

Elizabeth Ann Snell. GENOMIC ORGANIZATION OF THE *Hoxa4-a10* REGION FROM *MORONE SAXATILIS* (STRIPED BASS): MACROEVOLUTIONARY CHANGES OF THE *HOXA* CLUSTER IN VERTEBRATES. (Under the direction of Dr. Edmund J. Stellwag) Department of Biology, July 1998.

Studies of vertebrate *Hox* cluster evolution in the cephalochordate amphioxus and select vertebrates have shown an increase in cluster number among vertebrates, punctuated by lineage-specific gene loss events. I have sequenced a 31 kb region of the *HoxA* cluster from the teleost *Morone saxatilis* (striped bass), to provide a detailed physical map of this region, and to examine macroevolutionary changes of vertebrate *Hox* clusters. I identified five linked *Hox* genes: *Hoxa4*, *Hoxa5*, *Hoxa7*, *Hoxa9*, and *Hoxa10*, which are organized similarly to those of other vertebrates. Furthermore, I have documented the absence of the *Hoxa6* and *Hoxa8* genes within the 31 kb contig. Comparison of these results to those published for other vertebrates show that among the genes of the *Hoxa4-a10* region in vertebrates, genes *Hoxa5*, *Hoxa7*, *Hoxa9* and *Hoxa10* are common to all vertebrates. I have also determined that the absence of *Hoxa6* is a common characteristic of teleosts, whereas the absence of *Hoxa8* is common to vertebrates in general, with the possible exception of zebrafish. These results suggest that there are lineage-specific differences in the *HoxA* complement of vertebrates, and particularly among the teleosts. Further comparisons of the genomic sequences obtained in this study to PCR-derived *Hox* gene sequences from *M. saxatilis* and other teleosts provide evidence that PCR-based surveys of conserved gene families such as the homeobox-encoding genes, can be inconclusive. Comparisons of the *Hox* cluster intergenic regions among vertebrates revealed extensive conservation of the non-coding intergenic regions between closely related orders, but an overall absence of sequence

relatedness between phylogenetically disparate species. In addition to the lineage-specific gene alterations and variation within the intergenic regions, I also documented a marked reduction in the length of the *Hoxa4-a10* region between vertebrate lineages representative of teleosts and mammals. Comparative analysis of *HoxA* cluster organization among teleosts and mammals suggests that cluster length reduction and lineage-specific gene loss events are characteristic of *Hox* cluster evolution.

GENOMIC ORGANIZATION OF THE *Hoxa4-a10* REGION FROM *MORONE*
SAXATILIS (STRIPED BASS): MACROEVOLUTIONARY CHANGES OF THE
HOXA GENE CLUSTER AMONG VERTEBRATES.

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
Elizabeth Ann Snell

July 1998

2L

638

M678

S63

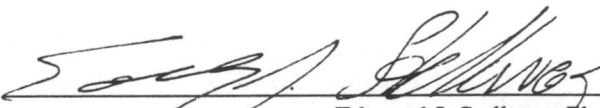
1998

GENOMIC ORGANIZATION OF THE *Hoxa4-a10* REGION FROM *MORONE*
SAXATILIS (STRIPED BASS): MACROEVOLUTIONARY CHANGES OF THE
HOXA GENE CLUSTER AMONG VERTEBRATES.

by
Elizabeth Ann Snell

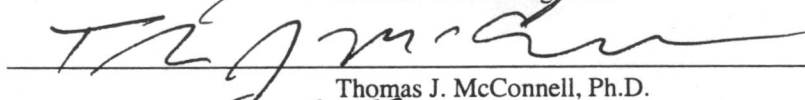
APPROVED BY:

DIRECTOR OF THESIS



Edmund J. Stellwag, Ph.D.

COMMITTEE MEMBER



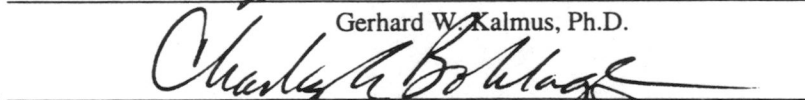
Thomas J. McConnell, Ph.D.

COMMITTEE MEMBER



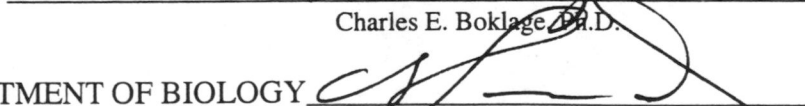
Gerhard W. Kalmus, Ph.D.

COMMITTEE MEMBER



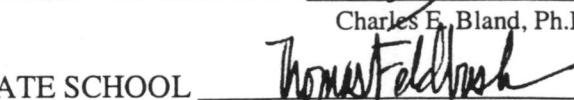
Charles E. Boklage, Ph.D.

CHAIR OF THE DEPARTMENT OF BIOLOGY



Charles E. Bland, Ph.D.

DEAN OF THE GRADUATE SCHOOL



Thomas L. Feldbush, Ph.D.

DEDICATION

This work is dedicated to my parents, Joe and Van, who by their own example of tenacity, integrity, dedication, self-sacrifice, and academic excellence have made it possible for me to achieve my own goals. For that, I am eternally grateful. It is also with their unconditional support, unbelievable patience, and love that I was able to complete this work. I am indebted to them for all they have done and continue to do.

ACKNOWLEDGMENTS

I owe a tremendous debt of gratitude to my mentor and thesis advisor, Dr. Ed Stellwag. I wish to thank Ed for his exceptional guidance, wisdom, and infectious enthusiasm in the lab, but I wish to thank him most of all for the sound advice given and for inspiring me toward greater challenges. The time spent in Ed's lab has been one of the greatest learning experiences (on many different levels), but also one of the most fulfilling experiences I have ever had.

I also am indebted to my "adopted co-advisor", Dr. Jean-Luc Scemama for his knowledge and support throughout my research. In addition, I would be remiss if I failed to mention his wonderful sense of humor, which always kept things in perspective.

I would like to thank my committee members, Dr. Tom McConnell, Dr. Gerhard Kalmus, and Dr. Charles Boklage for their thought-provoking discussions, and for all of their poignant advice on my future endeavors.

I thank my fellow lab-mates, Jeff McCallum (the lab squirrel) for showing me where all things hidden in the lab were, Scott Avery ("Uni, Junior, Boy, MonkeyBoy" – or whatever nickname I felt like calling him on any given day), and Yasmin Mohamoud (fellow proponent of the "Chicks Rule!" ideology). It has been a great pleasure working with you.

Also, I would like to give many thanks to my sister, Julia Carlson, for all of her support and patience, and for being my biggest cheerleader. I thank her for always being at the ready to help me in any way needed, and of course, for being my standing concert

date during my years here. I also wish to thank her family, Don, D.J., and Jack, for all of their support as well.

I owe a very special thank you to my fellow graduate student, lab partner, roommate, and great friend, Robin Gibson-Brown. Robin was always there many a late night to wax philosophical about school and life in general, as well as being there to give me the proverbial kick in the hindquarters in her matter-of-fact manner when I needed it the most. She has been a true and wonderful friend, and her dogged determination is truly inspirational.

TABLE OF CONTENTS

LIST OF FIGURES.....	ix
LIST OF TABLES	xi
LIST OF SYMBOLS AND ABBREVIATIONS	xii
INTRODUCTION	
Homeobox-encoding genes and animal development	1
<i>Hox</i> gene evolution and organization in amphioxus and vertebrates	3
Amphioxus and cluster duplication.....	4
Mammals.....	6
Pufferfish.....	8
Zebrafish.....	9
<i>Hox</i> gene surveys (PCR vs. genomic)	10
Objectives	13
MATERIALS AND METHODS	
Isolation of <i>Morone saxatilis</i> <i>Hox</i> gene-containing λ clones	15
Bacteriophage λ DNA isolation	17
Mapping, Southern blotting, subcloning and sequencing of λ clone inserts	19
Genomic Southern blots	23
DNA sequence analysis	23

RESULTS

Isolation and characterization of <i>HoxA</i> cluster λ clones	27
Confirmation of <i>Hox</i> gene assignment and linkage	32
Comparative analysis of <i>Morone HoxA</i> genes	35
PCR vs. genomic-based <i>Hox</i> gene assignments	39
Intergenic sequence comparisons	40
<i>Hox</i> cluster compression	43

DISCUSSION

The <i>Hoxa4-a10</i> region from <i>Morone saxatilis</i>	45
Comparison of <i>HoxA</i> genes and cluster evolution	48
PCR vs. genomic surveys	58
PCR amplification of non- <i>Hox</i> homeobox-encoding genes	60
Intergenic comparisons	62
<i>Hox</i> cluster compression	66
Conclusions	68

REFERENCES	71
------------------	----

APPENDIX A: COMPARATIVE ALIGNMENT OF THE *Hoxa9-a7* AND

Hoxa10-a9 INTERGENIC REGIONS FROM HUMAN,

FUGU RUBRIPES, AND *MORONE SAXATILIS*.....141

APPENDIX B: SYNTHESIZED OLIGONUCLEOTIDE PRIMER LIST

FOR THE *Hoxa4-a10* REGION.....153

APPENDIX C: DNA SEQUENCE OF THE *Hoxa4-a10* REGION FROM*MORONE SAXATILIS*156

LIST OF FIGURES

1. Schematic diagram of the phylogenetic relationships between amphioxus and selected vertebrates.....	78
2. Schematic diagram of <i>Hox</i> cluster organization in specific vertebrates and amphioxus.....	80
3. Restriction endonuclease maps of λ clones 7H and 8H.....	82
4. Genomic organization of the <i>Hoxa4-a10</i> region in <i>Morone saxatilis</i>	84
5. Unrooted evolutionary tree comparing vertebrate homeodomain sequences.....	86
6. Genomic Southern blots.....	88
7. Restriction endonuclease map for the <i>HoxA</i> 31 kb contig.....	90
8. Comparison of vertebrate <i>HoxA</i> cluster organization and amphioxus.....	92
9. PILEUP alignment of vertebrate <i>Hoxa10</i> exon1 amino acid sequences.....	94
10. PILEUP alignment of vertebrate <i>Hoxa9</i> exon1 amino acid sequences.....	96
11. PILEUP alignment of vertebrate <i>Hoxa7</i> exon1 amino acid sequences.....	98
12. PILEUP alignment of vertebrate <i>Hoxa5</i> exon1 amino acid sequences.....	100
13. PILEUP alignment of vertebrate <i>Hoxa4</i> exon1 amino acid sequences.....	102
14. PILEUP alignment of the amino acid sequences of exon1 from the <i>Morone saxatilis Hoxa4-a10</i> genes.....	104
15. PILEUP alignment of vertebrate <i>Hoxa10</i> exon2 amino acid sequences.....	106
16. PILEUP alignment of vertebrate <i>Hoxa9</i> exon2 amino acid sequences.....	108
17. PILEUP alignment of vertebrate <i>Hoxa7</i> exon2 amino acid sequences.....	110
18. PILEUP alignment of vertebrate <i>Hoxa5</i> exon2 amino acid sequences.....	112

19. PILEUP alignment of vertebrate <i>Hoxa4</i> exon2 amino acid sequences.....	114
20. PILEUP alignment of the exon2 amino acid sequences from the <i>Morone saxatilis Hoxa4-a10</i> genes.....	116
21. Pair-wise similarity plot comparing the <i>Hoxa9-a7</i> intergenic regions from human and <i>Morone saxatilis</i>	118
22. Pair-wise similarity plot comparing the <i>Hoxa9-a7</i> intergenic regions from human and <i>Fugu rubripes</i>	120
23. Pair-wise similarity plot comparing the <i>Hoxa10-a9</i> intergenic regions from human and <i>Morone saxatilis</i>	122
24. Pair-wise similarity plot comparing the <i>Hoxa10-a9</i> intergenic regions from human and <i>Fugu rubripes</i>	124
25. Pair-wise similarity plot comparing the <i>Hoxa9-a7</i> intergenic regions from <i>Morone saxatilis</i> and <i>Fugu rubripes</i>	126
26. Pair-wise similarity plot comparing the <i>Hoxa10-a9</i> intergenic regions from <i>Morone saxatilis</i> and <i>Fugu rubripes</i>	128
27. Comparison of intergenic distances among vertebrate <i>HoxA</i> clusters and amphioxus.....	130
28. Amino acid sequence alignment of <i>Hoxa4</i> homeodomain sequences and the zebrafish <i>Hoxx4</i> and twelve 3'- encoded diagnostic amino acid residues.....	132
29. Amino acid sequence alignment of <i>Hoxa9</i> homeodomain sequences and the zebrafish <i>Hoxx9</i> and twelve 3'- encoded diagnostic amino acid residues.....	134
30. Amino acid sequence alignment of homeodomain sequences of paralogous groups 5 and 6.....	136

LIST OF TABLES

1. Predicted and observed fragment sizes from genomic Southern blot.....	137
2. Location of gene exons and introns for the five <i>Hox</i> genes within the 31 kb contig.....	138
3. PCR vs. genomic gene assignments.....	139
4. Intergenic distances for genes 4-10 from amphioxus, and the <i>HoxA</i> clusters from human, <i>Morone saxatilis</i> and <i>Fugu rubripes</i>	140

LIST OF SYMBOLS AND ABBREVIATIONS

The following symbols and abbreviations are used throughout this thesis.

a.a. – amino acid

Abd-B – *Drosophila Abdominal B* gene

ANT-C – *Antennapedia* gene complex

Antp – *Drosophila Antennapedia* gene

bp – base pair

BLAST – Basic Local Alignment Tool

BX-C – *Bithorax* gene complex

C – degrees Celsius

DNA – deoxyribonucleic acid

G – force of gravity

g – gram

HOM-C – Homeotic complex

kb – kilobase (1000 base pairs)

L – Liter

M – Molar (moles per Liter)

mg – 10^{-3} gram

mL – 10^{-3} Liter

mM – 10^{-3} Molar

mm – 10^{-3} meter

MOI – Multiplicity of Infection

N – Normal

NCBI – National Center for Biotechnology Information

nM – 10^{-9} Molar

nt – nucleotide

PCI – phenol/chloroform/isoamyl alcohol

PCR – Polymerase Chain Reaction

p.f.u. – plaque forming units

RNA – ribonucleic acid

rpm – rotations per minute

UTR – untranslated region

μg – 10^{-6} gram

μL – 10^{-6} Liter

w/v – weight per volume

INTRODUCTION

Homeobox-encoding genes and animal development

Studies of *Drosophila* genetics at the turn of the century revealed specific mutations that caused the induction of major changes in body plans. The resulting phenotypes of these mutants demonstrated partial or complete segmental transformations, the incorrect establishment of positional axes in *Drosophila* larvae, or the ectopic expression of one body structure in an incorrect region of the body. One example of this type of mutation, called a homeotic mutation, was described by Calvin Bridges in 1915. Bridges described a mutant, which he called *bithorax*, that resulted in the transformation of the flight-balancing halteres of the third thoracic segment of a fly into a second pair of wings. Therefore, the *bithorax* mutant had two pairs of wings instead of one, and the third thoracic segment was transformed into an additional second thoracic segment. Another example of a homeotic mutation in *Drosophila* was the *antennapedia* mutant, in which the resulting phenotype was the transformation of the antennae on the head of a fly into legs. These homeotic mutations were later mapped to two distinct clusters, called the *Bithorax* (BX-C) and *Antennapedia* (ANT-C) complexes, on chromosome 3 in *Drosophila melanogaster* (Lewis, 1978; Kaufman *et al.*, 1990). Together, these two clusters comprise the Homeotic complex (HOM-C) of *Drosophila* (Akam, 1989).

Molecular analysis of the genes within Homeotic cluster revealed the presence of a conserved 180 base pair (bp) region, called a “homeobox”, within each of the genes. The homeobox encodes a 60 amino acid (a.a.) helix-turn-helix DNA binding motif,

termed a “homeodomain” (Scott and Weiner, 1984; McGinnis *et al.*, 1984a). The proteins encoded by the Homeotic cluster genes are localized in the nucleus and function as transcriptional activators of other developmental regulatory genes that are important during embryogenesis. These proteins function as transcriptional regulators by a mechanism in which the third helix of the homeodomain associates with the major groove of DNA, resulting in the regulation of the target gene (Laughon, 1991; Shen *et al.*, 1996).

Cross-species hybridization studies showed that portions of the 180 bp homeobox could hybridize to DNA from other invertebrates as well as vertebrates (McGinnis *et al.*, 1984b). The presence of homeobox-encoding genes in diverse metazoan species suggested that these genes had been conserved throughout metazoan evolution. Based on the conservation of these genes, it was hypothesized that the genes were part of an ancient developmental regulatory system that is crucial in all metazoan species (McGinnis *et al.*, 1984b).

Homeobox-encoding genes may be grouped into different classes or families based on sequence identities both within the homeodomain and the regions flanking the homeodomain. The largest group of homeobox-encoding genes is the *Antp*-class, but there are other classes such as the *POU*, *even-skipped*, and *engrailed* (Scott *et al.*, 1989; Krumlauf, 1992). A distinguishing feature of the *Antp*-class is that they are organized into clusters comprised of up to eleven genes uninterrupted by other genes (Acampora *et al.*, 1989; Akam, 1989; Graham *et al.*, 1989; Wang, *et al.*, 1993; Garcia-Fernàndez, 1993). These clustered *Antp*-class homeobox-encoding genes constitute the *HOM-C*

genes in invertebrates and the orthologous *Hox* genes in vertebrates. Several unprecedented and conserved features characterize *HOM-C/Hox* genes among metazoans. Firstly, the genes are expressed in a colinear manner such that their expression domain within the developing embryo reflects their position within the cluster. Therefore, the *HOM-C/Hox* genes located at the 3' end of the cluster are expressed in the most anterior region of the developing embryo, while the 5' genes are expressed within the posterior of the developing embryo. Additionally, the genes are expressed temporally according to their position such that the genes at the 3' end of the cluster are expressed earliest in development and the genes located at the 5' end of the cluster are expressed later (Duboule and Morata, 1994; Graham *et al.*, 1989). The position of each gene within the cluster is conserved among all metazoan clusters characterized, suggesting that the position of each gene, relative to the others within the cluster, is crucial for proper spatio-temporal expression throughout development. Lastly, the genes within the cluster are arranged such that the direction of transcription is oriented 5' to 3'.

***Hox* gene evolution and organization in amphioxus and vertebrates**

Hox genes have been identified in many invertebrate and vertebrate species (Fjose *et al.*, 1988; Pendleton *et al.*, 1993; Wang, 1993; Ruddle *et al.*, 1994a; Pavell and Stellwag, 1994). In this section, I will briefly discuss *Hox* gene organization in the cephalochordate amphioxus, and the vertebrates for which complete physical linkage maps of the *Hox* gene clusters have been established. The phylogenetic relationships of these organisms are diagrammed in Figure 1. Included in this section is a brief discussion

on the *Hox* cluster organization from the zebrafish as deduced from the linkage of *Hox*-specific polymerase chain reaction (PCR) amplification products to somatic cell hybrids. *Hox* gene nomenclature used throughout the discussion and the remainder of this thesis is according to the standard *Hox* gene nomenclature as described by Scott (1992).

Amphioxus and cluster duplication. Amphioxus (*Branchiostoma floridae*), also called the sea lancelet, is a member of the subphylum Cephalochordata, a sister group to the subphylum Vertebrata. It is thought to be an extant representative of the vertebrate progenitor due to the common characteristics of a dorsal hollow nerve cord and axial notochord. Therefore, its key phylogenetic position between the invertebrates and vertebrates makes amphioxus ideal for providing a basal reference point in the study of gene evolution in vertebrates.

The ten linked *Hox* genes (*AmphiHox*-1 through *AmphiHox*-10) from amphioxus are oriented 3' to 5' within a single cluster that spans a region of 270 kb (Garcia-Fernàndez and Holland, 1993) (Figure 2). The single *Hox* cluster, similar to the single *HOM-C* cluster among invertebrates, is thought to represent the archetypal vertebrate *Hox* cluster (Holland and Garcia-Fernàndez, 1996).

Based on molecular phylogenetic analyses of the homeobox and flanking regions, it is believed that the single cluster found in invertebrates and amphioxus arose from a smaller cluster comprised of three linked *Hox* genes that existed in metazoans before the divergence of the diploblasts from the triploblasts (Schubert *et al.*, 1993). It has been hypothesized that these genes of the original three-gene cluster underwent multiple tandem intra-cluster gene duplication events, resulting in a single cluster comprised of up

to 13 linked cognate genes (Holland, 1992; Schubert *et al.*, 1993; Holland *et al.*, 1994; Bailey *et al.*, 1997). This single cluster of 13 genes is believed to have undergone successive rounds of cluster duplication through polyploidization or genome duplication events followed by secondary gene loss within the clusters to generate the multi-cluster state characteristic of vertebrates (Ruddle *et al.*, 1994b). The genes within each cluster are designated as one of thirteen possible paralogous groups based on sequence homology to the genes of the HOM-C, and location within the cluster. Under this proposed mechanism in which cluster duplication occurred in two stages, the cognate genes of a particular cluster are more closely related to the corresponding cognate genes of different clusters than to the cognate genes within the same cluster. This is supported by the sequence similarity within the cognates of a paralogous group than among cluster cognates.

The two-stage cluster duplication mechanism beginning with a single ancestral three-gene cluster is further supported by the fact that genes from various paralogous groups can be divided into three distinct classes. The three classes are comprised of the 3' genes (paralogous groups 1-3), the central class genes (paralogous groups 4-8), and the 5' (*Abd-B*-like) genes (paralogous groups 9-13). However, once the proposed single cluster of 13 cognates similar to amphioxus was established, it is unclear how many successive rounds of intercluster duplications occurred in the vertebrate lineage. The *Hox* genes from most vertebrates, with the possible exception of zebrafish, are located in four clusters, suggesting that the four vertebrate clusters were generated either by two rounds of complete genomic duplication: $1 \rightarrow 2 \rightarrow 4$ clusters, or by an arithmetic method of

three rounds of single cluster duplication: $1 \rightarrow 2 \rightarrow 3 \rightarrow 4$ clusters. Based on maximum-parsimony and minimal-evolution tests using the homeodomain sequences and fibrillar-type collagen sequences, Bailey *et al.* (1997) recently suggested that the three-step scenario, placing the *HoxD* cluster as the root or original cluster, is the most likely cluster amplification mechanism. However, the hypotheses that would support the three-step amplification mechanism are more problematic. Since complete genome duplications would result in the duplication of entire clusters, the resulting number of clusters would be 2^3 , or 8 clusters. Therefore, this suggests the loss of four clusters had to have occurred at some point in vertebrate evolution. An alternative hypothesis is that cluster duplication occurred through successive single chromosome duplication events instead of complete genome duplication. It is also unclear when these duplication events occurred after the divergence of vertebrates from cephalochordates. These competing hypotheses should be resolved by analysis of the *Hox* cluster organization of vertebrates that are phylogenetically located between amphioxus and vertebrates, such as lampreys and cartilaginous fish, which might have more than one, but less than four *Hox* gene clusters.

Mammals. The 39 *Hox* genes from mice (Duboule *et al.*, 1989; Graham *et al.*, 1989; Zeltser *et al.*, 1996) and 38 genes from humans (Acampora *et al.*, 1989) are organized into four clusters, named A-D, located on four separate chromosomes (Figure 2). Each cluster is comprised of nine to eleven linked genes. Therefore, each vertebrate *Hox* gene has a specific alphanumeric designation based on cluster association and paralogous group assignment (Scott, 1992). Additionally, the genes of each cluster are expressed colinearly and oriented 3' to 5' relative to the ends of each of the four clusters.

Whereas it has been hypothesized that the four clusters in vertebrates arose via interchromosomal duplication events beginning with a single ancestral cluster comprised of 13 linked genes (Schughart *et al.*, 1989; Krumlauf, 1992; Holland *et al.*, 1994; Holland and Garcia-Fernández, 1996; Bailey *et al.*, 1997; Nadeau and Sankoff, 1997), the gene content of each of the four clusters from mammals is different, indicating specific gene loss has occurred within each cluster (Acampora *et al.*, 1989; Figure 2). For example, the *HoxA* cluster in mammals is missing the genes that correspond to paralogous groups *Hox*-8 and *Hox*-12, while the *HoxB* cluster is missing the genes that correspond to paralogous groups 10-12. The cognates for paralogous groups 1-3 and 7 are missing in the mammalian *HoxC* cluster and the *HoxD* cluster is missing the *Hox* cognates that correspond to paralogous groups 2 and 5-7. This suggests that after cluster duplication occurred, there was rapid secondary gene loss within each of the clusters (Holland *et al.*, 1994; Ruddle *et al.*, 1994b). This secondary gene loss is reflected in the *Hox* gene clusters that have been identified among other vertebrates, such as the pufferfish (Aparicio *et al.*, 1997) and zebrafish (Sordino *et al.*, 1996; Prince *et al.*, 1998).

As in *Drosophila*, mutational studies of *Hox* genes in mammals resulted in homeotic mutations within the axial skeletal structures of the developing embryo (Small and Potter, 1993; Condie and Capecchi, 1993, 1994; Rancourt *et al.*, 1995; Chen and Capecchi, 1997). Site-directed mutagenesis studies have shown that heterozygous mutants for *Hox* genes, particularly those of the central class, result in transformations of posterior ribs or vertebrae to those with a more anterior identity. Homozygous or double mutants within a paralogous group effect the phenotype in a dose-dependent manner,

resulting in the transformation of posterior axial structures to anterior structures, fused ribs or vertebrae, or in some extreme cases missing ribs, vertebrae or other axial structures. This demonstrates that, like the *HOM-C* genes in *Drosophila*, *Hox* genes play a role in defining positional axes in the developing vertebrate embryo as well.

Pufferfish. The genomic organization of the *Hox* genes in the Japanese pufferfish, *Fugu rubripes* (order Tetraodontiformes) was recently characterized by Aparicio *et al.* (1997) through a genomic survey in which only the homeodomain for each gene was sequenced. Based on the physical linkage maps constructed, the gene content of the four *Fugu Hox* clusters was determined to be different from the corresponding clusters in mammals (Figure 2). Most notably is the unprecedented absence of all members of paralogous group 7. This loss is of considerable interest since the origin for this gene can be traced to prevertebrate lineages in amphioxus (*AmphiHox-7*). The loss of an entire paralogous group suggests a systematic deletion of all *Hox-7* genes. Moreover, it suggests that the target genes of these *Hox-7* cognates are not required or are activated by different transcription factors during the development of the pufferfish. Given the conservation among clusters in which there is at least one representative of every paralogous group in other metazoans (Figure 2), it suggests that at least one representative is necessary for normal development. Furthermore, since the evolutionary origin of the gene is documented in a cephalochordate, and its presence is documented in other vertebrates, this discovery remains both intriguing and puzzling.

Relative to the corresponding clusters in mammals, the *Fugu HoxA* cluster is missing the *Hox-6* cognate and the *Hox-13* cognate from the *HoxB* cluster, and the 1,3, 4,

8, and 12 *Hox* cognates from the *HoxD* cluster. Additionally, three genes were identified that have no corresponding genes in the mammalian *Hox* clusters: two pseudogenes were reported in the *Fugu HoxC* cluster, a *Hoxc1* and *Hoxc3*, while a *Hox-2* cognate gene was reported from the *HoxD* cluster. These results demonstrate that gene retention and loss throughout vertebrate evolution might be lineage-specific. The number of missing *Hox* genes in pufferfish relative to mammals and the presence of the *Hoxc1* and *Hoxc3*, and *Hoxd2* genes, which are not present in mammals, suggest that the number and organization of *Hox* genes might reflect the dramatic morphological differences between these two lineages. Moreover, it demonstrates conservation of *Hox* gene organization within the clusters regardless of differences in gene content. This conservation of organization supports the idea that gene organization within the clusters is critical for proper spatio-temporal expression.

Zebrafish. Recently, Prince *et al.* (1998) deduced the organization of the *Hox* gene clusters from the zebrafish, *Danio rerio* (order Cypriniformes). The zebrafish *Hox* genes were isolated using a 3' RACE PCR methodology. For the sequences obtained, assignment to a particular paralogous group was based on sequence similarity, and assignment to a particular cluster was based on linkage to a cluster-specific zebrafish/mouse somatic cell hybrid (Ekker *et al.*, 1996). Similar to *Fugu*, the zebrafish is missing genes relative to the mammalian *Hox* clusters (Figure 2). Relative to mammals, the zebrafish *HoxA* cluster is missing the *Hox* cognates for paralogous groups 3, 4, and 6, while the *HoxB* cluster is missing the *Hox-13* cognate. The *HoxC* cluster is missing the *Hox-12* and -13 cognates, and the *HoxD* cluster is missing *Hox* cognates 1

and 8 relative to mammals. It should be noted that among the genes of the *HoxA* cluster, Prince *et al.* (1998) failed to identify a *Hox-9* cognate, however the *Hoxa9* cognate had been previously characterized by Sordino *et al.*, (1996). In addition, there are three genes identified in zebrafish that have no mammalian counterparts: *Hoxa8*, *Hoxb10*, and *Hoxc4*. There are also three other genes identified *Hoxx4*, *Hoxx9*, and *Hoxy6* that have no mammalian counterparts and are located outside the four vertebrate *HoxA-D* clusters. The presence of these genes outside the *HoxA-D* vertebrate clusters suggests the presence of an additional one to two *Hox* gene clusters, the *HoxX* and *HoxY* clusters, in zebrafish relative to other vertebrates. Therefore, Prince *et al.* (1998) hypothesized that there have been an additional *Hox* cluster duplication events in vertebrates, which are represented in these fish.

In summary, the variable *Hox* gene complement of vertebrates suggests that the evolutionary divergence of vertebrate lineages may be reflected in changes associated with *Hox* gene content. To date, the physical organization of the *Hox* genes within the clusters has been characterized in only a few species. Physical mapping of *Hox* gene clusters from additional vertebrate taxa are necessary for determining whether the *Hox* genes will provide useful markers for discerning the evolutionary relationships among vertebrates and deducing the likely pathway for *Hox* cluster evolution.

***Hox* gene surveys (PCR vs. genomic)**

Previously, PCR surveys of *Hox* genes were conducted in an attempt to determine the *Hox* gene content and cluster number in teleosts, lamprey, amphioxus, and striped

bass (Misof and Wagner, 1996; Misof *et al.*, 1996; Pendleton *et al.*, 1993; Pavell and Stellwag, 1994). The PCR-based method has the advantage of being a relatively quick and easy way to isolate the entire *Hox* complement from a particular organism. These surveys utilized short PCR amplification products (69-82 bp: 23-27 a.a.) corresponding to the conserved, central region of the homeodomain. Gene assignment was based on DNA and amino acid sequence similarity to other *Hox* genes characterized from divergent organisms. However, it has been shown that results obtained from *Hox* gene-specific PCR-based surveys can be inconclusive. One such example of inconclusive results obtained from a *Hox* gene PCR-based survey has been shown in amphioxus. Two different laboratories addressed the number of *Hox* genes and clusters in amphioxus by the use of degenerate PCR, however the conclusions reached differed. Pendleton *et al.* (1993) included amphioxus in a broad phylogenetic survey of *Hox* genes in which they concluded that there were two *Hox* clusters in amphioxus. Holland *et al.* (1994) conducted a similar search, but their results suggested that there was only one cluster. The discrepancy between the results from these two labs was later resolved with the characterization of the single cluster in amphioxus by an exhaustive genomic screening and sequencing of homeodomains. In addition, two genes were entirely sequenced to assure correct gene assignment, *AmphiHox-2* and *AmphiHox-6*, since they were difficult to assign from the short PCR-derived sequences (Garcia-Fernàndez and Holland, 1994).

In teleosts, PCR-based surveys have been used to determine *Hox* gene content and possible cluster numbers in the zebrafish, *Danio rerio* (Misof *et al.*, 1996), the killifish, *Fundulus heteroclitus* (Misof and Wagner, 1996), and the striped bass, *Morone saxatilis*

(Pavell and Stellwag, 1994). From these studies, Pavell and Stellwag determined there were at least three clusters present in teleosts, while it was determined that at least four *Hox* gene clusters were present in killifish and zebrafish. However, cluster content and gene number could not be determined because PCR-amplification products among the genes of the central class could not be assigned with certainty for the zebrafish or killifish. For these PCR amplification products, lack of diagnostic residues within the N-terminal region and the conserved nature of the homeodomain can make gene assignment ambiguous and inconclusive, particularly among the highly conserved genes of the central class. Another problem is the difficulty in distinguishing genes (loci) from sequence polymorphic differences of the same gene (alleles) for these short sequences (Holland and Garcia-Fernàndez, 1996), thereby potentially resembling additional genes, or perhaps clusters.

In contrast to the ambiguities and inconclusive results from the PCR-based method, physical mapping studies provide direct physical gene linkage data and additional DNA sequence data. Examples of these types of additional DNA sequences that help with gene assignment are exon1 and the N-terminal region of exon2 which are conserved within a particular paralogous group or with the corresponding region from other organisms. Furthermore, a genome-based survey provides physical linkage data between particular cognates, which provides physical evidence for assignment of genes to a specific paralogous group and cluster. Most importantly, genomic sequence data provide physical evidence that a particular cognate is present or absent within the genome.

Given the inherent problems with PCR-based studies for conserved gene families, a complete physical map of the *Hox* clusters from *Morone saxatilis* (order Perciformes) is being constructed in this laboratory. The genomic survey that forms the basis for this thesis research will provide gene linkage data and comprehensive DNA sequence information regarding the *Hoxa4-a10* region of the *M. saxatilis* genome, which will enable a more detailed examination of *Hox* gene evolution in vertebrates. This will be particularly helpful in examining lineage-specific events that occurred after the divergence of the actinopterygians (ray finned fish) from the sarcopterygians (lobe finned fish and mammals). Therefore, since only partial data exists for the genes and intergenic regions from *Fugu* and zebrafish *Hox* clusters, the information from *Morone* will complement and extend the ongoing research being conducted in these fish.

Objectives

The objective of this research was to characterize and sequence the entire region of the *HoxA* cluster that includes the *Hoxa4-a10* region from *M. saxatilis*. This was done to establish physical linkage of the genes, generate complete sequence data of this region for comparison of coding and non-coding regions to other vertebrates, and to examine macroevolutionary events that may have occurred within the *HoxA* cluster throughout vertebrate evolution. In order to accomplish this, I have isolated and sequenced five overlapping λ genomic clones that together comprise a 31 kb contig, which corresponds to the *Hoxa4-a10* region. Comparisons of the *Morone Hoxa4-a10* region with corresponding regions from other vertebrates, such as mammals and other teleosts,

suggest there are lineage-specific differences in the *HoxA* gene complement of vertebrates. In addition, the genomic data from the *HoxA* cluster genes were used in comparisons to the sequences generated from the previous PCR-based survey (Pavell and Stellwag, 1994) to evaluate the original PCR-based gene assignments. From these comparisons, I have determined that cluster and gene assignments of PCR-derived *Hox* sequences from PCR-based surveys are largely unreliable and inconclusive. In order to examine conservation of cluster sequences among non-coding regions, I have conducted comparisons of the intergenic regions from *Morone* to mammals and teleosts which document the divergence of the sequences between teleosts and mammals, but not between teleost representatives of closely related orders. Furthermore, my results provide the first documentation of the extensive reduction in the overall length of the *Hoxa4-a10* region between teleosts and mammals, and an even greater degree of cluster length reduction between amphioxus and vertebrates. These comparisons suggest that cluster length reduction and duplication may have played important roles in vertebrate *Hox* cluster evolution.

MATERIALS AND METHODS

Isolation of *Morone saxatilis* Hox gene-containing λ clones

A genomic λ EMBL-3 library was constructed from the DNA of a single, two-year-old specimen of *Morone saxatilis* (McCallum, 1997). A total of 1.0×10^6 recombinant DNA clones (the equivalent of 3 times the total genomic DNA content of *M. saxatilis*) were screened independently using four different ^{32}P -labeled DNA probes as follows: a genomic library phage stock was titered by performing 10-fold serial dilutions in SM buffer (0.1 M NaCl, 0.008 M $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$, 50 nM Tris•Cl pH 7.5, 0.01% gelatin). Afterward, 200 μL of *E. coli* XL-1 blue cells (grown to a density of 2.0×10^8 cells/mL in Luria Broth supplemented with 0.2% maltose and 10 mM MgSO_4) were infected with 100 μL of phage dilution preps for 30 min at 37° C, and plated in soft agar (0.7% agarose) on the surface of Luria agar plates. Phage plaques were counted to determine titer, the library was plated as before, and infected with the appropriate amount of phage stock to obtain 5.0×10^4 plaque forming units (p.f.u.) per 150 x 15 mm petri culture plates (approximately 15-20 plates) and grown overnight. The following day, the plates were put into the refrigerator at 4° C to harden the top agar for about 4 hours, allowing for greater ease of plaque lifting. To lift the phage particles, 137 mm circular Duralon UV™ nitrocellulose membranes (Stratagene, LaJolla, CA 92037) were placed on the surface of the plates for 2 min, and scored with orientation marks. The nitrocellulose filter was then placed phage side up on blotting paper soaked in 1.5 M NaCl/0.5 N NaOH for 2 min to disrupt the phage particles and denature the DNA. The

filter was then moved to another blotting paper soaked in 1.5 M NaCl/0.5 M Tris (pH 7.8) for 2 min to neutralize. Finally, the filters were moved to another sheet soaked in 1.5 M NaCl/0.5 M Tris (pH 7.8) for 2 min. After neutralization, the filters were air-dried and the DNA was bound to the filter by crosslinking with the Stratagene UV Stratalinker™ (Autocrosslink setting) (Stratagene, LaJolla, CA 92037). The filters were then prehybridized in pre-hybridization buffer overnight and probed with one of the following ³²P-labeled DNA probes as follows:

Probe (1): a heterogeneous PCR amplification product from genomic *M. saxatilis* DNA using homeodomain-specific primers *Hox48-71* and *Hox141-163* (Pavell and Stellwag, 1994); hybridized in 6X SSC at 55° C; washed three times in 2X SSC/0.1% SDS, once in 0.5X SSC/0.1% SDS, once in 0.1X SSC/0.1% SDS at 65° C.

Probe (2): a heterogeneous PCR amplification product from isolated λ clones containing *M. saxatilis Hoxa5*, *Hoxa7*, *Hoxb2*, and *Hoxb3* using *Hox*-specific primers *Hox48-71* and *Hox141-163*; hybridized in 2X SSC at 60° C; washed four times in 2X SSC/0.1% SDS at 65° C.

Probe (3): a 300 bp *Hoxa9* intron-specific *SalI*-*AvaII* digestion product from λ clone 8H; hybridized in 2X SSC at 65° C; washed four times in 2X SSC/0.1% SDS at 65° C.

Probe (4): a *Hoxa5*-specific PCR amplification product from isolated λ clones containing *M. saxatilis Hoxa5* using *Morone Hoxa5* homeodomain-specific

primers; hybridized in 6X SSC at 55° C; washed three times in 2X SSC/0.1%

SDS, once in 0.5X SSC/0.1% SDS, once in 0.1X SSC/0.1% SDS at 65° C.

The filters were placed on X-ray film (Kodak X-OmatAR, Eastman Kodak Company, Rochester, NY 14650) and stored at -70° C for 5 days on Kodak BioMax™ intensifying screens. After exposure, the X-ray film was developed for 2.0 minutes in Kodak developer, rinsed for 1.0 minute in water, and fixed for 5.0 minutes in Kodak fixer, rinsed for 2.0 minutes in water, and allowed to air dry. Once the X-ray films had been developed, the plates were oriented to the films and strongly hybridizing plaques were picked from the agar by removing an agar plug with the wide-bore end of a Pasture pipette, and placed in 1.0 mL of SM buffer with 20 µL of chloroform. The isolated phage from the primary screen was then subjected to plating, lifting, and probing (each time at reduced plaque densities), three more times, in which strongly hybridizing plaques were removed from the agar using the small-bore end of a Pasture pipette, until single positive clones were isolated. For the dual screening using *Probes* #2 and #3, phage particles on the plates were adsorbed by two filters that were oriented the same way to determine which clones were both *Hox*- and *Hoxa9* intron-specific.

Bacteriophage λ DNA isolation

Bacteriophage λ DNA for each of the six clones was isolated using a broth lysate methodology that was adapted from Yamamoto *et al.* (1970). A culture of *E.coli* cells (P2393 or XL-1 blue) was grown in Luria broth that had been supplemented with 0.2% maltose and 10mM MgSO₄ to a density of 2.0×10^8 cells/mL, and then infected with the

appropriate amount of titrated phage stock to generate a multiplicity of infection (MOI) of 0.01. The culture density of the lysate was monitored using the Klett-Summerson photoelectric colorimeter (Model 800-3, Klett MFG. Co., Inc., New York) with a #66 red filter until maximum lysis occurred, which was marked by a profound decrease of culture density. Once lysis had occurred, 1.0 mL of chloroform/500 mL of culture was added along with sodium chloride to a final concentration of 1.0 M, and was shaken for an additional 10 min. Cellular debris and unlysed bacterial cells were pelleted by centrifugation for 10 minutes at 4° C at 11,000 G (Sorvall GSA rotor @ 7,000 rpm). The supernatant was collected and polyethylene glycol (PEG) 8000 was added to a final concentration of 10% w/v and swirled until dissolved. The solution was then placed on ice for a minimum of 1 hour, or refrigerated at 4° C for up to 16 hours to facilitate phage λ precipitation. Phage particles were collected by centrifugation at 11,000 G for 15 minutes at 4° C (Sorvall GSA rotor @ 7,000 rpm). The supernatant was discarded and the phage particles were resuspended in 9 mL of SM phage buffer without gelatin and extracted twice with equal amounts of chloroform to remove any residual PEG (3,000 G for 15 min at 4° C; Sorvall HB-4 rotor @ 7,000 rpm). Phage particles in the aqueous layer were collected and isolated by layering onto a CsCl block gradient which was constructed using 1.0 mL 1.7 g/mL CsCl/SM without gelatin and 2.0 mL of 1.4 g/mL CsCl/SM without gelatin. Samples were centrifuged at 24,000 rpm at 4° C for 3 hours (SW-41 Ti rotor) in a Beckman L8-80M Ultracentrifuge (Beckman Instruments, Palo Alto, CA 94340). Phage particles were collected from the gradient interface using a 21 gauge needle and a 1 cc syringe in which a total volume of approximately 250 μ L was

retrieved and placed in a 1.5 mL Eppendorf Micro Centrifuge tube. The CsCl phage solution was treated with a 0.1 volume of a 2.0 M Tris•Cl pH 8.5, 0.2 M EDTA solution. The DNA was extracted by adding an equal volume of formamide and incubating for 30 min to denature the phage particles, then the DNA was precipitated by adding two volumes of 100% ethanol and the solution was placed at -20°C for 1-16 hours. The DNA/ethanol solution was centrifuged for at least 15 minutes at 5,000 rpm using an Eppendorf Micro Centrifuge (model 5415 C, Brinkman Instruments, Inc., Westbury, NY 11590). The supernatant was discarded and the DNA pellet was washed twice with a 70% ethanol solution, allowed to air dry, and suspended in Tris low EDTA (10 mM Tris•Cl, 0.1 mM EDTA pH 8.0). To remove any RNA from the prep, RNase A was added to a final concentration of 10 $\mu\text{g/mL}$, incubated for at least 30 minutes at 37°C , and then extracted twice with one volume of phenol/chloroform/isoamyl alcohol (PCI) (24:50:2) to remove the RNase. The aqueous layer was collected and the DNA precipitated by adding NaCl to a total volume (including the ethanol) of 0.3 M, and then two volumes of 100 % ethanol. The solution was stored at -20°C , centrifuged, washed with 70% ethanol and suspended in Tris low EDTA as above to a concentration of approximately 0.5-1.5 $\mu\text{g}/\mu\text{L}$.

