTTAG	CACATCATCG	ACACAGGTAA	4300
4301	ATAATAAGAG	TGGAGAACTA	TTTATTTATG	TTTGGGTAAG	TCATTGCCCT	4350
4351	CCTCTCTCTC	TCTCTCTCTC	TCTCTTCCCC	CTCTCTCTCT	CTCTCTCTCT	4400
4401	CTCTATCACT	TTGGGAAGGG	TTCACTTTCA	TTAGCGGACA	CCAATGCGTA	4450
4451	AAAGTCCACA	TTTTATATTA	TTTTCTTTTG	TTCTCAACAT	ATCTTGACTT	4500
4501	TTAACGTGGA	TACAGGGTAT	ATATTTGAAC	AAAAATGCAT	GTCCCAGAAA	4550
4551	TCTGAGCTGA	TGTGTATCAC	ACATTTTTGA	AAACAAC TGG	GTTTATTTGT	4600
4601	TTGAACCGTC	AGTGAGGGGA	GTCTAGGAAA	ACACACACAC	GAGACCCTTA	4650
4651	CAGACTGTTT	CAGCCAACT	GAAAACAAGC	TGCGTGTGTG	GCTGACCCCT	4700
4701	GCTACACTCT	CTGCTGTTCT	GGAAATAAAC	CTGTCATATG	GTTTTTAAAA	4750
4751	GTGCAGATTT	ACACAACTC	GTTTCAATC	TGGGCATCAT	TGTTGCACCT	4800
4801	AGAGTTTACA	TTTTGATGGT	TAAAGGTAAA	AAATAATAAT	AAAATCTTAT	4850
4851	TTTGAAGACG	GAAGGCCCTC	CAATCAAAGC	GTCTTTCGCC	AATGTGTGTT	4900
4901	CAAGTGAACA	TTCATATATA	AATATTTATT	TGTTATAGCC	AGTTTAAAAA	4950
4951	GACTTTCTGT	TTTGTATTAT	TATTATTTAT	CCTTCATGTA	TTTATATCGA	5000
5001	GAAAAAATAC	TGTACTTTTT	TAGTATTTAC	CTGTTATGAA	AGAAAAAAA	5050
5051	AACAACATTG	TGTTTCTTCT	TGTCCATTGT	ACTGAAGTTA	CATTTTTTAG	5100
5101	TTCTTGTATT	TTAGAAGAAA	ATAAGTAGTT	TTTATGTTCC	TATAATGACC	5150
5151	CTTGTACATA	AATGTCTTGG	CTTGTACAGG	GAAAAAGGCA	AGGGTGGGGA	5200
5201	TTGGTGACCC	ATCCGTGCAT	AGTCGGAAAA	TTTAGACAAA	CTATTAGCTG	5250
5251	GAAAAAAGAA	CGAGTCTTTT	TGGAAAAAAA	CAACTTCTTC	CTTTGTCCCT	5300
5301	TGTAGATCC	TTTTTCTTTC	ATTGTTGAGA	GAGTTGTCTG	TCAAACAATT	5350
5351	CTTTAATAAA	ATGTTATGTT	ATTCAGCGCT	ATCGTCAGTT	TTTTATGATT	5400
5401	CTCAGGAGTG	AAGAACTTGA	CCAGAATGTC	CCATCTGAAA	AGACACTCTG	5450
5451	TTAGGCAGTG	TTAGTGTGAG	G TAGAAACCT	GCCAACACTG	ATTCTATTGA	5500

5501	CTGTCATGGA	AACATGCAAA	CAACCCCTGG	CACTGCTGGG	TTCCTCTTGG	5550
5551	CTGCAAGCTG	TGGCAAAATT	CATAAACTAT	GAAAATATGT	AAATAATTGC	5600
5601	ACACCCTAGG	ATTAGGACTA	AACAAACTTT	TAAAGTCAGT	CGAGTTCTGT	5650
5651	GGTGGTGCAT	GTTGTTGTTG	ATTATCAACT	TTGAAGGATG	ACATATTCTT	5700