Mapping, Southern blotting, subcloning and DNA sequencing of λ clone inserts

Selected clones were restriction endonuclease mapped and Southern blotted as follows: briefly, approximately 1.0 μg of isolated phage λ DNA per reaction was

restriction endonuclease digested using enzymes *Bam*HI, *Eco*RI, *Hind*III, and *Sal*I in single and double digests *Bam*HI x *Eco*RI, *Bam*HI x *Hind*III, *Bam*HI x *Sal*I, *Eco*RI x *Hind*III, *Eco*RI x *Sal*I, and *Hind*III x *Sal*I (10 units of each). The digest products were electrophoresed on a 100 mL 0.7% agarose gel containing 0.05 mg/mL ethidium bromide for 16 hours at 30 volts. The gel was then incubated in 200 mL of 0.25 N HCl for thirty minutes, with the acid changed once after fifteen minutes. The gel was then rinsed in deionized H₂O and incubated in 200 mL of 0.5 M NaOH/1.5 M NaCl thirty minutes, changing the buffer after fifteen minutes. The gel was again rinsed in deionized H₂O and neutralized in 0.5 M Tris•Cl pH 7.0/1.5 M NaCl, changing the buffer once after fifteen minutes. The DNA was then transferred to a Maximum Strength Nytran® Nylon membrane using a Schleicher & Schuell Turboblottter™ Rapid Downward Transfer System (Keene, NH 03431) according to manufacturer's instructions, using 20X SSC (3.0 M NaCl, 0.3 M sodium citrate, pH 7.0) transfer buffer. After blotting overnight, the DNA was bound to the filter using the Stratagene UV Stratalinker™ (Autocrosslink setting).

The filters were pre-hybridized in pre-hybridization buffer at 65° C for twenty-four hours and hybridization was conducted in a 0.25X prehybridization solution with a ³²P-labeled heterogeneous homeobox-specific DNA probe (*Probe #1*, above) at 65° C. The filter was washed with a 2X SSC, 0.1% SDS solution three times at 65° C, allowed to dry, placed on Kodak X-ray film and exposed at -70° C for 1-4 days. The film was developed, revealing the location of homeoboxes within the λ clones fragments.

The inserts of λ clones 7H and 8H were “shotgun” subcloned into pUC-19 plasmid vector (GenBank accession #: M77789). Restriction endonuclease mapping suggested that the best approach to isolate the homeobox-containing fragments as well as the other regions of the inserts was to perform two single digests with *Bam*HI and *Hind*III, and to perform a double digest with *Eco*RI x *Sal*I. Therefore, for the single digests, approximately 2.0 μ g of λ clone 7H and 8H DNA and 1.0 μ g of pUC-19 plasmid vector DNA were restriction endonuclease digested to completion with either *Bam*HI or *Hind*III. Both digestion products were PCI extracted twice, ethanol precipitated, washed with 70% ethanol, and resuspended in 20.0 μ L of Tris low EDTA. The digestion products from either of the λ clones were then mixed with the pUC-19 plasmid vector that had been digested with the same enzyme in a molar ratio of approximately 3:1 and ligated with T₄ DNA ligase at 14° C for 14 hours.

For the double digests, approximately 4.0 μ g of λ clone 7H or 8H DNA and 2.0 μ g of pUC-19 plasmid vector were first digested with *Sal*I to completion and then the DNA was extracted twice with PCI, precipitated with ethanol, washed with 70% ethanol, and resuspended in 20.0 μ L of Tris low EDTA. The digestion products from both the λ clone inserts and pUC-19 plasmid vector were then restriction endonuclease digested with the second enzyme, *Eco*RI, PCI extracted twice, ethanol precipitated, washed with 70% ethanol, and suspended in 20.0 μ L of Tris low EDTA. The digestion products of the 7H and 8H λ clones were mixed with the pUC-19 plasmid vector that had been digested

with the same enzymes in a molar ratio of approximately 3:1 and ligated with T₄ DNA ligase at 14° C for 14 hours.

The ligated DNA was used to transform *E.coli* ER1453 using the Hanrahan protocol (1983). Following transformation, 100 µL of the transformed cells were plated on Luria agar plates with ampicillin (100 µg/mL), X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactoside; 50 µg/mL), and IPTG (isopropylthio-β-galactoside; 0.0001 M). The cells were grown at 37° C overnight and insert-containing transformants (indicated by white color) were selected and the plasmid DNA purified.

Inserts prepared from λ clones and subclones corresponding to the *Hoxa4-a10* region were completely sequenced via primer walking using the PRISM™ Ready Reaction Dyedexy Terminator Cycle Sequencing Kit (Applied Biosystems, Inc., Foster City, CA). Genes *Hoxa4-a9* and intergenic regions were sequenced repeatedly from multiple individual clones. Each sequencing reaction was conducted as follows: approximately 1.0 µg of template DNA (from λ clones or subclones), 9.0 pmol of DNA primer (Appendix B), and 8.0 µL of TaqFS Ready Reaction Mix in a total volume of 20.0 µL. Two drops of mineral oil were layered on the reaction mixture which was then cycled in a MJ Research MiniCycler™ (Watertown MA) as follows: denaturation at 96° C for 30 seconds, annealed at 50° C for 15 seconds, extension at 60° C for 4 minutes, repeated for 25 cycles. The sequencing reaction products were purified by removing the 20 µL mixture from under the mineral oil and adding it to a 1.5 mL Eppendorf tube containing 2.0 µL of 3.0 M sodium acetate, pH 4.6 and 50 µL of 95% ethanol and

precipitated on ice for 10 minutes. The sequencing products were then pelleted by centrifugation at 14,000 rpm for 20 minutes at 4° C, washed with 400 µL of 70 % ethanol, and air-dried at 37° C. The purified sequencing reaction products were suspended (6µL) in 5 parts formamide/1 part 50 mM EDTA, pH 8.0 loading buffer and analyzed on an automated DNA sequencer (Model 373A Stretch; Applied Biosystems, Inc.).

Genomic Southern blots

Genomic Southern blots were performed by restriction endonuclease digestion of approximately 2µg of *M. saxatilis* genomic DNA with *Eco*RI and *Hind*III, blotted as above, and probed with three independent ³²P-labeled DNA probes prepared from the entire isolated inserts of λ clones 2.4, 7H, and 3B (Figure 4), which span the 31 kb contig (hybridized in 6X SSC at 60° C; washed four times in 2X SSC/0.1% SDS at 65° C). The genomic maps generated by Southern analysis were compared to the predicted maps generated from the *Morone* sequence data using the Wisconsin Genetics Computer Group (GCG) 9.0 (GCG, 1996) MAPSORT program.

DNA sequence analysis

Overlapping DNA sequences generated via primer walking from selected clones were compiled into a contiguous 31 kb DNA sequence using the ABI AutoAssembler™ program (Applied Biosystems, Inc., 1995). To identify *Hox*-specific sequences (exon1 and exon2) for all isolated *Morone Hox* genes, overlapping DNA sequences of ≤ 500 bp

in length were compared to the National Center for Biotechnology Information (NCBI) databases using the Basic Local Alignment Search Tool (BLASTn and BLASTx) search programs (NCBI; Altschul *et al.*, 1990; Gish and States, 1993). On the basis of the BLASTn and BLASTx results, tentative gene assignments of *Hoxa4*, *Hoxa5*, *Hoxa7*, *Hoxa9*, and *Hoxa10* were determined based on DNA sequence relatedness to other vertebrate sequences. Intron-exon junctions were identified and tentatively localized using the Splice Site Prediction program by Neural Network (Brunak *et al.*, 1991). To confirm the absence of *Hoxa6* and *Hoxa8* genes in the *Hoxa5-a7* and *Hoxa7-a9* intergenic regions, I determined the lowest threshold of contiguous *Morone Hox*-specific nucleotides (25 bp) detectable in a 500 bp DNA sequence using NCBI BLAST search programs. Hence, 50 bp of *Morone Hoxa5* homeodomain sequence specific to the first helix of the homeodomain was embedded in 500 bp of *Morone* intergenic sequence and BLASTn and BLASTx searches were performed reiteratively with repeated 5 bp deletions of the *Hoxa5* sequence, until *Hox*-specific matches could not be detected (Advanced BLAST search using default parameters).

Unrooted evolutionary trees were constructed from amino acid substitution matrices using the p-distance values from the Molecular Evolutionary Genetics Analysis (MEGA) program (version 1.01; Kumar *et al.*, 1993). Complete homeodomain sequences (60 a.a.) were clustered by a neighbor-joining algorithm (Saitou and Nei, 1987) and a UPGMA algorithm (Sourdis and Krimbas, 1987), using the NJ Method and UPGMA Method, respectively, of MEGA. Homeodomain sequences used in comparisons: **amphioxus**- *AmphiHox-4* Z35144, *AmphiHox-5* Z35145, *AmphiHox-6*

Z35146, *AmphiHox-7* Z35147, *AmphiHox-8* Z35148, *AmphiHox-9* Z35149, *AmphiHox-10* Z35150; **Fugu-** *Hoxb4* U92575, *Hoxc6-c8-c9* U92572, *Hoxa9-a10* U92573; **human-** *Hoxa4* M74297, *Hoxb4* X16174, *Hoxc4* X07495, *Hoxd4* X17360, *Hoxa5* M26679, *Hoxb5* M92299, *Hoxc5* X61755, *Hoxa6-a7* AC004080, *Hoxb6* M30597, *Hoxc6* S82986, *Hoxb7* M30598, *Hoxb8* X16173, *Hoxc8* M16938, *Hoxd8* X15507, *Hoxa9* U82759, *Hoxb9* X16172, *Hoxd9* X15506, *Hoxa10* X58430, *Hoxd10* X59373; **mouse-** *Hoxa4* X66861, *Hoxb4* M36654, *Hoxc4* S62287, *Hoxd4* J03770, *Hoxa5* Y00208, *Hoxb5* M26283, *Hoxc5* U28071, *Hoxa6* M11988, *Hoxb6* M18401, *Hoxc6* X16510, *Hoxa7* U15972, *Hoxb7* X14570, *Hoxb8* M18399, *Hoxc8* M35603, *Hoxd8* X56561, *Hoxa9* M28449, *Hoxb9* S66855, *Hoxc9* M35603, *Hoxd9* X62669, *Hoxa10* S61771, *Hoxc10* D11330, *Hoxd10* M21046; **salmon-** *Hoxa5* M18904, *Hoxa7* M18903; **zebrafish-** *Hoxd4* M24085, *Hoxb5* X12802, *Hoxc5* X68324, *Hoxb6* X17267, *Hoxc6* X17266, *Hoxa9* Y07696, *Hoxa10* Y07697, *Hoxd10* U51421.

Exons of differing lengths were aligned by translating the DNA sequences with the TRANSLATE program into the peptide sequences, and then assembled using the GCG PILEUP program with parameters set for *In situ* to facilitate alignment with gaps. The 60 a.a. homeodomain sequences were aligned by translating the 180 bp DNA sequences into peptides using the TRANSLATE program and then assembled using the PRETTY program, with the *Drosophila melanogaster Antp* homeodomain sequence used as a consensus.

In Figure 11, the *Hoxa7* exon1 DNA sequences (frog M24752; human AC004080; mouse M17192; quail M59714) were translated using the GCG program

TRANSLATE and aligned using PILEUP (parameter set for *In situ* alignment). To align the presumptive *Fugu Hoxa7* exon1 (U92573), the DNA sequence was translated and compared in all three reading frames to the *Morone Hoxa7* exon1 amino acid translation using the GCG programs GAP and BESTFIT (GCG, 1996).

The intergenic regions were compared using the COMPARE program in GCG with default parameters for the two sequences, which generated a comparison matrix of plot coordinates. Harr plots were constructed using the output file from the COMPARE analysis and running the DOTPLOT program to generate a graphical interpretation of the matrix. Alignment of intergenic regions was done using the BESTFIT and GAP programs of GCG. The GCG REPEAT program was used (default parameters) to determine direct repeat regions within the *Hoxa9-a7* and *Hoxa10-a9* intergenic region sequences from *Morone* and *Fugu*.

RESULTS

Isolation and characterization of *HoxA* cluster λ clones

A *Morone saxatilis* genomic bacteriophage λ library (average insert size approximately 15 kb) was screened with a heterogeneous homeodomain-specific DNA probe (*Probe #1*; see Materials and Methods) generating 48 positive plaques, 14 of which exhibited strong hybridization. Three of these strongly hybridizing clones (named 7H, 8H, and 11H) were selected for further characterization. Additional screening of the genomic λ library was conducted using a heterogeneous *Hox*-specific probe followed by a 300 bp *Hoxa9* intron-specific probe, and a *Hoxa5*-specific probe generated a total of six overlapping λ clones (7H, 8H, 11H, 3B, 3.3, and 2.4).

Restriction endonuclease mapping of the λ clones 7H (15,617 bp in length) and 8H (13,080 bp in length) revealed an extensive overlapping region (Figure 3). Restriction endonuclease digestion of the two λ clones with *Bam*HI, *Eco*RI, *Hind*III, and *Sal*I as single and double digests generated the following fragments: 1 *Bam*HI/*Eco*RI fragment (1270 bp), 3 *Eco*RI/*Sal*I fragments (2158 bp, 479 bp, 368 bp), 1 *Sal*I fragment (1792 bp), 1 *Sal*I/*Hind*III fragment (1693 bp), and 2 *Hind*III fragments (273 bp and 1403 bp). These fragments, common to both clones, indicated a region of overlap, approximately 10 kb in length, within the inserts (Figure 3). Additionally, λ clone 8H generated a unique *Bam*HI fragment (2741 bp), and λ clone 7H generated 1 *Hind*III fragment (1358 bp), 2 *Eco*RI fragments (822 bp and 273 bp), 2 *Eco*RI/*Hind*III fragments

(532 bp and 111bp), and 1 *EcoRI/BamHI* fragment (2236 bp) that were unique to λ clone 7H.

Southern blot analysis for the 7H and 8H λ clones revealed the likely presence of at least two linked homeobox sequences within each of the overlapping inserts. The inserts were restriction endonuclease digested with *BamHI*, *EcoRI*, *HindIII*, and *SalI* as single and double digests, and Southern blotted using *Probe #1*. One of the hybridizing fragments that was generated from both the 7H and 8H λ clones was a 2.1 kb *EcoRI/SalI* fragment. Additionally, the 7H λ clone generated a strongly hybridizing 1.3 kb *HindIII* fragment, while the 8H λ clone generated a 1.0 kb *HindIII/SalI* fragment. However, the *SalI* site at the end of the 8H λ clone is not present in the genome, nor the 7H λ clone, since it corresponded to an endonuclease restriction site within the Multiple Cloning Site (MCS) of the λ EMBL-3 vector arm, just beyond the *Sau3a/BamHI* site into which the insert was cloned. Therefore, although the λ clone 8H fragment was truncated at the *Sau3a* cloning site, resulting in a fragment 300 bp shorter than the 1.3 kb *HindIII* fragment from λ clone 7H, they are identical. This is further supported by a *Sau3a* restriction site located 300 bp from the end of the 1.3 kb *HindIII* fragment from λ clone 7H. Furthermore, the 2.7 kb *BamHI* fragment, unique to the 8H λ clone, hybridized weakly to the *Hox*-specific probe indicating a total of three homeobox sequences within the 8H λ clone insert (Figure 3).

The *BamHI*, *EcoRI*, *SalI* and *HindIII*, and *EcoRI/SalI* fragments from both λ clones 7H and 8H were subcloned into the pUC-19 plasmid vector. This subcloning

strategy generated subclones that together represented the entire inserts of the 7H and 8H λ clones. The DNA fragments from each of the subclones were sequenced using primer extension either with pUC-19 universal primers or with subclone-specific primers.

DNA sequence data and predicted amino acid translations from each subclone were compared to nucleotide and amino acid sequences from the National Center for Biotechnology Information (NCBI) databases. DNA sequence data from the entire inserts of λ clones 7H and 8H revealed two complete genes, each comprised of two exons and one intron within the 10 kb overlapping region of the two clones. Intron-exon junctions were tentatively assigned using the Splice Site Prediction program by Neural Network (Brunak *et al.*, 1991). Based on these sequence comparisons, the homeobox-encoding region of the 2.1 kb *EcoRI/SalI* subclone from both the 7H and 8H λ clones was most closely related to *Hoxa7* homeobox sequences from other vertebrates. The homeobox sequence of the 1.3 kb *HindIII* subclone from λ clone 7H and the 1.0 kb *HindIII/SalI* subclone from the 8H λ clone was most closely related to *Hoxa5*.

In addition to the two complete genes within the λ clone 8H insert (*Hoxa5* and *Hoxa7*), the 2.7 kb *BamHI* subclone from the 8H λ clone revealed the presence of the presumptive 3' exon (the homeobox-containing exon2) of the *Hoxa9* gene. The 2.7 kb *BamHI* fragment also contained a 370 bp portion of the *Hoxa9* intron located upstream from exon2, which terminated at the end of the 8H λ clone insert. In order to characterize the entire *Hoxa9* gene, including the 5' exon (exon1), the *M. saxatilis* genomic λ library was re-screened to isolate clones containing the *Hoxa9* gene. To

ensure that the clones isolated from the screen were both *Hox*-encoding and *Hoxa9* specific, the library was screened sequentially; first with both a heterogeneous homeobox-specific probe (*Probe #2*), and then with a 300 bp *Hoxa9* intron-specific probe (*Probe #3*). Of the six λ clones identified using this screening method, one clone was selected (clone 2.4), the DNA was isolated, and the ends of the insert were sequenced using primers that were specific for the λ EMBL-3 left and right arms. Sequence data from λ clone 2.4 (14,004 bp in length) insert termini revealed a region of DNA sequence overlap with λ clones 7H and 8H (approximately 7 kb) at one end of the insert, and a sequence corresponding to exon1 from the *Hoxa10* gene located at the other end (Figure 4). The region of λ clone 2.4 that did not overlap with the 8H λ clone (approximately 7 kb) and extended 5' to the *Hoxa9* exon2, was sequenced via primer walking. This 7kb region of λ clone 2.4 contained exon1 of the *Hoxa9* gene and the complete *Hoxa10* gene comprised of exons 1 and 2 and a single intron.

Of the four clones isolated from the library screen that were not *Hoxa9*-specific, one clone was selected (clone 3.3), the DNA was isolated, and the ends of the insert were sequenced using the λ EMBL-3 left and right arm primers. The DNA sequence from the λ clone 3.3 (17,379 bp in length) insert ends revealed an overlap of approximately 12.9 kb with the 7H and 8H λ clones, and a region that extended approximately 4.4 kb beyond the 3' end of the 7H λ clone insert (Figure 4). The 4.4 kb region that extended 3' from the end of 7H λ clone insert was sequenced entirely via primer walking and revealed the presence of the entire exon1 from the *Hoxa4* gene and a 245 bp portion of the *Hoxa4*

intron. An additional screen of the *M. saxatilis* genomic λ library by Scott Avery using a *Hoxa5*-specific probe (*Probe #4*) generated λ clone 3B (16,672 bp in length). The ends of the λ clone 3B insert were sequenced, revealing an overlap with the 7H, 8H, and 3.3 λ clones. The non-overlapping region of λ clone 3B that extended 3' from the end of λ clone 3.3 was sequenced entirely, which revealed that it contained the remaining portion of the *Hoxa4* intron and the complete *Hoxa4* exon2. Interestingly, λ clones 2.4 and 3B were generated from the same *Sau3a* site within the genome, but extend in opposite directions from each other, and together span the entire *Hoxa4-a10* region. Lastly, the ends of the insert from λ clone 11H from the initial library screen (approximately 15 kb in length) were sequenced. Sequence data revealed that λ clone 11H also contained the complete *Hoxa4* gene, though the region of overlap with the other λ clones isolated was small (approximately 1.6 kb). The 11H λ clone extends 3' towards the predicted location of the *Hoxa3* gene within the genome, however, Southern blot analysis revealed only one homeobox, *Hoxa4*, is contained within the insert. DNA sequence data that was obtained from this clone was restricted to the 1.6 kb region overlap.

In all, a total of six overlapping λ clones were isolated, five of which were completely sequenced (7H, 8H, 3B, 3.3, and 2.4). Based on nucleotide sequence comparisons to other vertebrate *Hox* genes, the six clones contain five *Hox* genes (*Hoxa4*, *Hoxa5*, *Hoxa7*, *Hoxa9* and *Hoxa10*) that span the *Hoxa4-a10* region of the *M. saxatilis* *HoxA* cluster (approximately 31 kb in length) (Figure 4). Moreover, nucleotide sequence analysis of the intergenic regions between the *Hoxa5* and *Hoxa7* genes and between the *Hoxa7* to *Hoxa9* genes failed to provide evidence of *Hoxa6* or *Hoxa8* genes,

pseudogenes, or gene remnants. The assay used to search for the presence of pseudogenes or gene remnants was sufficiently sensitive to detect contiguous remnant sequences as short as 25 bp.

Confirmation of *Hox* gene assignment and linkage

Preliminary gene assignment of the five linked homeodomain sequences within the 31 bp contig as *Hoxa4*, *Hoxa5*, *Hoxa7*, *Hoxa9*, and *Hoxa10* was based on comparisons to NCBI nucleotide and amino acid databases. Additional support for gene assignments was provided via computational clustering in which the five homeodomain sequences from *Morone* were compared with related genes of paralogous groups 4-10 from other teleosts, mouse, human, and the cephalochordate amphioxus (Figure 5). In order to limit ambiguity of assignments, only complete homeodomain sequences of 60 amino acids were used in the comparisons, and partial homeodomain sequences, such as those generated by PCR surveys were not included (Pendleton *et al.*, 1993; Pavell and Stellwag, 1994; Misof *et al.*, 1996; Misof and Wagner, 1996; Prince *et al.*, 1998). Unrooted trees generated via computational clustering, using both the neighbor-joining and UPGMA algorithms, were comprised of two major branches corresponding to *Hox* central class genes (paralogous groups 4-8) or to members of the 5' (*Abd-B*-like) *Hox* genes (paralogous groups 9-10). The five linked *HoxA* genes from *Morone* clustered together with members of the same paralogous groups from other species. For example, the *Morone Hoxa9* homeodomain clustered together with the *Hoxa9* from zebrafish, *Fugu*, mouse and humans, and the *Hoxa9* cognate genes were linked most closely to

branches representing other *Hox-9* cognate genes. Branch topology comparisons between the trees constructed using the neighbor-joining and UPGMA algorithms were similar, providing further support for the preliminary gene assignments (Figure 5).

Physical linkage relationships and genomic organization of the *Morone HoxA* cluster genes represented in the 31 kb contig were confirmed by genomic Southern blotting (Figure 6). Genomic *M. saxatilis* DNA was restriction endonuclease digested with *EcoRI* and *HindIII*, Southern blotted, and probed with DNA prepared from the inserts of three independent overlapping λ clones (7H, 2.4, and 3B; Figure 7) that together constituted the entire region corresponding to the 31 kb contig. Hybridizing *EcoRI* and *HindIII* fragment lengths were predicted from individual λ clone maps of 7H, 2.4 and 3B, and by using the GCG program MAPSORT to calculate fragment sizes from the DNA sequence data of each of the λ clones, and the entire 31 kb contig (Table 1). These predicted fragment lengths were compared to the hybridizing fragments observed for both the *HindIII* and *EcoRI* genomic digests that were probed with the different λ clones. However, predicted fragments under 800 bp were not compared because there were no discernible fragments under 800 bp on any of the genomic Southern blots (Figure 6).

Each one of the *HindIII* fragments predicted from the 31 kb contig sequence data matched the fragments observed in each of the genomic Southern blots, including the predicted 12,500 bp fragment that spans the overlap between λ clones 2.4 and 7H, and the predicted 3895 bp overlapping fragment between λ clones 7H and 3B (Figure 7). The 7100 bp observed *HindIII* fragment from the genomic Southern blot that was probed with

the λ clone 2.4 indicates that the next *Hind*III restriction site within the genome is 4921 bp upstream from the *Sau*3a site at the end of the λ clone 2.4 insert terminus. The 103 bp *Hind*III fragment generated at the end of λ clone 3B did not have a corresponding observed fragment. Additionally, the *Hind*III digests that were probed with λ clones 7H and 3B, generated a single hybridizing fragment corresponding to the predicted 1403 bp and 1358 bp fragments. In both the λ 7H and λ 3B blots the 1400 bp bands were more diffuse, suggesting the 1403 bp and 1358 *Hind*III fragments had co-migrated and were therefore both identified. This was also true in the *Hind*III Southern blot that was probed with λ clone 3B. The hybridizing fragment that corresponded to the predicted 2313 bp and 2472 bp *Hind*III internal fragments was also more diffuse suggesting that both bands hybridized and were observed (Figure 6). Therefore, the predicted *Hind*III genomic map based on the DNA sequence of the 31 kb contig is supported by the genomic Southern blots.

The predicted fragments from the *Eco*RI restriction endonuclease digests matched the observed internal *Eco*RI fragments for each of the λ clones 2.4, 7H, and 3B. However, the 472 bp predicted *Eco*RI fragment that spans the λ clone 7H and 3B junction could not be identified on the λ 7H and λ 3B genomic Southern blots (Figure 7). The three other *Eco*RI fragments (7732, 822, and 3347 bp) that are located within the overlapping region between λ clones 7H and 3B were clearly identified in both the λ 7H and λ 3B genomic Southern blots. Furthermore, the 2637 bp *Eco*RI fragment that is located within the overlapping region between λ clones 2.4 and 7H was also clearly

identified in both of the λ 2.4 and λ 7H genomic Southern blots. The predicted 7732 bp *EcoRI* fragment that extends from the end of λ clone into the λ 2.4/7H junction was not observed in the genomic blot probed with λ clone 2.4, however, the overlap between the 7732 bp fragment from λ clone 7H and λ clone 2.4 is only 356 bp. Additionally, the fragment that extends the 918 bp fragment from the end of λ clone 2.4 within the genome was not identified, although it had no predicted length beyond 918+ bp. The 6730 bp fragment that extends from the end of λ clone 7H and is internal to λ clone 2.4 was not identified on the *HindIII* genomic blot that was probed with λ clone 7H. However, the *HindIII* predicted map is identical to the genomic map generated for the region, including the junction sites, and the predicted 2632 bp *EcoRI* fragment within the λ 2.4/7H, as well as the 7732 bp, 880 bp, and 3347 bp *EcoRI* fragments within the λ 7H/3B overlapping regions were identified. Therefore, these overlapping genomic maps confirm the arrangement of these clones within the genome.

Comparative analysis of *Morone HoxA* genes

In *Morone*, as in other vertebrates and amphioxus, the *Hox* genes are arranged along the genome such that the *Hoxa4-a10* regions are oriented 3' to 5' with respect to the ends of the cluster, in which the deduced direction of transcription is 5' to 3' (Figure 8). The positions and lengths of the exons, introns, and intergenic regions for the five linked *Morone Hox* genes within the 31 kb contig are given in Table 2.

The nucleotide sequences from each of the five *Morone Hox* gene exons were translated into amino acid sequences using the GCG program TRANSLATE and compared to the corresponding exons from other vertebrates using the GCG program PILEUP. Sequence relatedness of the *Morone Hoxa4-a10* genes (for both the nucleotide and translated amino acid sequences) to corresponding exons from other vertebrates was determined using the GCG programs GAP and BESTFIT (expressed as percentages).

The *Morone Hoxa10* and *Hoxa9* exon1 sequences show little sequence relatedness to mammals, but are conserved relative to *Fugu*. Comparison of the *Morone Hoxa10* exon1 (204 a.a. in length) to the reported *Hoxa10* exon1 from *Fugu* (72 a.a. in length), mouse (309 a.a. in length), and human (303 a.a. in length), show sequence relatedness (86% nucleotides; 89% amino acids) between the two teleosts, but much less relatedness between *Morone* and mouse (49% nucleotides; 36% a.a.) or human (48% nucleotides; 35% a.a.) (Figure 9). Furthermore, both *Morone* and *Fugu* have large deletions (~100 a.a.) within the N-terminal region of exon1 relative to both mouse and humans. It should be noted that the sequence length of 72 a.a. for the *Fugu Hoxa10* exon1 sequence is the length reported by Aparicio *et al.* (1997) (GenBank Accession #: U92573). However, comparisons of the translated amino acid sequence for the region encoded 3' to the reported end of the *Fugu Hoxa10* exon1 in all three reading frames to the *Morone Hoxa10* exon1 sequence, show that the *Fugu* amino acid sequence is similar to the *Morone Hoxa10* exon1 and is possibly comparable in length. Therefore, it appears that the *Fugu Hoxa10* exon1 length was reported incorrectly.

Similar to the *Hoxa10* exon1 results obtained from comparisons between the two teleosts, the *Hoxa9* exon1 from *Morone* (191 a.a. in length) and *Fugu* (189 a.a. in length) also show sequence conservation relative to each other (88% nucleotides; 87% a.a.). Comparisons of *Morone Hoxa9* exon1 show very little sequence relatedness (45% nucleotides, 54% a.a.) to mouse *Hoxa9* exon1 (194 a.a. in length) and human (193 a.a. in length; 60% nucleotide; 46% a.a.) (Figure 10).

The genes of the central class appear to be more conserved relative to other vertebrates compared, including mammals, than do the 5' genes. Comparison of exon1 from *Morone Hoxa7* (123 a.a. in length) to other vertebrates (frog 119 a.a., human 127 a.a, mouse 125 a.a., and quail 126 a.a. in length) shows that they are similar in size (Figure 11). Furthermore, sequence comparisons of the *Hoxa7* exon1 from *Morone* to human (66% nucleotides, 53% a.a.), mouse (63% nucleotides, 54% a.a.), frog (62 % nucleotides, 52% a.a.) and quail (63% nucleotide, and 54% a.a.) demonstrate higher sequence relatedness of *Morone* to other vertebrates than the 5' genes. The *Morone Hoxa5* exon1 (198 a.a. in length) sequence is similar in size relative to mouse (187 a.a. in length) and human (187 a.a. in length) and shows the most sequence relatedness of any of the *Morone HoxA* genes to mammals (human – 66% nucleotides; 65% a.a.; mouse – 65% nucleotides, 64% a.a.) (Figure 12).

Like the *Hoxa10* and *Hoxa9* exon1 comparisons, exon1 of *Morone Hoxa4* (139 a.a. in length) is shorter than the corresponding exons from other vertebrates: chicken (199 a.a. in length), mouse (170 a.a. in length), and human (205 a.a. in length) (Figure 13). Furthermore, the *Morone Hoxa4* exon1 shares the most sequence relatedness with

mouse (54% nucleotides, 48% a.a.) and human (50% nucleotides; 34%), but not with the *Hoxa4* exon1 from chicken (47% nucleotides; 32% a.a.).

Comparisons of all five *Morone HoxA* exon1-encoding sequences show that there are two conserved regions among all five exons (Figure 14). They are the hexapeptide (IYPWMR) regions at the C-termini of the central class genes, *Hoxa4-a7*, and within the conserved octamer (MSSYFVNS) sequence at the beginning of each sequence. The hexapeptide sequence is absent in the *Morone* 5' genes, *Hoxa9* and *Hox10*, which is consistent with the 5' *Abd-B*-like genes from other vertebrates that also lack the hexapeptide sequence (Benson *et al.*, 1995).

Despite the varying lengths of exon1 in the *Hox* genes from *Morone* relative to other vertebrates, the average length of the homeobox-containing exon2 (~85 a.a.) in the five *Morone* genes is conserved relative to the corresponding exon2 lengths from other vertebrate *HoxA* genes (Figures 15-19). The exception is exon2 from the *Morone Hoxa4* gene (111 a.a.; Figure 20) which is longer than the other *Morone HoxA* exon2 lengths characterized (80-103 a.a.). Despite being longer than other *Morone HoxA* exon2-encoding regions, its length is consistent with the size of the *Hoxa4* exon2 from chicken (110 a.a.), mouse (115 a.a.), and human (115 a.a.). A comparison of the exon2-encoding sequences from all of the *Morone HoxA* genes characterized shows that the sequences are diverged both upstream and downstream of the homeobox, but are ~75% conserved within the 60 a.a. corresponding to the homeobox (Figure 20). Intron lengths for the five *HoxA* genes from *Morone* are similar to the corresponding introns from other vertebrates, with the exception of the *Morone Hoxa9* intron (405 bp), which is similar to the intron

length of *Hoxa9* from *Fugu* (304 bp), but is shorter than the corresponding intron from human (1018 bp).

PCR vs. genomic – based *Hox* gene assignments

A previous PCR survey of *Hox* genes of the central class from *M. saxatilis* (Pavell and Stellwag, 1994) generated eight *Hox*-like sequences identified as *Hoxa3*, *Hoxa4*, *Hoxa6*, *Hoxa7*, *Hoxb3*, *Hoxb4*, *Hoxb5*, and *Hoxc4*. The *Hox*-like homeodomain-specific PCR amplification products of 69 bp (22 a.a.) were assigned to particular genes based on sequence relatedness to published mammalian sequences, particularly mouse homeodomain sequences. Pair-wise comparisons of these short *Morone* sequences to the *HoxA* genomic sequences generated from this study and to vertebrate *Hox* gene sequences published since Pavell and Stellwag's PCR survey (Sordino *et al.*, 1996; van der Hoeven *et al.*, 1996; Aparicio *et al.*, 1997), show that some of the previous *Hox*-like genes, such as *Hoxa6*, were mis-assigned. The results of these comparisons are summarized in Table 3. The *Hoxa6*-like sequence identified previously is unrelated to the *Hoxa4-a10* genes from *Morone* (most extensive sequence similarity is to *Morone Hoxa5*: 47/69 nucleotides), but is more related to *Hoxc6* sequences from other vertebrates, particularly *Fugu Hoxc6* (61/69 nucleotides). The *Hoxb5*-like PCR sequence is identical to the *Morone Hoxa7* gene characterized in this study, whereas the *Hoxa7*-like PCR-derived sequence is not related to the *Hox* gene sequences characterized from the *HoxA* cluster (most extensive sequence similarity is to *Morone Hoxa5*: 56/69 nucleotides). Moreover, of the eight genes identified in the PCR survey of *Morone*, only

the *Hoxb3*-like PCR sequence appears to have been correctly assigned based on comparisons to the DNA sequence of the *Hoxb3* gene (69/69 nucleotides) obtained from the genomic library screen and previously characterized by McCallum (1997).

Intergenic sequence comparisons

In order to examine sequence conservation in non-coding regions of the cluster, the *Hoxa9-a7* and *Hoxa10-a9* intergenic regions from *Morone* were compared in a pair-wise manner to the corresponding regions from human and *Fugu*, the only other species for which DNA sequence data of this region have been published. Harr plot analyses of sequence comparisons are shown in Figures 21-26 and all intergenic comparison data are presented in Appendix A. Figure 21 represents a pair-wise comparison between the *Morone* and human *Hoxa9-a7* intergenic regions. The random arrangement of points reveals an absence of regions of similarity between the two sequences point for point. However, using the BESTFIT program a 104 bp conserved region (82%) was identified that corresponded to the 5' UTR region of the *Morone Hoxa7* gene, which matches the region, that is 570 bp upstream from the translational start site of human *Hoxa7*. Comparison between the *Fugu* and human *Hoxa9-a7* intergenic regions generated similar results (Figure 22) in which a 103 bp region corresponding to the 5' UTR of the *Fugu Hoxa7* gene was 83 % related to the identical region in the human *Hoxa9-a7* intergenic region.

The *Hoxa10-a9* intergenic region from *Morone* was compared to the corresponding region in human, revealing a similar pattern reflecting an absence of

sequence relatedness (Figure 23). A BESTFIT comparison of the *Hoxa10-a9* regions from *Morone* and human revealed an 82 bp region of related sequences (74 %) located 3083 bp upstream from the *Morone Hoxa9* exon1, and 4105 bp upstream from the *Hoxa9* exon1. The pair-wise comparison of the *Fugu Hoxa10-a9* region to the corresponding region in human showed no apparent sequence conservation (Figure 24). However, when the sequences were compared using the BESTFIT program, a 123 bp region of moderate (66%) sequence conservation was revealed that corresponds to the region 2987 bp upstream from the *Fugu Hoxa9* exon1, and 4988 bp upstream from the human *Hoxa9* exon1. Therefore, the *Hoxa10-a9* sequence from *Morone* that is related to a region within the human *Hoxa10-a9* intergenic region is not the same region of *Hoxa10-a9* sequence relatedness from *Fugu* (Appendix A).

In contrast to the intergenic sequence divergence of teleosts relative to mammals, comparisons of the intergenic regions among the teleosts, *Morone* and *Fugu* revealed regions with considerable conservation. Sequence comparison of the *Hoxa9-a7* regions from *Morone* and *Fugu* (Figure 25) shows a section of moderate sequence conservation (539/700 nucleotides - indicated by a broken line of points) that corresponds to the region 200-900 bp downstream of the *Hoxa9* exon2. This region of moderate sequence similarity is followed by a box-like cluster of points which correspond to an extensive AC-rich repeat region of approximately 100 bp in length which is located approximately 950 bp downstream of the *Hoxa9* exon2 in both *Morone* and *Fugu*. Downstream of the AC-repeat region, the sequences are highly conserved (198/237 and 479/553 nucleotides) relative to each other (indicated by a straight line of points) for approximately 800 bp.

Toward the terminus of the intergenic region is another region of high DNA sequence conservation (241/289 nucleotides), approximately 200 bp upstream of *Hoxa7*, which corresponds to the 2800 bp region of the *Fugu* sequence. Interspersed among the highly and moderately conserved regions is a region of extreme sequence divergence (indicated by random points), which is shown within the 2000-3600 bp region from *Morone* and *Fugu* in Figure 25. Similar results were obtained when comparing the *Hoxa10-a9* intergenic regions from *Morone* and *Fugu* (Figure 26), including regions of high (401/443 and 219/247 nucleotides) and moderate (498/668 and 225/327 nucleotides) sequence conservation and sequence divergence. Moreover, comparisons of the *Hoxa10-a9* region to the *Hoxa9-a7* region within each species revealed no related regions with the exception of AC repeat sequences.

The pair-wise comparisons detailing the regions of similarity between the *Morone* and *Fugu* intergenic sequences are presented in Appendix A. Using the BESTFIT program from GCG, the *Hoxa9-a7* and *Hoxa10-a9* sequences from *Morone* and *Fugu* were aligned. The pair-wise alignment revealed large deletions within the *Fugu* intergenic sequences relative to *Morone* that corresponded to the highly variable regions of sequence dissimilarity from the Harr plot analyses. Further analysis of the *Morone Hoxa9-a7* and *Hoxa10-a9* intergenic sequences using the GCG REPEAT program identified numerous direct repeat sequences of at least eight nucleotides in length flanking the regions that corresponded to the deleted *Fugu* sequences. Examples of these direct repeat sequences are indicated at the 2753-2787 bp and 2788-2827 bp positions of the *Morone Hoxa9-a7* intergenic region in comparison with the corresponding region

from *Fugu*, and at the 3387-3413 bp position within the *Morone Hoxa10-a9* intergenic region in comparison to the corresponding region in *Fugu* (Appendix A). No repeat sequences were identified in either the *Hoxa9-a7* or the *Hoxa10-a9 Morone* intergenic sequences within the highly conserved regions relative to *Fugu*.

Hox cluster compression

The lengths in nucleotides of the region between and including genes 4-10 of the single cluster of amphioxus, and the comparable region of the *HoxA* clusters of human, *Morone*, and *Fugu* were compared (Figure 27). This region spans approximately 190 kb in amphioxus, 45 kb in humans, 31 kb in *Morone*, and 27 kb in *Fugu*. The intergenic regions of the *HoxA* cluster exhibit a general compression in vertebrates relative to the cephalochordate amphioxus of 4.2-fold in humans and 6.5-fold in teleosts. Furthermore, among the vertebrates, the *Hoxa4-a10* region of teleosts is compressed more extensively than the related region in human relative to humans (1.6-fold). Within this region of the *HoxA* cluster, the exons and introns of the *Hoxa4-a10* genes in *Morone*, *Fugu*, and human are of comparable lengths (as described above), however, the intergenic distances vary. Comparisons between vertebrates and amphioxus show that the intergenic regions in amphioxus are much larger in general (Table 4). The intergenic distance between *AmphiHox-9* and *AmphiHox-10* is 58 kb, while the region between *Hoxa9-a10* is 6.5 kb in human, 4.4 kb in *Morone*, and 3.8 in *Fugu*. The *Hoxa9-a7* intergenic region in human (7.1 kb) is larger than that of *Morone* (3.9 kb) and *Fugu* (3.0 kb), but is smaller than the *AmphiHox-9* to *AmphiHox-7* region which spans 46 kb and includes the *AmphiHox-8*

gene. The *Hoxa7-a5* intergenic distance in *Morone* (5.5 kb) is similar to *Fugu* (6.1 kb), but they are smaller than corresponding regions in human (9.3 kb) and amphioxus (34 kb), which include the human *Hoxa6* gene and *AmphiHox-6* genes, respectively. Lastly, the *Hoxa5-a4* intergenic regions in *Morone* (8.0 kb), *Fugu* (12.0 kb), and human (11.2 kb) are much smaller than the *AmphiHox-5* to *AmphiHox-4*, which spans 55 kb.

DISCUSSION

The *Hoxa4-a10* region from *Morone saxatilis*

The genomic organization of *Hox* clusters among vertebrates appears to be largely conserved. The macroevolutionary changes that have occurred within each cluster are primarily restricted to the deletion of one or more genes. In order to further examine *Hox* gene organization and evolution in teleosts, a region from the *M. saxatilis* *HoxA* gene cluster was cloned and characterized. Based on the results obtained from restriction endonuclease mapping, Southern blotting and DNA sequencing, I conclude that the region is 31 kb in length, and is comprised of five linked *Hox* genes. I have assigned these genes as *Hoxa4*, *Hoxa5*, *Hoxa7*, *Hoxa9*, and *Hoxa10* on the basis of comparisons to orthologous genes in other vertebrates, as well as physical linkage data.

The physical linkage of the five *Hox* genes represented in the 31 kb contig from the *Morone* genome was established based on the extensive sequence overlap among the six λ clones (7H, 8H, 11H, 3B, 3.3 and 2.4) and confirmed by genomic Southern blotting. The observed hybridizing fragments from the *Hind*III genomic Southern blot produced all of the expected fragment lengths within each of the genomic Southern blots, including the predicted 12,500 bp fragment that spans the overlapping region between λ clones 2.4 and 7H, and the predicted 3895 bp overlapping fragment between λ clones 7H and 3B (Figure 7). The only fragments in the *Hind*III digests that could not be identified are those that were under 800 bp in length, and the corresponding fragment that extends the 103+ bp *Hind*III fragment, which is located at the end of the insert for λ clone 3B, into

the genome. Since this fragment extends into the genome, the exact fragment size is unpredictable. However, even if the fragment is over 800 bp long, it might not be identifiable since the 103 bp hybridizing region might be too short to generate a detectable hybridization signal. Disregarding this region, the *HindIII* genomic maps generated from Southern blotting correspond with the individual λ clone maps and the predicted map generated from the sequence data, demonstrating that these overlapping clones accurately represent the genomic organization.

The *EcoRI* genomic maps created from the genomic Southern blots also reflected the predicted maps. The 472 bp predicted *EcoRI* fragment spanning the λ 7H/3B junction was not identified, however no other fragments that size were detectable. Most importantly, all three *EcoRI* fragments located within the λ 7H/3B overlapping region (7732 bp, 822 bp, and 3347 bp) were identified in both the λ 7H and λ 3B *EcoRI* genomic Southern blots. This demonstrates the overlapping nature of these clones and provides support that these clones reflect the genomic organization as well. The only fragments internal to the 31 kb contig that were not identified were the 6730 bp fragment that extends the end of λ 2.4 into λ 7H, and the 7732 bp fragment that extends λ 7H into the insert for λ 2.4. However, the 2637 bp predicted *EcoRI* fragment that is located within the overlapping region of λ clones 7H and 2.4 between the 7732 bp and 6730 bp fragments was identified in both the λ 7H and λ 2.4 genomic blots. This 2637 bp *EcoRI* fragment and the genomic map generated from the *HindIII* Southern blots, demonstrates the overlapping nature of these clones and provides support that these clones are represent this region of the genome (Figure 7). Within both λ 7H and λ 2.4, there are no

other predicted fragments that correspond to a length of 2637 bp, therefore it must correspond to the common fragment in the overlapping region between these two clones. It could be argued that the 2637 bp *EcoRI* fragment is the extension of the 918+ fragment of unpredicted length that extends into the genome. Since the 2637 bp *EcoRI* fragment was identified in both the λ 7H and λ 2.4 genomic Southern blots, it is most likely the common fragment shared in the overlapping regions. Furthermore, the region of the map that corresponds to the λ 7H/2.4 junction, which also corresponds to the location of the predicted 7732 bp and 6730 bp end fragments from the *EcoRI* blots and the common 2637 bp fragment, has no *HindIII* sites. In fact, the region spans 12,500 bp and overlaps all three λ clones, including the 7734 bp, 2637 bp, and 6730 bp *EcoRI* fragments. Since all three genomic Southern blots generated this 12,500 bp fragment, and the λ 7H and λ 2.4 blots had a 2637 bp fragment in common, these results show that the organization of this region of the genome is reflected in the overlapping genomic λ clones isolated. Therefore, I conclude that the organization of the five linked *Hox* genes represented in the 31 kb contig is an accurate representation of the gene organization within the genome.

The genomic arrangement of the clone inserts on the chromosome is further supported by the fact that the arrangement of the *Morone Hoxa4-a10* genes within the overlapping λ clones, which are oriented 3' to 5' with respect to the ends of the cluster, is consistent with the genomic organization of orthologous genes in other vertebrates (Acampora *et al.*, 1989; Njølstad *et al.*, 1990; Krumlauf, 1992; Sordino *et al.*, 1996; van der Hoeven *et al.*, 1996; Aparicio *et al.*, 1997). Moreover, similar to the *Hox* genes characterized from other vertebrates, the predicted direction of transcription is 5' to 3',

and each of the five genes consist of two exons and an intron. However, analysis of the *Hoxa5-a7* and *Hoxa7-a9* intergenic region DNA sequences revealed the absence of the *Hoxa6* or *Hoxa8* genes, pseudogenes or remnants.

Comparison of *HoxA* genes and cluster evolution

The total *Hox* gene complement of the 31 kb contig from *Morone* was compared with corresponding regions from other species for which physical maps have been constructed, including amphioxus (Garcia-Fernández and Holland, 1994), mouse (as reviewed by Krumlauf, 1992), human (Acampora, 1989), and *Fugu* (Aparicio *et al.*, 1997) (Figure 8). Among these vertebrates, the *Hoxa4*, *Hoxa5*, *Hoxa9*, and *Hoxa10* genes are always present, whereas the *Hoxa8* gene is consistently absent. However, a recent 3' RACE PCR survey of *Hox* genes from zebrafish (order Cypriniformes) reported the presence of a *Hoxa8* gene, but not a *Hoxa4* gene (Prince *et al.*, 1998) (Figure 8). Assignment of the partial zebrafish homeodomain sequences to a particular paralogous group was based on both the homeodomain sequence relatedness to other vertebrates, and comparison of the 12 "diagnostic" amino acid residues encoded 3' to the homeodomain (defined by Sharkey *et al.*, 1997). Cluster assignment was determined by linkage of the short homeodomain-specific amplification products to specific mouse/zebrafish somatic cell hybrids (Egger *et al.*, 1996).

The absence of *Hoxa4* and the presence of *Hoxa8* in zebrafish, but not in *Morone*, *Fugu* or mammals may be explained by several equally likely alternatives. It is possible that the cypriniforms, of which zebrafish is a representative, have diverged significantly

from *Morone* (order Perciformes) and *Fugu* (order Tetraodontiformes) through the loss of the *Hoxa4* gene and the gain or retention of the *Hoxa8* gene. If so, it suggests that convergent evolution has occurred among vertebrates for the *Hoxa8* gene, wherein *Hoxa8* was lost after the divergence of cypriniforms from perciforms and tetraodontiforms. This explanation requires that the loss of *Hoxa4* occurred independently of the *Hoxa8* gene retention and argues that the cypriniforms diverged from mammals independently of other teleost lineages, particularly the perciforms and tetraodontiforms. Given that actinopterygians are presumed to be monophyletic, and that the perciforms and tetraodontiforms are thought to have evolved more recently than the cypriniforms (Johnson and Patterson, 1993; Patterson, 1993; Nelson, 1994; Janvier, 1996), it is difficult to reconcile these conflicting results. However, given the lack of genomic linkage data for the central class genes from the zebrafish *HoxA* cluster, it is possible that the zebrafish *Hoxa8* gene is mis-assigned. The preponderance of evidence therefore suggests that the loss of *Hoxa4* is a singular event related to the divergence of cypriniforms from tetraodontiforms and perciforms, and that the loss of *Hoxa8* occurred prior to the divergence of mammals from teleosts.

Another likely alternative is that the *Hoxa4* gene is present within the zebrafish genome and was not identified by Prince *et al.* (1998), or was mis-assigned to another cluster. Prince *et al.* (1998) concluded, based on their 3'RACE PCR survey of zebrafish, that in addition to the four clusters identified in teleosts (Aparicio *et al.*, 1997) and mammals (Acampora, 1989; Krumlauf, 1992), there are one or two previously uncharacterized clusters, termed *HoxX* and *HoxY*, present in the zebrafish genome

(Figure 2). Comparisons of the *HoxX* sequences to *HoxA* homeodomain sequences from other teleosts, reveal that the amino acid sequences of the newly characterized *HoxX* genes from zebrafish are 97 % similar to paralogous genes from the *HoxA* cluster in *Morone* and *Fugu* (Figure 28). Most notably, comparisons of *Hoxa4* translated homeodomains from *Morone* and *Fugu* to the *Hoxx4* homeodomain sequence from zebrafish show sequence conservation among all three teleosts (38/39 a.a.), including the substitution of a methionine for a leucine residue at position #38, and a valine substitution for isoleucine at position #45. Within the homeodomain sequence, residue #59 demonstrates the most variation with glutamate in *Morone*, tryptophan in *Fugu*, and aspartate in zebrafish. Moreover, the 12 diagnostic residues encoded 3' to the *Hoxx4* homeodomain from zebrafish are identical to the 3'-encoded diagnostic *Hoxa4* residues from *Morone* and *Fugu*, with the exception of an asparagine in the first position of the *Fugu* sequence instead of a lysine residue, which is a conservative substitution. Prince *et al.* (1998) argued that the substitutions at positions #38, 45, and 59 within the zebrafish homeodomain sequence relative to mouse, coupled with the divergence of the 12 zebrafish diagnostic amino acids from the mouse *Hoxa4* gene, and the linkage of this sequence to a non-*HoxA*-specific zebrafish/mouse somatic cell hybrid, supported the assignment of this *Hox* gene to a novel *HoxX* cluster instead of the *HoxA* cluster. However, these substitutions in the zebrafish *Hoxx4* sequence are conserved in the *Hoxa4* genes identified from *Fugu* and *Morone*, suggesting that these substitutions might be teleost-specific. Therefore, the results indicate that the *Hoxa4* sequence from zebrafish was mis-assigned as *Hoxx4* and correctly belongs among the *HoxA* cluster genes. If so,

then the loss of *Hoxa8* was an independent event that occurred after the divergence of the actinopterygian and sarcopterygian lineages, and again after the divergence of the cypriniforms from the perciforms and tetraodontiforms. This seems a more likely scenario than the divergence of the *Hoxa4* gene and the convergent evolution of the *Hoxa8* gene within one cluster.

The hypothesis that the *HoxX* sequences from zebrafish are mis-assigned *HoxA* sequences is further supported by comparisons of the *Fugu* and *Morone Hoxa9* sequences to the zebrafish *Hoxx9* sequence, which was the only other sequence linked to the same non-*HoxA* zebrafish/mouse somatic cell hybrid as the *Hoxx4* sequence. In addition to the *Hoxa4* gene, Prince *et al.* (1998) also failed to identify a *Hoxa9* gene in their PCR study. However, the zebrafish *Hoxa9* homeodomain was previously characterized in a genomic survey of 5' *HoxA* genes by Sordino *et al.* (1996). Comparisons of the translated homeodomain sequences of *Hoxa9* from *Morone*, *Fugu*, and zebrafish to the partial zebrafish *Hoxx9* homeodomain sequence reveal remarkable similarities (Figure 29). The zebrafish *Hoxx9* homeodomain is identical to *Hoxa9* in *Morone* and *Fugu*, including the substitution of a phenylalanine residue at position #59 instead of the characteristic glutamate found in mammalian *Hoxa9* sequences. Comparison of the 12 diagnostic residues encoded 3' to the homeodomain reveals these residues are varied among the different teleosts. Whereas, unlike other vertebrate cognates in which Sharkey *et al.* (1997) identified paralogous group-specific conserved residues within the 3'-encoded region, variation within the 12 diagnostic residues encoded 3' to the homeodomain is characteristic of *Hox-9* cognates among vertebrates. Interestingly, while the *Hoxx9*

amino acid translation product is identical to the *Hoxa9* sequences from other teleosts, comparisons of zebrafish *Hoxa9* to zebrafish *Hoxx9* reveal differences at position #37 (serine for *Hoxa9*, threonine for *Hoxx9*) and position #59 (cysteine for *Hoxa9*, phenylalanine for *Hoxx9*). Given the extensive linkage data to support the *HoxA* assignments in *Fugu* and *Morone*, and the similarities of the *Fugu* and *Morone Hoxa9* homeodomains to the *Hoxx9* gene, this suggests that the zebrafish *Hoxx9* gene that Prince *et al.* assigned might be the zebrafish *Hoxa9* gene, which suggested that the *Hoxa9* characterized by Sordino *et al.* (1996) might be mis-assigned. Since only the homeodomain was characterized by Sordino *et al.*, this is a likely possibility. However, Sordino *et al.* (1996) also characterized the *Hoxa10*, *Hoxa11*, and *Hoxa13* homeodomains linked to the *Hoxa9* homeodomain, but did not generate any other sequence data outside of exon2 for these genes. This physical linkage data supports the assignment of Sordino *et al.*'s *Hoxa9* homeodomain. Therefore, physical mapping of the genes within the *HoxX* and *HoxY* clusters, including the 5' exons and introns, is necessary to reconcile these conflicting results.

Of the 32 different PCR amplification products obtained from the zebrafish PCR survey, only two mapped to the particular non-*HoxA*-specific (*HoxX*) cell hybrid – the *Hox-4* and *Hox-9* sequences. This suggests that the *HoxX* cluster is not the result of a complete *Hox* cluster duplication since most clusters are comprised of at least six genes (Figure 2). Instead, the extra gene copies could have arisen from single gene duplication and transposition of both the *Hoxa4* and *Hoxa9* genes to a different chromosome than the *HoxA* containing-chromosome. It is also possible that the extra *HoxX* gene copies are

artifacts of the creation of the zebrafish/mouse somatic cell hybrids, particularly given the fact that Prince *et al.* (1998) did not identify either the *Hoxa4* or *Hoxa9* homeodomains in their survey, and the remarkable similarities of the *HoxX* sequences to the *Hoxa4* and *Hoxa9* sequences from *Morone* and *Fugu*. Therefore, it appears that Prince *et al.* (1998) mis-assigned *Hoxa4* and *Hoxa9* gene sequences from the *HoxA* cluster as novel *HoxX* cluster genes.

The presence of additional clusters in zebrafish outside of the characteristic four from vertebrates would seem to support the three-step mechanism of cluster duplication suggested by Bailey *et al.* (1997), in which three successive rounds of genomic duplication resulted in eight clusters. It could be argued that the *HoxX* and *HoxY* clusters in zebrafish represent two remnant clusters that were retained after the original cluster duplication events. However, the presence of only four clusters in mammals and the more recently evolved teleost, *Fugu*, would suggest convergent evolution with the loss of the extra four clusters occurring after the divergence of teleosts from mammals, and again after the divergence of tetraodontiforms from cypriniforms. If this is true, then it is possible that representatives of the more basal actinopterygians, such as the sturgeon, as well as the sarcopterygians, such as the lungfish and coelocanth, would have additional clusters outside of the characteristic *HoxA-D* clusters. If so, then it is also possible that progenitor organisms, such as hagfish, lamprey or cartilaginous fishes, have four or more *Hox* gene clusters resulting from this genomic duplication. If these organisms do not have four or more clusters and the three-step mechanism of cluster duplication is accepted, then this would suggest cluster duplication occurred involving three rounds of

single cluster duplication, which resulted in four clusters established before the divergence of sarcopterygians and actinopterygians. This would also suggest that the additional clusters in the zebrafish genome deduced by Prince *et al.* (1998) are either an isolated duplication event in these fish, or a mis-assignment of these PCR-products. A test of this hypothesis awaits the characterization of physical linkage maps of the *Hox* clusters among extant representatives of ancient actinopterygians and sarcopterygians, as well as their progenitors.

Unlike *Hoxa4* and *Hoxa8*, which may have evolved separately in teleosts, *Hoxa6* is uniformly absent in teleosts examined to date, but present in mammals. This result suggests that the loss of *Hoxa6* occurred after the divergence of teleosts and mammals. However, since this premise is based on studies of teleosts that are considered to be more recently evolved, such as the perciforms, tetraodontiforms, and cypriniforms, I hypothesize that it is possible that a *Hoxa6* ortholog or gene remnant is present in the more ancient teleosts and therefore not uniformly absent in all actinopterygians. One such teleost with a possible *Hoxa6*-like gene is the Atlantic salmon. In 1988, Fjose *et al.* characterized two linked homeodomains from a single Atlantic salmon genomic λ clone. Originally, they had assigned these two linked *Hox* genes as *Hoxa7* and *Hoxb5* based on DNA sequence of the homeodomain (180 bp). In the time since this work was conducted, it has been shown that linked *Hox* genes could not originate from two different clusters (Krumlauf, 1992; McGinnis and Krumlauf, 1992; Scott, 1992). Therefore, in 1993, Fjose *et al.* (1988) reassigned these genes as *Hoxa7* and *Hoxa5*. Comparisons of the salmon *Hoxa7* homeodomain with the *Fugu* and *Morone Hoxa7*

homeodomains show that they are similar (55/60 a.a.), including the unique substitution of an aspartate for a glutamate at position #59, which is not present in other vertebrate *Hoxa7* homeodomains. Comparisons of the *Morone* and *Fugu Hoxa5* homeodomains, and the mammalian *Hoxa5* and *Hoxa6* homeodomains with the salmon *Hoxa5* homeodomain revealed that residues #8-60 are identical to *Hoxa5* and *Hoxb5* cognates from other vertebrates, with the exception of a methionine rather than an isoleucine at position #34 (Figure 30). However, positions #2-7 contain three residues that are characteristic for *Hox-6* cognates, particularly for *Hoxa6* and *Hoxb6* cognates (an arginine at position #2, glutamine at position #6, and threonine at position #7), and three residues that are not characteristic of any vertebrate homeodomain (glycine-arginine-glycine at positions #3-5). Furthermore, computational clustering of the salmon *Hoxa5* gene with the central class genes from other vertebrates places it on a solitary branch between the *Hox-5* and *Hox-6* cognates (Figure 5). One possible explanation for the unique *Hox-6*-like nature of the salmon *Hoxa5* sequence is that it is a remnant of a fusion between a *Hoxa6* and *Hoxa5* gene. Given the high conservation and relatedness of the genes of the central class, it is possible that a recombination event occurred between the two cognates resulting in a protein that is *Hox-6*-like at the N-terminus, and *Hox-5*-like at the C-terminus of the protein. Therefore, the resulting protein might be a fusion of exon1 and a portion of exon2 from *Hoxa6* to exon2 from *Hoxa5*. An alternative possibility is that the *Hoxa5* homeodomain characterized has been mis-assigned and should be designated as *Hoxa6*. Additionally, the unique sequence of the salmon *Hoxa5* gene might simply be due to errors in sequencing. It would be interesting to determine if the

remainder of the protein has *Hox-6*-like characteristics, particularly within the N-terminal region. Furthermore, it would be of interest to survey representatives of the more ancient actinopterygian orders to determine when the *Hoxa6* gene was deleted and, if possible, what morphological effects might have accompanied the gene loss. In addition, comparisons could be made with the corresponding region in an ancient sarcopterygian representative, such as the lungfish that is likely to have a *Hoxa6* present in the genome, to examine the possible molecular events surrounding the loss of a particular gene.

Recent work on the genomic organization of *Hox* genes in *Fugu rubripes* purported the loss of all members of paralogous group 7 (Aparicio *et al.*, 1997). However, a BLASTn search of the NCBI GenBank database using the DNA sequence of *Morone Hoxa7* exon1 revealed an extensive match (250/369 nucleotides, including inserted alignment gaps) with a *Fugu* DNA sequence. Amino acid translation of the *Fugu* DNA sequence in all three reading frames allowed the reconstruction of a complete *Hoxa7* exon1 from the published *Fugu* sequence using the *Morone Hoxa7* exon1 as an index (Figure 11). The putative *Fugu Hoxa7* exon1 sequence corresponds to the 3' region of a genomic phage λ clone that also contains the *Fugu Hoxa10* and *Hoxa9* genes (GenBank accession no: U92573), and is located 3 kb downstream from the *Fugu Hoxa9* homeobox. Therefore, the intergenic distance between the *Fugu Hoxa9* gene and the putative *Hoxa7* exon1 sequence is similar to the *Hoxa9-Hoxa7* intergenic distance of 3.9 kb in *Morone*. The likely presence of a *Fugu Hoxa7* gene, or gene fragment, is further supported by the considerable sequence relatedness shown in comparisons of the 5' UTR (100 bp from +1) from the presumptive *Fugu Hoxa7* gene to the corresponding region in

Morone (91/100 nucleotides), and between the *Hoxa7* intron sequences from *Morone* and *Fugu* (287/387 nucleotides, including inserted alignment gaps). The *Fugu* genomic λ clone terminates within the intron of the presumptive *Hoxa7* gene, or gene fragment, and therefore exon2, which contains the homeobox, is not present within the published sequence. Based on comparisons between the *Hoxa7* sequences from *Morone* and *Fugu*, which revealed an unexpected degree of sequence relatedness in the presumptive coding region, regulatory region, and intron, it appears that a *Hoxa7* gene, or gene fragment, is present in the *Fugu* genome.

The reported absence of all *Hox-7* genes in *Fugu*, and the presence of *Hox-7* genes in related orders, such as *Morone*, and in more ancient actinopterygian representatives, such as Atlantic salmon (Fjose *et al.*, 1988), killifish (Misof and Wagner, 1996) and zebrafish (Prince *et al.*, 1998) suggested that this thorough purging of *Hox-7* genes in *Fugu* was a specific event within the tetraodontiforms. However, the presence of the *Hoxa7* gene in the *Fugu* genome, identified by this study, is consistent with the observation that within the vertebrate *Hox* clusters identified to date, at least one member has been retained from each paralogous group throughout vertebrate evolution. This supports the suggestion that a complete complement of at least one representative from paralogous groups 1-13 is required in all vertebrates, including tetraodontiforms, in order to attain proper spacio-temporal expression of other *Hox* genes during development, and therein establishing the proper positional axes of the developing embryo.

Previous mutational studies of *Hox* genes in mammals, which revealed skeletal aberrations in developing embryos, have shown that *Hox* genes play a role in defining

axial skeletal morphology (Small and Potter, 1993; Condie and Capecchi, 1993,1994; Rancourt *et al.*, 1995; Chen and Capecchi, 1997). It has been suggested that the unique morphology of *Fugu* (including the absence of ribs and pelvic bones) might be attributable to the loss of *Hox* genes (~8 relative to mammals), particularly the loss of the entire paralogous group 7 (Aparicio *et al.*, 1997; Holland, 1997; Meyer, 1998; Prince *et al.*, 1998). However, I was not able to determine if the *Fugu Hoxa7* gene, or gene fragment, is either complete or functional. Therefore it is possible that an incomplete or non-functioning *Hoxa7* gene contributes to the unique morphology of the tetraodontiforms. However, if the *Hoxa7* gene is present and functional in *Fugu*, then it is still possible that the unique morphology of *Fugu* is related to the extensive loss of *Hox* genes, but is not attributable to the loss of the paralogous group 7 genes.

PCR vs. genomic - based *Hox* gene surveys

A *Hoxa6*-like sequence in *M. saxatilis* was reported by Pavell and Stellwag (1994) based on a PCR survey in which homeodomain-specific PCR amplification products of 69 bp (22 a.a.) were used to make gene assignments. In contrast, the results of my study on the *Hoxa4-a10* region failed to provide evidence for a *Hoxa6* gene. Further comparisons of the *M. saxatilis* PCR-derived sequences to *Hox* genes from other vertebrates that have been published since Pavell and Stellwag's study (Sordino *et al.*, 1996; van der Hoeven *et al.*, 1996; Aparicio *et al.*, 1997) show that seven of the eight *Hox*-like sequences, including the *Hoxa6*-like sequence, were mis-assigned (Table 3).

Based on these results and concerns raised by others from previous studies, such as the work conducted on amphioxus, (Holland and Garcia-Fernàndez, 1996; Misof and Wagner, 1996; Aparicio *et al.*, 1997), it can be concluded that gene assignments from *Hox*-based PCR surveys can be inconclusive and unreliable, particularly among the genes of the central class. The unreliable nature of the method appears to stem from several sources. Firstly, PCR-amplification generates relatively short sequences within the conserved region of the homeobox, which is used for analysis. The conserved nature of the homeodomain, and the paucity of diagnostic residues within the N-terminal region of the short homeodomain sequences can make gene assignment difficult. This is especially true for the *Hox-5* and *Hox-6* genes of the central class, which encode for identical amino acid sequences between positions #29 and #55 of the homeodomain. Secondly, there is the potential problem of distinguishing genes (loci) from allelic polymorphic differences of the same locus. This was a particular problem in the assignment of *AmphioxHox-2* from amphioxus, which revealed a sequence polymorphism that was only resolved through genomic cloning and sequencing (Holland and Garcia-Fernàndez, 1996). Without the genomic data, two different *Hox-2* cognate sequences would suggest that they were derived from two different *Hox* clusters. Thirdly, an additional problem of the PCR survey method is that *Hox* gene databases are representative of only a few well-characterized species, which among vertebrates are dominated by mammals. Comparison of DNA sequences derived from evolutionarily divergent species with database entries dominated by mammals biases gene assignments toward mammalian *Hox* genes. Such biases minimize the significance of sequence differences that result from evolutionary

clusters. The *Hoxa3*-like sequence is also related to the newly characterized *PDX* ortholog from amphioxus, *AmphiXlox*, (49/69 nucleotides; 19/22 amino acids), which is a member of the recently described *ParaHox* gene cluster (Brooke *et al.*, 1998). The *ParaHox* cluster is comprised of three homeobox-encoding genes – *Gsx*, *Xlox*, and *Cdx*, which are arranged on the genome 3' to 5' and are expressed in a colinear manner similar to the *Hox* genes, however *ParaHox* genes are expressed in different regions of the developing embryo. It has been suggested that the *Hox* gene clusters and the *ParaHox* cluster arose via the duplication of a single, ancestral “*ProtoHox*” cluster, which was comprised of three homeobox-encoding genes. After cluster duplication, the functions of the genes within each resulting sister cluster diverged such that the *Hox* genes are expressed within the developing mesoderm and ectoderm, while the genes of the *ParaHox* cluster are expressed within the developing endoderm. Comparison of the *ParaHox* homeodomain sequences to the homeodomain sequences of the *AmphiHox* genes shows that *Gsx* is most related to *AmphiHox*-1, *Xlox* is most related to *AmphiHox*-3 and *Cdx* is most related to *AmphiHox*-9 (Brooke *et al.*, 1998). Therefore, despite the divergence of the *ParaHox* cluster function, the clusters appear to be related. Due to the relatedness of *ParaHox* genes to *Hox* genes it is possible to amplify *ParaHox* sequences in a *Hox*-specific survey, and hence mis-assign a *ParaHox* gene to a particular *Hox* gene paralogous group. Therefore, comparison of the short *Morone* PCR-derived *Hoxa3*-like sequence to mammalian *Hox* sequences biased the assignment of a novel *Morone* *ParaHox* sequence to *Hoxa3*. This provides further support for the premise that PCR-based surveys of highly conserved sequences within gene families, such as the

assignment of the PCR-derived *Morone Hoxa3*-like sequence. Comparison of the *Hoxa3*-like sequence to the NCBI GenBank databases revealed that it is most closely related to the zebrafish *PDX* gene (60/68 nucleotides; 22/22 amino acids), which is a homeobox-encoding transcription factor, but it is not a member of the classical *Hox* gene clusters. The *Hoxa3*-like sequence is also related to the newly characterized *PDX* ortholog from amphioxus, *AmphiXlox*, (49/69 nucleotides; 19/22 amino acids), which is a member of the recently described *ParaHox* gene cluster (Brooke *et al.*, 1998). The *ParaHox* cluster is comprised of three homeobox-encoding genes – *Gsx*, *Xlox*, and *Cdx*, which are arranged on the genome 3' to 5' and are expressed in a colinear manner similar to the *Hox* genes, however *ParaHox* genes are expressed in different tissues of the developing embryo. It has been suggested that the *Hox* gene clusters and the *ParaHox* cluster arose via the duplication of a single, ancestral “*ProtoHox*” cluster, which was comprised of three homeobox-encoding genes. After cluster duplication, the functions of the genes within each resulting sister cluster diverged such that the *Hox* genes are expressed within the developing mesoderm and ectoderm, whereas the genes of the *ParaHox* cluster are expressed within the developing endoderm. Comparison of the *ParaHox* homeodomain sequences to the homeodomain sequences of the *AmphiHox* genes shows that *Gsx* is most related to *AmphiHox*-1, *Xlox* is most related to *AmphiHox*-3 and *Cdx* is most related to *AmphiHox*-9 (Brooke *et al.*, 1998). Therefore, despite the divergence of the *ParaHox* cluster function, the clusters appear to be related. Due to the relatedness of *ParaHox* genes to *Hox* genes it is possible to amplify *ParaHox* sequences in a *Hox*-specific survey, and hence mis-assign a *ParaHox* gene to a particular *Hox* gene

paralogous group. Therefore, comparison of the short *Morone* PCR-derived *Hoxa3*-like sequence to mammalian *Hox* sequences biased the assignment of a novel *Morone* *ParaHox* sequence to *Hoxa3*. This provides further support for the premise that PCR-based surveys of highly conserved sequences within gene families, such as the homeodomain-encoding gene family, can lead to ambiguous, inconclusive or incorrect results.

Intergenic comparisons

The *Hoxa9-Hoxa7* and *Hoxa10-a9* intergenic sequences of *Morone* were compared to the corresponding regions in human and *Fugu*, which are the only two organisms for which those data are available. The pair-wise comparison between the teleost *Hoxa10-a9* intergenic regions from *Fugu* or *Morone* with the corresponding region in human showed no apparent sequence conservation, with the exception of two different conserved regions for each fish relative to human (Appendix A). It is possible that these two regions have been conserved because they provide transcription enhancer sites. Such conserved regulatory sites have been identified in comparisons of the *Hoxa9-a7* intergenic regions between mouse and human (Knittel *et al.*, 1995). Knittel *et al.* discovered a 470 bp long control element (named AX 470), located 1.6 kb upstream from the murine *Hoxa7* gene that was 70% conserved relative to the human *Hoxa7* regulatory element, which is located 3 kb from human *HOXA7*. Through deletion experiments, they determined that this enhancer element specifies the anterior boundary of *Hoxa7* gene expression. Therefore, the islands of conserved sequence in the *Morone*-human

intergenic region may serve a regulatory function. The importance of these regulatory sequences for gene activation would explain their conservation throughout the 400 million years since the divergence of teleosts and mammals. Therefore, it is possible that these regions are crucial for transcriptional activation of specific *Hox* genes, such as the *Hoxa10*, *Hoxa9* and *Hoxa7* genes. A test of this hypothesis lies in deletion and expression studies in *Morone*, or the transgenic constructs of the *Morone* enhancer in mice.

In contrast to the regions of conserved sequences, there is a lack of overall sequence relatedness for the *Hoxa9-a7* and *Hoxa10-a9* intergenic regions between teleosts and mammals. This lack of sequence relatedness suggests that this region has not been subjected to the same evolutionary constraints as corresponding coding regions, a promoter region such as the *Hoxa7* 5' UTR, or an important enhancer element responsible for transcription regulation.

The comparisons of the intergenic regions between *Morone* and *Fugu* show that the sequences are generally conserved, suggesting that there has been a markedly reduced level of sequence divergence of non-coding regions between *Morone* and *Fugu* relative to the two teleosts and mammals (Figures 25-26). However, the regions that have diverged contain many deletions within the *Fugu* sequences relative to the *Morone* sequences. It is interesting that direct repeat regions in the *Morone* sequences flank the corresponding deleted regions in the *Fugu* sequences. This suggests that homologous recombination and the deletion of intergenic DNA between direct repeats occurred more readily within the *Fugu* genome than in *Morone*. This is interesting since the *Fugu* genome has been

proposed to be a model vertebrate genome based on the fact that it is compressed 7.6-fold relative to humans (Brenner *et al.*, 1993; Little, 1993; Aparicio *et al.*, 1995). Moreover, Brenner *et al.* (1993) have shown that the compressed nature of *Fugu* genome is due in some measure to shortened intergenic distances resulting from large deletions.

It has been shown in the eukaryote, *Saccharomyces cerevisiae*, that the genes of the RAD52 epistasis group are required for homologous recombination and DNA double-strand break repair (Tracy *et al.*, 1997). Included in the RAD52 group is Rad51, which has been identified as a homolog of the prokaryotic RecA protein that is essential for homologous genetic recombination in prokaryotes (Shinohara *et al.*, 1992). It is believed that the Rad51 protein plays a pivotal role in homology search and strand exchange while Rad52 is required for recombination between direct repeats which yield deletions (Shinohara *et al.*, 1998). Perhaps the members of the RAD52 group are regulated in a different manner in the tetraodontiforms relative to other organisms, allowing for the systematic deletion of intergenic DNA between direct repeat regions. Alternatively, it is possible that these direct repeat sites within the *Morone* sequences are permissive regions for genomic expansion to occur.

The extensive sequence relatedness of a non-coding region between *Morone* and *Fugu* is also interesting from a phylogenetic perspective. While the actinopterygians are considered to be monophyletic, the most recently evolved orders are found within the series Percomorpha. Considering the taxonomic grouping of the different percomorph orders, many taxonomists find the phylogenetic placement of the order Tetraodontiformes problematic (Johnson and Patterson, 1993). Presently there are three

different schools of thought on the evolution of the tetraodontiforms. The first, and most widely accepted, is that the tetraodontiforms are a sister order to the Perciformes, Pleuronectoformes, Scorpaeniformes, and Dactyloptiformes, which, based on fossil evidence and synapomorphies (morphological similarities – shared derived characteristics), diverged from a common ancestor sometime around 50 million years ago (Johnson, 1993; Patterson, 1993; Nelson, 1994). The second hypothesis is a controversial one proposed by Rosen (1984). On the basis of morphological similarities, he considered the tetraodontiforms to be a sister order of the Zeiformes, a more ancestral order within the Percomorpha. The Zeiformes have a fossil record that dates back to 150 million years ago. The third, and perhaps most vague, hypothesis was proposed by Tyler and Sorbini (1996) based on fossil evidence that places the divergence of an ancestral tetraodontiform superfamily prior to 70-95 million years ago. Tyler and Sorbini (1996) argue that the discovery of a primitive tetraodontiform fossil that predates the fossil record of other higher percomorphan orders (as defined by Johnson, 1993), such as the Perciformes, supports the zeiform/tetraodontiform sister order relationship and places these orders among the lower, older percomorphan orders. However, despite the taxonomic placement of the tetraodontiforms, the sequence similarities between *Morone* and *Fugu* within the non-coding regions (75% for the *Hoxa9-a7* region, 73 % for the *Hoxa10-a9* region) shows that little intergenic rearrangement and sequence alteration has occurred since the time the perciforms and tetraodontiforms diverged. Perhaps this degree of non-coding sequence relatedness is a characteristic of closely related teleosts, which might have implications for the phylogenetic placement of tetraodontiforms within

the percomorphs, however additional data from other members of the higher and lower percomorphs, as well as basal actinopterygians would be needed to test this hypothesis.

Hox cluster compression

Comparison of the region that extends from genes 4-10 in amphioxus and the corresponding regions of the *HoxA* clusters from human, *Morone* and *Fugu* shows that there has been an extensive reduction in length of the *HoxA* cluster that occurred subsequent to the divergence of the cephalochordates and vertebrates. Furthermore, it should be noted that among the vertebrates, teleosts appear to have undergone a more extensive reduction than mammals. Previously, Brenner, *et al.* (1993) reported that the *Fugu* genome is compressed approximately 7.6-fold relative to humans, and Aparicio *et al.* (1997) noted that the *Fugu HoxA*, C, and D clusters are shorter than the corresponding clusters in mammals. While it has been assumed that this degree of compression is unique to *Fugu*, the similarity in length of the *Hoxa4-a10* region in comparisons between *Morone* and *Fugu* suggests that the degree of *Hox* cluster compression evident in *Fugu* may be a property shared by these two teleost orders. The likely similarity in the degree of compression among teleosts in general is supported by the physical genomic mapping of 5' *Hox* genes from zebrafish, wherein it was noted that intergenic distances between zebrafish 5' homeodomains in the *HoxA* (Sordino *et al.*, 1996) and *HoxD* (van der Hoeven *et al.*, 1996) clusters are markedly reduced in comparisons to corresponding regions from mouse.

The degree of *Hox* cluster compression among vertebrates relative to cephalochordates is also demonstrated by comparisons of vertebrate *HoxB* clusters. The *HoxB* cluster has the most complete gene content and is the longest of all the vertebrate clusters, containing genes *Hoxb1-b9*, and (in mouse) the recently discovered *Hoxb13*. Due to lack of gene deletions within the *Hoxb1-b9* region, the *HoxB* cluster most closely resembles the single cluster from amphioxus, comprised of genes *AmphiHox-1* through *AmphiHox-10*. In amphioxus, the region comprising genes 1-10 spans a region of 270 kb while the *HoxB* clusters from humans (*Hoxb1-b9*) is 120 kb in length. The length of the *HoxB* cluster from mice is similar to mammals within the *Hoxb1-b9* region, however the *Hoxb-13* gene is located 70 kb downstream from *Hoxb9* (Zeltser, 1996). With the addition of the 70 kb *Hoxb9-b13* intergenic region, the *HoxB* cluster in mice spans a region 190 kb in length. Therefore, the entire *HoxB* cluster from mammals is also 1.5-fold compressed relative to the 1-10 region from amphioxus. Interestingly, while the *Fugu* A, C, and D clusters are shorter than the corresponding clusters in human, the *Fugu* *HoxB* cluster (*Hoxb1-b9*) is similar in length (130 kb) relative to the human *HoxB* cluster. Furthermore, among the 3' *HoxB* cluster genes, McCallum (1997) showed that the *Hoxb2-b3* intergenic distance for *Morone* (8 kb) and *Fugu* (~8 kb) is similar to both mammals (8.5 kb) and the corresponding region in amphioxus (9 kb). This suggests that the *HoxB* cluster in teleosts is not as permissive for genomic compression as the other clusters characterized, and that the intergenic distances established are crucial for gene regulation and therefore conserved.

The marked length reduction of the *Hox* clusters documented in this study, and the significant amplification in *Hox* cluster number in which a single cluster is hypothesized to have generated four clusters (Holland and Garcia-Fernàndez, 1996; Bailey *et al.*, 1997; Nadeau and Sankoff, 1997), seem to have played a role in the evolution of the *Hox* gene clusters. The reciprocal relationship between cluster amplification and cluster length reduction suggests these two events may be related. Although it is premature to infer a causal relationship between cluster length and number, it can be argued that duplication may have relieved the selective pressure responsible for maintenance of a complete *Hox* gene complement, as typified by the archetypal single cluster, and generated a suitable substrate for deletions. If there is a causal relationship, then it can be hypothesized that the extent of cluster amplification should be related to the degree of length reduction. A test of this hypothesis awaits the characterization of the *Hox* gene clusters in organisms with more than one, but fewer than four clusters.

Conclusions

Based on comparisons of the *Hoxa4-a10* region in *Morone* to corresponding regions in other vertebrates, I have shown that there are lineage specific differences among the genes of the *HoxA* cluster, particularly between the actinopterygians and sarcopterygians. These lineage-specific changes within the *HoxA* cluster, such as the loss of *Hoxa6* in actinopterygians, may serve as markers for the morphological differences demonstrated between these two lineages. Previously, it has been suggested that an increase in cluster number possibly allowed for greater complexity of organisms and

permitted morphological differences to develop (Holland, 1992; Pendleton *et al.*, 1993; Ruddle *et al.*, 1994b). However, since it has been shown that there are four clusters in vertebrates, and possibly more than four in zebrafish, this suggests that the hypothesis linking an increase in cluster number and gene content with increasing developmental complexity is not a likely possibility. Therefore specific gene loss events within the *Hox* clusters, as well as the regulation of novel developmental systems arising within the actinopterygians and sarcopterygians, might have contributed to the distinct morphological differences exhibited by the teleosts and tetrapods.

Comparisons of the *Morone Hoxa4-a10* genes with *HoxA* clusters from teleosts and the sequences previously isolated from *Morone saxatilis* PCR-based survey demonstrates the inherent problems with PCR-based gene surveys for conserved gene families. This signifies the need for careful gene assignment based on gene linkage and physical mapping.

This study has also demonstrated the lack of sequence relatedness within the intergenic regions between two disparate lineages, teleosts and mammals. This implies that while the coding regions have been conserved, as well as gene organization within the clusters, the intergenic regions have not been subjected to the same evolutionary constraints since the divergence of actinopterygians from sarcopterygians approximately 400 million years ago. This lack of sequence relatedness between ray-finned teleosts and mammals that is not reflected in comparisons between teleosts, such as *Morone* and *Fugu*, might also support the theory that tetrapods arose from lobe-finned teleost ancestors instead of the ray-finned teleost lineage. If tetrapods arose from the ray-finned

actinopterygians, than it is possible that the sequences would be more conserved between teleosts and mammals since they would have diverged from a more recent common ancestor than the estimated common ancestor that existed prior to the divergence of actinopterygians and sarcopterygians. It is interesting to speculate that if tetrapods indeed arose from the lobe-finned ancestors, the intergenic sequences of lobe-finned fish, as well as gene organization would be similar to the basal tetrapods, such as amphibians and perhaps even mammals. However, it is possible that the mammalian sequences have diverged greatly from the lobe-finned ancestor over time and would only be similar to other closely related groups, such as marsupials or monotremes. Nevertheless, the high degree of conservation of these non-coding intergenic sequences between two teleost sister orders, Tetraodontiformes and Perciformes, suggests that these regions might prove useful for examining phylogenetic relationships among vertebrates. Furthermore, the intergenic comparisons between *Morone* and *Fugu* in this study have also revealed possible recombination event sites, which might have contributed to the decrease in cluster length that occurred throughout vertebrate evolution along with cluster duplication.

REFERENCES

- Acampora, D., M. D'Esposito, A. Faiella, M. Pannese, E. Migliaccio, F. Morelli, A. Stornaiuolo, V. Nigro, A. Simeone, and E. Boncinelli. 1989. The human HOX gene family. *Nucleic Acids Research* **17**: 10385-10402.
- Akam, M. 1989. *Hox* and *HOM*: homologous gene clusters in insects and vertebrates. *Cell* **57**: 347-349.
- Altschul, S., W. Gish, M. Miller, E. W. Myers, and D. J. Lipman. 1990. A basic local alignment search tool. *Journal of Molecular Biology* **215**: 403-410.
- Aparicio, S., A. Morrison, A. Gould, J. Gilthorpe, C. Chaudhuri, P. W. J. Rigby R. Krumlauf, and S. Brenner. 1995. Detecting conserved regulatory elements with the model genome of the Japanese pufferfish, *Fugu rubripes*. *Proceedings of the National Academy of Science, USA* **92**: 1648-1688.
- Aparicio, S., K. Hawker, A. Cottage, Y. Mikawa, L. Zuo, B. Venkatesh, E. Chen, R. Krumlauf, and S. Brenner. 1997. Organization of the *Fugu rubripes* *Hox* clusters: evidence for continuing evolution of vertebrate *Hox* complexes. *Nature Genetics* **16**: 79-83.
- Bailey, W. J., J. Kim, G. P. Wagner, and F. H. Ruddle. 1997. Phylogenetic reconstruction of vertebrate *Hox* cluster duplications. *Molecular Biology and Evolution* **14**: 843-853.
- Benson, G. V., T. H. Nguyen, and R. L. Maas. 1995. The expression pattern of the murine *Hoxa-10* gene and the sequence recognition of its homeodomain reveal specific properties of *Abdominal B*-like genes. *Molecular and Cellular Biology* **15**: 1591-1601.
- Brenner, S., G. Elgar, R. Sandford, A. Macrae, B. Venkatesh, and S. Aparicio. 1993. Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature* **366**: 265-268.
- Brooke, N. M., J. Garcia-Fernández, and P. W. H. Holland. 1998. The *ParaHox* gene cluster is an evolutionary sister of the *Hox* gene clusters. *Nature* **392**: 920-922.
- Brunak, S., J. Engelbrecht, and S. Knudsen. 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *Journal of Molecular Biology* **220**: 49-65.