5701	CATTTATTTT	CAAGCTGCTA	AACCTTTTAG	AACCAGCAGT	GAAGGCCAAA	5750
5751	GCATGACTGC	TGTTACAAGT	GAACATTTTA	AAGATACCAG	ACATCGCCAT	5800
5801	TATTTTCATCC	AAATCCTTTC	AAAGTACAAT	AACTGTTTAA	ATCTTGACAA	5850
5851	AATCGTTTTG	TTTGAAGATG	TCATGTAATG	TGTACAGTAT	GGGGACCGTG	5900
5901	TATATTCTAG	CATCATGTGA	TGGTTGAGGA	CTCAAAGCGT	TGTTACAAAA	5950
5951	TCCCAGATCT	GATGTGACTT	CTTTCAAACA	GCGTGTCTCT	GCTTGAGGGA	6000
6001	TGACCTGACA	CTGATGCCAG	TATTGAATAT	CGGGTCAATA	CCTCAGGCTA	6050
6051	AAGCATGGCA	TCACTGTCAG	AAAACCTTCC	CAACAGTTCA	ATCCGAAGCC	6100
6101	TAGAGCCACA	TGTAACATGT	CAGAAAGAAA	AAGAAAAAAC	CTCTTCAACC	6150
6151	TTTATCAAAA	CTTTAAATGT	GCCTCTAATC	TCCTTTATGC	TTTACAGACT	6200
6201	AGATGGTTAC	TTGTTATTTT	TCTGGTCAGC	CTGATACATG	TTGCTCCTTG	6250
6251	ATTAGTTTTT	TTCTTAGAGA	CCTCTCTCTA	GGGAGGAAAC	ATTGACTTCA	6300
6301	GCTCTGATAA	CCCCAACTA	TTGGTCAGTG	ACCAGTTCAG	TAACAGTAAA	6350
6351	AGCTCAAAC	CACTCAGAAC	TACTCTAGAT	ATTTTAGTGT	CCATTGCACC	6400
6401	ATCACAGTAT	TGCATTTGAC	AGATTCTATT	AATGAGAGTC	AATTTCTCCT	6450
6451	CAAAAATAAG	ATACAATTTT	GAAATAATGA	ATGCAGGTCA	TGATTTGAAA	6500
6501	ACTGATAAAA	TAGATTATTT	TTAGTGTTTA	AGTTACTCGT	TATTGGAGCG	6550
6551	GCAAAGTAGC	ATACATGCAA	CTTGGGTTCA	GCCCTCCAAT	CCCATGCTGT	6600
6601	TGAAGTGTCC	TTGAGCAAGC	TGCTGAAACC	CTATCTGACC	TCTAAACTTG	6650
6651	GACGTTTCAC	GTGAGAAAGA	CCTGTTCCCTC	GCCATCATTG	CTCACTTTAT	6700
6701	TAAGTTCAAT	AGTGTTTAGG	AGCCAGTTCT	GTGGAAGTGC	AGCTCTGAAA	6750
6751	ATTGTGAGCA	AGAGGTTGAG	GTGTGGTTGA	AAGCTCTGGA	AAAATACAGT	6800
6801	GAGTTAAAC	TTTAAGTATG	AGTTAGCTTC	TTGTGACCAA	TTCATGATCC	6850
6851	CAGACCAGAA	CTGGAGAAAC	CCAGCAGATG	GTTGGAGTTA	TATAACACTC	6900
6901	AAAGCTCTGG	TCATTTCTTT	GACAGATGTC	TTGTACTGCT	TAGACTTACC	6950
6951	ATGCAAAAAG	AGAGTTTGGG	CTCTTTTTTT	ACTAGCTAAT	CCACTGTTTG	7000
7001	ATCATGTTAA	TTACTGACTG	TTTTTAATAA	TTGACTGTTG	ATGGCTGACT	7050
7051	GCCTGGATTG	ATGTTTTTTT	TGTTGTTTGT	TTTAGAAACT	TTTACTATTA	7100
7101	ATCAGAAAGT	TAATTTTCAC	TCTTGGGGGA	GGCAAACCTG	TTCCACATTT	7150