- Chen, F. and M. R. Capecchi. 1997. Targeted mutations in *Hoxa-9* and *Hoxb-9* reveal synergistic interactions. *Developmental Biology* **181**: 186-196.
- Condie, B. G. and M. R. Capecchi. 1993. Mice homozygous for a targeted disruption of *Hoxd-3* (*Hox-4.1*) exhibit anterior transformations of the first and second cervical vertebrae, the atlas and the axis. *Development* **119**: 579-595.
- Condie, B. G. and M. R. Capecchi. 1994. Mice with targeted disruptions in the paralogous genes *hoxa-3* and *hoxd-3* reveal synergistic interactions. *Nature* **370**: 304-307.
- Duboule, D., and P. Dolle. 1989. The structural and functional organization of the murine *HOX* gene family resembles that of *Drosophila* homeotic genes. *EMBO Journal* **8**: 1497-1505.
- Duboule, D., and G. Morata. 1994. Colinearity and functional hierarchy among genes of the homeotic complexes. *Trends in Genetics* **10**: 358-364.
- Ekker, M., M. D. Speevak, C. C. Martin, L. Joly, G. Giroux., and M. Chevette. 1996. Stable transfer of zebrafish chromosome segments into mouse cells. *Genomics* **33**: 57-64.
- Fjose, A., A. Molven, and H. G. Eiken. 1988. Molecular cloning and characterization of homeobox-containing genes from Atlantic salmon. *Gene* **62**: 141-152.
- Garcia-Fernàndez, J. and P. W. H. Holland. 1994. Archetypal organization of the amphioxus *Hox* gene cluster. *Nature* **370**: 563-566.
- Gish, W. and D. J. States. 1993. Identification of protein coding regions by database similarity search. *Nature Genetics* **3**: 266-272.
- Graham, A., N. Papalopulu, and R. Krumlauf. 1989. The murine and *Drosophila* homeobox clusters have common features of organization and expression. *Cell* **57**: 367-378.
- Hanrahan, D. 1983. Studies on transformation of *Escherichia coli* with plasmids. *Journal of Molecular Biology* **166**: 557-580.
- Holland, P. W. H. 1992. Homeobox Genes in Vertebrate Evolution. *BioEssays* **14**: 267-272.
- Holland, P. W. H., J. Garcia-Fernàndez, N. A. Williams, and A. Sidow. 1994. Gene duplications and origins of vertebrate development. *Development (Supplement)* **125**: 125-133.

- Holland, P. W. H., and J. Garcia-Fernández. 1996. *Hox* genes and chordate evolution. *Developmental Biology* **173**: 382-395.
- Holland, P. W. H. 1997. Vertebrate evolution: Something fishy about *Hox* genes. *Current Biology* **7**: 570-572.
- Janvier, P. Early Vertebrates. Oxford University Press, Inc., New York, 1996.
- Johnson, G. D. and C. Patterson. 1993. Percomorph phylogeny: a survey of Acanthomorphs and a new proposal. *Bulletin of Marine Sciences* **52**: 554-626.
- Kaufman, T. C., M. Seeger, and G. Olsen. 1990. Molecular and genetic organization of the *Antennapedia* gene complex of *Drosophila melanogaster*. *Advanced Genetics* **27**: 309-362.
- Knittel, T., M. Kessel, H. K. Myong, and P. Gruss. 1995. A conserved enhancer of the human and murine *Hoxa-7* gene specifies the anterior boundary of expression during embryonal development. *Development* **121**: 1077-1088.
- Krumlauf, R. 1992. Evolution of the vertebrate *Hox* homeobox genes. *BioEssays* **14**: 245-251.
- Kumar, S., T. Koichiro, and M. Nei. 1993. MEGA: Molecular Evolutionary Genetics Analysis, version 1.01. The Pennsylvania State University, University Park, PA 16802.
- Laughon, A. 1991. DNA binding specificity of homeodomains. *Biochemistry* **30**: 11357-11366.
- Lewis, E. B. 1978. A gene complex controlling segmentation in *Drosophila*. *Nature* **276**: 567-570.
- Little, P. 1993. Small and perfectly formed. *Nature* **336**: 204-205.
- McCallum, J. B. 1997. DNA Sequence Analysis and Characterization of a *Morone saxatilis* (Striped Bass) Genomic Clone Containing the *HoxB2* and *HoxB3* Genes. Master's Thesis, Department of Biology, East Carolina University, Greenville, NC 27858.
- McGinnis, W., M. S. Levine, E. Hafen, A. Kuroiwa, and W. J. Ghering. 1984a. A conserved DNA sequence found in homeotic genes of the *Drosophila Antennapedia* and *Bithorax* complexes. *Nature* **308**: 428-443.

- McGinnis W., R. L. Garber J. Wirz, A. Kuroiwa, and W. J. Ghering. 1984b. A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. *Cell* **37**: 403-408.
- McGinnis, W., and R. Krumlauf. 1992. Homeobox genes and axial patterning. *Cell* **68**: 283-302.
- Meyer, A. 1998. *Hox* gene variation and evolution. *Nature* **391**: 225-228.
- Misof, B. Y. and G. P. Wagner. 1996. Evidence for four *Hox* clusters in the killifish *Fundulus heteroclitus* (Teleostei). *Molecular Phylogenetics and Evolution* **5**: 309-322.
- Misof, B. Y., M. J. Blanco, and G. P. Wagner. 1996. PCR survey of Hox-genes of the zebrafish – new sequence information and evolutionary implications. *Journal of Experimental Zoology* **274**: 193-206.
- Nadeau, J. H. and D. Sankoff. 1997. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* **147**:1259-66.
- Nelson, J. Fishes of the World, third edition. John Wiley & Sons, Inc., New York, 1994.
- Njølstad, P. R., A. Molven, J. Apold, and A. Fjose. 1990. The zebrafish homeobox gene *hox-2.2*: transcription unit, potential regulatory regions and *in situ* localization of transcripts. *EMBO Journal* **9**: 515-524.
- Patterson, C 1993. An overview of the early fossil record of Acanthomorphs. *Bulletin of Marine Sciences* **52**: 29-59.
- Pavell, A. M. and E. J. Stellwag. 1994. Survey of *Hox*-like genes in the teleost *Morone saxatilis*: implications for evolution of the *Hox* gene family. *Molecular Marine Biology and Biotechnology* **3**: 149-157.
- Pendleton, J.W., B. K. Nagai, M. T. Murtha, and F. H. Ruddle. 1993. Expansion of the *Hox* gene family and the evolution of chordates. *Proceedings of the National Academy of Science, U.S.A.* **90**: 6300-6304.
- Prince, V. E., L. Joly, M. Ekker, R. K. Ho. 1998. Zebrafish hox genes: genomic organization and modified colinear expression patterns in the trunk. *Development* **125**: 407-420.

- Rancourt, D. E., T. Tsuzuki, and M. R. Capecchi. 1995. Genetic interaction between *hox-b5* and *hox-b6* is revealed by nonallelic noncomplementation. *Genes and Development* **9**: 108-122.
- Rosen, D. E. 1984. Zeiforms as Primitive Plectognath Fishes. *American Museum Novitates* **2782**: 1-45.
- Ruddle, F. H., J. L. Bartels, K. L. Bentley, C. Kappen, M. T. Murtha, and J. W. Pendleton. 1994a. Evolution of *Hox* Genes. *Annual Review of Genetics* **28**: 423-442.
- Ruddle, F. H., K. L. Bentley, M. T. Murtha, and N. Risch. 1994b. Gene loss and gain in the evolution of the vertebrates. *Development (Supplement)* 155-161.
- Saitou, N. and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**: 406-425.
- Sambrook, J., E. F. Fritsch, and T. Maniatis. Molecular Cloning: a laboratory manual, second edition. Cold Springs Harbor Laboratory Press, New York, 1989.
- Schubert, F. R., K. Nieselt-Struwe, and P. Gruss. 1993. The *Antennapedia*-type homeobox genes have evolved from three precursors separated early in metazoan evolution. *Proceedings of the National Academy of Science, USA* **90**: 143-147.
- Schughart, K., C. Kappen, and F. H. Ruddle. 1989. Duplication of large genomic regions during the evolution of vertebrate *hox* genes. *Proceedings for the National Academy of Science, USA* **86**: 7067-7071.
- Scott, M. P., and A. 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, USA* **81**: 4115-4119.
- Scott, M. P., J. W. Tamkun, and G. W., Hartzell III. 1989. The structure and function of the homeodomain. *Biochemistry and Biophysics Acta*. **989**: 25-48.
- Scott, M. P. 1992. Vertebrate homeobox gene nomenclature. *Cell* **71**: 551-553.
- Sharkey, M, Y. Grabba, and M. P. Scott. 1997. *Hox* genes in evolution: protein surfaces and paralog groups. *Trends in Genetics* **13**: 145-151.

- Shen, W.-F., C.-P. Chang, S. Rozenfeld, G. Sauvageau, R. K. Humphries, L. Ming, H. J. Lawrence, M. L. Cleary, and C. Largman. 1996. *Hox* homeodomain protein exhibits selective complex stabilities with *Pbx* and DNA. *Nucleic Acids Research* **24**: 898-906.
- Shinohara, A., M. Shinohara, T. Ohta, S. Matsuda, and T. Ogawa. 1998. Rad52 forms ring structures and co-operates with RPA in single-strand DNA annealing. *Genes Cells* **3**: 145-156.
- Small, K. M. and S. S. Potter. 1993. Homeotic transformations and limb defects in *Hox All* mutant mice. *Genes and Development* **7**: 2318-2328.
- Sordino, P., D. Duboule, and T. Kondo. 1996. Zebrafish *Hoxa* and *Evx-2* genes: cloning, developmental expression and implications for the functional evolution of posterior *Hox* genes. *Mechanisms of Development* **59**: 165-175.
- Sourdis, J., and C. Krimbas. 1987. Accuracy of phylogenetic trees estimated from DNA sequence data. *Molecular Biology and Evolution* **4**: 159-166.
- Tracy, R. B., J. K. Baumohl, and S. C. Kowalczykowski. 1997. The preference for GT-rich DNA by the yeast Rad51 protein defines a set of universal pairing sequences. *Genes and Development* **11**: 3423-3431.
- Tyler, J. C., and L. Sorbini. 1996. New Superfamily and Three New Families of Tetraodontiform Fishes from the Upper Cretaceous: The Earliest and Most Morphologically Primitive Plectognaths. *Smithsonian Contributions to Paleobiology* **82**: 1-59.
- van der Hoeven, F., P. Sordino, N. Fraudeau, J.-C. Izpisua-Belmonte, and D. Duboule. 1996. Teleost *HoxD* and *HoxA* genes: comparison with tetrapods and functional evolution of the *HOXD* complex. *Mechanisms of Development* **54**: 9-21.
- Wang, B. B., M. M. Muller-Immergluck, J. Austin, N. T. Robinson, A. Chisholm, and C. Kenyon. 1993. A homeotic gene cluster patterns the anteroposterior body axis of *C. elegans*. *Cell* **74**: 29-42.
- 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**: 734-744.
- Zeltser, L., C. Desplan, and N. Heintz. 1996. *Hoxb-13*: a new *Hox* gene in a distant region of the *HOXB* cluster maintains colinearity. *Development* **122**: 2475-2484.

Figure 1. Schematic diagram of the phylogenetic relationships among amphioxus, *Fugu rubripes* (order Tetraodontiformes), *Morone saxatilis* (order Perciformes), zebrafish (order Cypriniformes), and mammals.

Figure 1. Phylogenetic relationships of organisms for which the *Hox* gene clusters have been characterized.

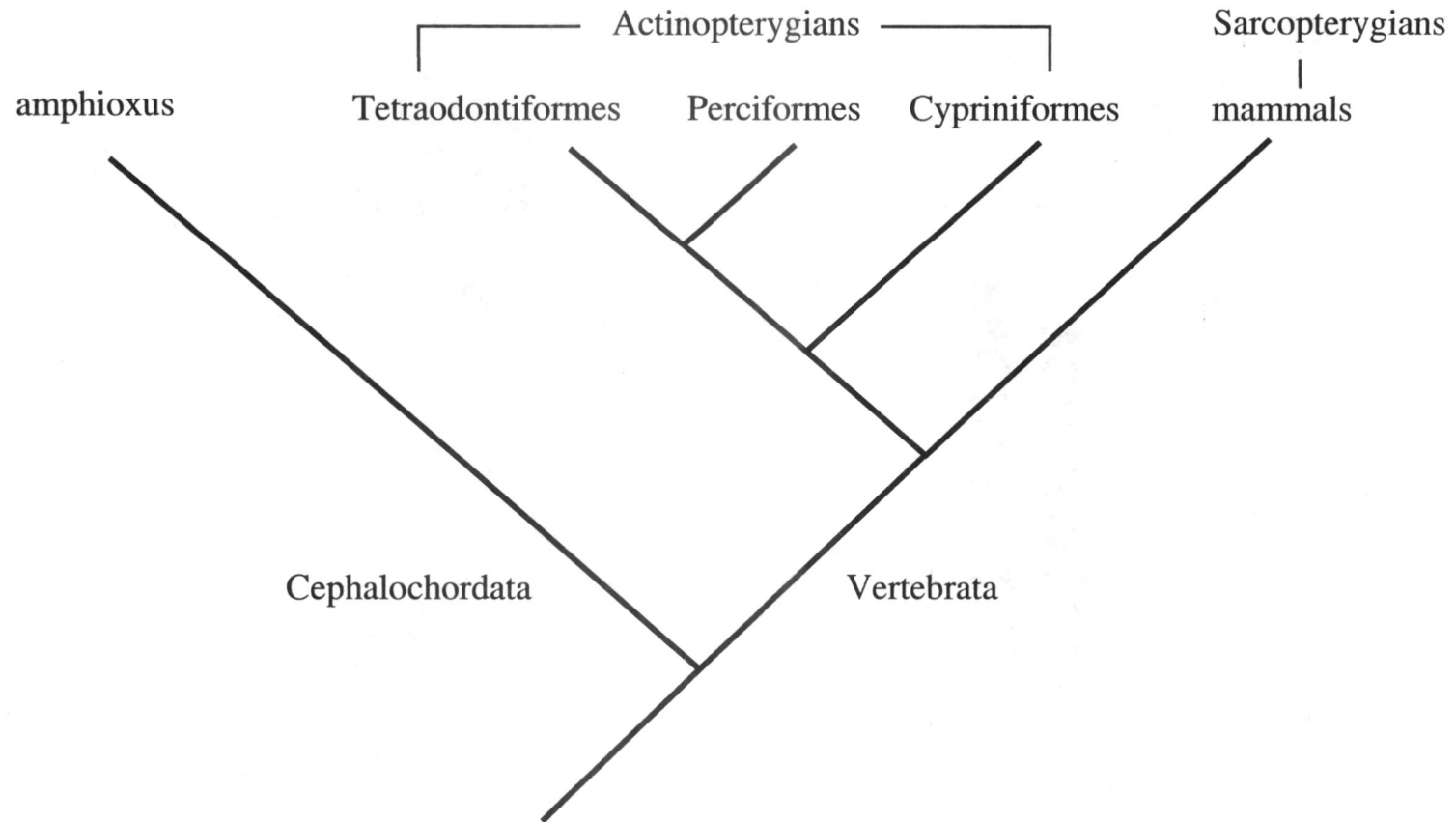


Figure 2. Schematic diagram of *Hox* cluster organization in amphioxus (*Branchiostoma floridae*), mammals (mice and humans), the Japanese pufferfish (*Fugu rubripes*), and zebrafish (*Danio rerio*). Hatched boxes represent genes in *Fugu* and zebrafish that have no mammalian orthologs. Stippled boxes represent pseudogenes. Ovals represent genes that are identified beyond the vertebrate *HoxA-D* clusters. The empty box represents the putative *Hoxa7* gene identified from *Fugu* in this study.

Figure 2. Schematic diagram of *Hox* cluster organization in specific vertebrates and amphioxus.

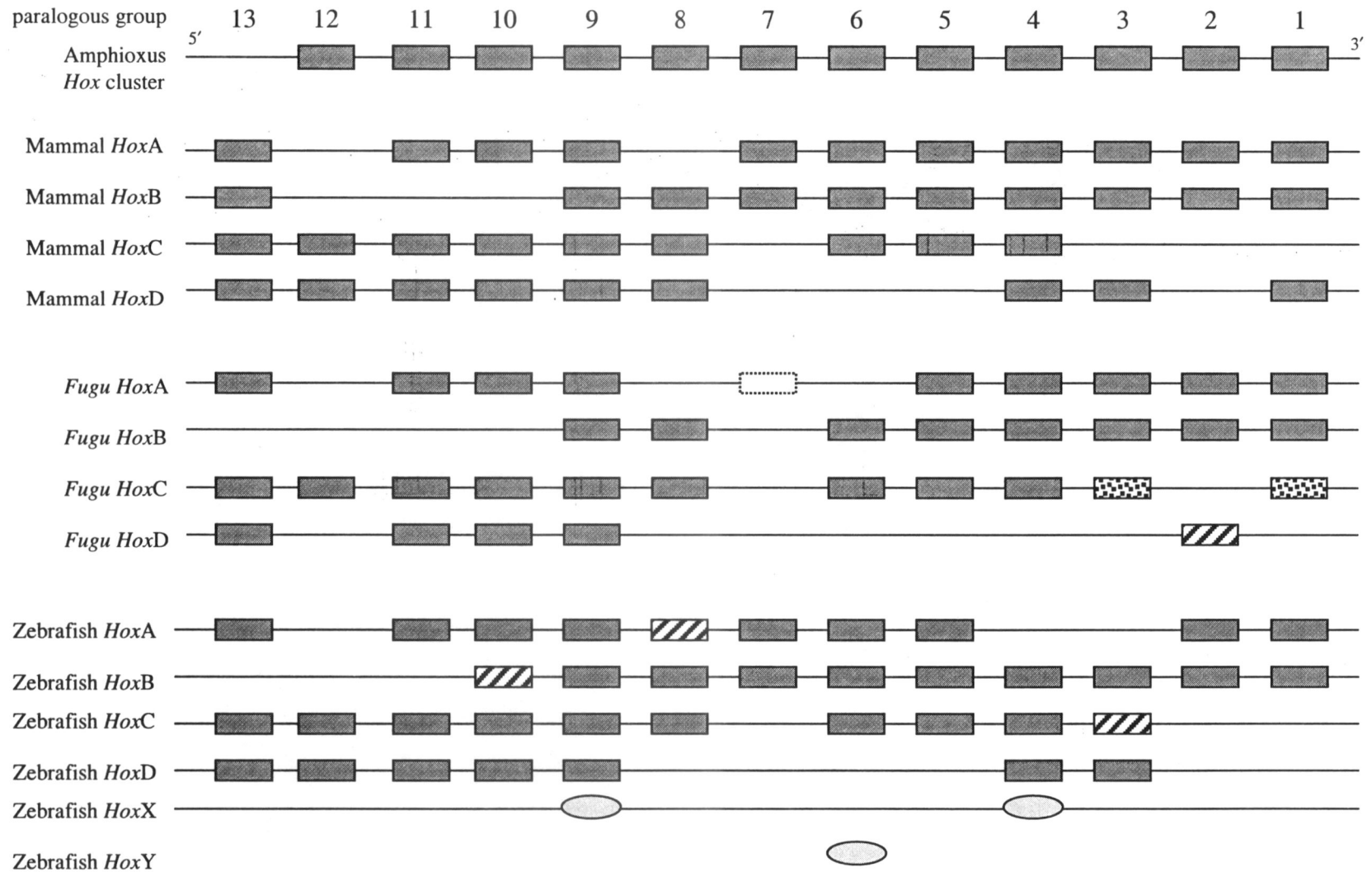


Figure 3. Restriction endonuclease cleavage map of λ clones 7H and 8H. Fragments that hybridized to a homeogeneous *Hox*-specific probe are indicated by a box. Fragments that hybridized strongly are indicated by a dark box, and the fragment that hybridized weakly is indicated by a lighter box. The dotted line indicates the *SalI* cleavage site that is part of the λ 8H cloning site, but is not present in λ clone 7H. The single *AvaII* site is shown to indicate the *SalI/AvaII* intron-specific fragment used for *Probe #3*. The scale bar indicates 5000 base pairs.

Figure 3. Restriction endonuclease maps of λ clones 7H and 8H.

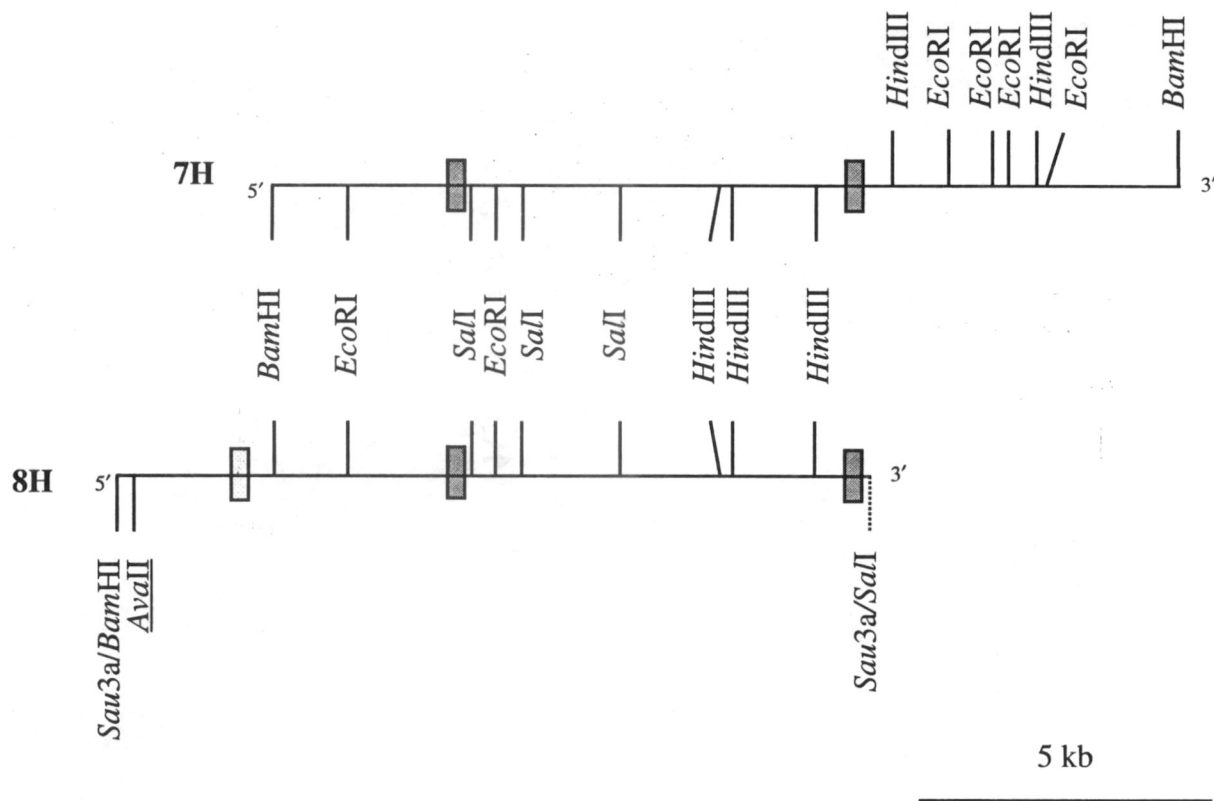


Figure 4. Schematic genomic organization of genes *Hoxa4-a10* in *Morone saxatilis*. For each gene, an empty box indicates exon1 and hatched boxes indicate the homeobox-containing exon2. Gene arrangement is drawn to scale with intergenic distances indicated in parentheses and intron distances given beneath. The positions of six representative overlapping phage λ clones are shown. The dotted lines and arrows indicate the region of phage λ clone 11H (approximately 15 kb in length) which extends toward the presumptive *Hoxa3* region of the genome.

Figure 4. Genomic organization of the *Hoxa4-a10* region in *M. saxatilis*.

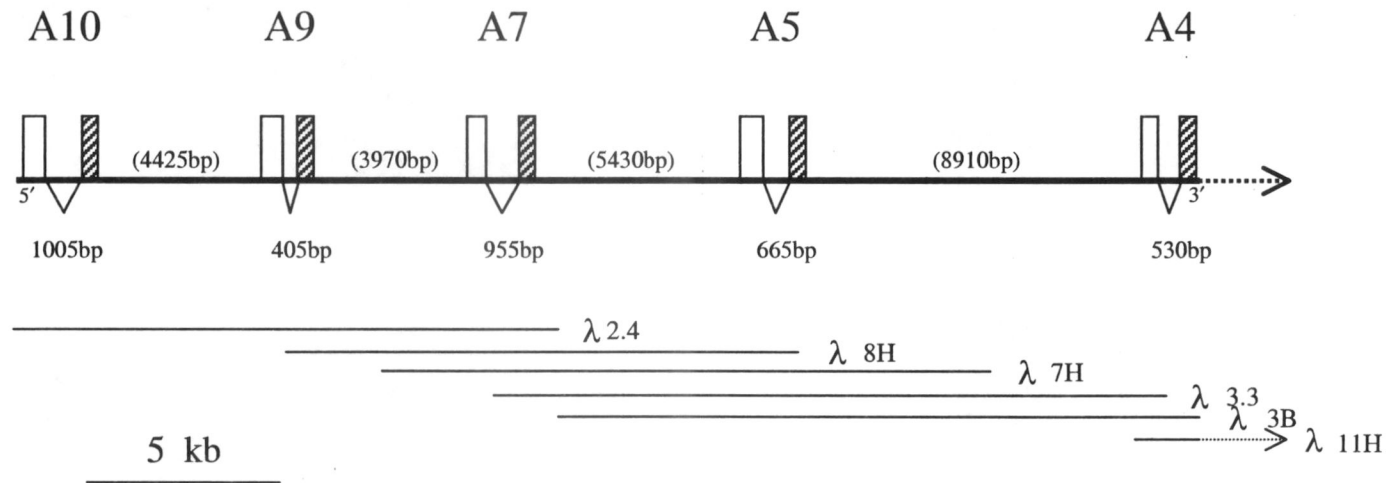


Figure 5. Unrooted evolutionary tree of complete homeodomain sequences (60 a.a.) from paralogous groups 4-10 of the single cluster of amphioxus, and clusters *HoxA-D* of *Morone saxatilis*, mouse, human, and other teleosts. The distance matrix resulted from calculation of pair-wise comparisons of amino acid differences. The unrooted evolutionary tree was constructed by applying the Neighbor-Joining program to the distance matrix. *Morone Hoxa4-a10* genes are shown in bold.

Figure 5. Unrooted evolutionary tree comparing homeodomain sequences.

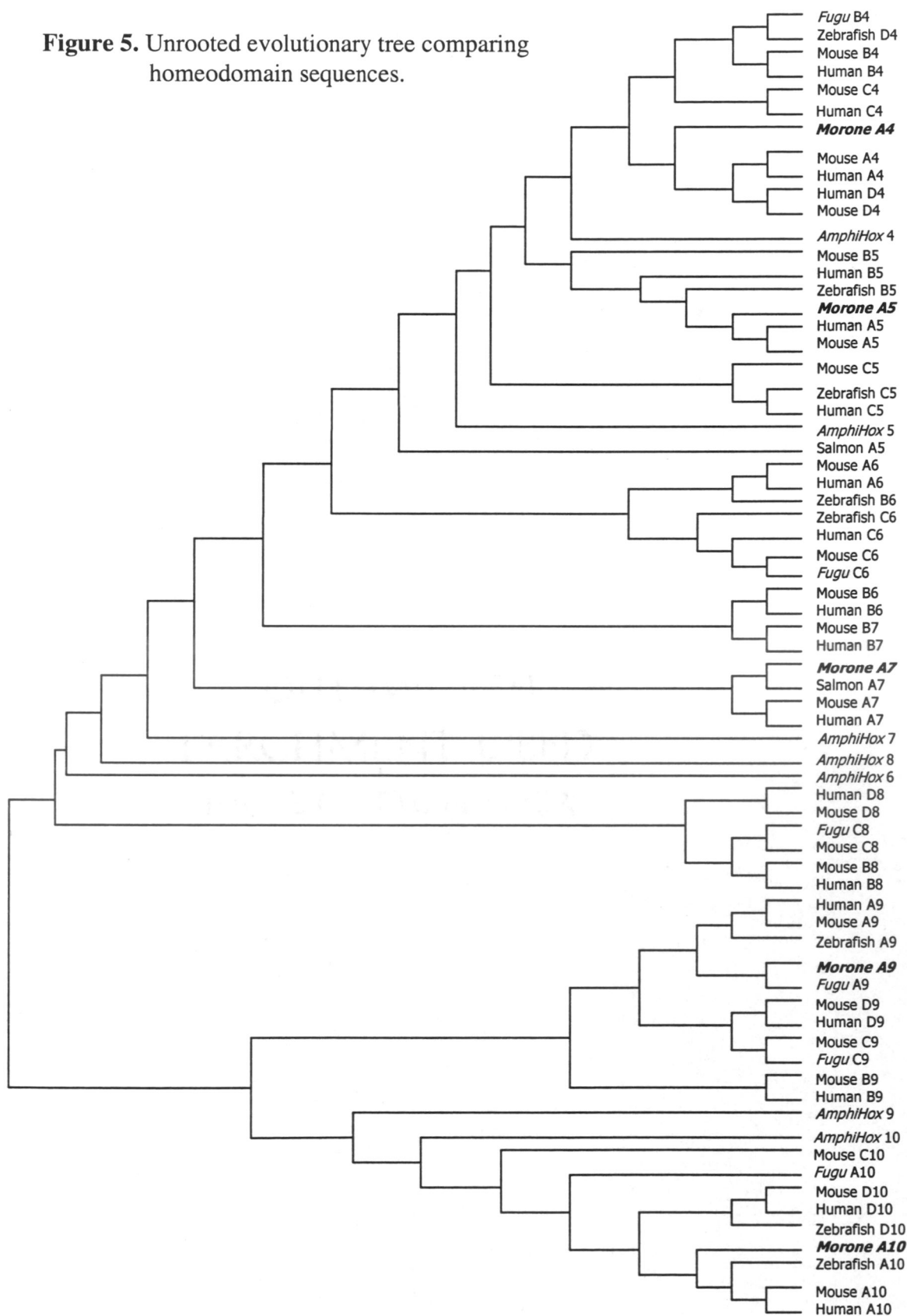


Figure 6. Genomic Southern blots from digestion with *Eco*RI and *Hind*III. Each blot represents genomic *M. saxatilis* DNA that had been digested and probed the different λ probes indicated. Hybridizing fragments are indicated in base pairs.

Figure 6. Genomic Southern blots

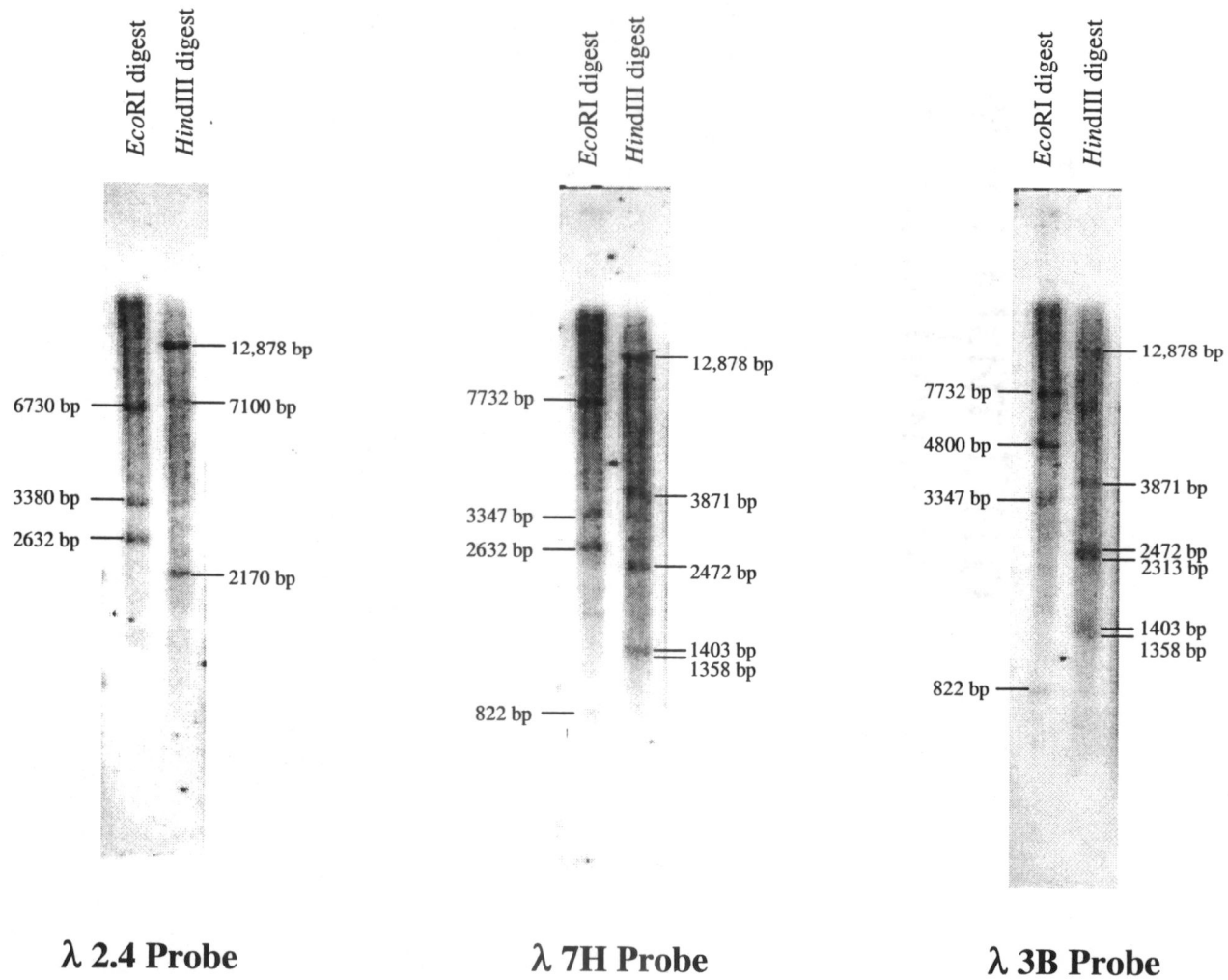


Figure 7. Predicted restriction endonuclease map of the 31 kb contig for *Eco*RI and *Hind*III. The positions of the three λ clones used as probes are shown. The scale bar represents 5000 base pairs.

Figure 7. Restriction endonuclease map for the *HoxA* 31 kb contig.

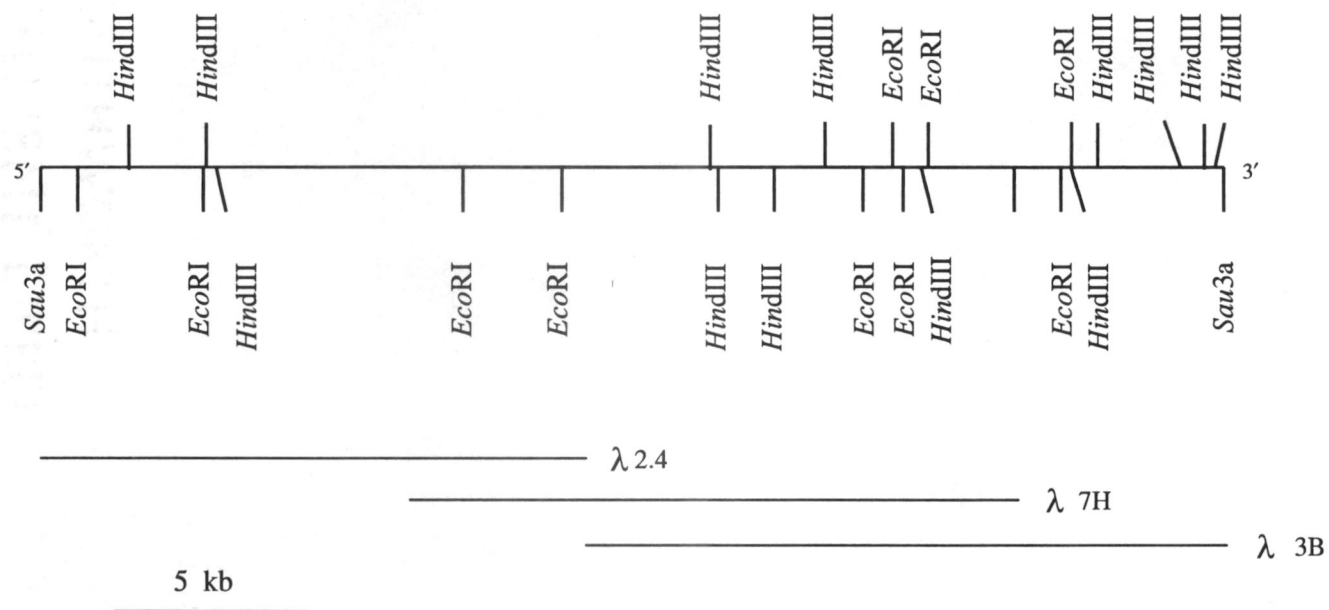


Figure 8. Comparison of the *Morone saxatilis* *HoxA* cluster organization (genes *Hoxa4-a10*) to corresponding regions of amphioxus and the *HoxA* cluster of other vertebrates. The empty box indicates the presence of a presumptive *Hoxa7* gene, or gene fragment, from *Fugu rubripes* as suggested in this thesis. The hatched box represents the zebrafish *Hoxa8* gene for which there is no vertebrate ortholog. The dashed line of the zebrafish *HoxA* cluster indicates that physical linkage of these genes has not been established.

Figure 8. Comparison of vertebrate *HoxA* cluster organization and amphioxus.

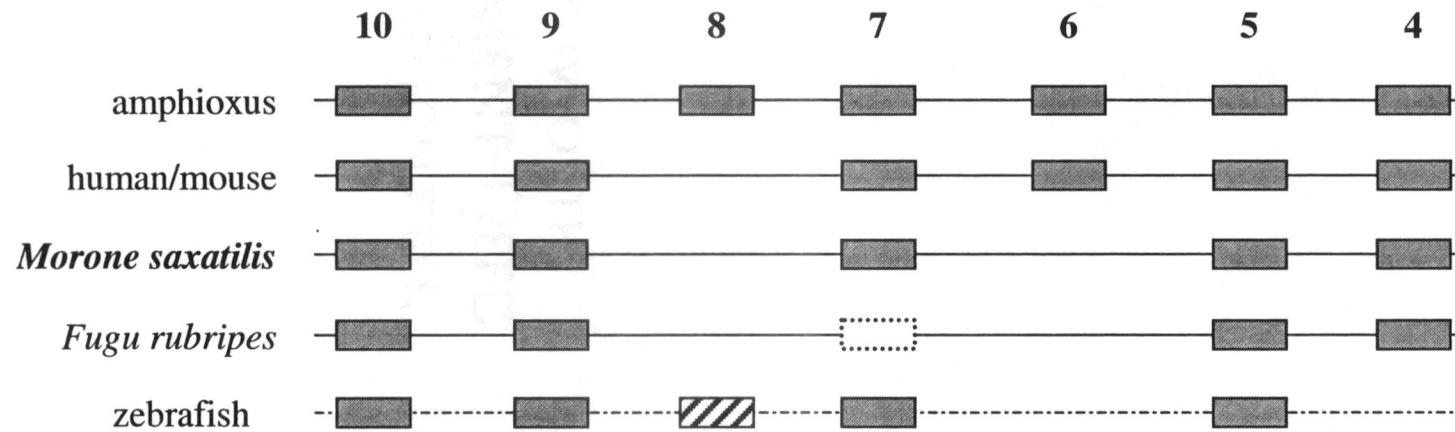


Figure 9. PILEUP alignment (parameter set for *in situ*) of the complete amino acid sequence of *Hoxa10* exon1 from *Morone saxatilis* and other vertebrates. Gaps that were inserted to maximise alignment are indicated by a dot (.). The *Morone saxatilis Hoxa10* exon1 is indicated in bold lettering.

Figure 9. PILEUP alignment of vertebrate *Hoxa10* exon1 amino acid sequences.

		1																	80
humanA10	exon1	MSCSESPAAN	SFLVDSLIS	GRGEAGGGG	GAGGGGGGGY	YAHGGVYLPP	AADLPYGLQS	CGLFPTLGK	RNEAASPGSG										
mouseA10	exon1	MSCSESPAAN	SFLVDSLIS	GRGEAGVGGG	SA.GGGGGGGY	YAHGGVYLPP	ASDLPYGLQS	CGLFPALGSK	RNEAPSPGGG										
MoroneA10	exon1	MSSME.....																	
FuguA10	exon1	MSSME.....																	
		81																	160
humanA10	exon1	GGG..GGLGP	GAHGYGPSPI	DLWLDAPRSC	RMEPPDGPPP	P...PQQQ..	..PPPPPQPP	QPAPQATSCS	FAQNIKEESS										
mouseA10	exon1	GGGGSGGLGP	GTHGYAPAPL	DLWLDAPRSC	RMEPPDGPPP	PQPQPQQQQQ	QPPPPPPQPP	QPQPQATSCS	FAQNIKEESS										
MoroneA10	exon1			.MWMDAQRSC	RMD.....		QE	PVGPQVAPCS	FPQNIKEEST										
FuguA10	exon1			.MWMDTQRSC	RMD.....		QE	PVGSQVAPCS	FTQNIKEEST										
		161																	240
humanA10	exon1	YCLYDSADKC	PKVSA....T	AAELAPFPRG	PP....PDGC	ALGTSSGVPV	PGYFRLSQAY	GTAKGYGSGG	GGAQQLGAGP										
mouseA10	exon1	YCLYDAADKC	PKGSA.....	AADLAPFPRG	PP....PDGC	ALGASSGVPV	PGYFRLSQAY	GTAKGFSGG	GGTEQL.ASP										
MoroneA10	exon1	YCLYEQ.PKC	PKASTAEDLT	YSRLTTTGLG	SSTCTVTDGG	GGGGGGSVPV	PGYFRLSQTH	AHSHKLYNTN	GPQPNPSSH										
FuguA10	exon1	YCLYEQ.PKC	PKASTTEDLA	YSRLTSASLG	STS.....														
		241																	320
humanA10	exon1	FPAQPPGRGF	DLPPALASGS	ADAARKERAL	DSPPPPTLAC	GSGGGSQGDE	EAHASSSAAE	ELSPAPSESS	KASPEKDSL										
mouseA10	exon1	FPAQPPGRGF	DPPPALASGS	TEAAGKERV	DSTPPPTLVC	TGGGGSQGDE	EAHASSSAAE	ELSPAPSENS	KASPEKDSL										
MoroneA10	exon1	FGLHPTAPAR	FHTPPTSTST	PATAQEGRNT	EEPVSSGNAD	LQPHAAPGTE	EARESSDAEV	SSADETEEHD	KERQAKSTG										
FuguA10	exon1																		

Figure 10. PILEUP alignment (parameters set for *in situ*) of the complete amino acid sequence of *Hoxa9* exon1 from *Morone saxatilis* and other vertebrates. Gaps that were inserted to maximize alignment are indicated by a dot (.). The *Morone saxatilis Hoxa9* exon1 is shown in bold lettering.

Figure 11. PILEUP alignment (parameters set for *in situ*) of the complete amino acid sequence of *Hoxa7* exon1 from *Morone saxatilis* and other vertebrates. Gaps that were inserted to maximize alignment are indicated by a dot (.). The conserved hexapeptide sequence is highlighted in black with white lettering. The *Morone saxatilis Hoxa7* exon1 is shown in bold lettering and the presumptive *Hoxa7* exon1 from *Fugu rubripes* is shown in italics. Dashes (-) indicate reading frame shifts in the alignment of the *Fugu Hoxa7* exon1.

Figure 11. PILEUP alignment of vertebrate *Hoxa7* exon1 amino acid sequences.

```

                                1                                80
quailA7 exon1 MSSSYVNAL FSKYTAGASL FQNAEPTSCS FAS.NSQ.RS GYGPGAGAFA SSMPGLYNVN STIYQS..PF SSGYSLGSDA
frogA7 exon1 MTSSYYVSAL FSKFTAGASL FPNPEPTSCS SLPNNSQ.RG AYSGGTGFPF SSVPSLY..S SPLYQN..PF PGYSLASDS.
humanA7 exon1 MSSSYVNAL FSKYTAGTSL FQNAEPTSCS FAP.NSQ.RS GYGAGAGAFA STVPGLYNVN SPLYQS..PF ASGYGLGADA
mouseA7 exon1 MSSSYVNAL FSKYTAGASL FQNAEPTSCS FAP.NSQ.RS GYGPGAGAFA STVPGLYNVN SPLYQS..PF ASGYGLGADA
MoroneA7 exon1 M.SSYVDGL LSKYTAGSSL FPNVERASCS IGPSGEDYGS ARTTAIAFA PSLSGVYSVS DAVYQSRPVF TSGYGQGQDA
+1 FuguA7 exon1 M.SSYNVDAL FSKYTTAWLL          ASCR IGPNGEDYGS ARVALADA.. .SLCGLHNIN NAVYHSRTL
+2 FuguA7 exon1                - FPNAER-
+3 FuguA7 exon1

                                81                                131
quailA7 exon1 Y.NLHCSSFD QNIPVLCNDL TKPSCEKAE E RQPAQPGRGQ FRIYPWMRSS G
frogA7 exon1 Y.NLHCSSFD QNIPLLCNEL .....PKADE AALHHQPD SH FRIYPWMRSS G
humanA7 exon1 YGNLPCASYD QNIPGLCSDL AKGACDKTDE GALHGAAE AN FRIYPWMRSS G
mouseA7 exon1 Y.NLPCASYD QNIPGLCSDL AKGACDKADE GVLHGPAE AS FRIYPWMRSS G
MoroneA7 exon1 H.TLHCSLFD Q.....SRL FTDSCCHPEP GPLTSPPDKQ YRMYPWMRAS G
+1 FuguA7 exon1
+2 FuguA7 exon1 H.S...NLSD R.....IRL I.DSCCQSQP GP
+3 FuguA7 exon1                -ITSPPAKR THIYP*MRAS G

```


Figure 12. PILEUP alignment (parameters set for *in situ*) of the complete amino acid sequence of *Hoxa5* exon1 from *Morone saxatilis* and other vertebrates. The conserved hexapeptide sequence is highlighted in black with white lettering. Gaps that were inserted to maximize alignment are indicated by a dot (.). The *Morone saxatilis Hoxa5* exon1 is shown in bold lettering.

Figure 12. PILEUP alignment of vertebrate *Hoxa5* exon1 amino acid sequences.

	1																		80
humanA5 exon1	MSSYFVNSFC	GRYPNGPDYQ	LHNYGDHSSV	SEQFRDSASM	HSGRYGYGYN	GMDLSVGRSG	SGHFGSGERA	RSYAASASAA											
mouseA5 exon1	MSSYFVNSFC	GRYPNGPDYQ	LHNYGDHSSV	SEQFRDSASM	HSGRYGYGYN	GMDLSVGRSG	SGHFGSGERA	RSYAAGASAA											
MoroneA5 exon1	MSSYFVNSFC	GRYPNGADFQ	LHNYGDHSTA	NDQYRDSTAM	HSSRYGYGYN	GMDLTVGRAG	TGHFVGNERT	QGYSPSHSAA											
	81																		160
humanA5 exon1	...PAEP.RY	SQPA....TS	THSPQPDPLP	CSAVA.PSPG	SDT...HHGG	KNSLSN..SS	GASADAGSTH	ISSREGVGTA											
mouseA5 exon1	...PAEP.RY	SQPA....TS	THSPPPDPLP	CSAVA.PSPG	SDS...HHGG	KNSLGN..SS	GASANAGSTH	ISSREGVGTA											
MoroneA5 exon1	TTPSVEPVRY	TQSANSTGTS	SLSPPPDPLP	CSSVASSSPV	TETQSQHRAV	KNSITTPCST	PCSSSNGGTL	LLNRDCVSKA											
	161																		201
humanA5 exon1	SGAEEDAPAS	SEQASAQSEP	SPAPPAQPQI	YPWMRK	LHIS	H													
mouseA5 exon1	SAAEEDAPAS	SEQAGAQSEP	SPAPPAQPQI	YPWMRK	LHIS	H													
MoroneA5 exon1	SPLEEEKPAG	SAPTTPQNV.	..SDSTQPQI	YPWMRK	LHIN	H													

Figure 13. PILEUP alignment (parameters set for *in situ*) of the complete amino acid sequence of *Hoxa4* exon1 from *Morone saxatilis* and other vertebrates. The conserved hexapeptide sequence is highlighted in black with white lettering. Gaps that were inserted to maximize alignment are indicated by a dot (.). The *Morone saxatilis Hoxa4* exon1 is shown in bold lettering.

Figure 14. PILEUP alignment (parameters set for *in situ*) of the complete amino acid sequences of exon1 from the *Morone saxatilis Hoxa4-a10* genes. The conserved hexapeptide sequences of the central class genes are highlighted in black with white lettering, and the conserved 5' octamer sequences are indicated in bold lettering. Gaps that were inserted to maximize alignment are indicated by a dot (.).

Figure 14. PILEUP alignment of the amino acid sequences of exon1 from the *Morone saxatilis* *Hoxa4-a10* genes.

[illegible]

Figure 15. PILEUP alignment (parameters set for *in situ*) of the complete amino acid sequence of *Hoxa10* exon2 from *Morone saxatilis* and other vertebrates. The conserved 60 amino acids of the homeodomain are highlighted in black with white lettering. Gaps that were inserted to maximize alignment are indicated by a dot (.). The *Morone saxatilis Hoxa10* exon2 is shown in bold lettering.

[illegible]

Figure 16. PILEUP alignment (parameters set for *in situ*) of the complete amino acid sequence of *Hoxa9* exon2 from *Morone saxatilis* and other vertebrates. The conserved 60 amino acids of the homeodomain are highlighted in black with white lettering. Gaps that were inserted to maximize alignment are indicated by a dot (.). The *Morone saxatilis Hoxa9* exon2 is shown in bold lettering.

Figure 16. PILEUP alignment of vertebrate *Hoxa9* exon2 amino acid sequences.

	1								80			
humanA9 exon2	DNPAANWLHA	RS	TRKKRCPY	TKHQ	TLELEK	EFLF	NMYLTR	DDRYEVARLL	NLTERQVKIW	FQNR	RMKMKK	INKDRA.KDE
mouseA9 exon2	DNPAANWLHA	RS	TRKKRCPY	TKHQ	TLELEK	EFLF	NMYLTR	DDRYEVARLL	NLTERQVKIW	FQNR	RMKMKK	INKDRA.KDE
MoroneA9 exon2	DNPVSNWLHA	SAT	TRKKRCPY	TKHQ	ILELEK	EFLF	NMYLTR	DDRYEVARLL	NLTERQVKIW	FQNR	RMKMKK	FNKDGAPKDE
FuguA9 exon2	DNPVSNWLHA	SAT	TRKKRCPY	TKHQ	ILELEK	EFLF	NMYLTR	DDRYEVARLL	NLTERQVKIW	FQNR	RMKMKK	FNKDGAPKDD

Figure 17. PILEUP alignment (parameters set for *in situ*) of the complete amino acid sequence of *Hoxa7* exon2 from *Morone saxatilis* and other vertebrates. The conserved 60 amino acids of the homeodomain are highlighted in black with white lettering. Gaps that were inserted to maximize alignment are indicated by a dot (.). The *Morone saxatilis Hoxa7* exon2 is shown in bold lettering.

Figure 17. PILEUP alignment of vertebrate *Hoxa7* exon2 amino acid sequences.

	1								80
quailA7 exon2	GPDRKRGRQT	YTRYQTLELE	KEFHFNRYLT	RRRIEYAHA	LCLTERQIKI	WFQNRRMKWK	KEHKEESSST	PAPNESRQQQ	
frogA7 exon2	.PDRKRGRQT	YTRYQTLELE	KEFHFNRYLT	RRRIEIAHA	LCLTERQIKI	WFQNRRMKWK	KEHKEESDQT	PDAGEESTAP	
humanA7 exon2	GPDRKRGRQT	YTRYQTLELE	KEFHFNRYLT	RRRIEIAHA	LCLTERQIKI	WFQNRRMKWK	KEHKDEGPTA	AAAPEGAVPS	
mouseA7 exon2	GPDRKRGRQT	YTRYQTLELE	KEFHFNRYLT	RRRIEIAHA	LCLTERQIKI	WFQNRRMKWK	KEHKDESQAP	TAAPEDAVPS	
MoroneA7 exon2	DPTKRGRQT	YSRYQTLELE	KEFHFNRYLT	RRRIEIAHA	LCLSERQIKI	WFQNRRMKWK	KDHKEEPVST	PSGEKCDIQ	

	81			117
quailA7 exon2	RLWRKWKMMK	KRRIRESHSR	QQSLEITYAR	DSCKSSL
frogA7 exon2		...TTTA...		DDDKDKE
humanA7 exon2		...AAATAAA	DKADEEDDDE	EEEEDEEE
mouseA7 exon2		...VSTAADK	ADEEEEEEEEE	EEEEEEEE
MoroneA7 exon2		...PVSEAAE	PKAVNNPERR	FGVRN.G

Figure 18. PILEUP alignment (parameters set for *in situ*) of the complete amino acid sequence of *Hoxa5* exon2 from *Morone saxatilis* and other vertebrates. The conserved 60 amino acids of the homeodomain are highlighted in black with white lettering. Gaps that were inserted to maximize alignment are indicated by a dot (.). The *Morone saxatilis Hoxa5* exon2 is shown in bold lettering.

Figure 18. PILEUP alignment of vertebrate *Hoxa5* exon2 amino acid sequences.

	1		80
humanA5 exon2	DNIGGPE	GKR ARTAYTRYQT LELEKEFHFN RYLTRRRRIE IAHALCLSER QIKIWFQNR	MKWKKDNKLK SMSMAAA.GG
mouseA5 exon2	DNIGGPE	GKR ARTAYTRYQT LELEKEFHFN RYLTRRRRIE IAHALCLSER QIKIWFQNR	MKWKKDNKLK SMSMAAA.GG
MoroneA5 exon2	D.LAGPE	GKR ARTAYTRYQT LELEKEFHFN RYLTRRRRIE IAHALCLSER QIKIWFQNR	MKWKKDNKLK SMNMAAAGG
	84		
humanA5 exon2	AFRP		
mouseA5 exon2	AFRP		
MoroneA5 exon2	GYRP		

Figure 19. PILEUP alignment (parameters set for *in situ*) of the complete amino acid sequence of *Hoxa4* exon2 from *Morone saxatilis* and other vertebrates. The conserved 60 amino acids of the homeodomain are highlighted in black with white lettering. Gaps that were inserted to maximize alignment are indicated by a dot (.). The *Morone saxatilis Hoxa4* exon2 is shown in bold lettering.

Figure 19. PILEUP alignment of vertebrate *Hoxa4* exon2 amino acid sequences.

	1												80
chickA4 exon2	VNPNYSGGEP	KRSRTAYTRQ	QVLELEKEFH	FNRYLTRRRR	IEIAHTLCLS	ERQVKIWFQN	RRMKWKDHHK	LPNTKMRSSN					
humanA4 exon2	VNPSYNGGEP	KRSRTAYTRQ	QVLELEKEFH	FNRYLTRRRR	IEIAHTLCLS	ERQVKIWFQN	RRMKWKDHHK	LPNTKMRSSN					
mouseA4 exon2	VNSSYNGGEP	KRSRTAYTRQ	QVLELEKEFH	FNRYLTRRRR	IEIAHTLCLS	ERQVKIWFQN	RRMKWKDHHK	LPNTKMRSSN					
MoroneA4 exon2	VNASYTGCV	KRSRTAYTRQ	QALELEKEFH	FNRYLTRRRR	VEIAHTMCLS	ERQVKIWFQN	RRMKWKKEHK	LPNTKIRSSS					

	81						115
chickA4 exon2	QPSL.GQQQA	KAQTQG...	H	PRPLDGAA.P	NAAAL		
humanA4 exon2	SASASAGPPG	KAQTQSPHLH	PHPHPSTSTP	VPSSI			
mouseA4 exon2	TASAPAGPPG	KAQTHSPHPH	PHPLPGASTP	IPSSI			
MoroneA4 exon2	SASSSASGAQ	QQQIKTGQ..	QLVPTP	CTAGL			

Figure 20. PILEUP alignment (parameters set for *in situ*) of the complete exon2 amino acid sequences from the *Morone saxatilis Hoxa4-a10* genes. The conserved 60 amino acids of the homeodomain are highlighted in black with white lettering. Gaps that were inserted to maximize alignment are indicated by a dot (.).

Figure 20. PILEUP alignment of the exon2 amino acid sequences from the *Morone saxatilis* *Hoxa4-a10* genes.

[illegible]

Figure 21. Pair-wise similarity plot comparing the *Hoxa9-a7* intergenic regions in human (x-axis) and *Morone saxatilis* (y-axis).

Figure 21. Pair-wise similarity plot comparing the *Hoxa9-a7* intergenic regions from human and *Morone*.

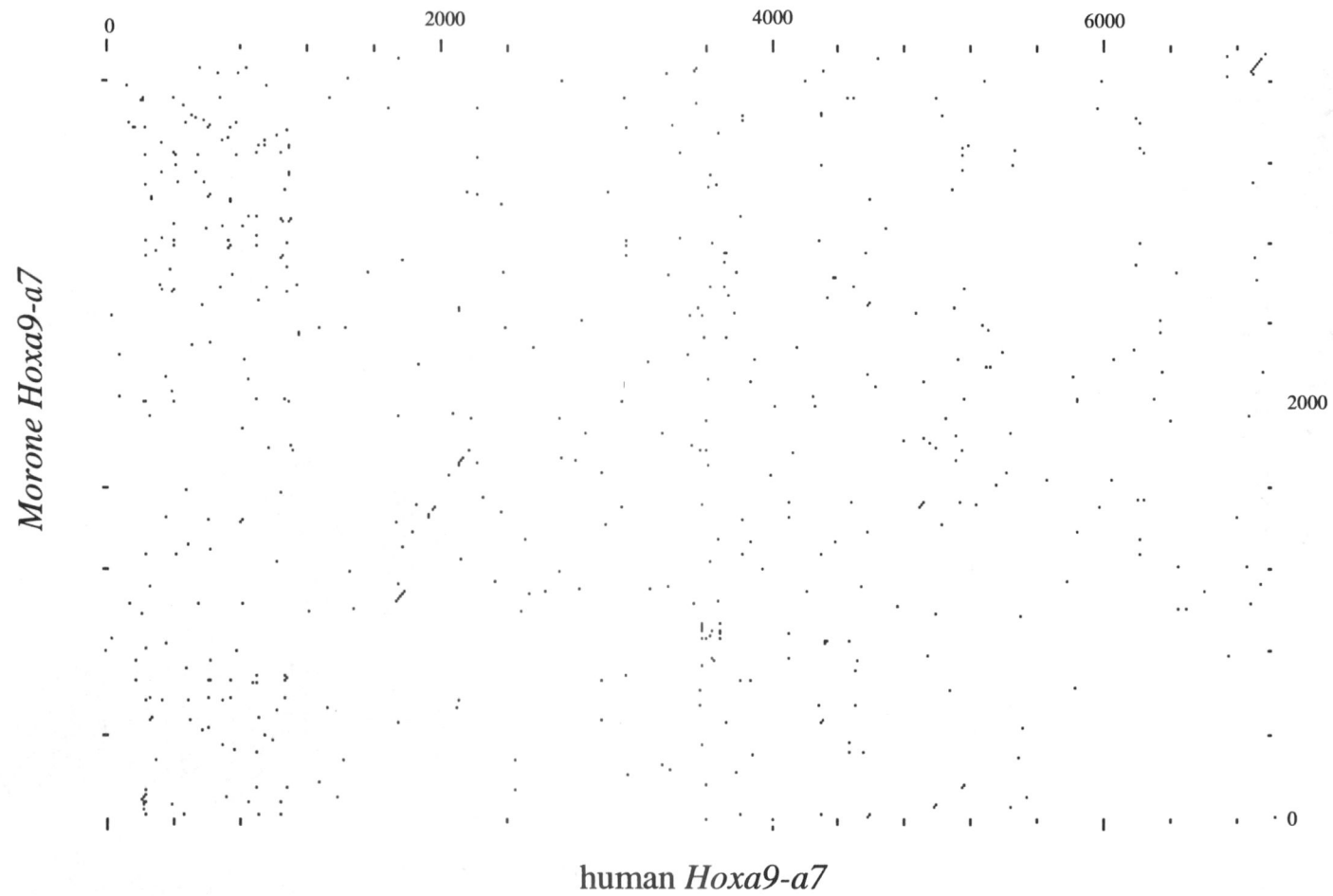


Figure 22. Pair-wise similarity plot comparing the *Hoxa9-a7* intergenic regions in *Fugu rubripes* (x-axis) and human (y-axis).

Figure 22. Pair-wise similarity plot comparing the *Hoxa9-a7* intergenic regions from human and *Fugu*.

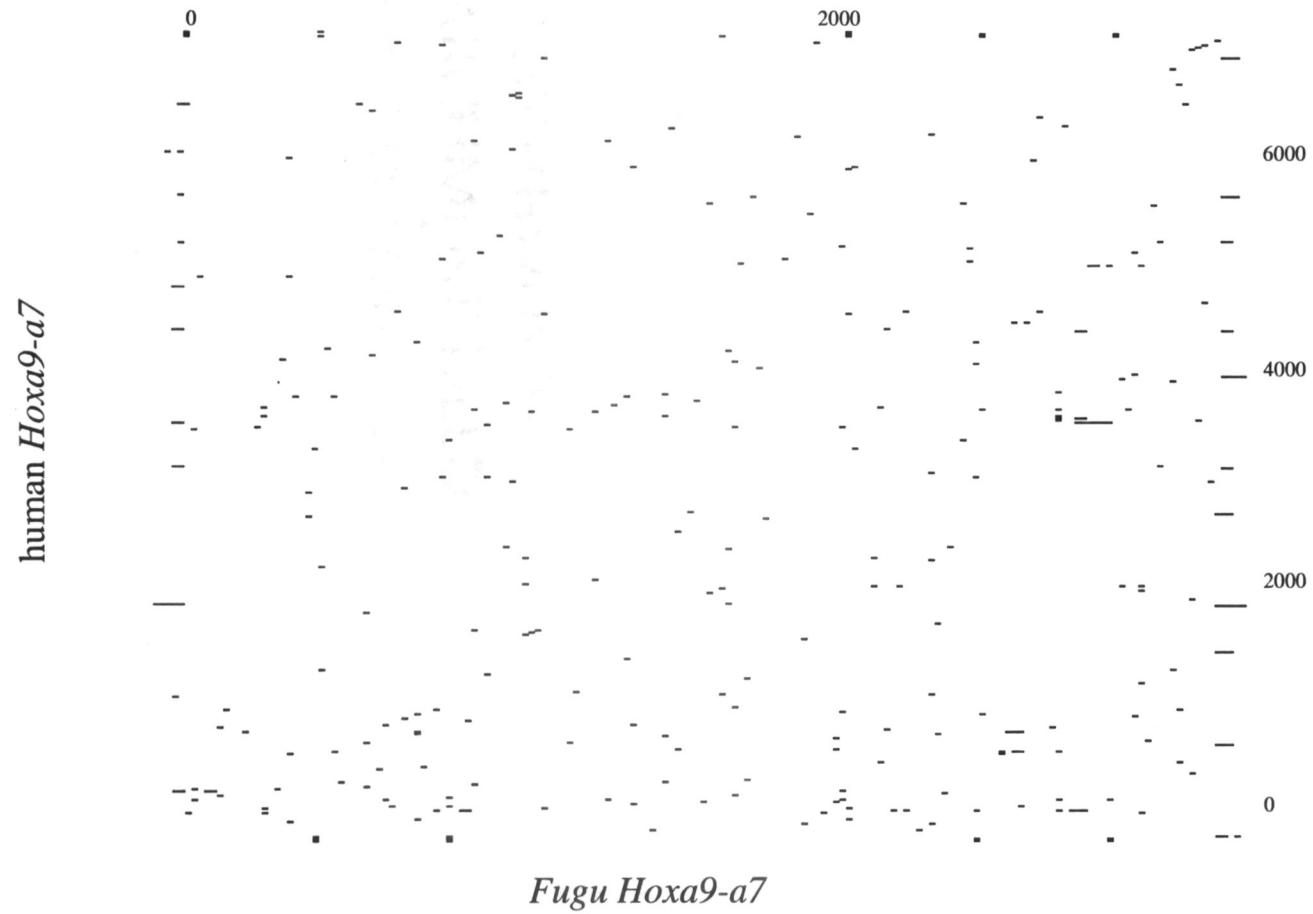




Figure 23. Pair-wise similarity plot comparing the *Hoxa10-a9* intergenic regions in human (x-axis) and *Morone saxatilis* (y-axis).

Figure 23. Pair-wise similarity plot comparing the *Hoxa10-a9* intergenic regions from human and *Morone*.

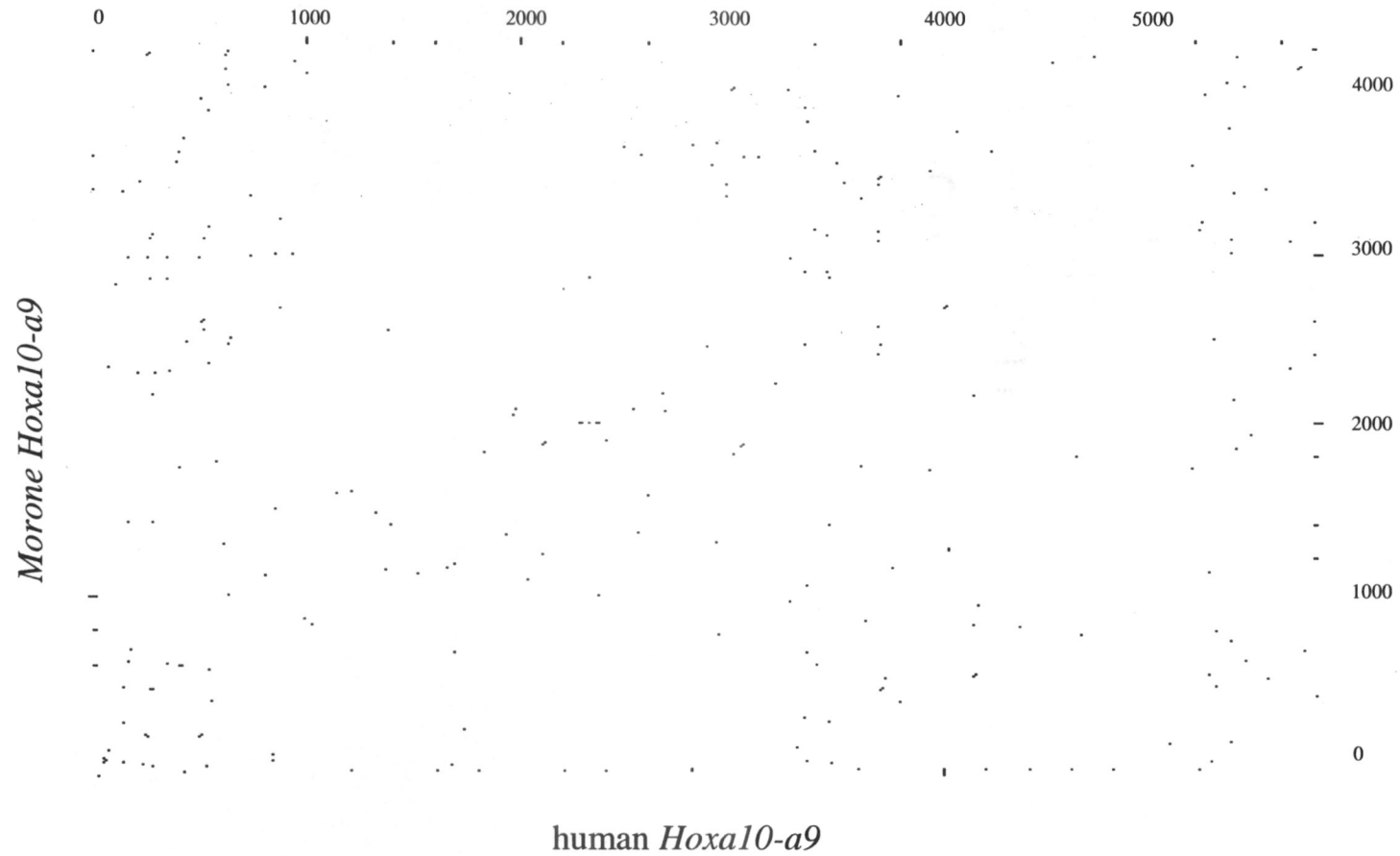


Figure 24. Pair-wise similarity plot comparing the *Hoxa10-a9* intergenic regions in *Fugu rubripes* (x-axis) and human (y-axis).

Figure 24. Pair-wise similarity plot comparing the *Hoxa10-a9* intergenic regions from human and *Fugu*.

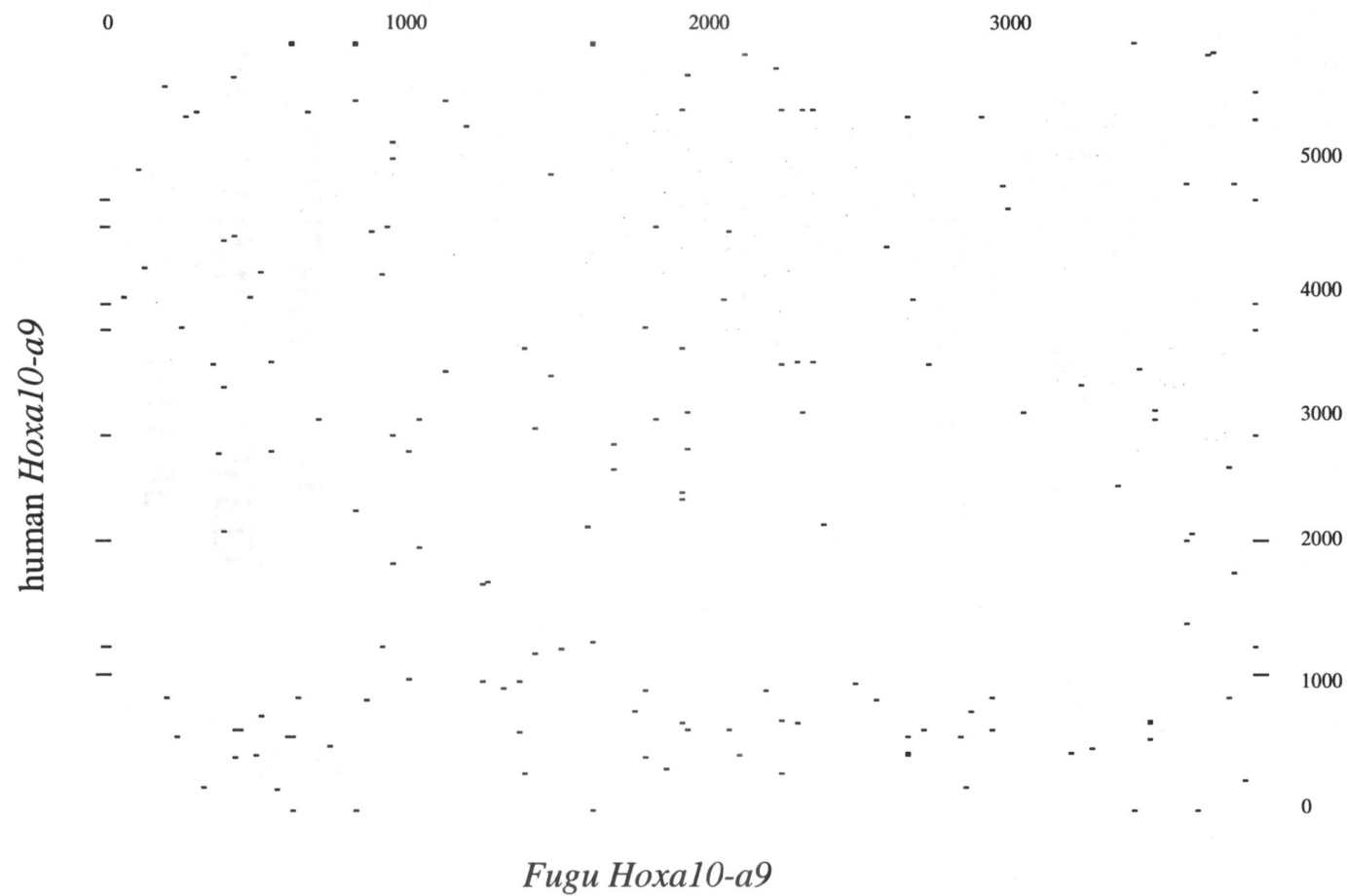


Figure 25. Pair-wise similarity plot comparing the *Hoxa9-a7* intergenic regions in *Fugu rubripes* (x-axis) and *Morone saxatilis* (y-axis).

Figure 25. Pair-wise similarity plot comparing the *Hoxa9-a7* intergenic regions from *Morone* and *Fugu*.

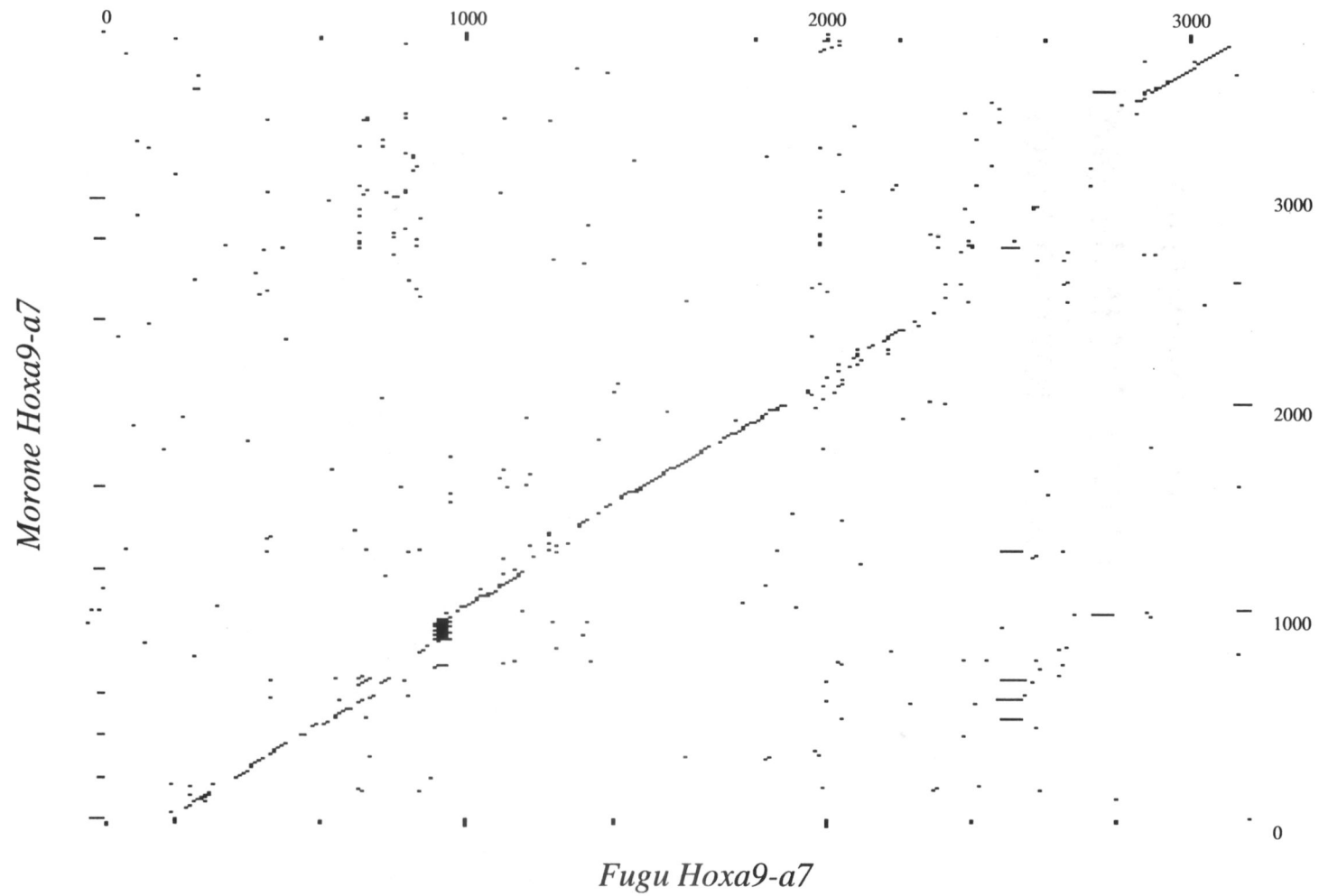


Figure 26. Pair-wise similarity plot comparing the *Hoxa10-a9* intergenic regions in *Fugu rubripes* (x-axis) and *Morone saxatilis* (y-axis).

Figure 26. Pair-wise similarity plot comparing the *Hoxa10-a9* intergenic regions from *Morone* and *Fugu*.

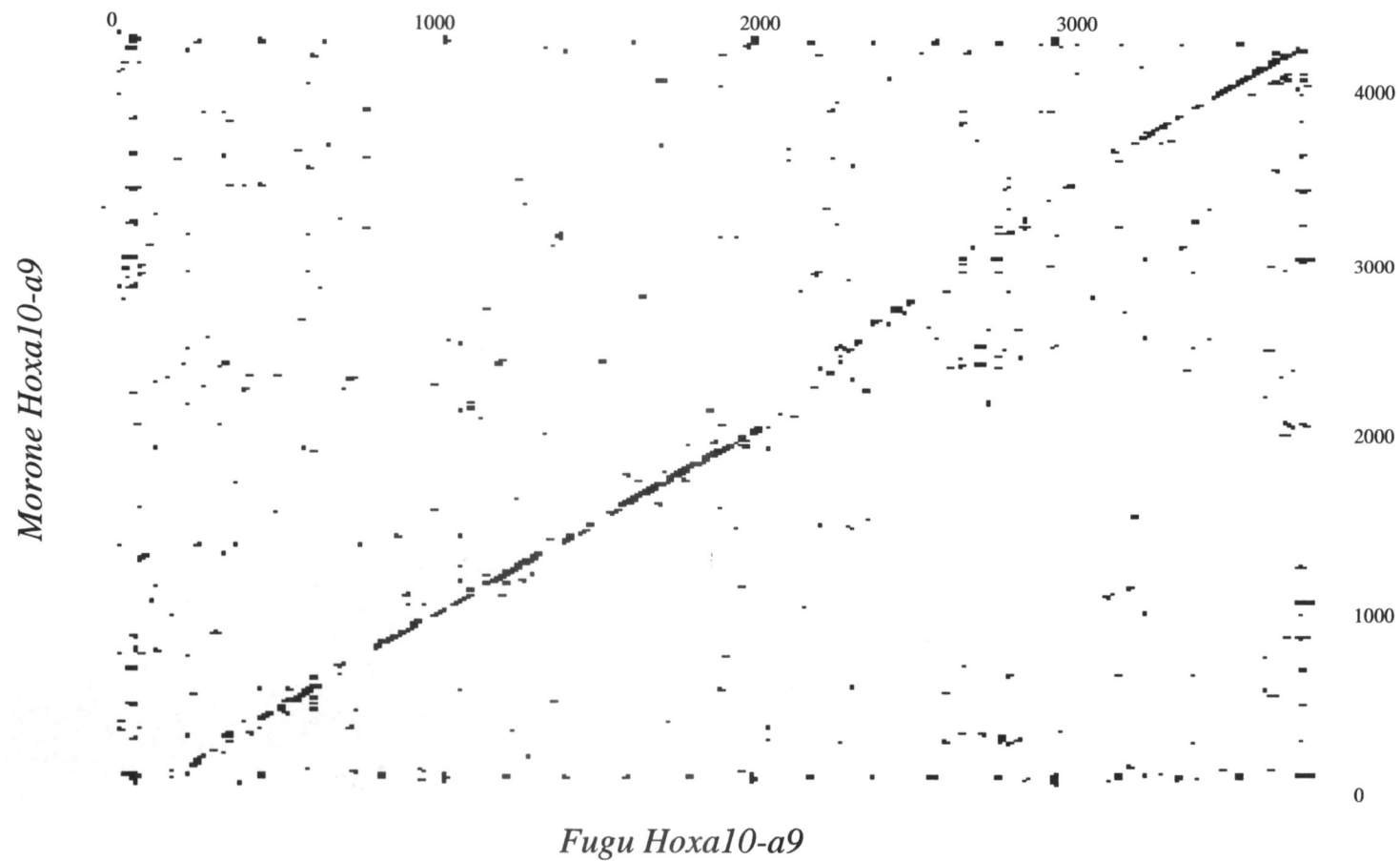


Figure 27. Schematic comparison of intergenic distances for the region containing *Hox* genes 4-10 in the single cluster of amphioxus and corresponding region from the *HoxA* clusters of human, *Morone saxatilis* and *Fugu rubripes*. Gene arrangements are drawn to scale to demonstrate genomic compression. The empty box indicates the presumptive *Hoxa7* gene, or gene fragment, in *Fugu rubripes*.

Figure 27. Comparison of intergenic distances among vertebrate *HoxA* clusters and amphioxus

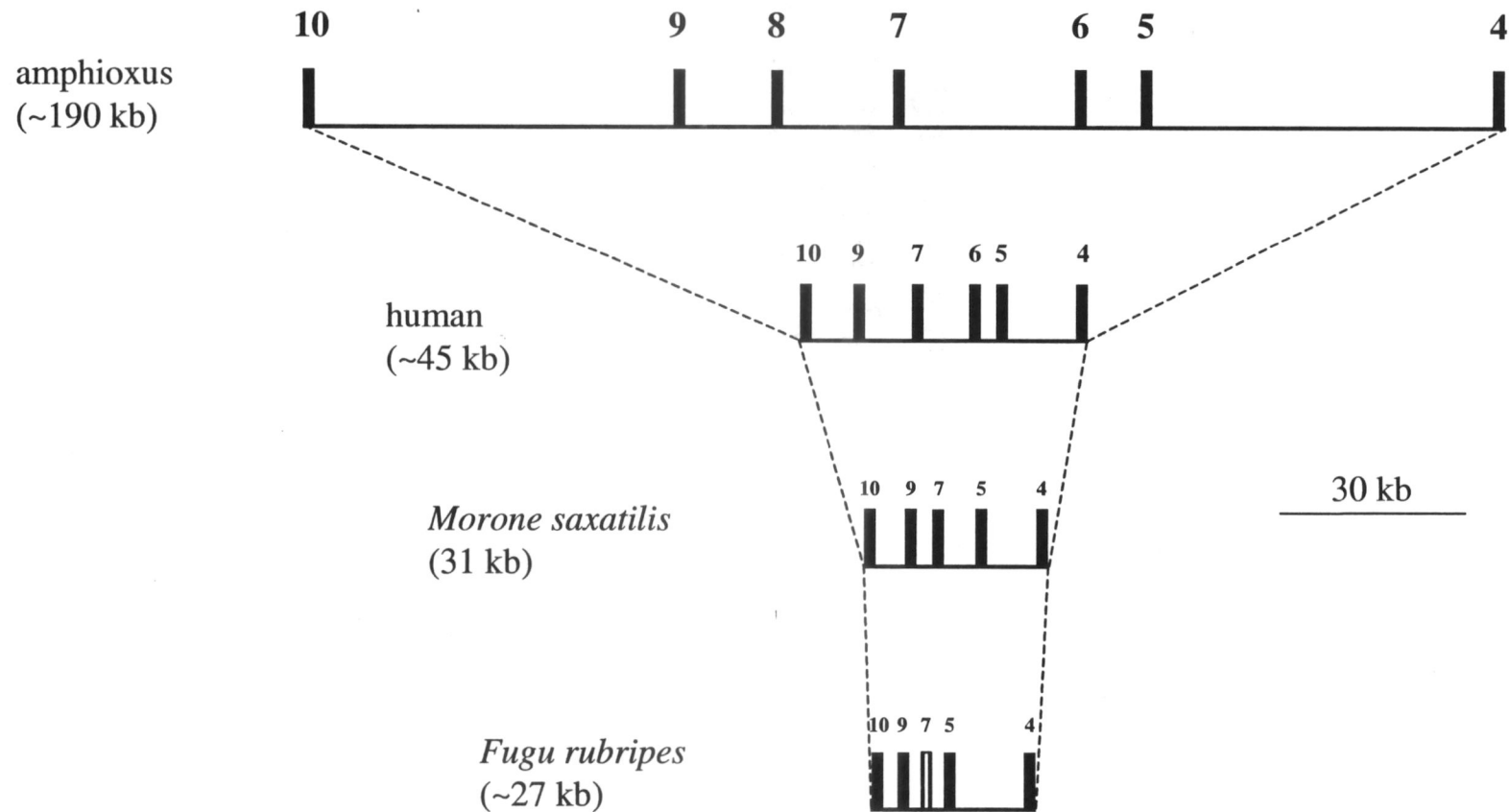


Figure 28. Amino acid sequence alignment of *Hoxa4* homeodomain sequences and the zebrafish *Hoxx4* and the 12 diagnostic residues. The *Drosophila Antp* homeodomain was used as a consensus sequence for alignment purposes. Dashes (-) indicate residues that are identical to the consensus. A dot (.) indicate residues for which the data was not available.

Figure 28. Amino acid sequence alignment of *Hoxa4* homeodomain sequences and the zebrafish *Hoxx4* and twelve 3'-encoded diagnostic residues.

		1							60							
<i>Dros. Antp</i>		RKRERQTYTR	YQTLLELEKEF	HFNRYLTRRR	RIEIAHALCL	TERQIKIWFQ	NRRMKWKKEN	1234567890	12							
<i>Morone Hoxa4</i>		P--S-TA---	Q-A-----	-----	-V---TM--	S---V-----	-----H	KLPNTKIRSS	SS							
<i>Fugu Hoxa4</i>		P-KS-TA---	Q-A-----	-----	-V---TM--	S---V-----	-----WH	NLPNTKIR..	..							
<i>Human Hoxa4</i>		P--S-TA---	Q-V-----	-----	-V---T---	S---V-----	-----D-	LPNTKMRSSN	SA							
<i>Mouse Hoxa4</i>		P--S-TA---	Q-V-----	-----	-V---T---	S---V-----	-----D-	LPNTKMRSSN	TA							
<i>Zebrafish Hoxx4</i>			Q-A-----	-----	-V---TM--	S---V-----	-----DH	KLPNTKIRSS	SS							

Figure 29. Amino acid sequence alignment of *Hoxa9* homeodomain sequences and the zebrafish *Hoxx9* and the 12 diagnostic residues. The *Drosophila Antp* homeodomain was used as a consensus sequence for alignment purposes. Dashes (-) indicate residues that are identical to the consensus. A dot (.) indicate residues for which the data was not available.

Figure 29. Amino acid sequence alignment of *Hoxa9* homeodomain sequences and the zebrafish *Hoxx9* and twelve 3'-encoded diagnostic residues.