7151	TTCAACAGTC	ACCTGTGACC	CAAACGCAGA	GATCCCCGCT	CTGTCTGTTG	7200
7201	CTTGCTCTCT	GTAACACACA	CAGGCTGATC	CTCCTGTTAG	CTATGTGTCC	7250
7251	TGAGTTGATG	TTTATGACTA	AAAAAAGCCC	CGAGGCCTGC	GCCTTGGCCA	7300
7301	GGTCACGTGG	TCTCTAACCC	CCCCAACCGT	ACCTCTCTAA	CTCTCCTCGC	7350
7351	TCTCTGTAGC	ACGCGGTTGT	ACTCGGAGGC	CTGAGAGCAG	CATAAAACTT	7400
7401	GGAACATAAA	TCAGGCTCCC	ATAAATAATG	GCCGTACTTC	CCTTTACATC	7450
7451	TCCGTTGGAC	TAACAGACAA	TAAAAACACC	AATAAGCAGA	GATTGGATCT	7500
7501	AATTTATTGC	CTCATAAAAA	CAGACGGAGA	GCTGGGCTCG	GCTTCAGCAC	7550
7551	GCAGTCTGCT	GAGAACAGTA	CAGTTATATA	CAGTGAGACT	GTAGAGCAGC	7600
7601	TCACAGCCAG	GTACACTCAT	CTGCCATTAT	AACCGACATT	TAATATGGCT	7650
7651	ACTACACTAG	GCAGGAAAGA	TGTTCTCTGT	CAGATGTCAA	CAAATGTCTA	7700
7701	ACAGTAGCTA	TCCATTTGAC	TAAACTGGAA	TAATAATAAT	AATGAAAACA	7750
7751	ACAACAAAAG	TGCTAAGAAA	AATCCAGGTT	ATTTTTACTC	AGTAGACCAC	7800
7801	AAAATAAATA	AATAAATAAA	CGAGGGAAAA	CAATTCATGA	ACTACATCTG	7850
7851	AAGCTTTTTT	TTTTCTTTTT	TTTCTTTTGG	TTTATTTATT	TATTTATTTA	7900
7901	TCCTTACCGG	AAAGAGGTTT	TGGACCAATA	AAGACTATTG	TGCTAAATAT	7950
7951	CAGTAATTAG	TAATAACAAT	TTAAATTGAA	TAAAGTGAAA	ACGTGGTTAC	8000
8001	TGGAAGGAAA	AAAGAAACGG	TTTACCCTAC	CACTGTCACT	AACCCCTTAA	8050
8051	CAAAAGAAGT	GACTCCTGAA	TAGAATAAAA	TAAAAAATAA	AATAAAATAG	8100
8101	AATGAATTAA	AAAGTCACTA	ATCGTCATCT	TCAGTCGACC	AATCCACGCA	8150
8151	TCCCACGTAA	AGACTTAAAA	AAAACCTGCTT	CTGCTCTCCA	CCAAAACCTC	8200
8201	AGAATTCCAC	TCGGAATTTT	ACAAAAAATA	AAAAAACATA	TTTAAAAGAA	8250
8251	AATCCTGACA	TGGTAAAAAT	AGGCCGAAGA	GGAGAGAGCA	GTGTGTTGTG	8300

8301	AAGCTGCAGC	TCAGTCAGCA	GTGTGTGTGT	GCTGTGTGTG	TGTGTATGCT	8350
8351	GTGTTTGGGC	TGTGTGAGCC	TGTGCAGCCT	CGGCCTACTC	ACTGCCTGCC	8400
8401	GCTTTCTACT	GCGCTCAATG	GCCCTCACTC	CCACACACAT	TTCTGCTTTT	8450
8451	TCTGCCTCTC	TTTCAGAGCC	AGGGGGCCAA	GCCAAATGGC	ATGGGAAAAA	8500
8501	CAGGGAGAAG	GGGGTGCGAG	CTAAATAGCT	TATATCCATT	TGTCTCTGGT	8550
8551	ATCTGAGTCG	GTTATTTATG	TATTTTCTCC	CAGCAGAGGA	AACATTTCTG	8600