	1	60
Dros. Antp	RKRERQTYTR YQTLLELEKEF HFNRYLTRRR RIEIAHALCL TERQIKIWFO NRRMKWKKEN	1234567890 12
<i>Morone Hoxa9</i>	TRKK-CP--K H-I----- L--T---D- -Y-V-RL-N- ----V-----	----M--F- KDGAPKDE.. ..
<i>Fugu Hoxa9</i>	TRKK-CP--K H-I----- L--T---D- -Y-V-RL-N- ----V-----	----M--F- KDGAPKDD.. ..
<i>Zebrafish Hoxa9</i>	TRKK-CP--K H----- L--M--S-D- -Y-V-RL-N- ----V-----	----M--C-
<i>Human Hoxa9</i>	TRKK-CP--K H----- L--M---D- -Y-V-RL-N- ----V-----	----M--I- KDKF.....
<i>Zebrafish Hoxx9</i>	 ----D- -Y-V-RL-N- ----V-----	----M--F- KNETKED....

Figure 30. Amino acid sequence alignment of homeodomain sequences of paralogous groups 5 and 6. The *Drosophila Antp* homeodomain was used as a consensus sequence for alignment purposes. Dashes (-) indicate residues that are identical to the consensus. A dot (.) indicate residues for which the data was not available.

Figure 30. Amino acid sequence alignment of homeodomain sequences of paralogous groups 5 and 6.

		1	60														
Dros. Antp		RKRERQTYTR	YQ	TLELEKEF	HF	NR	YL	TRRR	RI	EA	HALCL	TER	QIKIWFQ	NRR	MKWK	EN	
Morone	Hoxa5	G--A-TA---	-----	-----	-----	-----	-----	-----	S-----	-----	D-----						
Fugu	Hoxa5		-----	-----	-----	-----	-----	-----	S-----	-----	D-----						
Mouse	Hoxa5	G--A-TA---	-----	-----	-----	-----	-----	-----	S-----	-----	D-----						
Human	Hoxa5	G--A-TA---	-----	-----	-----	-----	-----	-----	S-----	-----	D-----						
Fugu	Hoxb5	G--A-TA---	-----	-----	-----	-----	-----	-----	-----	-----	D-----						
zebrafish	Hoxb5	G--A-TA---	-----	-----	-----	-----	-----	-----	S-----	-----	D-----						
Human	Hoxc5	G--S-TS---	-----	-----	-----	NN--	N-----	-----	-----	-----	DS-----						
Human	Hoxa6	GR-G-----	-----	-----	-----	N-----	-----	-----	-----	-----	-----						
Mouse	Hoxa6	GR-G-----	-----	-----	-----	N-----	-----	-----	-----	-----	-----						
Fugu	Hoxb6	GR-----	-----	-----	-----	-G-----	-E-----	-----	-----	-----	EER-----						
zebrafish	Hoxb6	GR-G-----	-----	-----	-----	-----	-----	-----	-----	-----	-----						
Mouse	Hoxb6	GR-G-----	-----	-Y-----	-----	-----	-----	-----	-----	-----	H-----						
Fugu	Hoxc6	-R---I-S-	-----	-----	-----	N-----	-----	-----	-----	-----	S-----						
Mouse	Hoxc6	-R-G--I-S-	-----	-----	-----	N-----	-----	-----	-----	-----	ES-----						
Salmon	Hoxa5	GRGRG-----	-----	-----	-----	M-----	S-----	-----	-----	-----	D-----						

Table 1. Predicted and observed hybridizing fragment lengths after the digestion of *Morone saxatilis* genomic DNA with *EcoRI* and *HindIII* and probed with the labeled inserts of λ clones 2.4, 7H and 3B. Observed fragments that confirm overlap between λ clones are shown in bold.

Predicted and Observed Fragment Sizes from Genomic Southern Blot						
	λ 2.4 predicted	λ 2.4 observed	λ 7H predicted	λ 7H observed	λ 3B predicted	λ 3B observed
<i>EcoRI</i>	918+	*				
	3380	3400				
	6730	6500	* 1270+ (6730)			
	2632	2750	2632	2600		
	356+ (7732)	*	7732	7400	7376+ (7732)	7200
			822	880	822	880
			273	---	273	---
			643	---	643	---
			3347	3250	3347	3300
			323+ (472)	---	472	---
					3370+	4800
<i>HindIII</i>	2179+	7100				
	2170	2100				
	273	---				
	9381+ (12,878)	12,500	7760+ (12,878)	12,200	3496+ (12,878)	12,000
			273	---	273	---
			1403	1400	1403	1450
			1358	1375	1358	1400
			2472	2350	2472	2300
			3607+ (3871)	3895	3871	3750
					611	---
					2313	2200
					463	---
					309	---
					103+	---

“---” indicates a fragment length that was too short to resolve on the Southern blot.

“*” indicates a predicted fragment for which no corresponding observed fragment was identified.

Table 2. Location of gene exons and introns for the five genes *Hox* genes within the 31 kb *Morone saxatilis* *HoxA* cluster contig. All distances and positions given in bp and oriented 5' to 3'.

<i>M.sax</i> gene	Position of Exon1	Length of Exon1	Position of Intron	Length of Intron	Position of Exon2	Length of Exon2	Intergenic lenth
<i>Hoxa10</i>	175-763	588	763- 1767	1004	1767- 2047	280	4426
<i>Hoxa9</i>	6473- 7024	551	7024- 7430	406	7430- 7671	241	3970
<i>Hoxa7</i>	11,641- 12,009	368	12,009- 12,964	955	12,964- 13,272	308	5428
<i>Hoxa5</i>	18,700- 19,294	594	19,294- 19,957	663	19,957- 20,196	239	8911
<i>Hoxa4</i>	29,107- 29,524	417	29,524- 30,056	532	30,056- 30,382	326	294...

Table 3. Results obtained from comparing the PCR-derived *Morone saxatilis* Hox-like DNA and amino acid sequences to the National Center for Biotechnology Institute databases and the genomic *M. saxatilis* obtained.

PCR vs. genomic gene assignments			
<i>M.sax.</i> PCR sequence (# of clones)	Nucleotide BLASTn match	Amino Acid BLASTx Match	Genomic sequences from <i>M.sax</i>
<i>Hoxa3</i> (15)	- 60/68 zebrafish <i>PDX1</i> - 57/69 human <i>PDX1</i> - 49/69 amphioxus <i>Xlox</i>	- 22/22 zebrafish <i>PDX1</i> - 19/22 amphioxus <i>Xlox</i> , human <i>PDX1</i>	None
<i>Hoxa4</i> (1)	- 61/69 <i>M.sax</i> PCR A6- like - 59/69 frog <i>C6</i> - 57/69 <i>Fugu</i> <i>C6</i> - 56/69 mouse <i>D4</i>	- 22/22 zebrafish <i>Y6</i> - 21/22 human <i>D4</i> , mouse <i>D4</i> , frog <i>B5</i>	None
<i>Hoxa6</i> (27)	- 61/69 <i>M.sax</i> PCR A4- like - 60/69 <i>Fugu</i> <i>C6</i> - 59/69 frog <i>C6</i> - 54/69 zebrafish <i>B6</i>	- 22/22 human A6, mouse <i>C6</i> , human <i>C6</i> , zebrafish <i>C6</i> , <i>Fugu</i> <i>C6</i> , frog <i>C6</i> , mouse A6	None
<i>Hoxa7</i> (2)	- 61/69 sheep A7 - 60/69 human B7, human <i>C6</i> , sheep <i>C6</i>	- 22/22 mouse A7, mouse <i>B6</i> , human A7, human <i>B6</i> , human <i>B7</i> , <i>Drosophila</i> <i>Ant-p</i>	- 56/69 nt genomic A5
<i>Hoxb3</i> (5)	- 64/69 mouse <i>D3</i> - 63/69 mouse A3 - 62/69 human <i>D3</i> - 61/69 human B3	- 22/22 human B3, frog B3, chick B3, mouse B3 - 21/22 human <i>D3</i> , mouse <i>D3</i> , human A3	- 69/69 nt genomic B3 - 22/22 aa genomic B3
<i>Hoxb4</i> (20)	- 64/69 frog B5 - 57/69 mouse C5, human <i>C5</i> - 54/69 frog C5	- 20/22 zebrafish <i>D4</i> , <i>Fugu</i> <i>B4</i> - 18/22 human <i>D4</i> , chick <i>D4</i> , mouse <i>D4</i> , chick A4, frog B5	None
<i>Hoxb5</i> (29)	- 59/69 mouse A5, mouse <i>C5</i> , human A5	- 22/22 zebrafish B5, chicken <i>B5</i> - 21/22 human A7, mouse A7	- 69/69 nt genomic A7
<i>Hoxc4</i> (1)	- 58/69 <i>M.sax</i> PCR B5- like - 56/69 zebrafish <i>D4</i> , frog	- 22/22 zebrafish B5, chick <i>B5</i> , mouse B5, human A5, mouse A5	- 68/69 nt genomic B5 - 59/69 nt genomic A5

Table 4. Intergenic distances in kilobases (kb) for genes 4-10 from the single cluster of amphioxus, and the corresponding regions in the *HoxA* clusters of human, *Morone saxatilis*, and *Fugu rubripes*.

organism	<i>Hox10-9</i>	<i>Hox9-8</i>	<i>Hox9-7</i>	<i>Hox8-7</i>	<i>Hox7-6</i>	<i>Hox7-5</i>	<i>Hox6-5</i>	<i>Hox5-4</i>
Amphioxus	58.0	26.0	---	20.0	24.0	---	10.0	55.0
Human	6.4	---	7.1	---	7.2	---	2.1	11.2
<i>Morone</i>	4.5	---	3.9	---	---	5.5	---	8.9
<i>Fugu</i>	3.9	---	3.0	---	---	6.1	---	12.0

“---“ indicates distances that cannot be measured due to gene content of the cluster.

APPENDIX A: COMPARATIVE ALIGNMENT OF THE *Hoxa9-a7* AND *Hoxa10-a9* INTERGENIC REGIONS FROM HUMAN, *FUGU RUBRIPES*, AND *MORONE SAXATILIS*.

BESTFIT comparison of *Morone saxatilis Hoxa9-a7* to human *Hoxa9-a7* intergenic region.

Percent Similarity: 81.905 Percent Identity: 81.905

```

Morone 3642 AAAATGGGGTTTGGTGTAAATCT.AGGGTGTATTGCTGTCATATATCACA 3690
          ||||||||||||||||||||| ||||| || | |||||||||
Human   6880 aaaatgggggttggtgtaaatctgggggtgtaatgttatcatatatcact 6929

Morone 3691 CTACCTCGTAAAAACGACACTGAGGATTCTGGGCCAACAAATCAGAGAGA 3740
          ||||||||||||||| ||||| | || ||||||||| ||
Human   6930 ctacctcgtaaaaccgacactgaaagctgccggacaacaaatcacaggtc 6979

Morone 3741 AAAAAT 3746
          |||||
Human   6980 aaaatt 6985

```

BESTFIT comparison of *Fugu rubripes Hoxa9-a7* to human *Hoxa9-a7* intergenic region.

Percent Similarity: 82.524 Percent Identity: 82.524

Fugu 3022 aaactgagggtttggtgtaaactct.aggctgtattactgtcatatatcaag 3070
 ||| || |||||||||||||||| || |||| | | |||||||||

Human 6880 aaaatgggggtttggtgtaaactctgggggtgtaatggttatcatatatcact 6929

Fugu 3071 ctacctcgtaaaaacgacact.aaggattctggccaacaaatcagagaga 3119
 |||||||||||| |||||| || | | | ||||||||| || |

Human 6930 ctacctcgtaaaaaccgacactgaaagctgccggacaacaaatcacag.gt 6978

Fugu 3120 caaaat 3125
 |||||

Human 6979 caaaat 6984

BESTFIT comparison of *Morone saxatilis Hoxa10-a9* to human *Hoxa10-a9* intergenic region.

Percent Similarity: 73.494 Percent Identity: 73.494

```

Morone 1152 ACTGTCGTGTGGTTTAGGTAGTTTCATGTTGTTGGGGTCCATTTCTAACT 1201
          |||||      ||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
human  1648 actggtcgggtgatttaggtagtttcctgttggtgggatccacctttctct 1697

Morone 1202 CTGCAACATGAAACTGTCTTAATTGCCCCAGTT 1234
          |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
human  1698 cgacagcacgacactgccttcattacttcagtt 1730

```

BESTFIT comparison of *Fugu rubripes Hoxa10-a9* to human *Hoxa10-a9* intergenic region.

Percent Similarity: 66.393 Percent Identity: 66.393

```

Fugu  800  gccgtcttgctgcatgccagccggctattgtagaatTTtatgactggct 849
          ||| | ||| | | | | ||||| |||| | ||||| ||| |||
Human 765  gccctggtgccgtaaggTTTgtcggctTTTgttcagTTTtatggcttgct 814

Fugu  850  gctatatgaggTTTtccctctcctctaacagcagcttcagtcgag.gga 898
          ||||| || | | | | | | | | ||||| ||| |||
Human 815  agtatatctggattgtggctgtcttggaccagtGATTTcagttgagaggg 864

Fugu  899  gagccggagagacggagaggaggcc 923
          |||| | |||| ||||| |||
Human 865  gagctacatagac..agaggaagcc 887

```

BESTFIT comparison of *Fugu rubripes Hoxa9-a7* to *Morone saxatilis Hoxa9-a7* intergenic regions - *Fugu*, (top sequence/lower case) *Morone*, (bottom sequence/capital letters). Examples of direct repeat sequences flanking a deletion site are underlined.

```

184 tgttgattctacat.....ttctgtgtgtagccctttcttgccctcgccacgtgcttgtt..tg 242
    |||||
11  TGTGTGATTTTAAATAATACCTCGTTCTGTGT.....ATTAGTTGGCCTGGTCCCTGTGCTTGTTTGTG 73
    |||||
243 tatggatgtttctatagatgtgcgtgttcttggtttctagctatattgtgttc.tatacttcctgccag 311
    |||||
74  TATGGATGTT..TTTAGCTGTGTGTGTCTGGTTTGTAGCTATATTGTTGTCTGTATAGT..... 131
    |||||
312 ctcagcgtcttgtaaaacttta..catttt...tagtctatccc.....tctcttatgac 360
    |||||
132 .....TTGTAAATGTATTAATTTTATTGTTGTTTCTCCCTTACTGCTAGAACAAATGACTGATGAC 192
    |||||
361 cgtgaaattgaatcgattaatggaaacagtagaaaa.....aagcaagagcagaaaa 411
    |||||
193 CATGAGTTTGAGTCTAACATTGG.AACCGTAGAAAAACAGACTGATCGCATTTTGAAGTAGGCGCATATAA 261
    |||||
412 cctcacaacacctcttgaaacatgctatgatactaaatctaaa.ataatatagaagatata.....acac 473
    |||||
262 CCTCAGAAACCTCTTGAAATGCTAAATCCTAAATTAAAAGATGTTATAGAAGTTATATTACATGACAC 331
    |||||
474 aattgactaccta..tctttcaactacctaacaataacgt.tttccccagtgccgattggctgcccata 540
    |||||
332 GATTGACTACCTATGTCTTTCAACTACCTACAATAACCTATTTTTTCAATGGCGGATTATAG..TAGAT 399
    |||||
541 ttgcttgattagatgctgaacmtatggatttc..agggagtaacaaactgaaagtagctacatgatatt 608
    |||||
400 TTGCTTGCAATTAGATGCTGAA..CTATGAATTTCCAGGGAAGTAACATTTTGAAAG....TACATAATGTT 464
    |||||
609 tgttagcgtg.agaaattaccataaaacactgtaggttacttattgaatacctcgcggtctgatgaccgt 677
    |||||
465 TGTTAGTGAGAAgAAAAATTAATAAAACACTGTA.GCTACTTATTGAATACCTCACGGTCTGCCGACCGT 533
    |||||
678 ccgcctttttgttcacaaatgtaattt.tattttatttattaaataaaacaaattatattctgctaattgcg 746
    |||||
534 CTGACTTTTGGTTCACAAATGTAATTTGTACTTTATTTTATTAAATAAAACAAATTATATCTGCGAATGCA 603
    |||||
747 tgc.....gcattgcgg.....gtttatttttttctaaaaataatcaaactga 788
    |||||
604 GGCaCGTTtACTTCcTATTTcATTTTGGCCTCACATCGAAATGAATTTTGTTTAAAAAATAAAATGTA 673
    |||||
789 agtaatttggtt.....tgtatt...attactgcggcttttaaac.....aa 827
    |||||
674 TGTTAATTTGTGTGCTTCACAAAAAGTCTTTATTGGCAAAAAGGTTGAATTTCAACACACACGCACACGAA 743
    |||||
828 actagaaattgtaaatgctcatggcctatgtatattt.....ttcaatgtatttacgca 881
    |||||
744 AATACATAAAGAAAAATCCAAGATATTTGGTACTTGAGCCTAGTCCACATCAACAATTTTTTTACACA 813
    |||||
882 ttt....taaatgtctaggattt..... 900
    |||||
814 TTTTTTATAAATATCAAGCATTTTAATCTGCAGAGTGCACACTCTCTCACACACACGCACTCATTTACAC 883
    |||||
901 .....tcagtggaacacacacacacacacacacacacacacacacacagcggttc 962
    |||||
884 ACACACACTCATTTACACACACACACACACACACACACACACACACACACACACACGAATTA 953
    |||||
963 a.....tccgtatttctgtattgggggacagagctgggtcttggc 1002
    |||||
954 AAACGCACACACCTATAGGTGAATATCTGTGTGTGTGTTTCTGGACT.GGGGACAGAGCTGCGTCTTGGC 1022
    |||||

```

```

1003 atagaatgctgttggtgtatataccgccataaaaagacaattaccgccataaccttttatggggtgcaaagc 1072
1023 AGAGGATGTTGTTAGCTGTATACAGCCATAAAAGACAATTACCGCTATAACCTTTTATGGGGTGCAAAGC 1092
1073 gctgcgaggcgagaggacacaaaacaaaaaagacatcgagaacgacgacagctgaatcccaaaataatg 1142
1093 GCTGCGAGGCGAGAGGACACAAAACAAAAAAGACATCGAGGGCTTCGACAGCTGAAGCCCAAAATAATG 1162
1143 tgcccgatgtgtttctcggctactcagatttggccactacagcccgact.aagctctgcccatt.....t 1205
1163 TAACCGACGTGTTTCTGTTCAGCTGGGATATGACAAGGCCGCTCAGCCTAAAGCTCCGTCATTAGCGCCT 1232
1206 catgttcagctaatcttcaattgacagatgcacacatggtttattagcatcgggacat.....ccttt 1268
1233 CATGTCCACATAA...CCAATTTACAACAAGAGATTTTGTAACTAAATATTGTAATATATAAAAGCATTT 1299
1269 ttaatttatcttgtcaacgag..... 1289
1300 TGGATTTTAAACGCAACGTGACCTCAAATAGTTTATCGAAAAAGATTTTAGAGAAGATCTCTAGTTGTA 1369
1290 .....tcaacgggcggtg.ttggttgatggaataacttggcacaataacactaaata 1339
1370 AAATATGGCCTCTGCGTTTTCAAAAGTCTCTGCATGTTGGAGAAAATATTGGCGCAAAGACAGAAAATA 1439
1340 aagagcgaa.....ggaggtgcagagctgatgagtc...ttggggatcatggccagacacggg 1394
1440 CGCAGCGAAATCTAATGTGTGCTATGTGCAAAGCTGCTGTGTGATGGCGAATCAGGGCTGGGACGGT 1509
1395 tctgttacaaaattcaggccacgtaagtttgacttaaagctgcagagaaaagcatacatgaaagtagata 1464
1510 TCTGCCACACACCCAGCTCATTAGGTCTCATTAAAGGCTGCACAGAAACGGATACATTATTGTAGTTA 1579
1465 gaaagtgagtcggacttttagtccagctggagtgtgtgtgacgcttagggagccgagatggactcaaa 1534
1580 GAAGGTAAGTCGACTTTAGTCCAACCTGAGCGCTGTTGTGACGCGTTAGGGGACGCAGATGGACACAAA 1649
1535 cttcaaag..acggccagagacagcgcaataaaaacgcctggtcggtgtactgtctggcattccagttt 1602
1650 CTTCAAAGGCTCGACCAGAGACAGTGCAAT.AAAACGCCTGGTCTGCTATACTGTCTGGCATTCCAGTTT 1718
1603 aaatggttttatggcgtccagacacaattagccggtttccagaatggcacccatttgtttttctcctc 1672
1719 TAATGGCTTTATGGCCGTCCAGACACAATTAGGCCGTTTCCAGAATGGCACCCATTGTTTTCTCTCTC 1788
1673 tttctgtgagacaacgctctggacaaaaggccagcggaatgatcagtttta.tggat..tccccgacg 1739
1789 TTTCTGTGAGACTGCGCTCTGGACAAAAGGCCGAGCGGAAATGATCAGCTTTATTGGATCCTCCCGACG 1858
1740 gggacgcatgactcgtggtcatttgtatagaccttgtttctacctgggaaccttccgccctctgct 1809
1859 GGGACGCACAGGACTCACGGTCATTGGAGCGATCCTTTGTTTCTACCTGGGAACCTTCCGCCCTCTGCT 1928
1810 cctttgcttagccagtcctaaagcaaaagaagcaataaagccaacagctagatataaaaaacagtggtagg 1879
1929 CCTTTGCTGTGCCCGTCTAAAGGAGAAGGAGCAATAAAGCCAGCAGCTAGTTTCAAACACAGTGGCAGG 1998
1880 cctgaagcctatgtatct.....taaaagcaagctgcatgttgac.ggtatgtatcagaccatgttc 1941
1999 CCTATAGATGGTGTGTTGTTTGTGTTTTTTCAAAATACATCTGGACAGGTCTGTTAAAGACATCAGC 2068
1942 a.tctgcttgaagatatcaggacaaga.....tttctaatgtttattc 1984
2069 ACTTTGCTTGAA.ATCGCAATCCAAGAAACGTACGTTACCTTTGAGCGCCATTCTCTwACTTTTAGCT 2137
1985 ttttttgttttcttaaaacaaacaggaaaaaagacaaagcaaaacaaaaacacagctacgtgc 2054
2138 GGTATTTTATCCGGCATTGCAGCAGGAAGA.....CAAAACGCCTGCGCACTGCTGCCTCCACCCCTGC 2202

```



```

2836 .....ttaattccct 2845
3387 CTTCTAGATAGAGTATTATTTGTAAAGCAGTAGTAAAATAAGTGATATAAAACAGTGACTTTAAATTTTCG 3456
2846 acaataatttctacatcagtggttg.....gtgtgtgtttatgctcttcgtgtgtttat 2901
3457 TAAATTACCTCTCGATCATTGGGATTGTCTGAGTTTCTCTCGTGTGTGTTTATG.....TGTGTGTTTAT 3521
2902 atgtgtggaatgttaataaccactctgaagtcctgcgccataaagaaatccacatctccccgctgattgga 2971
3522 GTGTGTGAAGTGTTAATCCCACTCTGCAGTCCTGCGCCATAATGAAATACATGTCTGCTTGCTGATTGGA 3591
2972 cgagagctgcccaggccttagtcacatgggtgagtcctttggtcctgcgtaaaactgaggtttggtgtaaa 3041
3592 TGAGAGCGGCCAGGCCGAGTCACGTGGGTGAGTTCTTTGGTCCAGAGTAAAAATGGGGTTTGGTGTA 3661
3042 tctaggctgtattactgtcatatatcaagctacctcgtaaaaacgacactaaggattct.ggccaacaaa 3110
3662 TCTAGGGTGTATTGCTGTCATATATCACACTACCTCGTAAAAACGACACTGAGGATTCTGGGCCAACAAA 3731
3111 tcagagagacaaaat 3125
3732 TCAGAGAGAAAAAAT 3746

```

BESTFIT comparison of *Fugu rubripes Hoxa10-a9* to *Morone saxatilis Hoxa10-a9* intergenic regions - *Fugu*, (top sequence/lower case) *Morone*, (bottom sequence/capital letters). An example of direct repeat sequences flanking a deletion site is underlined.

```

180 tgtttctgtcaccgttggcatatggaagcatagt.....gttaaatcaacaca..... 227
    |||||
55 TGTTCCTGTCACCTCTGTCGTATGGTTGGATAGTGTCTAGTGGAGTTAAATTACCACTTAAAAAACA 124
    |||||
228 .acttcttctct...aaagtggcataaatcgt...gtttttccc.....tttgca 269
    |||||
125 CATTATGTCTCTGGAAGGTGGCTTCAATTGTCCAATTATCCCCCCCACCACTAGCAAATAATCTTTTTTAT 194
    |||||
270 caataaaattaaatcaaattggctattttttatttgaattggaacgatgtagtgactcaaacttttctta 339
    |||||
195 AAATTTAATCAAACAAATATGAATTTACTTTTATTTTAATTGGA...ATTGCATGGTTTAAAGATTT... 257
    |||||
340 cattttacagatttttggacatttctggaatatatgcccatgta.....aggagggagtacgcgcata 402
    |||||
258 TCTTTTAGCAATTTT.AACCCACCGTGAATA.ATGCCCATATATCACCAGGGGAAGAATAATTCACCTA 325
    |||||
403 tgaccttttagatgttgtctatagtgtacaacaacatataggcctctcctgaagtatatcagttaaatgtgt 472
    |||||
326 TGACCATTAGATG.....TAGTCTAC..CATGATATAGGCCTCTTCT...ATGTATGAGCTAAATGTAT 384
    |||||
473 atgcctacaatgaaataac.ccgaggacgttttacgcaa.....gtgcaaattgaagacggacggggcat 535
    |||||
385 AAATTGAATCTTTATTTCACATTGAGGACGTTTGCGCGAAAACGGTGCGCAAAGTGAAGGAGGA.ATGTTT 453
    |||||
536 ttaatgtatatattttgtgtatatatgacaatgtg...tttgtgtaagtataataaattatcagcacacat 603
    |||||
454 TTATGTATATTTCTGTGTACATATGACATTGTGTTTTTGTGTAATGCTTAATAAATTATCAG.TCGCAT 522
    |||||
604 tgag.....agttttg.....catatctgcgc.....tcttgc 631
    |||||
523 TGAGGAGTAGCTGCTTGTATTTTGTACTGTAGGTTTAAACATACATATTTAATCAGTGTGTATCATAC 592
    |||||
632 tgaagccgaaatga.cctttgcttgatgtttcaaatacaagcaacatttcctttgagtaactgttgctt 700
    |||||
593 AGACGCAGAAAAAAGGCATTGTTTGAAATTCAAATCCAAA..ATAATATGGGTTGTCTGCTGCAGGTT 660
    |||||
701 tatacgttggagccacttgttgttttttaggaaaaataataacattgcatgcag.ctttatacaaaagagt 769
    |||||
661 CATTAGT.....TTTTTGTGAACCACCTTAGAAATGTTTCTAGAGGCTATTGTAATGGCC 715
    |||||
770 gtccacacaaatgaagatatcgggtccacgcgctcttgtgcatgccagccggctattgtagaatttt 839
    |||||
716 GTCGGAAATAATTAAGATATTAGGCTATAAGCCGTCTAGCTGCATGCCAGCCGGCTATTGTAGAGTTTT 785
    |||||
840 atgactggctgctatatgaggtttctcctctcctctaacagcagcttcagtcgagggagagccggagag 909
    |||||
786 ATGACTGGCTGGTATATGAGGTTTCTCCTCTCCTCTAAGAGCAGCTTCAGTCGTGGGAGAGCCTTTTCTAG 855
    |||||
910 acggagaggaggcccgccactcccgggactcccggtccaaat.....ttgaatccatgcacgttggata 974
    |||||
856 ACGGAGAGGAGGCCCGGCACT.....CAGCTACAAATCTTAAGTACAGATACACTGTGGATA 914
    |||||
975 aaacawtgcctcagctctgttttgggaataatgtaattaactctttcgaggagccgggtgagtcattt 1044
    |||||
915 AAACGTTACGTTCTGCTTTAGTTTGGGAATACTGGGATTAAC.ATTTCCAGGA.TCGGGTAAGT.CATTT 981

```

```

1045 atcattgtgtgattgactattttgcgccttggtgcagcactttctctttattgcgga.....ccgca 1108
    |||||
982 ATCAATGTGCATTGTGCCATTTGCGCCTTGTTGCAGGGGCTTCACTATAGTGTGTACAATACAGTTGTA 1051
    |||||
1109 cttctctggactctagcgggtt.tgggacacttgatgatcgaggagc.....aagggtgtaaaatgtatt 1171
    |||||
1052 CTTTCTCAATACAGACGATTTGTTGGACACTTGATGA..CTGGACCGGCGACCACGGTGCATTATGAATT 1119
    |||||
1172 ggctgatcggtgtttcacacgagggccggaccgactgtcgagtggttttaggtagtttcatgttgggggt 1241
    |||||
1120 GGCTGATTGTGTTTCACTCGAGGCCCGGATCGACTGTCGTGTGGTTTAGGTAGTTTCATGTGTTGGGGT 1189
    |||||
1242 ccatttcaaactctgcaacatgaaac.gtcttaattgccccagttacatacaacgaagccaaatggaaag 1310
    |||||
1190 CCATTTCTAACTCTGCAACATGAAACTGTCTTAATTGCCCCAGTTACACACATCGACACCAAAGGAAAG 1259
    |||||
1311 cctaaatgtgacgggataggggtgtagtggaacttatatagaggacgtttg.....ggtttttt.....c 1369
    |||||
1260 CAAATCAGGGAC...ATAAAG.ATGGTGGGACTT....GATGACATTGTGGAGTGGTGTGTTTGGGAAGC 1320
    |||||
1370 ttttctggaatcttctggtttatgactcgggg.aggggctgactggctg..gcgcataattgctaaatac 1436
    |||||
1321 ATGTTGTAAACCTCCCTGGTTCATATACGGGAAGGGGGTGACTGGCTGTTGTGTATATTCCTAACCTG 1390
    |||||
1437 tctcccgatattgttgcctttacctttccac.....tcgtcgcggttaaaagcaggcaggagaggacat 1499
    |||||
1391 TCTCCTGTAATCGTTTCTTTAGCCTTCTCCGCTTGTTTTTCGGCGTTAAAA.TAAGCCCGAGCGAGCGT 1459
    |||||
1500 aggattggagtttaaacacga.....agggttgcgctatgcctgcgtggttagacttgggatttaggta 1562
    |||||
1460 GTGCT..GAATATAAACACGATCAAAGCTGAGCTGCAGTAT..ATGCGTCGTGGACTTGGGATTAAGGAA 1525
    |||||
1563 ....cgcgagtgatt..cggcataagcagtgacgcagacacgcccagacaaggcctcggttacgaccactg 1626
    |||||
1526 AAGTGCGAGTGTTTTCCGTGAGTAGTAGTGACGCAGTCACGCCAGACAAGG.CTCGGTTACGACCACTG 1594
    |||||
1627 cagtgagggggaaggggaaggtgaatggggcaacagcgacctccgtgggcaatggagaccggatgtttcc 1696
    |||||
1595 CAGTGGAGGGGAAGGGGAAGGTGAATGGGGCAACAGCGACTTCCGTGGGCAAT.GAGACCGGATGCTTCC 1663
    |||||
1697 gctgcgcaatagcgctcagacaaaaggaggggaagagctttatggcgatatttagatccgccagacagg 1766
    |||||
1664 GCTGCGCAATAGTGGCTCAGACAAAAGGAGGGGAAGGGCTTTATGGCGATATTTAGATCCGCCAGACAGG 1733
    |||||
1767 ctgttatctgcaaaggcaaaagaagtgaatagcccctgcggttaaactgggtgaaaaaacaccctgattt 1836
    |||||
1734 CTGTTATCTGCAGAGGCAAAAGAAGTGAATAGCCCTGCGGTAAACTGGGTG.AAAAAACACCCTGATTT 1802
    |||||
1837 aggtttacagcctgggagccattggcaaaagtggcgtcttttctgctccaaagggtgtttgtctttttctt 1906
    |||||
1803 AGGTTTACAGCCTGAGAGCCATTGGCATAGTGGCGTCTTTTCTGCTCTGGAGGTGTTGTCT....AT 1867
    |||||
1907 ttttcttttagtgacaaaataggcgacggcgcttc..tttctactcctcggttgtttgatctctgt 1974
    |||||
1868 TCTAACTTTTAGTGACAAAATAGGCGCACAGCCACTCCATTTCCCCCTCCCATGCTTGTGTTGATGCTGT 1937
    |||||
1975 tggacaaaataat.tgc.actaccagtgggtgagagattggaacgatggtgcacattacaggaagcatgg 2042
    |||||
1938 TGGCCAAATAATGTGCAAGTACCAGTGAGGTGATAGCTTGGAACATAATGTACATTATGGACAGCTTGT 2007
    |||||
2043 gccctt.....tttggtacttttagcccaaacataattactgcaga....aaaatagagct.agtggttaa 2102
    |||||
2008 GTCTTTTTTCTTTGCTCTCTTCTAAGTGCAAGAGTAACAGCAGAGTGGAAGAATTCTGCGTGTGA 2077
    |||||

```


2881tattttaagaaatatactatagaatacaacttt.....gcta..... 2917
 3192 CATGTTTATTGATTTACTTTTCAACTCTAAGATATGTACTACATAGTCCTGCATTATGGTAGTTAAAACG 3261
 2918 gtgcatttagaataaagtatat.....ttgcttttagccaatgaatgagagatgcggc 2969
 3262 GTGCAGATGTTGAAATGTAGATAATTAGATGTTAGGGATGTAGCTCCACATACAGGGACGTGAAACGTC 3331
 2970 cttttctaaatcacatgtttatttttattgaat..tattgaaacgtgtaattgttttatg..... 3025
 3332 CATGTCTAAATCACAAAAAGTGGTTAATGTATAAAAAATAAGACCAAATAAGCATTATGCTTATATAATCA 3401
 3026tcacgtaggtttttaaagtctaaataaggttagattatgccatattccaagtaagtctcagttta 3091
 3402 GTATGCGTATATGTTTTATATTTTTTAGATGAG.....CAACTATCTTGTGTGGCTAAATGTT 3459
 3092 agagagaaaaat...gattttaaccgagaaatgggaaacac.....gtgggga 3135
 3460 AAAGTGAATAACAATTTTAGCCTA.ATATAAGATACACCCCGCATGTATTTTGGTGTATTTGTCAGTT 3528
 3136 tgggtgggagacgtcggttg.....aaaacacggacgtgcctacatat..... 3178
 3529 TGGTTTGTAATAATATGTTTGAAACTTGAAGCTGAAAAACAGGAAATATACACACATTGGAAGATGTCTG 3598
 3179aaaaataacattaaaaataatctgcgta..tcctaatttagttgttgacgccaggta 3236
 3599 ACACAGCCGAAAAGGAGCAAAATAAAAAATGAAAAAGTAATTTCTAGTTGAGGTCTGCATG..... 3660
 3237 ctgtcagcgttgagctgctgacttttcccaatattgtgtgtggttaaaatgtgatttggtcagca..... 3300
 3661 .TTTCAGTGTGTGCTGCTAAGTTTCTTTAATACAGTGTGGTTAAATGTGTTTTGGTGAGCAGCAATG 3729
 3301 gtaggagccttatgtgtcctataaaatagctcgtaaaacaaggcttccaaaagcataatagtatcgtttt 3370
 3730 GTAGGGGGCTTACGTGTCTTATAAAGTACCTCGTAAATAAGGTTTCCAAAACAT.ATTGTATCGTTTT 3798
 3371 ggggcggtgaaaaggcgtttaacagcatcaatagactcagacagcatactgtaaaaaca.tttttattc 3439
 3799 GGATAACCAAGAAGTGATTTTAACAGCAGTAGTAG....GGTAACACACAGTAAATCACATTTTATTTC 3863
 3440 aaatatttttactttccatctcaagacaac.....acaaatttgcaactcttacattttaa.. 3496
 3864 AGTCATTCATATTTTCTGTCTCACAGATAACAGATCTAACTAATATCTTGATAATTCTTAGCTTTAAAC 3933
 3497 tggggattgttataggggagaggtgagttatgcccagagggattgatttatccctttattgttcag 3566
 3934 TAGAGGTTGTTATAGGAGAGCGGGTGAGTATTTAATGACCAAGGGATTGATTATCCTTTATGTTCAG 4003
 3567 ccttgatgggtcacgtgtgctgctgccccaatgacatcgcgcttgactccccattacaagccactgtag 3636
 4004 CCTTTATGATCACGTGTGTCTGTACCCAATAGCATGGCAGCCTGACTCCCCATTACTAGCCCACTGTAG 4073
 3637 ttctctgtggggccaagttgctacttgatttctccacattgttattttgtgagactgtgtttactgcatg 3706
 4074 TTCTCTGTGGGGCCAAGTTGCTACTTGATTTCTCCACATTGTTATTTTGCGAGGCTGGGTTTACTGCGTG 4143
 3707 tgtatgtgtgtggtgtgtgtgctgctgctgtgttttactttt..atgcacctctgtat.....act 3765
 4144 TGTATGTGTGTGTGGTGTGTGTGTGTGCGTTTACTTTTGCATGTATCTGTTGTATGACTGCGTACAT 4213
 3766 gtgaacttagaatttttctgc 3786
 4214 GTGAACGCATAATTTCTCTGC 4234

**APPENDIX B: SYNTHETIC OLIGONUCLEOTIDE PRIMERS
SPECIFIC FOR THE *Hoxa4* TO *Hoxa10* REGION
FROM *MORONE SAXATILIS*.**

Primer Name	Designation	Position	Sequence 5'-3'
pUC-19 forward	pUC-F	n/a	ccc agt cac gac gtt gta aaa cc
pUC-19 reverse	pUC-R	n/a	agc gga taa caa ttt cac aca gg
Lambda Right Arm	Lam RA	n/a	taa cga gtg gat ctg ggt cg
Lambda Left Arm	Lam LA	n/a	gca gac aaa ctg cgc aac tcg tg
Lam2.4 R230	1F	184-204	atg gag atg tgg atg gac gc
Lam2.4 R625	2F	555-575	tgc tcg ctt cca cac gcc ac
Lam2.4 R725	3F	653-672	tgc agc ctc acg ctg cac c
Lam2.4 R817	4F	1031-1050	tca gtc agg gat tat agg g
2.4Bam F9030	5R	1185-1205	tac tag tac att tta cgc ag
Lam2.4 R1375	6F	1361-1380	tac tga ctg ttt cta ctc c
Bam2.4 F4800	7R	2268-2287	tac aga aac acg aca tca g
Bam2.4 F4300	8R	2765-2785	aat aca agc agc tac tcc tc
Bam2.4 F3670	9R	3387-3407	aac cac acg aca gtc gat cc
Bam2.4 3215	10R	3851-3871	tgt tgc ccc att cac ctt cc
2.4Lam R4280	11F	4280-4299	aag agt aac agc aga gtg g
2.4Bam F2634	12R	4441-4461	aat atc caa cat cta agc cc
Bam2.4 F2085	13R	4984-5003	ttt ctc gct taa cct gcc g
Bam2.4 F1568	14R	5497-5517	tac att tca aca tct gca cc
Bam2.4 F1175	15R	5887-5907	aca gct gaa aca tgc aga cc
2.4Bam F720	16R	6226-6246	taa agg ctg aac aat aaa gg
2.4Bam F405	17R	6668-6687	aac tgg gca gga acg tgg c
8HBam F2820	18R	7105-7125	atc tgt tca gac gcg cac cc
B1.1 R418	19F	7431-7450	aga taa ccc agt gtc caa c
Bpp1.1 F420	20R	7507-7526	act ctt tct cca gct cca g
B1.1 R848	21F	7848-7867	act acc ttt tat att gca c
Bpp1.1 R1385	22F	8309-8327	att aga tgc tga act atg
Bpp1.1 R1476	23F	8487-8508	ata tct gcg aat gca ggc cg
B1.1 F863	24R	8862-8881	aca gat att cca cct ata g
Bpp1.1 F897	25R	8910-8928	atc ctc ctg cca aga cgc
B1.1 F437	26R	9293-9311	tct cca aca tgc aga gac
Bpp1.1 R218	27F	9524-9543	tta ggg gac gca gat gga c
S22 #2.2EX (1R)	28R	9987-10,006	CCA AGA AAC GTA ACG TAA C
S22 REX (2F)	29F	10,165-10,184	TCC TCC CTA ATG AAC CTG C
S22 #2.2 (3R)	30R	10,430-10,449	GCA GGG TTA TTA AGG CTA T
S22 #1.2AN (4F)	31F	10,514-10,533	ACC TGT TAC TTT CCT CTG G
S22 #1.2 (5R)	32R	11,027-11,046	GGA TTA TTT ATC ATG TTA T

E4PR1.2 (6F) (bad)	33F	11,351-11,370	TCC GTA AAT TAC CTC CTC G
E4PR1.2ANEDAN (7F)	34F	11,379-11,398	ATT GTC TGA GTT TCT CTC G
E4PR1.2AN (8R)	35R	11,454-11,473	GCG CCA TAA TGA AAT ACA T
E4PR2.2 (9F)	36F	11,719-11,738	GGG CGT CTT GCA GTA TTG G
E4PR3.3 (10F)	37F	12,022-12,041	GAT ATT GTC TCT TAT GTG C
E4PR3.3AN (11R) (bad)	38R	12,451-12,471	GGA AGT CAA GGG GAT TAG G
E4PR3.3EX (12F)	39F	12,599-12,618	GAG GAA TGA GCC AAA CAC C
E4PF2.2 (13R)	40R	12,910-12,929	GCA AGG AAA AAA AAT GCA T
H722AN.1 (41F)	41F	13,120-13,139	CTG GTT TCA GAA CCG CAG G
E4PF1.2 (14R)	42R	13,265-13,284	GGT GTC AGA AAC GGC TGA A
E4PFAN (15F)	43F	13,598-13,617	TCC TGA ATT AAA AAC TAC C
S10R#1AN (16)	44F	14,258-14,277	TTA GGT TTA GTG CCC ACG G
S10R#1EX (18F)	45F	14,514-14,533	TGG TCA ATA TTA GGT TGC C
S10R#1AN (17R)	46R	14,4550-14,569	GCA GTC CAT TAT GTG GGT T
Luck 1	47F	14,811-14,832	ggg tca taa atc cgt tgt tcg
S10F#2AN (19F)	48F	14,899-14,918	AGC TGT GTC ACG TGA CAC C
S10F#2 (20R)	49R	15,088-15,107	CGG TCG GGC AGA GGT TCC A
S10F#1AN (21F)	50F	15,336-15,355	TAA GAC AGC TGA AAC TTG C
S10F#1 (22R)	51R	15,476-15,495	CCT GTG GAA GAG ATG GTT T
E ₂ 12R#3 (23R)	52R	16,165-16,154	GCA CAG ACA GGT TTA TTT T
E ₂ 12#2AN2 (25F)	53F	16,193-16,212	GAA TCA AAT ATA ATG TTC G
E ₂ 12#2EX (24R)	54R	16,242-16,261	CCG GGT TTT TTC TCT TTG C
E ₂ 12#2AN1 (26F)	55F	16,593-16,612	TGG GTG TGT CTC TGG CTC G
E ₂ 12R#2 (27R)	56R	16,673-16,692	GGA GCA TGG CTG TTT AGC C
E ₂ 12#1AN (28F)	57F	17,020-17,039	AAA CCA AAC GGA GAG CGC C
E ₂ 12R#1 (29R)	58R	17,112-17,131	CCT TCT CAG AAT GAA CAT A
PCR1CE#1.1EX (30F)	59F	17,435-17,455	CGA ATA GCA TAT CAC AAA GC
#4EX (31R)	60R	17,752-17,771	GCG TAA TAA AGT CGG GTT A
#2.1EX (32F)	61F	17,855-17,874	CGT GCC TTC ATA TAT TGC C
#3AN (33F)	62F	18,122-18,141	GCA CAA CAG ATT CTG ATC G
PCR1CH#3EX (34R)	63R	18,206-18,225	GGA GCC CGT TGG TTT TTG C
Luck 2t	64R	18,422-18,443	gtg cac gag ttt acc tct gga
#2ANEDAN (35F)	65F	18,609-18,628	CAC AAA TCA AGC ACA CAG G
PCR1CH#2EX (36R)	66R	18,711-18,731	GCT CTA TTT TGT GAA CTC A
7H 36R	67R	18,802-18,821	cca ata cag aga ctc aac a
7H F9370	68F	19,001-19,020	aca tcg tct cta tcg cca c
PCR1CH#1EX (37R)	69R	19,179-19,198	CCA AAG CTT CCC CCC TCG A
H7pR (38F)	70F	19,571-19,592	ATT TAA TAA CAT AAC AAA AGA
H723AN.1 (39F)	71F	19,668-19,687	GGT AAG ACA GCC CAC ATT T
H722.EX.1 (40R)	72R	20,115-20,134	GGT TTC AGA ACC GCA GGA T
H7pF (42R)	73R	20,176-20,196	GGC TGC AGG CGG GGG AGG T

H12FAN (43R)	74F	20,814-20,833	CGA CTC GGT CTA CAG TGG C
H12FEX (44F)	75F	20,814-20,833	CGA CTC GGT CTA CAG TGG C
H125-1AN.1 (45R)	76R	21,500-21,519	GCG CAC AGA CTC CTG GGC T
H1210-1AN.1 (46R)	77R	22,319-22,338	GCT GAT GAC TTT TGG AAG T
H1211-2EXT (47F)	78F	22,779-22,798	TCC AGA CCA AAT ATG CAC G
ES4-8EX (48R)	79R	23,494-23,513	GGT TCA TCT GGA GGT CAG C
ES4-4AN (49F)	80F	24,533-24,552	AAT TGA GCC CCC GGA CCC C
ES4-0AN1 (50F)	81F	25,153-25,172	TTT TAT TTG TAT GAA CTG G
7H 51F	82F	25,459-25,477	taa ttc acc tta cag tgc
A4 52.2F	83F	26,008-26,028	tgc aat ctg tca cta tcg tc
7H 52F (bad)	84F	26,099-26,119	act tcc cag tta tat gct cc
A4 53F	85F	26,479-26,499	tga gct tcc atg ata atg tg
A4 54F	86F	26,853-26,872	tta cag caa cag tgt gct c
A4 55F	87F	27,201-27,222	tga ccc cga gct aag tgc tcc
A4 R2750	89R	27,989-28,009	gac caa gtg ttt ctg cag ta
A4 F1480	90R	28,332-28,352	gac aaa cag aga cca aaa at
A4 R930	91R	28,865-28,885	ggc ttt gct tta tat act gt
A4 F590	92F	29,430-29,450	atg cca gca aag act gca gc
A4 R365	93R	29,433-29,452	cca gca aag act gca gcc t
A4/3B F1740	94F	29,689-29,709	att gcc tgt gtt cgt ctg tc
A4 F1170	95F	30,008-30,028	tca gca ctg cat gta aca cc
OTHER PRIMERS DESIGNED BY T. GIESLER			
E4PR23.2			CTG AAC CAG TGG CGT TTG C
PRC1CE#1EX			TTA TCC ATT AAT CGT TTC G
S22#1.1			TCC TCT AGA GCT GAC CTG C
PCRICH#2AN			AAC AAT ACT ACA ACG GAG C
H12FEXEDAN			AAA CAG CAG CTC CCC TGG C
H12 FANEDAN			CCA AGC AGA GTA TGT AAA GCC

APPENDIX C: DNA SEQUENCE OF THE *MORONE SAXATILIS* *Hoxa4-a10* REGION. (5' TO 3')

1	GATCAGTGGC	CGCACCGAGA	GCGGCGGAGG	TCACTACACT	GGGAGCGCGG	50
51	TGTGCGTGCC	CCATAGCACC	GCTACAGAAG	TGCCTTACGG	ACTACACAAC	100
101	TATGGATACT	TCCCAGGTAT	GGCTAAGCGG	AACGAAGTGG	GTCCCCAAAA	150
151	CATGGTTTCC	GCCTCAGGCG	CGTACATGTC	CAGCATGGAG	ATGTGGATGG	200
201	ACGCGCAGAG	GTCATGTCGA	ATGGACCAAG	AGCACCCAGT	GGGTCCCCAG	250
251	GTCGCGCCCT	GCTCGTTCCC	GCAGAACATT	AAGGAAGAAA	GCACCTACTG	300
301	CCTGTATGAG	CAGCCGAAGT	GCCCTAAAGC	CTCCACCGCC	GAAGATCTTA	350
351	CATATTCGAG	GTTAACGACA	ACCGGCCTGG	GCTCGAGTAC	TTGTACAGTT	400
401	ACCGATGGCG	GCGGCGGCGG	CGGCGGGGGC	TCAGTGCCTG	TGCCGGGCTA	450
451	CTTCCGCTTG	TCTCAGACGC	ACGCACATTC	TCATAAACTT	TACAACACCA	500
501	ACGGCCCCCA	GCCGAACCCT	TCCCATTCTC	ATTTTGGCCT	ACACCCCACT	550
551	GCTCCTGCTC	GCTTCCACAC	GCCACCAACA	TCGACTTCTA	CTCCAGCCAC	600
601	GGCGCAGGAA	GGGAGGAACA	CCGAGGAGCC	CGTGTCTCTG	GGAAACGCAG	650
651	ACCTGCAGCC	TCACGCTGCA	CCGGGGACCG	AGGAGGCACG	CGAGTCTCTG	700
701	GATGCAGAA	TCTCTTCAGC	CGACGAGACA	GAAGAACATG	ACAAAGAGAG	750
751	ACAAGCAAAA	TCCACTAAAG	GTATTCACCC	TTCCTTATAG	GTGTGGTCTT	800
801	GCACTCTTAT	TTTGTTTGCA	TGCAGCACAG	GTAGAGTGAG	TTTTTTTGTT	850
851	TTTTTCTTAA	TGACACCTGC	CAAGTTTCAT	TTGGCTGGTT	AGACTGCAGT	900
901	AACAAGTCAA	GGCACGGGCA	CGGAATTCTT	ACATTAACAT	ATGGGCTCCA	950
951	AGTCTAATGA	TGCTAGTTTT	TGTTGTATAT	ATGCGTTAAA	GTGATGTACA	1000
1001	TGTGCATGAA	TACGCCCCGTA	CAATTTAAAA	ATCAGTCAGG	GATTATAGGG	1050
1051	CAAAACTAGA	TTAGGTTTAA	TTCATCAGTT	TTACGCTAAT	ATGTGAAATT	1100
1101	ATATTAGTTA	TATTTATATT	TTATATTTAA	ACATACAATA	TGCATTTTGA	1150
1151	ACACCTATGG	CCTTAGCCAG	TATTTATTTT	AGGCGCTGCG	TAAAATGTAC	1200
1201	TAGTACATAG	GCCATGTTTT	GGCTTACATG	TTTTTTCTGT	GCCATATCTG	1250
1251	CTTCATATTC	ATTCACTCTT	TATGACCCAA	AAACTGTTTT	GTTTCCCTAA	1300
1301	ACTGTTTGTG	TACAATCTTA	CAGGCATTCT	TGTTTATGCA	TATGTCATGT	1350
1351	TTTCTAACTA	TTACTGACTG	TTTCTACTCC	AACTCTGGCG	ACTCTTACAA	1400
1401	AGAAGCCATA	CTTTACGCAT	AAGACACACA	CACACACACA	CACACAAGTT	1450
1451	ACTGTTACTG	TAATGTGTGT	ATAACACAAC	CATAAACACA	AACACACTAC	1500
1501	TAGATACAAC	TAGATACACT	GTCTCTCTGT	TTCTCTCTCT	CTCTCTCTCT	1550
1551	CTCTCTCTCT	CTCTCTCTCT	CTCTCTCTCT	CTCTCTCTCT	CTCTCTCTCT	1600
1601	CTCTCTCTCA	CACACACATA	CACACACTCA	CACACTCTCT	CTCTCTCTCT	1650
1651	CTCTCTCACT	CTCTCACTCT	CACACACACA	CACACACACA	CACACACACA	1700
1701	CACGCACACA	CACGTAAACA	AACACACACC	CTCATGTTGC	TGTGTGTAAA	1750
1751	ACCATGTATC	TGCCATTTTC	TTTAGGAGAC	ACAAAAAGTG	AAAACACGGC	1800
1801	CAACTGGCTG	ACAGCAAAAA	GCGGTCGGAA	GAAACGCTGT	CCATACACCA	1850
1851	AACATCAGAC	ACTGGAGCTG	GAGAAGGAGT	TCCTCTTCAA	CATGTACCTG	1900
1901	ACCCGAGAGC	GCCGCCTAGA	GATCAGTAAA	AGCGTCCACC	TGACGGACAG	1950
1951	ACAGGTAAAA	ATCTGGTTCC	AGAACC GCAG	GATGAACTG	AAGAAAATGA	2000
2001	CGCGGGAAAA	CCGGATCCGG	GAGCTCACCT	CCAACACGG	ATTTTCATGA	2050
2051	TGAACGAGGT	GATCACATTG	AAATAGAAGA	AATTTTGGCG	AAATCAAAGA	2100
2101	CGCCCTATAG	TCCTCACCAC	TTTGTCCACA	TAATGACTTA	TTTCACTGTG	2150
2151	GGGTATAACT	TTCTTTTATT	TCCAGACTAA	GCTTTAAAGT	CAATGCATCG	2200
2201	TCCTCTGTGA	GCTACTGAGA	AAAGAAAAGG	AAAAAATAG	AATGTTATTG	2250
2251	GTTCTATTTG	ATTGTTTTCT	GATGTCGTGT	TTCTGTATTG	TCGACCTGTT	2300
2301	TCTGTCACCT	CTGTCTGTATG	GTTGGATAGT	GTCAGTGGAG	TTAAATTACC	2350
2351	ACACTTAAAA	AAAACACATT	ATGTCCTGGA	AGGTGGCTTC	AATTGTCCAA	2400
2401	TTATCCCCCC	CACCACTAGC	AAATAATCTT	TTTTATAAAT	TTAATCAAAC	2450
2451	AAATATGAAT	TTACTTTTATT	TTAATTGGAA	TTGCATGGTT	TAAAGATTTT	2500
2501	CTTTTAGCAA	TTTTAACC CA	CCGTGGAATA	ATGCCCATAT	ATCACCAGGG	2550
2551	GAAGAATAAT	TCACCTATGA	CCATTAGATG	TAGTCTACCA	TGATATAGGC	2600

2601	CTCTTCTATG	TATGAGCTAA	ATGTATAAAT	TGAATCTTTA	TTCACATTGA	2650
2651	GGACGTTTGC	GCGAAAACGG	TGCGCAAAGT	GAAGGAGGAA	TGTTTTTTATT	2700
2701	GTATATTTCT	GTGTACATAT	GACATTGTGT	TTTTGTGTAA	TGCTTAATAA	2750
2751	ATTATCAGTC	GCATTGAGGA	GTAGCTGCTT	GTATTTTTGT	ACTGTAGGTT	2800
2801	TAAAACATAC	ATATTTAATC	AGTGTGTATC	ATACAGACGC	AGAAAAAAGG	2850
2851	CATTGTTTGA	AATTCAAAAT	CCAAAATAAT	ATGGGTTGTC	CTGCTGCAGG	2900
2901	TTCATTAGTT	TTTTGTGAAC	CACCTTAGAA	ATGTTTCTAG	AGGCTATTGT	2950
2951	AATGGCCGTC	GGAAATAATT	AAGATATTAG	GCTATAAGCC	GTCTAGCTGC	3000
3001	ATGCCCAGCC	GGCTATTGTA	GAGTTTTATG	ACTGGCTGGT	ATATGAGGTT	3050
3051	TCTCCCTCTC	CTCTAAGAGC	AGCTTCAGTC	GTGGGAGAGC	CTTTCAGACG	3100
3101	GAGAGGAGGC	CCGGCACTCA	GCTACAAATC	TTAACTGTAC	AGATACACTG	3150
3151	TGGATAAAAC	GTTACGTTCT	GCTTTAGTTT	GGGAATACTG	GGATTAACAT	3200
3201	TTCCAGGATC	GGGTAAGTCA	TTTATCAATG	TGCATTTGTC	CATTTTGC GC	3250
3251	CTTGTTGCAG	GGGCTTCACT	ATAGTGTGTA	CAATACAGTT	GTACTTTCTC	3300
3301	AATACAGACG	ATTTGTTGGA	CAC TTGATGA	CTGGACCGGC	GACCACGGTG	3350
3351	CATTATGAAT	TGGCTGATTG	TGTTTCACTC	GAGGCCCGGA	TCGACTGTCG	3400
3401	TGTGGTTTAG	GTAGTTTCAT	GTTGTTGGGG	TCCATTTCTA	ACTCTGCAAC	3450
3451	ATGAAACTGT	CTTAATTGCC	CCAGTTACAC	ACATCGACAC	CAAAAGGAAA	3500
3501	GCAAATCAGG	GACATAAAGA	TGGTGGGACT	TGATGACATT	TGTGGAGTGG	3550
3551	TGTTTTGGAA	GCATGTTGTA	AACCTCCCTG	GTT CATATAT	CGGGGAAGGG	3600
3601	GGTGACTGGC	TGTTGTGTAT	ATTCCTAACC	TGTCTCCTGT	AATCGTTTCT	3650
3651	TTAGCCTTCT	CCGCTTGTTT	TTTCGGCGTT	AAAATAAGCC	CGAGCGAGCG	3700
3701	TGTGCTGAAT	ATAAACACGA	TCAAAGCTGA	GCTGCAGTAT	ATGCGTCGTG	3750
3751	GACTTGGGAT	TAAGGAAAAG	CTGCCAGTGT	TTTCCGTGAG	TAGTAGTGAC	3800
3801	GCAGTCACGC	CAGACAAGGC	TCGGTTACGA	CCACTGCAGT	GGAGGGGAAG	3850
3851	GGGAAGGTGA	ATGGGGCAAC	AGCGACTTCC	GTGGGCAATG	AGACCGGATG	3900
3901	CTTCCGCTGC	GCAATAGTGG	CTCAGACAAA	AGGAGGGGAA	GGGCTTTATG	3950
3951	GCGATATTTA	GATCCGCCAG	ACAGGCTGTT	ATCTGCAGAG	GCAAAAGAAG	4000
4001	TGAATAGCCC	CTGCGGTAAA	CTGGGTGAAA	AAACACCCTG	ATTTAGGTTT	4050
4051	ACAGCCTGAG	AGCCATTGGC	ATAGTGGCGT	CTTTTCCTGC	TCTGGAGGTG	4100
4101	TTTGTCTATT	CTAACTTTTA	GTGCACAAAT	AGGCGCACAG	CCACTCCATT	4150
4151	TCCCCCTCCC	ATGCTTGTTT	GATGTCTGTT	GGCCAAATAA	TGTGCAAGTA	4200
4201	CCAGTGAGGT	GATAGCTTGG	AACTACAATG	TACATTATGG	ACAGCTTG TG	4250
4251	TCTTTTTTCC	TTTGCTCCTC	TTCTAAGTGC	AAGAGTAACA	GCAGAGTGA	4300
4301	AAGAATTCCCT	CCGTGTTGAA	TTGCCTGTGG	CCTTGTGGGG	TGTTGGCCAA	4350
4351	GCTTTCCGTT	TATTTGCTGG	TAATAAATAA	AACACGCGGC	GTAATGTGCG	4400
4401	AGAACTCAAG	GTTGCTGCTT	TGTGGACGTT	TTCTGCAATG	TGGGCTTAGA	4450
4451	TGTTGGATAT	TATGGACGCT	GATACGCACA	CAGCCAGCAC	GTTTAGTAAA	4500
4501	AGCCACCTTG	AGTTTAAAGT	ACTGGTGTAT	TTCTGTTATA	TAGGCTACGC	4550
4551	AAATTCTCAG	TATACTGTGA	ATATTGTCTT	ACAACAATAT	ATTTACAATT	4600
4601	AAATTGTATT	GTATTAAATT	GAAGCTTCAA	CAAATACAAA	GGTATCATTT	4650
4651	ATTTAATTTA	ATCTTCACAC	CTCTAGCGGG	CTATTTTCTG	AATGTTTATT	4700
4701	AATTAATTAT	ATTTAATTTT	TTTGTTTAAT	TGTATAATCT	CAACTTTCAC	4750
4751	TACATATTAT	TAAAGTGTGC	TTTCGTGTCT	CGGTGTTCTC	CATGCTCGC	4800
4801	ACCATTTAAT	TTTAAAAGGA	GGAAACCACA	AACACCACAT	TAGGCATATT	4850
4851	TTTAGCTCTA	TTTGAGTGTA	ACGTTAAACC	GAGGCAC TGG	CTGCTAGCAG	4900
4901	GCATTAGAGT	TAATTTGTTG	ATGGTGGTGA	TGATAGCAAC	ATATAGAGGC	4950
4951	CCATACAAGT	GATAACGGCG	GTGTAAATAA	TAAGCGGCAG	GTTAAGCGAG	5000
5001	AAACCGGCAA	TCCGCCATCA	GTATGTTATT	TGGAAAATAT	TTAATAAATA	5050
5051	TTCGGCTACC	GTGGTTCAGT	TCAGCAGGTT	GATAACAAAT	CTTTAACCAC	5100
5101	CTTTATTTAA	TTTTCTTTGC	TCGTTTTAAT	GGATGTAATT	CTATATTTTA	5150
5151	ATTTCTATGT	TTGTGCCCCAT	TTTGTAACAC	ACTTAGTACT	TTATATTTCA	5200
5201	TAACGTCAAG	GTCATACGCT	ATCTAGGAAT	TGTTATTATT	TTATTTGTTA	5250
5251	CATAGACCAG	TGGTTTTTACA	CAGTAATTAT	AGTCAAACCA	ACAATAGTCT	5300
5301	GAAACTACAA	CTTAGAAGGT	AAATTTGGGA	AAAAAATAGT	CCTGCTTCTT	5350
5351	TTTTCTTTTT	GTTAATCTGA	TTGATTGTTG	AAATTTTCAA	AATAAAAAATA	5400
5401	TTCCTTAAAC	ATTTTAAAGT	AATATTCTCA	TGTTTATTGA	TTTACTTTTC	5450

5451	AACTCTAAGA	TATGTACTAC	ATAGTCCTGC	ATTATGGTAG	TTAAAACGGT	5500
5501	GCAGATGTTG	AAATGTAGAT	AATTAGATGT	TAGGGATGTA	GCTCCAACAT	5550
5551	ACAGGGACGT	GAAACGTCCA	TGTCTAAATC	ACAAAAAGTG	GTTAATGTAT	5600
5601	AAAAATAAGA	CCAAATAAGC	ATTATGCTTA	TATAATCAGT	ATGCGTATAT	5650
5651	GTTTTATATT	TTTTAGATGA	GCAACTATCT	TGTGTGGCTA	AATGTTAAAC	5700
5701	TGAAAATAAC	AATTTTAGCC	TAATATAAGA	TACACCCCGC	ATGTATTTTG	5750
5751	GTGTATTTGT	CAGTTTGGTT	TGTAATAATAT	GTTTGAAACT	TGAAGCTGAA	5800
5801	AAACAGGAAA	ATATACACAC	ATTGGAAGAT	GTCTGACACA	GCCGAAAAGG	5850
5851	AGCAAAATAA	AAAATGAAAA	AAGTAATTTT	TAGTTGAGGT	CTGCATGTTT	5900
5901	CAGCTGTGTG	CTGCTAAGTT	TCTTTAATAC	AGTGCTGGTT	AAAATGTGTT	5950
5951	TTGGTGAGCA	GCAATGGTAG	GGGGCTTACG	TGTCCTATAA	AGTACCTCGT	6000
6001	AAAATAAGGT	TTCCAAAAAC	ATATTGTATC	GTTTTGGATA	ACCAAGAAGT	6050
6051	GATTTTAAACA	GCAGTAGTAG	GGTAACACAC	AGTAAAATCA	CATTTTATTTC	6100
6101	AGTCATTCAT	ATTTTCTGTC	TCACAGATAA	CAGATCTAAC	TAATATCTTG	6150
6151	ATAATTCTTA	GCTTTAAAAA	TAGAGGTTGT	TATAGGAGAG	CGGGTGAGTA	6200
6201	TTTAATGACC	AAAGGGATTG	ATTTATCCTT	TATTGTTTTCAG	CCTTTATGAT	6250
6251	CACGTGTGTC	TGCTACCCAA	TAGCATGGCA	GCCTGACTCC	CCATTACTAG	6300
6301	CCCACTGTAG	TTCTCTGTGG	GGCCAAGTTG	CTACTTGATT	TCTCCACATT	6350
6351	GTTATTTTGC	GAGGCTGGGT	TACTGCGTG	TGTATGTGTG	TGTGGTGTGT	6400
6401	GTGTGTGTGC	GTTTTACTTT	TGCATGTATC	TGTTGTATGA	CTGCGTACAT	6450
6451	GTGAACGCAT	AATTTCTCTG	CCATGTCGAC	ATCCGGAACA	CTGACTAGTT	6500
6501	ACTACGTCGA	TTCCCTCATT	CTGCCCAGAG	GCGAGGAGCT	CTCTGTGCCG	6550
6551	AGATACCCCT	CCGGTCCTGG	GCTCCAGCAC	GCTAGACAGT	CCGCTCCCAT	6600
6601	AAGCGACCAC	AGTGAGCTTG	GGACCTGCAC	CTTCCCGTCC	AAACCTCCTG	6650
6651	TATTCGGGCC	GTCGTGGAGC	CACGTTCTCT	CCCAGTTCCC	AGGCACCGTC	6700
6701	TCTCTGTGT	ATCATCACC	CTATGGGCAT	CCCCAAGGCC	CGGTCGGTGC	6750
6751	AGATACAGAT	GGCCGCTATC	AGTCATGGCT	GCTGGAGCCC	ATGTCTGGTT	6800
6801	CCCTTCCCAT	GGCCGGATTA	CCGACGACCC	ATCACTACGG	AATCAAACCG	6850
6851	GAGGGTCTGG	GGACGCGAGC	TGACGGCGCG	CTGCCCGGCT	CCCACACGGC	6900
6901	GCTGCTGCTC	TCTGATTACG	CCAACGGGAC	CGTGGCCACG	GCGTCACCGG	6950
6951	TGGAAGAGGA	CACACTGTCC	GGCCAGGCAG	GGGACGTGAG	CGGGGAAGCT	7000
7001	GAAGAGAAAC	CGGGCCTTGA	TCCAAGTAAG	TCGCTGCAGA	GTCCGTTATG	7050
7051	TTTATAGCCT	ACTATTTATA	CCGCATAAAG	CTGCGAGGAT	TGGGGTGCGT	7100
7101	GATGGGGGTG	CGCGTCTGAA	CAGATGCCAT	TCTGAACGTG	CCATTTCCTT	7150
7151	ATGGGGACTG	TTAATGGCCC	AGCTCCCAT	GCTATATTAG	ATAACCTTTT	7200
7201	GCAGCTTTTA	AAGTGAACGT	GAGCACAGAG	AGGAAAAGGG	GGAGTTGTGT	7250
7251	TTGGGTGCGT	ACATGTGCAG	CCATTTTCTC	TTCACAATAA	GGGGGGGGTT	7300
7301	TGTTTTTGTG	TCTCTTTTTT	TTCTTTTTTT	TTTGGTCTGA	GCCAGATTAT	7350
7351	GAGTGATGTT	TGCTTTTACA	TTTGGTCACT	CGGTACCGGG	CCCCTCACAC	7400
7401	CAGTGTGCGC	CCTTCCCCTC	CCGCTCTTTC	CAGATAACCC	AGTGTCCAAC	7450
7451	TGGCTCCACG	CGAGTGCCAC	GAGAAAAAAG	CGCTGTCCGT	ACACCAAGCA	7500
7501	CCAGATCCTG	GAGCTGGAGA	AAGAGTTCCT	CTTTAACACC	TACCTGACCA	7550
7651	GGGACCGTAG	GTACGAGGTT	GCGAGGCTGT	TAAACCTCAC	CGAGAGACAG	7600
7601	GTTAAAATCT	GGTTCCAGAA	CCGCAGGATG	AAGATGAAGA	AGTTTAATAA	7650
7651	AGACGGCGCA	CCGAAAAGAC	AGTGACTGAA	CCGTAGATAT	ACAGTATTTA	7700
7701	GGCTGTTAGC	GAGCTTTGTG	TGTGAAAGCT	CTTAACTCTG	CTGGACCTTG	7750
7751	TTTTTAAACT	TCATTGTCTT	ATTGTTCTAT	ATAACAGAAG	CTATTAGGCT	7800
7801	TATTTGTCTT	GTTAAAGAAA	AAGCATGTTG	GGCTTCCAGT	TTGATGCGAC	7850
7851	TACCTTTATA	TTGCACAATT	TTAACAGTGC	CCACTCCGTT	TTTGTATTTG	7900
7901	TTCTTTTCTC	ATGTTGATTT	TAAATAATAC	CTCGTTCTGT	GTATTAGTTG	7950
7951	GCCTGGTCCC	TGTGCTTGTT	TGTGTATGGA	TGTTTTTAGC	TGTGTGTGTT	8000
8001	CTGGTTTGTA	GCTATATTTG	TTGTCGTATA	GTTTGTAATA	TGTATTAATT	8050
8051	TTATTGTTGT	TTCTCCCTTA	CTGCTAGAAC	AATGACTGAT	GACCATGAGT	8100
8101	TTAGTCTAA	CATTGGAACC	GTAGAAAAAC	AGACTGATCG	CATTTTGAAG	8150
8151	TAGGCGCATA	AACCTCAGAA	ACCTCTTGAA	AATGCTAAAA	TCCTAAATTA	8200
8201	AAAGATGTTA	TAGAAGTTAT	ATTACATGAC	ACGATTGACT	ACCTATGTCT	8250
8251	TTCAACTACC	TACAATAACC	TATTTTTTTT	AATGGCGGAT	TATAGTAGAT	8300

8301	TTGCTTGCAT	TAGATGCTGA	ACTATGAATT	TCAGGGAGTA	ACATTTTGAA	8350
8351	AGTACATAAT	GTTTGTTAGT	GAGAGAAAAT	TAAATAAAAC	ACTGTAGCTA	8400
8401	CTTATTGAAT	ACCTCACGGT	CTGCCGACCG	TCTGACTTTT	GGTTCACAAA	8450
8451	TGTAATTTGT	ACTTTATTTA	TTAAATAAAA	CAAATTATAT	CTGCGAATGC	8500
8501	AGGCACGTTA	CTTCTATTTT	ATTTTGGCCT	CACATCGAAA	TGAATTTTGT	8550
8551	TTAAAAAAA	TAAAATGTAT	GTTAATTTGT	TGCTTCACAA	AAAGTCTTTA	8600
8601	TTGGCAAAA	GGTTGAATTT	CAACACACAC	GCACACGAAA	ATACATAAAG	8650
8651	AAAAATCCAA	AGATATTTGG	ATACTTGAGC	CTAGTCCACA	TCAACAATTT	8700
8701	TTTTACACAT	TTTTTATAAA	TATCAAGCAT	TTTAATCTGC	AGAGTGCACA	8750
8751	CTCTCTCACA	CACACGCACT	CATTTACACA	CACACACTCA	TTTACACACA	8800
8801	CACACACACA	CACACACACA	CACACACACA	CACACACACA	CACGAATTAA	8850
8851	AACGCACACA	CCTATAGGTG	AATATCTGTG	TGTGTGTTTC	TGGACTGGGG	8900
8901	ACAGAGCTGC	GTCTTGGCAG	AGGATGTTGT	TAGCTGTATA	CAGCCATAAA	8950
8951	AGACAATTAC	CGCTATAACC	TTTTATGGGG	TGCAAAGCGC	TGCGAGGCGA	9000
9001	GAGGACACAA	AACAAAAAAA	GACATCGAGG	GCTTCGACAG	CTGAAGCCCC	9050
9051	AAATAATGTA	ACCGACGTGT	TTCTGTTTCAG	CTGGGATATG	ACAAGGCCGC	9100
9101	TCAGCCTAAA	GCTCCGTCAT	TAGCGCCTCA	TGTCCACATA	ACCAATTTAC	9150
9151	AACAAGAGAT	TTTGTAACATA	AATATTGTAA	TATATAAAAAG	CATTTTGGAT	9200
9201	TTTAAACGCG	AACGTGACCT	CAAATAGTTT	ATCGAAAAAG	ATTTTAGAGA	9250
9251	AGATCTCTAG	TTGTAAAATA	TGGCCTCTGC	GTTTTCAAAA	GTCTCTGCAT	9300
9301	GTTGGAGAAA	ATATTTGGCG	CAAAGACAGA	AAATACGCAG	CGAAATCTAA	9350
9351	TGTGTGCTAT	GTGCAAAGCT	GCTGTGTCAG	ATGGCGAATC	AGGGCCTGGG	9400
9401	ACGGTTCTGC	CACACACCCC	AGCTCATTTA	GGTCTCATTA	AGGGCTGCAC	9450
9451	AGAAACGGAT	ACATTATTGT	AGTTAGAAGG	TAAGTCGGAC	TTTAGTCCAA	9500
9501	CTGGAGCGCT	GTTGTGACGC	GTTAGGGGAC	GCAGATGGAC	ACAAACTTCA	9550
9551	AAGGCTCGAC	CAGAGACAGT	GCAATAAAAC	GCCTGGTCTG	CTATACTGTC	9600
9601	TGGCATTCCA	GTTTTAATGG	CTTTATGGCC	GTCCAGACAC	AATTAGGCCG	9650
9651	TTTCCAGAAT	GGCACCCATT	TGTTTTTTCT	TCTCTTTCTG	TGAGACTGCG	9700
9701	CTCTGGACAA	AAGGCCGAGC	GGAAATGATC	AGCTTTATTG	GATCCTCCCC	9750
9751	GACGGGGACG	CACAGGACTC	ACGGTCATTT	GGAGCGATCC	TTTGTCTCTA	9800
9801	CCTGGGAACC	TTCCGCCCTC	TGCTCCTTTG	CTGTGCCCCG	CCTAAAGGAG	9850
9851	AAGGAGCAAT	AAAGCCAGCA	GCTAGTTTCA	AACACAGTGG	CAGGCCTATA	9900
9901	GATGGTGTGT	TTGTTTTTGT	TTTTTTTCAA	ATACATCTGG	ACAGGTCTGT	9950
9951	TAAAGAGCAT	GACCACTTTG	CTTGAAATCG	CAATCCAAGA	AACGTAAACG	10000
10001	TACCTTTGAG	CGCCATTTCT	CTAACTTTTA	GCTGGTATTT	TATCCGGCAT	10050
10051	TGCAGCAGGA	AGACAAAACG	CCTGCGCACT	GCTGCCTCCC	CACCCTGCAG	10100
10101	TCCCTGCATG	TCACCAGAGG	TGGCGACACT	GTTTCACAAC	ATCAACCCAC	10150
10151	CGCTTTTAAC	ACTCCTCCCT	AATGAACCTG	CGCTGGGGAA	CAAAGTGTAG	10200
10201	TAACATTTTA	AGAGTCCAGC	AAAATAATTT	CACACCTTTT	TACAGTCTCA	10250
10251	GTGCAGAAGC	TGCTTCTCCA	GCCTGCGGCT	TTGTGGCGAA	GGATGAAACA	10300
10301	GGCAGTCTCT	CAGTGATTGC	TGGGTTACTG	CTTTCACCTT	CAGACATGGC	10350
10351	TCTCTGTTCA	GAATTTGTCC	TTCTAGGATG	TTTTTATATG	ACTTTATCGA	10400
10401	TGCAAATGAT	TTAAACATTG	TTTGATAGCA	GGGTATTATA	GGCTATCATG	10450
10451	AAATATGAAC	AGTTTTTTCT	TATTGCATTT	ATGAATGTAG	ATTATGTAGA	10500
10501	TTTTAATGTT	AACCTGTTAC	TTTCCTCTGG	AAAATGTCAA	TAGGCAAACA	10550
10551	CAATCGGCGG	ATAGAGCTGC	ACAAATATAA	TCAATCGTTT	TCCATATTCA	10600
10601	CAAATAATTT	AACGAATTAG	GGGTGTATTG	ATGATATCAT	TTTTAAAATT	10650
10651	CTTATTATAA	TTCTCAATAT	TATTCTTATT	ATCATTATTA	TTATTAGAAG	10700
10701	TATATTTATT	GTCATTATTA	TTTTTAGTAT	TATTAGTAAT	AATAATAGTA	10750
10751	ATAGTAATAA	TAGTAGCCTA	CATAAATTCC	GACCATTTCG	TATGTTTTAT	10800
10801	CGTATTGTAT	TTTTTGTGCG	CATGCTATCT	TTAATGCTAA	TTTATTTAGG	10850
10851	TAACATTGTA	ATGCAGTAAA	AACAATGTTT	GGAGGATGAA	TTACGAGATG	10900
10901	GTTGTTATTA	TTAAATAAAA	CTAAAAAAC	TAAAAAAAC	TATTAAGGC	10950
10951	ATTTTATTTA	TCTTCTGCC	AAATTCTCAC	CTCATAAATC	CATTGTTACA	11000
11001	AATCTGTGTG	AATTCAATAC	ACGTGGATTA	TTTATCATGT	TATATATTGT	11050
11051	TATGTGTCTT	TGTTTTATGA	CACTAGACCG	TATGTAAGTA	TTGAGTTTTA	11100
11101	AAGCAAGATA	AACTAAAACA	TTTTCTGTTG	GAATATTTTT	TTTTTAGTGA	11150

11151	TGATAATCGC	TGATTAGTTA	ATTATTATCT	ATTGTAGACT	ATTATTGAGT	11200
11201	GAATGTAGAT	TATCATTATA	ACCTATAGGC	CTATAGGTTA	AAGACATTCC	11250
11251	ACGAACATAT	ATTCAATTTA	AAAAAACAAA	ACTTCTAGAT	AGAGTATTAT	11300
11301	TTGTAAAGCA	GTAGTAAAAT	AAGTGATATA	AAACAGTGAC	TTTAAATTTT	11350
11351	GTAAATTACC	TCTCGATCAT	TGGGATTGTC	TGAGTTTCTC	TCGTGTGTGT	11400
11401	TTATGTGTGT	GTTTATGTGT	GTGAAGTGTT	AATCCCCTC	TGCAGTCCTG	11450
11451	CGCCATAATG	AAATACATGT	CTGCTTGCTG	ATTGGATGAG	AGCGGCCAGG	11500
11501	CCCAGATCAC	GTGGGTGAGT	TCTTTGGTCC	AGAGTAAAAA	TGGGGTTTGG	11550
11551	TGTAAATCTA	GGGTGTATTG	CTGTCAATATA	TCACACTACC	TCGTAAAAAC	11600
11601	GACACTGAGG	ATTCTGGGCC	AACAAATCAG	AGAGAAAAAA	TATGAGTTCT	11650
11651	TATTATGTGG	ACGGTCTTCT	TAGCAAATAT	ACGGCGGGTA	GTTCACTCTT	11700
11701	TCCAAACGTG	GAGCGGGCGT	CTTGCACTAT	TGGACCCAGT	GGAGAGGACT	11750
11751	ATGGCTCGGC	TCGGACTACT	GCGGCCATCG	CGTTTGCACC	GTCCCTGTCT	11800
11801	GGAGTTTACA	GCGTCAGTGA	TGCAGTGATC	CAGAGCCGCC	CCGTGTTTAC	11850
11851	CTCGGGCTAT	GGCCAGGGGC	AGGACGCGCA	CACACTGCAC	TGCAGCCTGT	11900
11901	TCGACCAGAG	CCGCCGTGTT	ACCGACAGCT	GCTGCCATCC	GGAGCCGGGA	11950
11951	CCCCTCACCT	CACCGCCGGA	CAAACAATAC	CGGATGTATC	CCTGGATGAG	12000
12001	AGCGTCAGGT	ATGTGCTGAT	ATTGTCTCTT	ATGTGCTATT	TCAAAGCCCC	12050
12051	TCGGGTTCGAT	GCTTTTACAA	AACATTGGGA	TACTTACATT	GGCCGTTAGC	12100
12101	GGTTCAATTA	ATTTCTTGAA	GCCAGTGGCG	TTGCATGGAA	AGAATCGTAA	12150
12151	AACTTTTATT	GCGCAACTAA	TGAGTTCTGC	AGCCATAAAT	TCATTTGGCC	12200
12201	AGCGCTCCTG	TGTGTTCACT	ACGCAGAGAC	CTTGCCGCTG	GGGGTTTTCT	12250
12251	CTGCACTCCT	ATTACTCTTA	CAGCCTCCTC	TCCCCTTTGT	GGTTGTAAAA	12300
12301	AAGACCGGCC	TGTTGTGTAA	ATCAGTTCTC	TTGCACTCTA	CAGCTCCTTT	12350
12351	CTAATTGGGT	CTGCACTGCA	GAGGCTCTGG	TATTTTCTGA	AACTCGTTCC	12400
12401	CACACATTGC	TGGATCAATA	TGCAGCAGTA	TTGTCTATTA	CATAAAGAAG	12450
12451	TCAAGGGGAT	TTAGAATACA	AGCATTTAGA	AACATTGGTA	TTTGTATATCA	12500
12501	TTTTGAAAGA	GAATGTGTCT	AATTACTTAT	GGCCCCCTTC	GTGAGCAGAA	12550
12551	TATCGCTGGC	TTTTAATTGG	AAAGGCATGC	AGAGTATCAA	AGGAGGAATG	12600
12601	AGCCAAACAC	CCCAGCAGGA	CTACTTCAAA	TACACACAGA	GAACTTAAAA	12650
12651	AAAACTTCT	TTCACCTAGA	GATTTGATTG	TGGGATGTGT	GTGCGTGCGT	12700
12701	GTGTGCATGT	GTGCGTGTGT	GCATGTGTGC	ATGTGTGCGT	GGGTGGGTGG	12750
12751	GTAGGGGGTA	ATGTAGAAAT	ATGAGCTATA	ATTAATTGCG	CAATCACAAA	12800
12801	TAAAAGAATG	AGTCTCACGC	TGATTCCAAC	TACAAGTTTT	TATTATCAGT	12850
12851	ACCAAGTTTGA	ATACCATCAA	TTATTGCTTC	ATGTAGTATC	ATATTTTATA	12900
12901	CATGCAAGGA	AAAAAATGTC	ATTTTCTGCA	AATTATATAG	AGCTAGTGCA	12950
12951	TCTTCTCTTC	CACAGATCCA	ACCCGAAAGC	GGGGGCGGCA	GACCTACTCC	13000
13001	CGGTACCAGA	CCCTGGAGCT	TGAGAAGGAG	TTCCACTTCA	ACCGGTACCT	13050
13051	GACCCGCAGG	AGGAGGATCG	AGATAGCCCA	TGCCCTCTGC	CTCTCAGAGA	13100
13101	GACAGATCAA	AATCTGGTTT	CAGAACCSCA	GGATGAAATG	GAAGAAAGAC	13150
13151	CACAAGGAGG	AACCGGTGTC	GACCCCTTCC	GGAGAAAAGG	ACTGTGATAT	13200
13201	CCAGCCGGTG	TCAGAGGCCG	AAGAGCCCAA	AGCTGTGAAC	AACCCGGAGC	13250
13251	GCCGGTTTGG	TGTCAGAAAC	GGCTGAAAAG	TAACGAAAGA	GATGGACAAG	13300
13301	AGCATTAAT	GAAAGGAACG	CACTATTAAG	TGTTAGGGAT	ATACCTGAGA	13350
13351	TGTTACTCAC	ATACCATTCA	TGCACTAAGA	GCTTTGTAGG	CTGAATAAGT	13400
13401	GAATGACAGC	TGAAGGCAAT	TATGTAACAA	GCATTTAACA	GTTGTAGGCT	13450
13451	GTAGCTCTGT	ATTTTCATATT	TAAACCTAAA	AAGCATAAAA	AGGCTCCGCA	13500
13501	TGAAACAGCG	ATATAACTAC	CTATCCAAAC	TATATTTTTA	TATATGCAAT	13550
13551	ATAACATGTC	ATTTAGCCTG	ACTGAACTAC	TGAATCAACA	TTCTTGAATT	13600
13601	AAAACTACC	TAAAATGTAC	CTGCTAATAT	TAAACACTGA	CACAAGGAAT	13650
13651	TCAACGCAAG	TTGAAATTGA	CTGTAAATTG	ATAATCTACC	TAACGGAATA	13700
13701	CATTTTCTCC	ACAGACGGTT	GTATTTTATA	AAGGGAACAA	ATACAGCAGT	13750
13751	CTGCGTAAAT	GTGCCTTATT	TGGAATAAAG	GCTAATTAAA	AGTGAATGTT	13800
13801	TTGTTAGTTA	AATGCTTTTA	ACTGTGATTT	AGGTGCTTCT	TTAGATTTTA	13850
13851	ATAATTGTGC	CACGTGTGTT	TTATTTGATT	TTATTATCAC	TGCTTATTAC	13900
13901	TTAGAATGTT	AATATGCCTG	TTTTACTTAT	TTACCCTATA	TACATCCCAG	13950
13951	GATTAATAAA	AAACATTTGT	TTAGTTTCGG	ATGAACCTAT	GTAACCTACC	14000