8601	CGTGACATGC	GCGTATTTGC	TGCGCGTAAA	AGCAGTGTA	AAAAAAAATA	8650
8651	TACTAAAAAC	ATGGCGAATA	ATGGAGCTGA	AAGGAGAGAG	ATATAGGGAG	8700
8701	TAAGACTGGA	GCGTTTCTTG	TTATTTAGGG	AGAGGGAGAG	AGGGGCTGCT	8750
8751	GTCAGGCTCT	TTGCAGGGAG	AGAAAGATTG	ATCACCGCAC	TTTTTTTCCC	8800
8801	ACTCGGGGCT	CCTCGGGCCC	TCCCCAAAT	ATCCTCTGCT	GAACCTCCGA	8850
8851	GTTCTGATC	TTTATAAAAA	TCTACATGTG	CGCCCTGTTG	AGTCTGCAGG	8900
8901	CCACTCAGCG	CACTCTCACT	TACTGTTAGC	GTCCACATTA	TCGCTGCTCT	8950
8951	ATTATTATTT	TTAATATTAG	GAGGAACGGG	AACGGGAGGC	CTGCGCTCTC	9000
9001	CCCTCCTGTC	TGTTTCCAGT	TGCACGGAGC	GACCTCGCTG	TGAACTCTCT	9050
9051	CTCTCTCTCT	CTCTCCCTCT	CTGCGTAAAC	ACACCGAACA	TTTTAACACA	9100
9101	CACACACCAC	ACTGGAGAGA	GAGAGAGAGA	GAGAGAGAGA	GAGAGAGAGG	9150
9151	GCAGTTTGAA	AGAAGAATTC	TGGAGCCTGT	GTGCTGAGAC	TATTTTAGTG	9200
9201	AGATTATATT	TAGACAGTGT	TTCCGTTTGT	CTCGCAGACA	CACGCATCGA	9250
9251	GGAGAGATCC	AACCACTATT	TCAGAAACAG	CACAACCCCA	AATACAAGTC	9300
9301	TGTCAAATAC	AAAAATAAAA	ATGTGCACTT	ACACACTCTC	TGATTCTCTT	9350
9351	CACAGCTAAA	TCTAAGTGTC	TGCCTCTCTG	CTCTGTTAGA	AGCCAATCTG	9400
9401	CTCCATCTCC	ACTACCGCTG	AGTGGAATC	TTCATCATCC	TCTCCTCCTG	9450
9451	AGCTGCTGCG	CCATCTAAAA	CCTTCTCCAA	GTTTCCTGTG	TGGAAGTTGG	9500
9501	GGATTTTCAG	CCTCTGTCAC	ACTTTATAAG	ATTAATGAGC	TGAGTGCAGT	9550
9551	AAACAGTGCT	TTTTTTCTCT	CTCAAGGAGC	CATATTGTAA	ACGGTGCTGG	9600
9601	CTGACTCCGC	TCACCTAATG	TTAACCGTCA	TTACATCACC	CAGAGGAAAA	9650
9651	CGAAAAATG	TCCGAGAGAT	GCTCTATTGT	TTGTGCGTTA	CACCGGGCCG	9700
9701	GGATGCGCCA	ACCCGCCGCC	CACACGGCGC	ACATGGCGAC	CGGGCGCCTT	9750
9751	CCCGTTTTTT	CTCCCCAGAG	CTGACCCCCC	GTTAAGACAA	AACAATGGCA	9800
9801	GCTAGCCTGT	GATACAGCCC	CTCGTTGCCA	CAGCATTGTT	TGGAAAAATA	9850
9851	ATAATGCTGC	TTCTCACAGC	CATCGCCGCC	GCCAGGCCCG	GACCCCCGAT	9900
9901	TGGCTGTTTT	GGGTCACATG	ACCAGGAATT	GGCTGCAATT	TCGCCCCATA	9950
9951	GTTCTTCAGT	TTAGCCTACC	CAGGTCCCCA	TACGCTAGTA	ATATCACCAA	10000
10001	TATACACAAT	TATTATATAG	CCACCGCATT	TACGCCATTG	CGGTGCGGAG	10050
10051	CCTACATGCT	CGTGCGTTTT	TTCTTTTCGT	TTTATTTTTT	TAAGAAGCCA	10100
10101	CATTGTGTTG	TGTGCGTGCG	CGAGCGTGCA	TGCGTGCGTG	TATGTGTTTT	10150
10151	GTTTTTTTTT	TTTTGGGGAA	GGGATTGTTC	GGGGGTCTCG	CCGTTTGAGG	10200
10201	GTGAAGCCGG	TCTGAGGGAG	CTATGAATTT	TGAATTTGAG	AGGGAGATCG	10250
10251	GTTTTATAAA	CAGCCAGCCG	TCCCTAGCAG	AGTGCCTGAC	GTCTCTCCCC	10300
10301	GCCGTCTTGG	AGTCATTTCA	AACTTCATCA	ATCAAGGAGT	CGACAGTAAT	10350
10351	TCCGCCTCCC	TTCGAGCACA	CCATCCCCAG	CCTCAGCCCC	TGCACAGCCA	10400
10401	GCCAGCCGAG	ACCGACTGCT	AGGAGCCAGC	AGAACCGGAG	AACCTCCGCC	10450
10451	ACTAATGGCC	TCCAGACGCA	CCACCGAGGC	GCCCAGCAGC	CCTGTCCGCC	10500
10501	CGGTACAGGC	CCGGCCACGG	CCACGGCAGC	CCCGGCCGGT	CCGCTGGCCC	10550
10551	ACGAGTTCCC	CTGGATGAAG	GAGAAGAAGT	CATCAAAGAA	GTGCCCAAG	10600
10601	CCTGGGAGCA	GCTCCAGCGG	TGGCGGATCC	ATCATCCCCC	CTGCCTCATC	10650
10651	CTCCCCGTCG	CCGACCGCCT	CCGGGTACGC	TTCGGCGGGG	ATCGAGTCGC	10700
10701	CCACAGGTCG	GTATGCTGGG	ATGCGGGAGC	CAGTATGTGT	GTATGAGTGT	10750
10751	GTGTGGTGGG	AGAGTGAGGG	GTGTGCAGCC	TGTGCTTTAA	TTTTCTGTCT	10800
10801	GCATCTTCAT	TTCATTCCCA	CTTCAGAAAT	TACCTAGCAT	TCATTCCACA	10850
10851	CAGCATGGAT	GTGCTTAGAT	GACATTGTCT	TTTTCTTCCT	CGGTGTGTGT	10900
10901	GTGTGTGTGT	GTGTGTGTAT	TGTGCAGTGT	GCTGTGATGC	TGTGTCAACA	10950
10951	GGGAAGCCAT	ATCTAGACCA	CTGCATTATG	CTGCAGTGCT	TTGCTCACTC	11000
11001	TCTCATCATC	ATGCATTTAA	ATATGCACCC	ATTAATATAT	CTGCCATTTT	11050
11051	TAGAGGATCA	GTATGCAGAA	TTCAGCTTAA	CAGTGAAGTC	AGACCAAGTTC	11100

11101	ACACCAGAAT	CCTCTTAGTT	CTGTATCATA	TAGCCCACAT	TTTCTAAATG	11150
11151	CATCTTCTGC	ATATGGAATT	ATTTATCGTG	CGTAAAGACG	CACAGAGGTT	11200
11201	TTTTTTTTGT	CTTTATCCAT	CCTGCAGTTG	ATTGCTGCTG	CAGCCTGTAT	11250
11251	CATTCTGTTC	ACTGGGTGGT	GTGATTTGAT	ATCAGCAGCA	GCATCACCTT	11300
11301	GGTATTATAT	ATTCATAACA	CAACTGAGAT	GGTAGCCATA	TTGGATGGAG	11350
11351	GTAGGAAATA	TAATGGCGAG	GCAAGGCGAC	CGGACAAAAT	ATGTTGGGGT	11400