14001	TCAGATCCCA	GAAAGTCGAC	ATGCTTATAA	TATTGTAAAC	AGGAATACTC	14050
14051	CAACGTAAAT	GTTTTGAAGT	CTTAGCCGTG	TACTGTGAGA	ATTTTCTTTT	14100
14101	ACACAGATTG	TCAGGCTCAA	AATGTCATAC	GTGTTTTTAA	ATGCCAGCGA	14150
14151	TAAATGCATG	CCCTTTTATT	TAGACAGAAAT	TTACATTTCC	GTTTTTTTAGT	14200
14201	GTAAAATGAG	TAAGGTATTG	TAATCCAGCA	AGTGCAAATA	CTGTGATATA	14250
14251	TTTAGGTTTA	GTGCCCACGG	TTTCAATGTT	GAATCTATAA	GTAATATGTG	14300
14301	AACAAATCTT	AAAGAGTCTT	AAATCTCTGG	GAATGTCTCA	TACCTGTAAA	14350
14351	TGGAAATATA	TAAATTGAAT	AAAATAGCTT	TTACCTGACT	ATTTGTTTCC	14400
14401	ATTCAATATT	TTGCCAACTT	AATATTGTGC	AACCACCGCA	ATCACAATAA	14450
14451	TAATCACAGC	AATTACGCGC	CGGTTGAATA	ATATGTTCCA	CTAATTGCCG	14600
14501	TCCACTGTGG	TCAATATTAG	GTTGCCATGT	AATGAAACAA	GTTGCAGTCC	14550
14551	ATTATGAGGG	TTAGGTAAAC	ACTAGAGTTG	GGGAATGTAA	ACTATGAGGC	14600
14601	TGCGAGATGG	GGCAGGCTGC	AGATGAAGGA	AAATTACTGT	TGAACTTGCA	14650
14651	ATTCCTCTCT	TCCCTCCTGT	ATGCAGTAGG	ATGGACGCCA	TGATCGGTGA	14700
14701	AACATGCCAC	GGCCCCCTCC	CCATTGGCCC	CCTCGCATTC	TTTAGTCAGT	14750
14751	GTTTTACACC	CCCCCCCCCT	TATTTTTTTT	CCCCGTATCT	CTCCCTCCAG	14800
14801	CGATGGGTCA	TAAATCCGTT	GTTTCGTTTAG	TGAAAAATTA	CAACAGTGGT	14850
14851	CAATACAAC	CTTAACGACC	CGGCTCGACA	TCATATTGGC	CCAGCTGTGT	14900
14901	CACGTGACAC	CGAGCCGTGA	ACATGAACTT	TTCTATATAT	AATTTCCAAT	14950
14951	AGAGGAGATA	GAGCAGCCTT	TCATTTAACG	GTTTCGGGCT	TCCATGCGGC	15000
15001	TGTTTTACAC	AGCAACCTTC	TACATTTTTA	CACCGCCAAG	GGGCAAATTC	15050
15051	TCTTATTATT	TTTAATTTTT	TATTTTTTCTC	TCGGTCGGGC	AGAGGCTCCA	15100
15101	AAGCGCCAGG	GGCTTTATTA	CAGAACATGA	GTGGATGATT	TTGGGTATCG	15150
15151	GTGTCGTGTA	CATGTTTGGG	CAAGATGAGG	GGAAGGAGCC	CCGGTGCGTC	15200
15201	ACTGACCCCC	GGGAGCCCCA	GCTTCCGTCC	GGGGACTCAG	CTGGATACTT	15250
15251	CCTCTCCGTT	ACTGGCCGTC	TGCACCTGCC	AACTCCCTCC	CCTCACTTCT	15300
15301	CCTGTTTACA	AGCAGAGACC	TCCACAAGGT	AAGACAGCTG	AAACTTGCAT	15350
15351	GTTTGCTCTT	TTTGTCAATT	GCAATGGCAA	ACAACCTAAG	CAAATAATAA	15400
15401	CCTTTTTTGT	GTGTGTGTTT	AAATATCACG	CCTCATTTGT	CTTTTGTTTA	15450
15451	AAAATCCAGT	TGTTTCAATC	CTGTGGAAGA	GATGGTTTCA	TGTTACAGTT	15500
15501	TGTTTGGTCG	TGATTTTAAT	GACAGTTGCT	ACGTTGTAAA	TCAATCTGAG	15550
15551	TTAGATAACA	TTTAGCACCG	GTTTGGCTTT	GCGTCTTGGT	CTTCTGTGTTG	15600
15601	GTTTTTGCAT	GCTGTTGACA	CCTTATTGTT	TTGTTGTTGG	ATTTGAAATG	15650
15651	GGATGGGCTT	TTGCGCTGAA	AAATTCAATTG	GATGGTGGCC	AGAAGAGGTT	15700
15701	CTCTTTGACG	CCCTCCTGTG	GTTAAACTGG	GGTGTGTGTTG	TTGCAGCTCT	15750
15751	GCCTGTTTTG	CTGCAAATGA	CCATCAGCCA	AGTTCGACCT	CTATTTCTAC	15800
15801	CTGTGTGCCC	ATTCATAATG	CTTCAAGGTT	TCATACATGT	AGGCCCCGAC	15850
15851	TCACTGTACA	TTGAATTGAT	AATTAAGATA	GTTTACAGTG	AAGAAATCAA	15900
15901	TACCATTTC	CCGTTTGAAT	GACTTTAATT	GACACACAGG	AATGTCCCTG	15950
15951	TCGATGGAAT	TTATAAATGC	GACATATTTT	TGCTCTCTGT	TTGCACATCG	16000
16001	GTGTTGTCCA	CGGTGTTGGA	CAACATGCAC	TCAACAGCCT	AACTATATTG	16050
16051	TGCACTGCA	GTGATTTGTT	TATTGCTTTT	AATTATTTGT	CAGTCAGGCA	16100
16101	ACCAATAACT	GAAGATTAGA	AATGGAAAGC	ACAGACAGGT	TTATTTTACC	16150
16151	AAGATATCAA	TCGTGCAAAA	GTGCATGTGG	TTAAATGAAT	CAAATATAAT	16200
16201	GTTTCGTTTT	ATTCTGCATG	TAACATTTGC	TCTTTCCGGG	TTTTTTCTCT	16250
16251	TTGCAAAAAT	AGTTTCACAT	CAATGTGCTG	TGGAAAAAAC	CGTTGCTGTC	16300
16301	TTCATATGTT	TCATATGCTC	TGCTCATCGT	ATTCCACAAA	AAGACATAAC	16350
16351	TGTCCTAAAT	GTGACCACTG	AGAGTCGTTT	ACAGTCGCGC	GTAATGATAT	16400
16401	TTAAGCAGAA	AATAAATTAA	GATAATACGA	AATTTTATTT	GTATCCTCCG	16450
16451	AGTGCTGGCG	TGTGAGGGTG	CCTGCAGAGT	GCACTGCCCT	CCTGTTTTCC	16500
16501	TGCCCCCTCG	CTCCCAGTCC	GACATGTTTG	ACACTAGTGC	GCACTCACTT	16550
16551	CTTTCCTCTC	CTTTCCTTTG	GTCCTTATTT	AGTTCATGGG	TGTGCTCTCTG	16600
16601	GCTCGACTCG	GGCAAATAAT	TTTCGTGGGA	ACTTTGTTGC	CATTGGCCAG	16650
16651	CCATAAATCA	GCTGTTGGAG	CATGGCTGTT	TAGCCAACAC	ACACACACAC	16700
16701	GCACACACAC	ACACACACAC	ACACACACAC	ACACACACTC	AGTTAAGTCT	16750
16751	TCCCTTATTG	TGGGCCCCACT	GGATGTGGGA	GCCCGGTATT	CACCTTTCCT	16800
16801	CACATCGTGC	CTTCTGGGCG	TCTTTTACC	CAAAGGGCAC	GCATACCTGG	16850

```

16851 GACAGAATAA ATAATCAATA TCTCACTGTT TTTTGTGAGT AACACATGAT 16900
16901 TGGGCTTGTTG CTGTGCAAAG GGGCTGTTAT TGGGAAATAG CATGTTGGGG 16950
16951 GAAAGCAGAG GACAAGTCGC TGCCTTCGAG AGACGAAGGT GCCATAAAAA 17000
17001 TTCCACAAGG AGAAAACCAA ACGGAGAGCG CCTGACGAGG CTTTCATGCC 17050
17051 CTATTTAGGC TTCTTGCCCTG GACAGGCTTT GTGCGGATCT CCCCGTTTTT 17100
17101 GTATTCCTTC TCAGAATGAA CATAAACAAAT GGTAGATTTA CACGATGAAA 17150
17151 TGAGAGCATT GGTGAGGTTA CAAAAAGAGA GAGTGCGTCA GGGTTTTTTT 17200
17201 TAAACGTAAG CAACGATTTC TCTAAAAGCA AAAAGGTTAA ACTACTTTAA 17250
17251 AAAAAATATA GCCAATCATA TTTACTTTCT GCCCGTGCTG CATTACAATT 17300
17301 GATCGATTAG ACAGATTGTT TTAAACTCC AACTTTAATA TCACGTAGAA 17350
17351 TAGCCCATAT TCTTATTTAT TTTTCTAACA GAAGAAACAG AAAATCATAG 17400
17401 GACGCTCTTT GTAGCCTAAT GTTAAAAACG AATAGCATAT CACAAAGCAG 17450
17451 AAATAAAAAA CAGGAATTAT TATTTTACAA TTAAACATTT TAAATATATA 17500
17501 AGCTTCATTT TATCCATTAA TCGTTTCCAG TGTTTTTGCT TCTTTTTTTG 17550
17551 TTGCTTTATT TAGCCTAAAT TGCAGAGTGA CTGTACAGAA CTCCACAATG 17600
17601 TGTAGCTGCT TCTAGAAAAA AAGATTTTGT GTTGTCTCTG ATCTGTGTGT 17650
17651 GATATCTAAC GAACATAGAA AACAAATAAA GCTGCATTTG ATAAGATTAT 17700
17701 GACTGTTTAA AAGTCAATAT GGACGATGAT TCCCTTTAAA CGAGTGCGTA 17750
17751 ATAAAGTCGG GTTACATGCA AAAAGCTTGT TAATATCATC TGATGGCTTC 17800
17801 TTGTTTGGAA TCGTACAATT TTAGTAAATA GACAGTAACG CGTATTAACG 17850
17851 TGCCTTCATA TATTGCCTAT ATTTAAACT ATTAATTTTT CGGTGTGGCA 17900
17901 TAAACCACAA GCATTTAAAT GGCATGCCTC TGTCATAATT CTGGACAATG 17950
17951 TAATGTTTGG TTTTATCTT TCCACCAAAA CATTTATTCA TTAGACTTTT 18000
18001 CCCACATGAG TGGTTGTTCA CTTGTTTTAT GATTATGCGA AAAAGCCCT 18050
18051 GTTTTGAATG GGCTTTTATT CTATTAAGTG CGCAGACTAT TTGCACTCAT 18100
18101 CGCCTATCAT AACATGCACA ACAGATTCTG ATCGAATTAT TTCTGTACAC 18150
18151 TATCCTCGCA GAGATTGCAC TCTCAGCATG TCACTTGTTG TATTGTTTAG 18200
18201 GAGCCCGTTG GTTTTTGCGG CAATGGTCGT TAACCTCTGA ATGACCCCGT 18250
18251 GTGGAGCCCA CGCCTCACAA AGCGCCGCAG ACAGGCAGTG TAACAGTTGA 18300
18301 GCCAGAGGAT GGCGCACCAC GACAATTCAA AACACCCAAG CCCTGGCTCT 18350
18351 GAGTTTGCCC TAGCCACGAG CGCATGGAGG CGCGCACCTC TCACATCCCA 18400
18401 CCCAGTATTT CCTCTGTGCA CGAGTTTACC TCTGGAGGTC ACCAAGCAGG 18450
18451 ATTTACGACT GGTCAACAAA AGCACGTGAT TCTCCGCCGT ACCCCATATT 18500
18501 TGGGTGCCTA CGTAAGAGAG AATCAAGTCC ATGTCCCACT CATTTCATA 18550
18551 ATTCATCATA AATTGTGCAA GGGTGCTATA GGACGCGCTA AACCATAAGA 18600
18601 GCCACAAATC AAGCACACAG GTTTATGCTA TTTTACCTCT AAAAAACAATA 18650
18651 AAAAAACAAC AACAATACCT ACAACGAGCA AACGGTGGAA TTGTCAACAA 18700
18701 ATGAGCTCCT ATTTTGTGAA CTCATTCTGC GGTGCTATC CCAATGGCGC 18750
18751 GGACTTCCAG TTACATAATT ATGGAGACCA TAGTACGGCA AACGACCAAT 18800
18801 ACAGAGACTC AACAGCCATG CATTCCAGCA GGTACGGCTA TGGCTACAAC 18850
18851 GGGATGGACC TAACAGTCGG GCGGGCCGGT ACCGGACACT TTGTGGGCAA 18900
18901 CGAGCGGACG CAGGGCTACT CCCCAGTGCA CTCAGCGGCG ACAACACCTT 18950
18901 CGGTGGAGCC GGTGAGGTAC ACCCAGTCCG CCAACAGCAC CGGGACATCG 18950
18951 TCTCTATCGC CACCACCTGA CCCACTACCG TGCTCTTCGG TCGCCAGCTC 19000
19001 GTCCCCGGTC ACCGAGACTC AATCTCAACA CCGAGCCGTG AAAAACTCTA 19050
19051 TAACGACCCC GTGCTCAACC CCGTGCTCGA GCTCGAACGG AGGCACGCTG 19100
19101 CTGCTCAACC GGGACTGTGT GTCCAAAGCT TCCCCCTCG AGGAAGAGAA 19150
19151 GCCTGCGGGA AGCGCACCGA CAACTCCTCA AAATGTCAGC GACTCAACAC 19200
19201 AGCCTCAGAT ATATCCGTGG ATGAGGAAAC TGCATATAAA TCATGGTAAA 19250
19251 CTTCTTTTCT TTGTAATGAT CTTGTATTGA TTTTAGACAC CTTGCGCTGT 19300
19301 GTTATTGTTT AGTGAAGGTG TGATCATGTT TATAGCTCAT TCAGTGAATT 19350
19351 TTAACTATC AGAAAAACTG GTCAAATTTG CACTATATGA TATATATATC 19400
19401 GGTAAAAAAC AAAAAACAAA CACGTTTACT CTTTAAAGGA GCATATGGCG 19450
19451 CCATATGCGC CTCTCACCTG TTAGGTGCTT TTCAGTACTG TCAACATAGA 19500
19501 CCAGGTACGG GATGATTTAA TAACATAACA AAAGAGTTTA AGGTAGTTA 19550
19551 AAATGGTATT ACATATATAC AGGATATATG GGTGTGTTG TTTGTTATAC 19600
19601 TGTAGTTTTA TGGTAAGACA GCCCACATTT ACATCATTTT ACCGTGTTAA 19650

```


19651	TTCCTCTCCG	CAACATGCAA	TGGAAAAGGT	GTGTGTCTCA	TTTTGTTACA	19700
19701	TGTTGTTTTT	AATGATTTTT	GGTTCTATTT	AAGCAACATT	TAAAACCTTC	19750
19751	TCGTTTACCA	ACGTTTACGC	CATTCCTACC	TACAGAAACC	ATATAGGCAA	19800
19801	TATAAAAGAA	ATGAAAGAAA	TGTAGTCCAA	AAAGTTTTTG	CACAGATGTT	19850
19851	TTAAACATCT	ACAGTATCTC	ATGTGTTTTCC	ACTCTCTTTT	TTGTACAAGA	19900
19901	TTTGGCAGGA	CTTGAAGGGA	AGAGAGCGAG	AACTGCATAC	ACCCGCTATC	19950
19951	AGACCTTGGA	GCTTGAGAAG	GAGTTCCACT	TCAACCGGTA	CCTGACCCGC	20000
20001	AGGCGGAGGA	TCGAGATAGC	CCATGCCCTC	TGCCTCTCAG	AGAGACAGAT	20050
20051	CAAAATCTGG	TTTCAGAACC	GCAGGATGAA	GTGGAAGAAG	GACAACAAGC	20100
20101	TGAAGAGCAT	GAACATGGCG	GCTGCAGGCG	GGGGAGGCTA	CAGGCCCTGA	20150
20151	TATTCACCTC	TCTCTACCCC	TGATAGCGAG	CACTATGAAA	ATAAAACAAA	20200
20201	GATTGTCTATA	TCAACTCCTA	TTTCAAAAAA	GCGCAAAGAC	TTGGTTAACG	20250
20251	GGCTGTGTGC	GGGACATGGG	GACACCCAC	CTACGCCAG	TGTATGTCCA	20300
20301	ACAACCGTCC	CATTATATTC	CTTATTAATA	TTATTTCCCTT	TTTATTTCCA	20350
20351	CCGTGGTTAT	GCTTATGTGA	GTAAAAATGC	ATTATGTACA	GTTCTGTCTA	20400
20401	GATTTAGAGG	AAAAAGAAAC	ACAAGAAAAT	AAATGTTTGT	TTAAATTATT	20450
20451	TCTTGTTCAG	ACGAAAAAAA	AATAAAAACG	TCAAAGCTTG	AAATTGCTAC	20500
20501	CTACCGCAAA	TAAATCCAGT	AAAAGTTGTG	TGATGCACTC	AGAACACTAA	20550
20551	CAGCCTATTA	AATAGAGGGT	TGGGGTCGAG	CTGCCTTTGT	GTTTAGCCTT	20600
20601	CGTTTTCTAT	GTATGTACTG	TATATGTATG	TATTTATTTG	TTTGATGTAT	20650
20651	TGTACCAACT	TGTCAGATAA	CTTGCTTCCA	AAAGTAGAGC	CAGCAGAGTA	20700
20701	TGTAAAGCCT	TGTGTGGTTA	TTGTGTGCGT	TGTGAACACA	TGTATGGACA	20750
20751	GAGCTGCCGA	CTCGGTCTAC	AGTGGCTGTG	CCTTGTATCT	TTTTTGTTTG	20800
20801	TTTGTGTTGG	GGTTTTTTTTT	TTTTTTTTTTG	CTGTGCCTTG	TATCTTTTTT	20850
20851	GTTTGTTTGT	TTGGGGGTTT	TTTTTTTTTTT	TTTTTTTTTTT	TTGCCAGGGG	20900
20901	AGCTGCTGTT	TGCGTTGTTT	CAAACCTGGG	GCGCCACCAT	CTTTTAAAAA	20950
20951	GGTTTTTCTT	TGTTCTTTGG	TTTCCCAATA	AGCTGTTTTT	AATGTGGGAG	21000
21001	GGTCCCACAG	TAAATCCCAT	TGCCTAATGT	GTATAAACTC	CATCGTGGAT	21050
21051	TAGCTTTTTG	TAAATGTTTT	TCAAAGACTG	GAAGTCAAGA	ATATACTATA	21100
21101	ATAATAATGA	TGATGATGAT	GATGATGATG	ATGATGATGG	CCATTATAAT	21150
21151	AACTTTTCTA	CTGGAACGGC	AACATTACAG	TGAGTGTACG	TTTTTGTAAG	21200
21201	ATATTTAGGA	TTCATTTGCG	CTGTTGCTGC	TGCATGGAAG	ACAACAATTT	21250
21251	GCTGTTTGGT	CTAAAATCTT	TGAGTCACTT	GTATGAGCTG	TCGAAAAATG	21300
21301	TCAAAGTTTC	CTCAAGTCAA	ATCTCGTTGA	ATTCAACAGG	TCCCATTGAA	21350
21351	TGAAAGTTTC	ATATTTTACT	CACCTTGCTG	TGGAGCTGGG	GGGCTTTTGG	21400
21401	GGGTCTTTGG	ACACTGACGG	GCTCCAGGGC	CGGGCGTCCC	CTTGCGCACA	21450
21451	GACTCCTGGG	CTATTCCCGG	AGAATCAACT	CGACTTTTCA	AGCTGTAAAC	21500
21501	TTTATTAGGG	CCGATTCTGG	CTCTCTGACA	TTTGGGTGCT	AAATGAATGG	21550
21551	GGGGTTTTGT	CTATGAATTA	AATCGTAAAA	ATCATCCGGA	GCGCGGCCAG	21600
21601	ATAGGCTCAC	TGGCCATAAA	CGGTCACGTG	GTAGCCATTA	AAGTAAGTTT	21650
21651	TATGGTTTTG	GGGAGTTGAC	AGTATATTGC	ACATAACATA	TAATCGCACT	21700
21701	GACGGCGAGG	CTTCGTCTGA	CTCTCTCTTT	GCGGCTGTAA	GTGTTACTTG	21750
21751	ACCGGCACAG	ACCGTCTCTG	TTGCTCAGTC	TGCCTCAGTA	TGGACCGGAC	21800
21801	GCTCTCCAGG	ATGAGGACCG	CGTCTCTGGG	TAAGTTTCTG	CAGACTATCC	21850
21851	GGTAGCCCGA	GAACAGCCAG	CCTGTGTTAA	ACAGTCTTAT	TATTGTATTC	21900
21901	CCTCCCTCTA	TCCGTCCCTA	TTAGACCATT	ACACTAATTT	GTCAGTTGTT	21950
21951	GCAGTCTGTA	ATACGGGTCT	CGGCAGATGT	TGCGCAGGGT	TGGGTCCGGC	22000
22001	TCATGGCTAT	TTAATTGCTT	CCTTCCCCGT	GTAAGTTGTT	TCCCTCCATC	22050
22051	TGTGGTGTA	CTCATAATCC	ATTTTTTACAT	GTTGGATAAT	TATCTCTGTG	22100
22101	TCCATTTCCA	CCCCACAAA	CTTCTAGTTG	TTGTGCTTTA	TTACATTGGA	22150
22151	GAATTCACAG	ACCGAAGCCA	AAATTCCAGG	TTTTGGACTC	TAATTCCTGA	22200
22201	TAAAGTTAGC	GCAAATGCAG	CCTACAATTA	CAGCTAGCTG	CCTTTGATTG	22250
22251	GATTCTGTAT	GTGCTGATGA	CTTTTGGAAG	TTGTAGGTTT	ACCCTTGAAA	22300
22301	AATTGCAGTT	GTTTTGGTGA	GCGCTGCGCT	CCACATGTTG	GCACCATGCA	22350
22351	TTAGGACAGG	CTGTAAACAG	TGTCAAGTAA	ACACAACATG	GGAACAAAACA	22400
22401	GCAGCAGGGA	GACTTTTGAA	AAGGAATTCC	GTTTTGTGTC	TGCCACGTCG	22450
22451	TTTTGTGTTT	GGGGTTGAAA	AGTTCAAATG	ACCTTTCAAA	AGACACTCTC	22500

22501	GCTTTGTGGT	GACATCTTGT	CCGCTCAATG	CGGGGCGAAA	CTGGCCACAA	22550
22551	ACTGACCCCT	GAGTCCAAAA	AATGCACTGG	AAATTGAAAA	TTAAAAATGAT	22600
22601	ATAGTTTTTT	GACTAATTTT	TCACAATATT	TTTGGACACA	TTGGCCCTTT	22650
22651	CTTTTTCTTT	TTTTAAATCA	TTTTTATCAT	TAAGAGTGTG	TGTGTCGATA	22700
22701	CGCCCGCTTC	TTATTTGAAA	TTTCCAGACC	AAATATGCAC	GAAAAGATCT	22750
22751	TTAATAAAAA	CGGATATCAC	CATTAGAGTC	GTATGAGTTT	TGGCAGAAAA	22800
22801	AACACCCTGC	AAATAACCAT	TCCGTTTTTT	AAAGTTTGTG	CCTGCGCGGA	22850
22851	TTTTCAAAGG	TTACATTGTC	CGCCAGCAAT	CATAACTTAC	CAAGGTTTTA	22900
22901	TTTTCGGTGC	AGAGGCCTAC	ATCACACTGT	CTAGAACCAT	TAAAGGTTAT	22950
22951	TTCTGAAGCT	TTCTGCAAA	AAATATATGT	ATTTATTAGC	GTAGAACCTA	23000
23001	TTAAATCCAT	TTCCGCCACT	GCTGAAATGA	TCACCTTTAA	AAAATAGATT	23050
23051	CTGGACAGAC	AAGGTGGAAT	TCTTCCACTT	TGATTTTATT	TCTTCCTGTT	23100
23101	TTTAAACAG	AAATCATAAA	GTTTTGAGGT	TGCACCGATT	GTCGTTCTCA	23150
23151	CCAGAATAAT	CCAAACCGCA	TTGCTTACAT	CCTTCAGTTT	ATTGCTTAAC	23200
23201	AAATTCCGCT	CATTGTCATG	CAAAGATGTC	TCCCTCTTAA	CAAAATACGA	23250
23251	ACTTACACTC	TTTCCAGTTT	GATTTTTTAGA	AGCACTAAGA	ATACAACAGG	23300
23301	GCGCACTGAG	AATGCGGCTA	TTTGGAAATG	AAAACCTTTG	CCAATATTTT	23350
23351	AAAAGGAGAA	ATTTGACAG	GGCCCCATGT	GCGGCTCACT	GCCAGGCAGA	23400
23401	GAGAGAGAGA	GAGAGAGAGA	GAAAGAGAGG	AGCTGTAGGT	TCACTGGGAG	23450
23451	GTCAGCCACA	GCACCAAGAT	AAATCTGCAT	CTTCTCAGAG	CCACCAGCAG	23500
23501	AGCTCGCTTT	AGGCCAAGTT	CACTGTCACC	CTGCTGTATA	GAGAGAGGGA	23550
13551	AACCGCTCAC	ACATCCACGG	TACCGCGCGC	ATACACACAC	ACACACACAC	23600
23601	ACACACACAC	ACACACACAC	ACACACACAC	ACACACAAGT	TCACCGTTAA	23650
23651	TGACACGTTT	ACTTGGCGTT	GCCTATAAGC	ATAACGCATA	AATGAATTTA	23700
23701	CCGAGGGACG	CAGAGCCTGT	GGTAACATCC	TGCACCAGAG	GAAAAGACTG	23750
23751	TGAAGAGAGA	GGGGGAGTAT	AACTTTAAGT	GTTGCCTCCA	ACAAAAGACTT	23800
23801	TTAGCTGCTT	TGTTTGCTCT	ATAACGCTAA	TGAATGATTT	GCATATTTTA	23850
23851	CGGTCCTATT	AGTCTTGAGT	GGGAGCCATG	GATTTATCTT	TTCAATCCCT	23900
23901	CGTTGGAAAC	ATCATAAAGC	TACAATTGCA	TTTTAGTTTT	GTGCTGGCAT	23950
23951	CTTCTCTAAT	TATGAAATTT	GGGAAATGA	TCACCGTGAT	GGCAAGAGAG	24000
24001	ATACACATTG	TTAAATCAA	TTTTTTTCGT	TTCAAGAGGG	CCAGATATCT	24050
24051	CACCGGGCTG	GGACAGATGA	GAGGGAGATA	GCCGCAGAAA	GAGGAGGAAA	24100
24101	AATTGGGGTC	AAAAGTTTAC	GCGCTGAACA	AGTCTCCGTC	ATCTGTTCAC	24150
24151	TCGGAGCTTG	ATGCCAGCCT	TATAATAGCG	ATCTTGACTC	TCCGCACAAA	24200
24201	AGACAGAATA	GCTTTGAATT	ACATATGTTG	CACGGTGCAC	TCCAGGTGAA	24250
24251	CCCTGAAAGC	CGGGTGACCT	CGCGTTTGGG	ACCGGGGGAG	GGGAGCCCCC	24300
24301	ACCCGTCCCC	AGCAACAAGA	TTGAGTGGCT	GGCGTTTTAT	TAGGCGGGTT	24350
24351	GATTGGTGAA	TTTCTTTTGA	ATAAATTGCA	TTTGATATGT	TTGACGAGGG	24400
24401	ACGGAGGCAC	AATTTGCATT	TTGAGGATTG	ATACTCTCGA	CCCGGTCACA	24450
24451	GGGTAATGGG	AAGGCCCCAG	CCCCCAAATT	GAGCCCCCGG	ACCCCTTAAA	24500
24501	AAACATTTAG	CCCCCTCTTT	CCTGATCAGC	TGAATAGAAG	CCTTTTAAAA	24550
24551	AAGCAGAAAT	GGGACAGATG	TTTTTGTGAC	TTGCAGAGAG	AAAGTTTATA	24600
24601	AAACATTTAT	TTGTGGAAAT	ATTTAGGCCA	ACATATATGA	ATTTTATTTT	24650
24651	TCAGTAGGCC	TATTTAATCA	CAGCCTATTG	GTTTCCCTTA	TTTAATCTGG	24700
24701	AAAATGAACG	CAGCTATTTT	TTAAAGCGGA	CTCAAATATT	TAACCTTTGT	24750
24751	AAAGGCCCAA	TGAGAAATAA	AAGAATTTAG	TGAAAAATTA	CGTCAGATAA	24800
24801	CAAACCCAAA	TCGACCCAAA	AAACAAAATA	TTTTTATTTT	ATTATGCTTA	24850
24851	AAAAGCCTAA	AACTGGGTAA	ATCTGATAAC	AACGTATGTG	CAACCGGCAA	24900
24901	CACACACGAT	CACAACTACA	TTTTGTTGGA	AAAAATAAAA	TTTTCAATTT	24950
24951	AAACCACAAT	GAAAACACTT	GATAAAAAAG	GGCCAATGAA	ATCTGTCTAT	25000
25001	ATTGACATTT	TCGATTTTAT	CAGCTTTCTA	TTGTCATCTT	AACACAAAAT	25050
25051	GAATAGAAAA	GACATACTTC	ATTTTAGAAA	TGTTCACAAT	AAACAGTTTT	25100
25101	ATTTGTATGA	ACTGGACATA	TTTGTATAAA	CACAACACAC	AAATGCAAGC	25150
25151	ATAATAGTTT	GCTTTTTTTT	TTGCACCTTT	TTCTACCACA	AAAGCGTTAC	25200
25201	ATTTAAAAAG	AGTCGAGTGA	CCGCATCTGT	TCATTCACAA	CTTTGTTTAT	25250
25251	TGAATGCAAA	AGCTCGTCAA	TGAAGGCTCA	GTTTCATCAA	AGTTACAGCG	25300
25301	CGGGATCCTT	ATTATTTTCT	ATAGTTTTTT	GTTTTTTTAT	TAATTACACT	25350

25351	ATGATTTCCA	ACAGCACGAA	TCTGACAGCA	GGTAATCTGA	CTGTTGGCAC	25400
25401	TTTAATTCAC	CTTACAGTGC	ACAGTACAAC	TTATTTTCAGC	TTTTTGAAATC	25450
25451	TATCTTTAAC	GGACTTCAGT	CTGACGCTTT	TTCGTTTTGG	AAACATTAGT	25500
25501	TGATTCTTTG	TCTAATATAT	TTTGTCTCAG	TACTTTCAAT	TACAACATCC	25550
25551	TCTTTACTTA	AAGGAAAAAG	GAA GTTGTTT	ATGCACTTTC	TAAAGGATGT	25600
25601	AGCTGCTTTT	GTCTCTTCTA	ACCATCGTTG	AAGGCTTGAA	CGTCACTTTG	25650
25651	TTAGTGGTCT	AATCAAAGTT	TGCAGCTGGT	GATAAAAGCT	GTCCATTCTC	25700
25701	TGTTACTGTG	GAGGCTCATT	CTGTCGCCTA	TATCCTTGCT	GTAAAGACCA	25750
25751	AGTGGTGACT	CACTGGGAAGA	AAAAAAATAG	AAAAAAGGGG	GAAAAATATTT	25800
25801	TATTATGAGT	GTTTACCTAT	ATCAAAATAT	TTCCATAAAG	GTATGAAAGG	25850
25851	TGTTTTCTAG	CTAGAATCCT	TGGCACATTT	TGAATGCAGA	AAGGATTTTTT	25900
25901	TATTCCGCTG	TCAAGGCTCC	GTATAGACAA	GAGACCTTGT	GAGAGCCATC	25950
25951	CTGCAATCTG	TCACTATCGT	CACAGCACTT	CTCTGATAAT	CTAAATTTTAC	26000
26001	AGTAAACTGT	CAGCATGATG	TATATTGGTT	AAATAGACCT	TCACTTCACA	26050
26051	GTATATGCTC	TGAAAGGGAA	CACCCGTGAT	CTGGCTCTTA	TTTTGTCCCT	26100
26101	AAATAAAATC	TCTAGCACGA	GGCTGTAAGG	TTTTTTTTTT	GTGCGCATTT	26150
26151	AGAGCAAAT	AGCTCTTCAA	ATTTTATGGA	AAATGCTTGA	GATAGTGAAA	26200
26201	GCTAGTGTA	CTTTGTTTTG	AAATGCTCCC	TTTCAGTCCA	ATGGTCAATT	26250
26251	GGATGATCAT	TATCCTATTA	GGCCTGTATG	TTGGGCATGC	TGCATTGTTC	26300
26301	AAAAGAGATT	GCCAGATGCT	TTATTGCATT	CAACATACA	GCAGCACCTT	26350
26351	ACAGGCCTGG	ACTAGATAAC	GGAAGTATTA	CCAGGACTGT	TATCTAAAAT	26400
26401	GCCTCTTTTA	CTGGAATTCT	TGAGCTTCCA	TGATAATGTG	TTTTAATAAAA	26450
26451	GTTATTACAC	GCTGAAACAT	TAAACAAAAA	GTATCCATAG	CAACAAGTAG	26500
26501	AATGACCCTG	TAGCCGTTAT	TCATGCATGG	ATAATTTAAG	AGTAAGTGGA	26500
26551	TAATAGCCAA	CGTGATCTGG	CCCATATTTA	TTCAGGAACT	ATATATAATA	26600
26601	GAATGGCAGC	GGAAGTGTAA	TTTAGATATC	ATGTGAAGTC	ATCTTACCCT	26650
26651	TGACATTATG	TTGACACATT	GCCTCTCAGG	TGGTCTCTTG	GCCGTACGCA	26700
26701	TACAGTGGAT	ATATGGCGGA	AAATACATCA	GCTGCCATTT	ACAATGAGTT	26750
26751	AGGTGCGAGA	AAGCCTGTAA	GATACACAAG	ATAGTAGCTC	TGCCTTACAG	26800
26801	CAACAGTGTG	CTCACATGGG	GAAAAAAGC	TTTTTTTTTGA	AGAACAGAAA	26850
26851	ACTGAGGCCCT	GCCCTCTGGT	TGTTGCTGAA	AAGAAGAATT	CACCTGTTTT	26900
26901	AAGTCGTCTG	CATCAGTGTA	TCCAAAACTA	GAGGTTCAAT	GATGGAGGTC	26950
26951	AGCAATGTGA	AAAAAAAATT	CTACTATATT	TGGTGTTAAA	AATATTAGCC	27000
27001	CATCTGCAGC	AGATCTCTGT	AGTTGTAAGA	TCCCCTTTTC	CACAGATGCA	27050
27051	TATTCGAAC	CACAGTTGGA	CGTGTTGGTAT	TCTGGATAGT	TGGAGGTCTT	27100
27101	TTGCACATTT	TTAAAAATCC	TTTGGGGACC	TTCTGGGGC	GCTGACCCCG	27150
27151	AGCTAAGTGC	TCCCCCAAC	TCCCCCAGAT	TTGAACGGAC	CTGTAAAAC	27200
27201	GAATCCAAGT	CCGGGGTGAA	CCCAGGTCAC	TGCGTCTAAC	AGATATGAAA	27250
27251	ATGTCGCCTC	TTTGGA AAAA	AAGGGCACGC	CTTGTTGTTT	AACAAAGACT	27300
27301	GTCAATGGGC	AAGATTAAATC	AGAAACAAAA	TGGAAACGGT	GTCACGTTGG	27350
27351	GGTCAGGCAG	AAGTTATCTC	TGTTTTAGAGG	AGCGAATAAA	TAGAGGGGGG	27400
27401	ACCCTGGAAA	AAAAAGAGAA	AAAAAGGACA	GAAAAAGAAAG	CTTTCAATAT	27450
27451	TTCCGAAGTT	TGGCCTCTTG	TTTCTCCATG	TAGAGGCCTT	GTTGACGTGG	27500
27501	ACAGTGTGGA	TATGCTTTGT	GAAGAGCTGC	TGCATCTTCC	AGGGACACAG	27550
27551	ATACAGTGTG	CCACCTCGGA	CATGGAGAGG	AAACACGACC	TGGACGTGAT	27600
27601	GGCTGATGGC	GGGGAGTACG	GTCAGAGACA	TTTATGTTTT	TATCATGTGA	27650
27651	ACATGCATTT	AAAAAATATA	CATTTTTTTT	TTCATTATAT	ATAAATGTAT	27700
27701	GCGAGCCTGG	AAGGATGATC	CTGCATTTGT	GCAGTTTATT	TCTCTAAAAT	27750
27751	AATGGCATT	CATTTTCTTA	TTATAATTAC	AATAGAGGTT	ATAATTGTTA	27800
27801	TTATTATTAT	GTGTTGTTGT	AGTTGTACAC	TGGAATTTTT	GCCCCTTTAC	27850
27851	ACTTTTAGTA	CCCTTCTCTT	CAACACATAT	TTTCATTGTT	GCAATACATG	27900
27901	ACAAGTTTTT	ATCTATAGCC	TTGGTTGTTG	GACCAAGTGT	TTCTGCAGTA	27950
27951	AAGTGTTAAC	TCCATTAGCA	CAACTTAGCA	TGTTGCAGAT	TTGTATTTTA	28000
28001	ATGTTGGCAA	ATTGTTTGTG	GAAAAAGATT	TAATCTTTCC	AAATTCTGGA	28050
28051	AAGTTGTATT	AGCTGTGAGA	TTATTTCTTC	TGTGCTGAGC	AATTTGTGACA	28100
28101	TTTACAACAA	ATGCTCCATC	CTGGTCAAAA	ATCTCACGCT	TTGTTTATAC	28150
28151	TATAATGCAA	GATACGTGTC	CTGCTTGTGT	TTGGATGACT	GAGGCAGTTA	28200

28201	ACACAGGATA	ACTGATCATA	CGTGATGGCC	GTGTGTGATC	AGTGCCAGTA	28250
28251	AACTTTCATG	GGGATATTTA	TAGGACAAAC	AGAGACCAAA	AATAGGCTTG	28300
28301	AAATGATTGG	CTGTGTCCTG	GCAAATTGCA	AAAAAAAAAT	AATTGCAAGG	28350
28351	TCTTATGCAT	GTACACGGTT	TGTGCAACGT	GAGAGATGAG	CGTTTTCCCTG	28400
28401	TTGTTTATGT	GTGTGCCATA	AAGCAGGTTT	TACTGATGCA	CAAGTCATAT	28450
28451	GCTTTATATG	CAGGCTTGTG	TGTCCGCAGA	GCAGCGGCAG	AGGGCTTGTG	28500
28501	CACTATTGCT	GTGTTTTTTT	TTTTCTCGAT	CATTTCAATC	GGCTAACAGA	28550
28551	ACAATACGAT	TGGTATTGTT	GTGTTTTTTT	TGTTTCACTG	CTCATCTCTT	28600
28601	TATTTAGATT	CAACACGTAT	TAAACTCCAA	AAACAGCCAT	TCACGTCACT	28650
28651	ATTAGATACA	GCCTTTTGTT	TGTGATAACC	GCATTTTCTC	AGGGCTTGGA	28700
28701	TTTCATATCC	ACTTCTCTTT	CCAGAGGAGA	TGGATTTTTT	TTCACATGCA	28750
28751	GAAAAGCCTC	ACCCACATAA	GTTTGGCCCC	ATGCAGTAAA	CAAGTTAGTT	28800
28801	GGATTGGCTT	TGCTTTATAT	ACTGTGATTT	AATAAGGAAA	ACACCAACGT	88250
28851	AAGCCAGTGC	GGAGCAGTGT	TGGTGGTACC	AGTGCCATGC	TTTCCATTGG	28900
28901	TTCTTGTTTA	CATGATGCCC	ACAGGACACG	CGGTGATTGG	TGGCGCTTCA	28950
28951	CACGTGACCA	GGCAACTTTG	TACATTTGAC	AGGGAGTAGG	AGGGTTTTTG	29000
29001	GGAGATCAGA	AAAACGACAG	CGCGATAAAA	ATTAGTATTG	TTGCACTTCA	29050
29051	CAAATTAATG	ACCATGAGTT	CCTACTTGAT	AAACTCCAAC	TATATCGAGC	29100
29101	CTTCTTTTCC	TCCATGCGAC	GAATATCAGC	AGAGCGGATA	CATCCCCAAC	29150
29151	CCTGGTGACT	ACTACGAGCG	GCCAAAAGAT	ACGGGCTTCC	CCCATCACGA	29200
29201	CGAGCCATCC	TATCCGAGGT	CAAACACAC	AGAATCGGGC	TACGATTACG	29250
29251	GTAACGTCCC	CGCCACCGGA	CTCGATGATT	TCGGCGACGG	GCACCACGCA	29300
29301	CAACCGCAAC	CGGTTCCACA	GAGCCACGGT	CCTCGCCTCA	CCGCGGCCCC	29350
29351	AGATGGTGGC	GCAGGGGCAA	ATGCCAGCAA	AGACTGCAGC	CTCGCCAGTG	29400
29401	AGGTATATCC	CGGCGTAGCG	AAGGGCAAGG	AGCCGGTGGT	CTATCCTTGG	29450
29451	ATGAAAAAGG	TCCACGTTAG	CACCGGTGAG	TTATCGTCAC	GGTTAAATAG	29500
29501	CAGAACCACT	GAGAAACAAC	CACTGGAACA	CACGCAAGCC	TATTGAGGCA	29550
29551	GCACTAGTAG	CAGCCCTTAC	AACAGCAATT	TACGACTATC	GAGCAGCAGG	29600
29601	GTCGGTAATT	AAGGACCCTC	GTAAATTTTA	TTGCCTGTGT	TCGTCTGTCTG	29650
29651	GGGCCATGTG	CCTTTGTTGC	TTTTTAACCC	AAACTACCTC	CACCACTGGA	29700
29701	GCCGAGCCAA	ACTGTAATAT	CCAACCCCGC	TGTTTTGTTA	ATCATTTCTA	29750
29751	AGCTTGATCT	GTCCCTGGAT	GCGAGTGCTT	GTACACTCGA	CATCTCCACA	29800
29801	TTAAGGTCAT	GTCCGTAAAC	TGAAACAAGA	CCGAGAGTAA	AACAAACAGG	29850
29851	GCGGCTGTGA	TTTCCCACCA	AACCCAGAAA	TGTTTTATAG	TTTTTCAACC	29900
29901	TTATAATGTG	TCCAGTGGAG	GGACTAAATA	GCTTTGGAAT	AAACGCTATC	29950
29951	AGCACTGCAT	GTAACACCTG	CTTGCTTGCC	TGTGTATGGC	TTTTTTTTCCC	30000
30001	CTCCAGTCAA	TGCCAGTTAC	ACTGGAGGAG	TGCCCAAGAG	GTCCCGCACT	30050
30051	GCCTACACCC	GCCAGCAGGC	TCTGGAGCTG	GAGAAGGAGT	TTCACTTCAA	30100
30101	CCGGTACCTG	ACCCGGCGCA	GGCGGGTGGA	GATTGCACAC	ACCATGTGTC	30150
30151	TGTCCGAGCG	CCAGGTCAAG	ATCTGGTTTC	AGAACCGGAG	GATGAAGTGG	30200
30201	AAGAAGGAGC	ACAAGCTTCC	CAACACTAAG	ATCCGCTCCT	CCAGCTCGGC	30250
30251	CTCGTCTCA	GCCTCGGGGG	CCCAGCAGCA	GCAGATCAAA	ACAGGCCAGC	30300
30301	AGCTCGTCCC	CACGCCGTGC	ACCGCGGGTC	TATAGTGCTT	GAATGAACCC	30750
30351	CACCCACAA	AAAACCCAAT	CCAGACTGAG	ATGGCAATTA	ACACAAGTGG	30400
30401	AATTTGATGG	ACCGATTCTC	AGTATTTTAC	ACACATGGTG	TTCTCTCTTT	30450
30451	CCCCTTGCCT	CTGCCATTGG	GCCCTGTCA	CGTCCTTTGT	ACCCTTTCCC	30500
30501	TCTCCGCTCG	CCCTCTCATA	TAAGCTTGAA	ATTACCTTT	AATTGCCACG	30550
30551	GTGGCAACTG	AAGTGTTTAT	ATATTTGTTT	ATTATCATGA	AAGAAGGATA	30600
30601	ACGCACATTC	AATAGGACTT	GGATC			