11401	GGGGGTGGG	AGGGGGGGTA	TAAGCTCCAT	CCACACAGAA	TACAGACTGC	11450
11451	ACACAATCTC	ATTTTTCTTG	GGTTTTGGTC	ATTTCTGTCT	TTTACACTGA	11500
11501	CCTCTATAAC	AATCACTGCA	TTTTTAAAG	TTTTTTTCTC	CCCCTTCTTT	11550
11551	ATTCACACA	CAATCCCACC	TCCACTCATT	TTTAATGGAA	TTCTTGTTTT	11600
11601	CTCGTCTTTC	TTTCCTGCAC	AATTTCTGTC	CTTTATCGCA	TCCATGCAGC	11650
11651	CATTACAGGA	TTTACTTATG	CTTCTGACAG	ATACTGCCAC	ATTGGTTTCA	11700
11701	CATCGACATT	CCCACATTTT	CCTGCACATT	TCCTCCCCAT	GCCCTGATAT	11750
11751	ATTTATTTTC	CTCTCTTTGC	ATTTCTCTGA	TCACAGGCTG	CAGCCTCTCC	11800
11801	CAGCAGGCTC	ACAGAAGAGT	AAAACCGGCT	TTGACATTAT	GATCATCTGT	11850
11851	CTTTATCTGT	GTTTTTATTT	TGGGATTTTT	TTTACCAGCA	CCAAATTGGT	11900
11901	CCGCTCTTAC	AGTTATGGCT	TTGGCTGCTA	CTGTAATTGG	AAGGAGTATT	11950
11951	TTTTATGAGA	CTTTTTGTGT	TTTATTACTG	TTTTATTACA	GACAGAAGAG	12000
12001	CCGAAGGGGA	TTTCTGCTCC	ATGCTGAAGC	AGCTTATTTT	TTGGCTGGAA	12050
12051	ATCTTTGTAC	GCAATAAAAT	TGTAGAGGTT	ATAACAGGGC	TCTCGTGAAG	12100
12101	TAGATTACAA	TTTGAAGTTT	AAGAGAAATA	TATAAAATTA	AAATAAAAAA	12150
12151	TGAGAATGGC	TTCAAAAGAT	TTGCGCGCAA	TAAATCATCA	CAGACATAAT	12200
12201	CTCTCTCTAT	CTCCCTCCTT	TATTTGTCTT	TGTCTCCCCT	GCAGACCCGA	12250
12251	GCCAGGCCAG	TCTGGATGCT	GGAGGAGCCG	GTCCCGCCCG	CTGCGCACCG	12300
12301	CATACACCAA	CACACAGCTG	CTGGAGCTGG	AGAAGGAGTT	CCACTTCAAC	12350
12351	AAGTACCTAT	GCCGGCCCCAG	GCGGGTGGAG	ATCGCCGCCC	TCCTGGATCT	12400
12401	CACCGAGCGG	CAGGTCAAAG	TGTGGTTCCA	GAACCGGCGG	ATGAAACACA	12450
12451	AGCGCCAGAC	TACTCCCCAC	CGGGACGGAG	GTGGTGGAGG	CGGCCATGAA	12500
12501	TCCAGCGACC	CGGGCGGGTT	TGAGCCGCTC	GAAGGCGCAG	ATGCCTCATC	12550
12551	CCCGTACTCA	AGCCAGCCGC	TGGCAACATC	TGGCACCGGG	TGGCACCACAT	12600
12601	CGGAGAGCGA	GCCGGGGAGC	GGCAATCTGT	TTTCGGCCGT	TGCTCCCACC	12650
12651	GCTTACGCCA	ATGACAGCGG	GGACAATGCA	CAGCCCACGC	CGGAGGAGCG	12700
12701	AGGCTGTTTG	AGCCGCCCCG	AACGCCACC	GCAGCCGGGG	CCACTTGGCT	12750
12751	CATCAGACGT	TGTTGCTTCC	CACCACTCCC	CGCTGCGTTC	AGCGCCACGA	12800
12801	ACTCCGCAGA	CAACCCGGGT	TCTGATGCAG	CTCAGCGAGG	CCGGGCCCGC	12850
12851	AGCAGCAGCG	TCCGATCCGT	CATTCTCTGA	GCAGCGGGAC	TCTTCCATCA	12900
12901	CCTCCACTTC	TTCTTCGGCT	CCTCCTGCCG	CTTCTTCATC	CGCGCTCCCC	12950
12951	GACCTGACTT	TCTTTGCCGG	GGACGCCTGC	CTGTCACCCA	GCCTGCAGAG	13000
13001	CTCACTGGAT	AGTCCGGTGG	ATTTCTCCGA	GGAGGACTTC	GACCTGTTCA	13050
13051	CAAGCACACT	GTGCACCATA	GACCTGCAGC	ACCTCAACTT	CTGAGGCCGG	13100
13101	CTGCAACCTT	TAAACATGAA	GGGAAACTAA	AAAGGAGTGA	AGGAGTGGAG	13150
13151	GGGCACATGC	AGGCCTGTGG	ATGAACAATC	GGGCGGAGGA	GACGGCACTC	13200
13201	GGCGAGGACT	CTGTTGTTTA	AAAAGCATAA	AAGCTTCTCG	GCAACATTCC	13250
13251	TGAAACATTT	CTGAATGAAA	ATTATGTTCT	TTTAAAAATA	TATTTTTTTCG	13300
13301	TCTTCTTGCT	GGTTTTACTT	TGCGGCTGTA	ATGTAGTTTA	AGAGAGATGC	13350
13351	AGTTCATTTT	ATTTTTTTTAA	TTTTTAAACAT	CGAAGGGAAA	TTGTTTTTCC	13400
13401	ACGATCGAAA	CGTTGTGGAT	TTTTTCTCTC	TCCCCTTTCC	CAAATTTTCAG	13450
13451	GTATTTATTG	CCTTCTTGTT	TTTTGATCGT	TTTTGTAAA	ACCCTTATCA	13500
13501	GTATGCGTGG	ATTTGTGTTT	TCTAGTCCAT	TCTGGTGCAG	CTTCATGGAT	13550
13551	CCTTAATAAC	TTCTCAGGAA	CTGTTACAGG	AGTCTAGGTG	CAGATAGCCG	13600
13601	GACTGCTCTT	AGAAAATCTG	AACCTGAAGAA	TGAATTTTCT	TCTCTCTCTG	13650
13651	AACGAGACAA	GACGAGAGAG	GAGCCCTGCT	GATGACCCGA	GTTTTAGACC	13700
13701	CTAGTTTATT	TATTGAGATT	CTTTAATAGT	GAAAATCCAA	ATTATTTATT	13750
13751	GTACATATAT	TTTCATAGTG	GAATAATAAT	AATAAAAAAC	ACACAAACAT	13800
13801	TTACGATGTG	GCCCATGTTT	TTTTGTTTGA	TAGCTATATG	AATTTAAGGA	13850
13851	GTGCTGAATA	TGGAATAATA	ATATTTAAAC	AAATGGTATG	CTTTAAATAC	13900

13901	ATTTTTATGC	AAATAACATT	AATAATACAG	TGCAATTTTA	TAATAGCCTA	13950
13951	AATATAAAAC	AAAGTGAGCA	TTTTATTAGT	AGGATAATTT	TGTATAACGA	14000
14001	TTTGGCTGCT	GTTACTTGTA	ATAATGAAAA	TTATTTTATT	TATTTACAG	14050
14051	TAGCCTATAT	ATAATTTTTA	CAATACATTT	ATCTATTGCA	ATTAGCCTAT	14100
14101	AATTTCCAGG	TTTGTTTTGT	CCATAATTTT	AAAGATACAA	TTTTTCATTA	14150
14151	TATAGGATAA	TAAAATATTG	ATATAATTTA	TAATATTTTC	ATCTCCTAAA	14200
14201	ATATATTTTA	ATTCTGGCTC	CTTGTTTCAA	GTTTAAATGC	ATGGCTCCTC	14250
14251	CATCTAGTTT	TCCCACCTTA	TTGAACATAA	ATATGTGGCC	TGATATAGCC	14300
14301	TGGTATGGCC	CAGTGTGGTT	CAGCTCATGT	GTGTTTACAC	AGAGATAACC	14350
14351	CCTGAGAAAT	GACTGAGAAA	GAATGAGATA	TTTTCCAGAG	CAGATAAAAA	14400
14401	GAGGAGAAAC	AGTCTGAGGC	CACAGCAGGA	TCATTGTTTT	CCTCTGTTAG	14450
14451	CTCAGCACTG	GTAGCCCTCT	GCCCCGAGTA	GGAGCTAATG	GTTAGCATAG	14500
14501	CAGCTTTCAC	CATGTGAGAA	AAGCAGAGTA	ACTACAATGT	GGCCTTTTAC	14550
14551	TGAAAGTGTG	TGAGAGCTAA	CCACAGTGTA	GCACCACTCA	CCCTCTGTTT	14600
14601	GTAGTGTGTG	AGCACAGGCG	TTAACTGTTG	CTTGGCACTG	CACATTCCCT	14650
14651	CCCATTATCA	AGCTAGAAGA	GGACTGTGGA	GTAAAGCCTG	AGTTAGCATG	14700
14701	GCTAGCTCCC	TGCCATTAGT	GTGTGTGGAT	ATGTTGCTAA	CTAGCTGCTC	14750
14751	CTCCTCCAC	AGAGTCAAAA	TGGCGCTGAA	GACAAAGCAG	GATGTGAACA	14800
14801	ATGAGCAGTG	GCAGCCATTA	CAGGATGACT	GGCTCACTGT	GACAGGCATG	14850
14851	GCGTGTGTGT	GTTTGTGTGT	GTGTGTGTGT	GTGTGGGAGC	GATAGGGAGA	14900
14901	GAGAGAGAGA	GAGAGTGTGT	GTGCCTGTAA	TAGGCAAAGA	GAAAGAAATT	14950
14951	ACTTGTTTTT	TTAGGCTACA	GATTAAAGAC	AGGTTTATGG	CAGATTGCTT	15000
15001	TAGCCCATAA	TAATACTAAT	ACTAATAATA	CTAATACTAA	TACTAATACT	15050
15051	AATAAAAATA	ATAATAATAA	TAATAATAAT	AATAATTCCT	CTTATGATAA	15100
15101	ATATATTTAG	ACAATATCAA	ATTGTAAAAA	AAAAATATAT	TTTCAAATAT	15150
15151	AATGATTAGG	GCAAAATAGG	AATAATGACA	ACGATCTAAC	AACAATATTT	15200
15201	CATTTATTAT	ATTGCAGTAT	TGTAGTCAAT	ATAATGGCCT	ATTGCCCAAT	15250
15251	GCATGACTAT	AGTGAAACAA	ACTTAAAGTC	TGTAAAAATG	GAAACAATAA	15300
15301	CCCAGTTTAG	AGTCCTAATC	CACAGTTTGA	TAAACCCTAC	TGAGTGTCCA	15350
15351	ACTGCAGCTC	CATCACCTTT	CATAAGAAAA	AACGTTTGCA	GCTCAATCAA	15400
15401	TATTGCATTA	AAGAAAAACA	CATACAAAAC	ATTAAGAGTT	CAAATCCATA	15450
15451	CAGAAGAGGC	TAAAGTCCTG	TTGGTCTGAC	TTTCTTTTTA	ACCACAAACA	15500
15501	CTTCTTTATC	TGGGACTCAC	TTAAACATTA	CATTGCTTAC	TATTTCTATG	15550
15551	AGATCCGTTG	ACCTGCAGGT	CGAC			15574